ISIPTA '07

Proceedings of the Fifth International Symposium on
Imprecise Probability: Theories and Applications

ISIPTA '07

Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications

held at Charles University, Faculty of Mathematics and Physics
Prague, Czech Republic, 16-19 July 2007

Edited by
Gert DE COOMAN
Jiřina VEJNAROVÁ
Marco ZAFFALON

Published by Action M Agency for SIPTA (Society for Imprecise Probability: Theories and Applications).
http://www.sipta.org

Printed by Reprostředisko MFF UK, Prague, Czech Republic.

ISBN 978-80-86742-20-5

ORGANISATION

LOCAL ORGANISATION
Jiřina Vejnarová
Milena Zeithamlová

ELECTRONIC ORGANISATION
Serafín Moral

STEERING COMMITTEE
Gert de Cooman
Fabio Gagliardi Cozman
Serafín Moral
Teddy Seidenfeld
Jiřina Vejnarová
Marco Zaffalon

PROGRAMME COMMITTEE

Programme Committee Board
Gert de Cooman, Belgium
Jiřina Vejnarová, Czech Republic
Marco Zaffalon, Switzerland

Members
Joaquín Abellán, Spain
Thomas Augustin, Germany
Salem Benferhat, France
Dan Berleant, USA
Jean-Marc Bernard, France
Veronica Biazzo, Italy
Luis M. de Campos, Spain
Andrea Capotorti, Italy
Frank P. A. Coolen, England
Inés Couso, Spain
Fabio Gagliardi Cozman, Brazil
Dieter Denneberg, Germany
Thierry Denoeux, France
Serena Doria, Italy
Didier Dubois, France
Love Ekenberg, Sweden
Scott Ferson, USA
Terrence Fine, USA
Angelo Gilio, Italy
Michel Grabisch, France
Peter Grunwald, The Netherlands
Rolf Haenni, Switzerland
Jim Hall, UK
Joseph Y. Halpern, USA
David Harmanec, Czech Republic
Manfred Jaeger, Denmark
Jean-Yves Jaffray, France
Radim Jiroušek, Czech Republic
Gabriele Kern-Isberner, Germany
Etienne E. Kerre, Belgium
Gernot Kleiter, Austria
George J. Klir, USA
Igor Kozine, Denmark
Rudolf Kruse, Germany
Isaac Levi, USA
Thomas Lukasiewicz, Italy
Glen Meeden, USA
Enrique Miranda, Spain
Serafín Moral, Spain
Sujoy Mukerji, UK
Michael Oberguggenberger, Austria
Endre Pap, Montenegro
Renato Pelessoni, Italy
Henri Prade, France
Erik Quaeghebeur, Belgium
Marco Ramoni, USA
Giuliana Regoli, Italy
David Rios Insua, Spain
Jose C.F. da Rocha, Brazil
Fabrizio Ruggeri, Italy
Mark J. Schervish, USA
Teddy Seidenfeld, USA
Prakash P. Shenoy, USA
Michael Smithson, Australia
Wynn Stirling, USA
Choh M. Teng, USA
Matthias C. M. Troffaes, England
Lev Utkin, Russia
Barbara Vantaggi, Italy
Paolo Vicig, Italy
Frans Voorbraak, The Netherlands
Kurt Weichselberger, Germany
Nic Wilson, UK

SUPPORT, SPONSORSHIP AND ORGANISATION

Elsevier
www.elsevier.com

Action M Agency
www.action-m.com

Czech Society for Cybernetics and Informatics
www.cski.cz

Institute of Information Theory and Automation
www.utia.cas.cz

CHARLES UNIVERSITY PRAGUE

faculty of mathematics and physics

Charles University in Prague, Faculty of Mathematics and Physics
www.mff.cuni.cz

TABLE OF CONTENTS

PREFACE

These are the proceedings of the *Fifth International Symposium on Imprecise Probability: Theories and Applications* (ISIPTA '07). ISIPTA meetings are a primary forum for presenting and discussing new advances in *imprecise probabilities*. This research field has a very wide scope, at least as wide as that of probability itself, and it encompasses a number of theories and applications that do not feel comfortable with using precise numbers to model chance or uncertainty in general.

These approaches to imprecise probabilities may be characterised by different types of mathematical calculus and are moreover based on a diversity of philosophical standpoints: a number of them regard precision as an idealisation and imprecision as a more fundamental component of probability, while others regard imprecision more simply as a convenient notion for sensitivity analysis, or for making robust inferences; some approaches interpret probability as subjective, while others are focused on physical interpretations; some are founded on the requirement of self-consistency for probabilistic inference, and still others do not stress this aspect and rather propose *ad hoc* methods. Many other distinctions are possible, and the list could become very long.

What is perhaps more important here is that the ISIPTA meetings so far—and ISIPTA '07 is no exception to this rule—have welcomed all of these different approaches to imprecise probabilities. Indeed, it has been one the aims of past and present ISIPTAs to increase awareness of the existence of different approaches to, and languages for, dealing with imprecise probabilities, as well as to foster discussion among the different communities of researchers in this field. These aims stem from a widespread belief that there currently is no single theory of imprecise probability that is consistently superior to all the others in addressing the multiplicity of problems and controversies that pervade this complex field. This is why SIPTA (`http://www.sipta.org`)—the international society that manages the ISIPTA conferences—has decided to change the original name of the conferences from ISIPTA '07 onwards: to emphasise that there are *theories* of imprecise probability, rather than a single theory.

We have come to realise that this openness to different approaches also has its drawbacks: it can for instance make it difficult to classify the ISIPTA meetings into one of the more usual categories for conferences. This was forced upon us by one of the reviewers for ISIPTA '07, who at a certain point in the review process wanted to know what the ISIPTA meetings really are: "[are they conferences] where anything that is technically sound and may generate discussion can be presented, no matter how premature [...]? Or [are they conferences] where 'accept' means 'this paper is not just sound, but also interesting, and it tells a coherent story; a longer version would also be accepted for a journal'?"

Let us formulate our point of view on this matter, by indicating what we believe should be characteristic features of an ISIPTA meeting. One of these is *open-mindedness*, for the reasons given above. *Quality* is another feature we feel strongly about. For ISIPTA '07 we have tried to select papers that are well-motivated, significant, original, and serious. To achieve this, we have relied on a large group of Program Committee Members, who have subjected the submitted papers to a very careful refereeing process. This process included—for the first time since the inception of the ISIPTAs—electronic discussion between the reviewers of a paper (usually three, sometimes four in number) that lasted for two weeks. Our reviewers have contributed very generously to this discussion, and have thus helped us make up our minds about some of the more difficult papers. This is the right place to thank them for all their careful work, which has proven invaluable to us. We also thank the contributors for their diligence in preparing submissions, and for their patience with our selection process.

We want also to give our views on the important issue of *selectivity*. The well-established format of the ISIPTA meetings has made the review process quite selective: there are no parallel sessions, and each paper is presented both in a plenary session and as a poster. This has limited significantly the number of accepted papers, and it has obliged us to reject a few valuable submissions. In spite of this, it should be stressed that we do not regard selectivity as important *per se*, because quality and selectivity do not always see eye to eye. Moreover, we feel that stretching selectivity may have an adverse effect on the diversity of the accepted papers, and more generally on their originality. We take this seriously as we have a tradition of attaching great importance to originality. That is why in a number of ISIPTA proceedings you may find papers that are not fully mature yet, or that are based on an ongoing research program, or that perhaps could be refined at the level of didactical presentation. Moreover, these proceedings may well contain controversial papers, on which the reviewers themselves had diverging views: we believe such contributions may sometimes be particularly important vehicles for stimulating critical discussion at our meetings.

We believe that the 48 papers included in these proceedings show that at least some of the above-mentioned aims have been realised for ISIPTA '07. More generally speaking, we hope that we have succeeded in making ISIPTA '07 an interesting and stimulating conference to attend. We have been very much helped in our efforts by the invited speakers (Terrence Fine: *In the realm of probability: Limits to standard probability*; and Glenn Shafer: *Game-theoretic probability: Theory and applications*) and the people responsible for the tutorials (Scott Ferson: *Risk analysis: Rough but ready tools for calculations under variability and uncertainty.*; George Klir: *Generalised information theory*; Enrique Miranda: *An introduction to the theory of coherent lower previsions*; Teddy Seidenfeld: *Decision theories for imprecise preferences and imprecise probabilities*). They have contributed their time and many talents in a very generous fashion. Copies of their contributions are included in the electronic version of these proceedings (`http://www.sipta.org/isipta07`).

Finally, and as with all the previous ISIPTAs, the Programme Committee Board is exceptionally grateful to Serafín Moral, who has overseen the electronic management of these papers, their submissions, reviews, and discussions, unselfishly spending much of his time to make these proceedings possible.

—Gert de Cooman,[1] Jiřina Vejnarová and Marco Zaffalon[2]

6 June 2007

[1]Gert de Cooman was supported by the Flemish BOF grant 01107505.
[2]Marco Zaffalon gratefully acknowledges support by the Swiss NSF grants 200020-109295/1 and 200021-113820/1.

Credal networks for military identification problems

Alessandro Antonucci
IDSIA
Galleria 2
CH-6928 Manno (Lugano)
Switzerland
alessandro@idsia.ch

Ralph Brühlmann
Armasuisse (W+T)
Feuerwerkerstrasse 39
CH-3600 Thun
Switzerland
ralph.bruehlmann@ar.admin.ch

Alberto Piatti
IDSIA
Galleria 2
CH-6928 Manno (Lugano)
Switzerland
alberto.piatti@idsia.ch

Marco Zaffalon
IDSIA
Galleria 2
CH-6928 Manno (Lugano)
Switzerland
zaffalon@idsia.ch

Abstract

Credal networks are imprecise probabilistic graphical models generalizing Bayesian networks to convex sets of probability mass functions. This makes credal networks particularly suited to capture and model expert knowledge under very general conditions, including states of qualitative and incomplete knowledge. In this paper, we present a credal network for risk evaluation in case of intrusion of civil aircrafts into a no-fly zone. The different factors relevant for this evaluation, together with an independence structure over them, are initially identified. These factors are observed by sensors, whose reliabilities can be affected by variable external factors, and even by the behavior of the intruder. A model of these observation mechanisms, and the necessary fusion scheme for the information returned by the sensors measuring the same factor, are both completely embedded into the structure of the credal network. A pool of experts, facilitated in their task by specific techniques to convert qualitative judgments into imprecise probabilistic assessments, has made possible the quantification of the network. We show the capabilities of the proposed network by means of some preliminary tests referred to simulated scenarios. Overall, we can regard this application as an useful tool to support military experts in their decision, but also as a quite general imprecise-probability paradigm for information fusion.

Keywords. Credal Networks, Information Fusion, Sensor Management, Tracking Systems.

1 Introduction

In the recent times, the establishment of a no-fly zone surveyed by the Air Force around important potential targets has become usual practice, also in neutral states like Switzerland, because of the potential danger of terror threats coming from the sky. In this paper we refer in particular to the Swiss case, where no-fly zones are usually established to protect international conferences, like the World Economic Forum in Davos, or to protect strategic buildings, like for example nuclear power plants and dams.

A no-fly zone for the protection of a single strategic object usually consists of a circular-shaped region with a radius of several kilometers around the target to defend. All the aircrafts flying in this region without the required permissions are considered *intruders*. The no-fly zone is usually divided in two concentric regions: the external no-fly zone is a large region, with many sensors, devoted to the identification of the intruder, while the internal no-fly zone is a small region, containing the object to protect, where fire is eventually released if the intruder is presumed to have bad aims.

But not all the intruders have the same intentions: there are intruders with bad aims (or *renegades*), intruders with provocative aims, and erroneous intruders. Since only renegades represent a danger for the protected object, the recognition of the intruder's aim plays a crucial role in the following decision, which, if it is wrong, is clearly critical. This is the recognition problem we address in this paper.

This problem is complex for many reasons: (i) the risk evaluation usually relies on qualitative expert judgments; (ii) it requires the fusion of the information coming from different sensors, and this information can be incomplete or partially contradictory; (iii) different sensors can have different levels of reliability, and the reliability of each sensor can be affected by exogenous factors, as geographical and meteorological conditions, and also by the behavior of the intruder. A short review of the problem and some detail about these difficulties is reported in Section 2.

Nowadays, the problem is faced by military experts without the support of any mathematical model. The reason is partly the difficulty of finding a suitable mathematical paradigm for this kind of problems.

In this paper, we propose *credal networks* (Section 3)

as a mathematical paradigm for the modeling of military identification problems. Credal networks are imprecise-probability graphical models representing expert knowledge by means of sets of probability mass functions, which are particularly suited for modeling and doing inference with qualitative, incomplete, and also conflicting information.

More specifically, we have developed a credal network for the considered identification problem. This is achieved by a number of sequential steps: determination of the factors relevant for the risk evaluation and identification of a causal structure between them (Section 4.1); quantification of this qualitative structure by imprecise probabilistic assessments (Section 5.1); determination of a qualitative model of the observation mechanism associated to each sensor, together with the necessary *fusion scheme* of the information collected by the different sensors (Section 4.2); quantification of this model by probability intervals (Section 5.2). An analysis of the main features of our imprecise-probability approach to information fusion is indeed reported in Section 6.

The credal network is finally used to evaluate the level of risk, which is simply the probability of the risk factor conditional on the information collected by the sensors in a given scenario. A description of the approximate procedure used to update the network, together with the results of a preliminary test, is reported in Section 7.

Summarizing, we can regard this model as a practical tool to support the military experts in their decisions for this particular problem. But, at the same time, this credal network can be regarded as a prototypical modeling framework for general identification problems requiring information fusion.

2 Military Aspects

This section is focused on the main military aspects of the identification problem. In particular, we explain: (i) what are the possible intentions of the intruder, (ii) what are the factors that are observed to determine the intention of the intruder, (iii) what are the sensors used to determine these factors.

We consider only civil aircrafts; military aircrafts and flying weapons like rockets or cruise missiles are not taken into consideration. For the possible intentions of the intruder, four categories can therefore be considered: *renegade, agent provocateur, erroneous intruder* and *damaged intruder*. A *renegade* is an aircraft that has entered the no-fly zone with the purpose of attacking the protected object using itself as weapon; terrorists belong to this category. The purpose of an *agent provocateur* is the provocation of the protection structure for demonstrative purposes. An agent provocateur usually knows exactly what it is doing and does not want to die. An *erroneous intruder* has no particular purpose: it has entered the no-fly zone by mistake, because of bad preparation of the flight or due to a bad level of training of the pilot. Finally, a *damaged intruder* is an aircraft without bad aims that is incurring an emergency situation due to technical problems. A damaged intruder enters a no-fly zone because it cannot avoid it or because it is in a situation of panic. In our model, the intention of the intruder is modeled as a random variable, called the *risk factor*, whose possible values are the four cases described above.

The intruder is assumed to be observed for a sufficiently long time window, when it is flying in the external no-fly zone. The factors observed during this period to determine its intention can be divided into two categories: factors describing the *flight behavior* and factors describing the *reactions*. For this first category we consider: *height*, *changes in height*, *absolute speed*, *flight path* and *type of aircraft*. These factors are observed in a passive way, without any interaction with the intruder. The factors belonging to the second category are the *transponder (mode 3/A)*, the *reaction to radio communication with the civil Air Traffic Control (ATC)*, the *reaction to radio communication with the Air Defence Direction Center (ADDC)* and, finally, the *reaction to interception*. The common point of these factors is that they require an interaction (code emission, radio communication or visual contact) between the intruder and the civil or military control.

All these factors are regarded as random variables, taking only a finite number of possible values. Variables which are not intrinsically categorical, are discretized. For instance, regarding the height above the ground maintained by the intruder during the observation period, we are not interested in the precise elevation of the aircraft, but on its flight level. According to military practice, the airspace is divided in four levels: VERY LOW (0-150m), LOW (150-3'000), HIGH (3'000-7'000m) and VERY HIGH (above 7'000m).

Many sensors can be used to determine the factors described above. In our application the ADDC works as a centralized decision center receiving all the information collected by the sensors in order to evaluate the intention of the intruder. The network formed by the ADDC and all the sensors is called the *identification architecture*. The sensors in the identification architecture are divided in four main categories:

- *Signals intelligence*. Sensors belonging to this

category detect signals emitted by the intruder. In our application, the only sensor of this type is the secondary surveillance radar (SSR), that detects the Mode 3/A (identification code) and the Mode C (height) emitted by the intruder.

- *Radar intelligence*. Sensors belonging to this category are all the radars. In our application we have three types of radars: 3D radars, detecting the 3D position of the intruder in the airspace; 2D radars, detecting the 2D position but not the height of the intruder; and tracking radars, detecting the 3D position of the intruder but only at low heights and with a limited range.

- *Imagery intelligence*. Sensors belonging to this category record TV or infrared (IR) images of the intruder using cameras.

- *Human intelligence*. Sensors belonging to this category are sensors where the information is elaborated by humans before being transmitted to the ADDC. In our application there are two sensors of this type: ground-based observation units, where humans observe the intruder using optical instruments and communicate their observations to the ADDC, and interceptors, where the pilot observes directly the intruder and communicate the observations to the ADDC.

The identification architecture is a complicated non-homogeneous structure. In fact, not all the sensors are present at the same time in each point of the no-fly zone. The *presence* and the *reliability* of a sensor for observing a given factor of the intruder depend on the position of the intruder (in particular on its height), on the position of the sensors in the architecture and on the meteorological and geographical situation. In Section 4.2 we explain in detail how presence and reliability are modeled by our network.

3 Mathematical Aspects

In this section, we briefly recall the definitions of *credal set* and *credal network* [4], which are the mathematical objects we use to model expert knowledge and fuse the different kinds information in a single coherent framework.

3.1 Credal Sets

We use uppercase letters to denote random variables. Given a random variable X, we denote by Ω_X the possibility space of X, with x a generic element of Ω_X. Denote by $P(X)$ a mass function for X and by $P(x)$ the probability of x.

We denote by $K(X)$ a closed convex set of probability mass functions over X. $K(X)$ is said to be a *credal set* over X. For any $x \in \Omega_X$, the lower probability for x according to the credal set $K(X)$ is $\underline{P}(x) = \min_{P(X)\in K(X)} P(x)$. Similar definitions can be provided for upper probabilities, conditional credal sets, lower and upper expectations. Note that a set of mass functions, its convex hull, and its set of *vertices* (also called *extreme mass functions*) produce the same lower and upper expectations and probabilities.

Conditioning with credal sets is done by elements-wise application of Bayes rule. The posterior credal set is the union of all posterior mass functions. Denote by $K(X|Y = y)$ the set of conditional mass functions $P(X|Y = y)$, for generic variables X and Y. We say that two variables are *strongly independent*, when every vertex in $K(X, Y)$ satisfies stochastic independence of X and Y.

A set of *probability intervals* over Ω_X, say $\mathbb{I}_X = \{\mathbb{I}_x : \mathbb{I}_x = [l_x, u_x], , 0 \leq l_x \leq u_x \leq 1, x \in \Omega_X\}$, can be regarded as a specification of a credal set $K(X) = \{P(X) : P(x) \in \mathbb{I}_x, x \in \Omega_X, \sum_{x\in\Omega_X} P(x) = 1\}$. $\mathbb{I}_X$ is said to *avoid sure loss* if the corresponding credal set is not empty and to be *coherent* (or *reachable*) if $u_{x'} + \sum_{x\in\Omega_X, x\neq x'} l_x \leq 1 \leq l_{x'} + \sum_{x\in\Omega_X, x\neq x'} u_x$, for all $x \in \Omega_X$. $\mathbb{I}_X$ is coherent if and only if the intervals are tight, i.e., for each lower or upper bound in $\mathbb{I}_X$ there is a mass function in the credal set at which the bound is attained [12, 3].

3.2 Credal Networks

Let $\mathbf{X}$ be a vector of random variables and assume a one-to-one correspondence between the elements of $\mathbf{X}$ and the nodes of a *directed acyclic graph* $\mathcal{G}$. Accordingly, in the following we will use *node* and *variable* interchangeably. For each $X \in \mathbf{X}$, Π_X denotes the set of the *parents* of X, i.e., the random variables corresponding to the immediate predecessors of X according to $\mathcal{G}$.

The specification of a *credal network* over $\mathbf{X}$, given the graph $\mathcal{G}$, consists in the assessment of a conditional credal set $K(X_i|\pi_i)$ for each possible value $\pi_i \in \Omega_{\Pi_i}$ of the parents of X_i, for each variable $X_i \in \mathbf{X}$. The graph $\mathcal{G}$ is assumed to code strong dependencies among the variables in $\mathbf{X}$ by the so-called strong Markov condition: every variable is strongly independent of its nondescendant non-parents given its parents. Accordingly, it is therefore possible to regard a credal network as a specification of a credal set $K(\mathbf{X})$ over the joint variable $\mathbf{X}$, with $K(\mathbf{X})$ convex hull of the set of joint mass functions $P(\mathbf{X}) = P(X_1, ..., X_n)$ over the n variables of the net, that factorize according to $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\pi_i)$. Here π_i

is the assignment to the parents of X_i consistent with $(x_1, \ldots, x_n)$; and the conditional mass functions $P(X_i|\pi_i)$ are chosen in all the possible ways from the respective credal sets. $K(\mathbf{X})$ is called the *strong extension* of the credal network. Observe that the vertices of $K(\mathbf{X})$ are joint mass functions $P(\mathbf{X})$. Each of them can be identified with a Bayesian network [9], which is a precise probabilistic graphical model. In other words, a credal network is equivalent to a set of Bayesian networks.

3.3 Computing with Credal Networks

Credal networks can be naturally regarded as expert systems. We query a credal network to gather probabilistic information about a variable given evidence about some other variables. This task is called *updating* and consists in the computation, with respect to the network strong extension $K(\mathbf{X})$, of $\underline{P}(X|E = e)$ and $\overline{P}(X|E = e)$, where E is the vector of variables of the network in a known state e (the evidence), and X is the node we query. Credal network updating is an NP-hard task [5], for which a number of approximate algorithms have been proposed [8, 2].

4 Qualitative Assessment of the Network

We are now in the position to describe the credal network developed for our application. According to the discussion in the previous section, this task first requires the qualitative identification of the conditional dependencies between the different variables involved in the model, which can be coded by a corresponding directed acyclic graph.

As detailed in Section 2, the variables we consider in our approach are: (i) the *risk factor*, (ii) the nine variables used to assess the intention of the intruder, (iii) the variables representing the observations returned by the sensors, (iv) for each sensor two additional variables representing presence and reliability of the sensor. In the following, we refer to the variables in the categories (i) and (ii) as *core variables*.

4.1 Risk Evaluation

Figure 1 depicts the conditional dependencies between the core variables according to the military and technical considerations of the Expert.[1] As an example, the arcs connecting the nodes *type of aircraft*, *height*, and *risk factor* with the *speed*, correspond to the following Expert's remarks: *there is a strong relation between the height above the ground and the corresponding speed of an aircraft* (technical considerations); *a renegade is expected to fly as fast as possible* (military consideration); *an intruder flying with a light aircraft, because of the limited maximal speed of this type of aircrafts, would necessarily flight very slowly.* The specification of this part of the network has required a considerable amount of military and technical expertise that, due to space limitations, cannot be explained in more detail here.

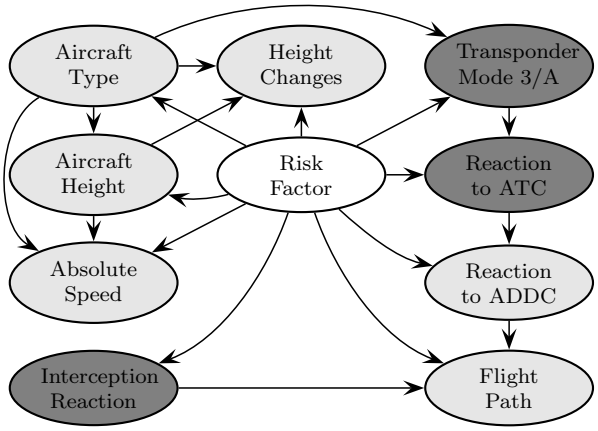


Figure 1: The core of the network. Dark gray nodes are observed by single sensors, while light gray nodes are observed by set of sensors for which an information fusion scheme (see Section 4.2) is required.

4.2 Observation and Fusion Mechanism

We distinguish between *latent variables*, that are assumed to be unobservable, and *manifest variables*, which are actually observed. The *core variables* in Figure 1 are regarded as latent variables that, to be determined, usually require the fusion of information coming from different sensors, with different levels of reliability. Nevertheless, in the case of the identification code emitted by the intruder (*Transponder Mode 3/A*), the *reaction to interception* observed by the pilot, and the reaction to civil air traffic control (ATC) observed by the controllers through SSR, the observation mechanism is immediate; thus we simply identify the latent with the corresponding manifest variable, adding the value MISSING, as possible value of the variable. This value can have particular meanings (eg., a missing Mode 3/A probably means a switched off transponder) and will be also added to the possibility space of the other manifest variables.

Clearly, if the risk factor was the only latent variable, the network in Figure 1 would be the complete network needed to model the risk evaluation. But, because we are dealing with latent variables observed by many sensors, a model of the observation and a fusion

[1] In this paper we briefly call *Expert* a pool of military experts from the Swiss Air Force, we have consulted during the development of the model.

mechanism has to be added to the current structure.

Observation Mechanism We begin by considering observations by single sensors, and then we explain the fusion scheme for several sensors. Consider the following example: suppose that an intruder is flying at low height and is observed by ground-based observation units in order to evaluate its *flight path*. For this evaluation, the intruder should be observed by many units. If our identification architecture is characterized by too a low number of observation units, it is probable that the observation of the flight path would be incomplete or even absent, although the meteorological and geographical conditions are optimal. In this case, the low quality of the observation is due to the scarce presence of the sensor. Suppose now that the architecture is characterized by a very large number of observation units but the weather is characterized by a complete cloud cover with low clouds, then the quality of the observation is very low although the presence of units is optimal. In this case the low quality of the observation is due to the low reliability of the sensor under this meteorological condition. This example motivates our choice to distinguish between *reliability* and *presence* of the sensors in the network.

Figure 2 illustrates, in general, how the evidence provided by a sensor about a latent variable is assessed. The manifest variable depends on the relative *latent variable*, on the *presence* of the sensor and on its *reliability*. Both *reliability* and *presence* are categorical variables with three possible values, HIGH, MEDIUM and LOW for the *reliability*, and PRESENT, PARTIALLY PRESENT and ABSENT for the *presence*.

The *reliability* of a sensor depends on the meteorological and geographical situation and on the height, while the *presence* of a sensor depends only on the identification architecture and on the height of the intruder. The dependence on the latent variable *height* can be explained considering the technical limits of the sensors. There are sensors that are specific of the low and very low heights, like tracking radars and TV or IR cameras. There are other sensors, like the 3D radars of the fixed military radar stations, that are always present at high and very high heights, but are not always present at low and very low heights.

The *meteorological and geographical conditions* do not affect the presence of a sensor, but only its reliability. It is important to point out that these conditions are always observed and we will not display them explicitly as variables in the network, being already considered by the Expert during his quantification of the reliability.

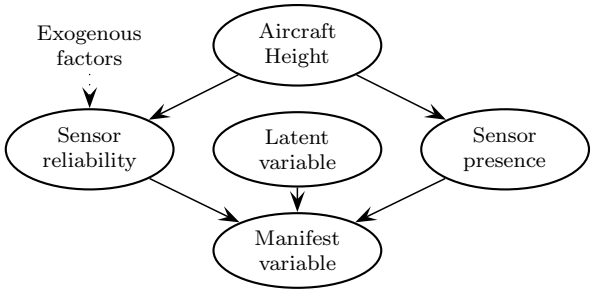


Figure 2: Observation mechanism for a single sensor. The *latent variable* is the variable to be observed by the sensor, while the *manifest variable* is the value returned by the sensor itself.

Sensors Fusion We can finally explain how the information collected by the different observations of a single latent variable returned by different sensors can be fused together. Consider, for example, the determination of the latent variable *type of aircraft* depicted in Figure 3. The *type of aircraft* can be observed by four types of sensors: TV cameras, IR cameras, ground-based observation units and air-based interceptors. For each possible sensor, we model the observation using a structure like the network in Figure 2: there is a node representing the *presence* of the sensor and a node representing the *reliability* of the sensor, while the variable *height* influences all these nodes. This structure permits the fusion of the evidence about the latent variables coming from the different sensors, taking into account the reliability of the different observations in a very natural way and without the need of any external specification of explicit fusion procedures. Similar approaches have already been used for Bayesian networks [6].

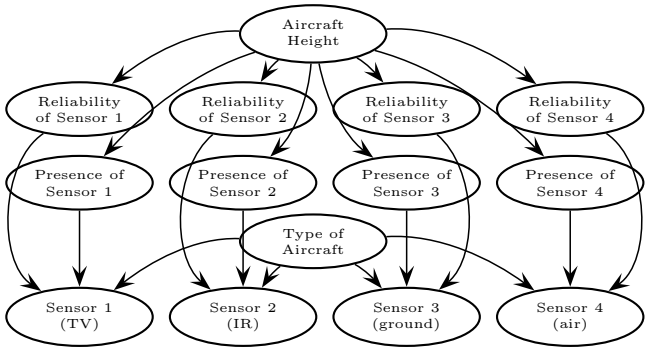


Figure 3: The determination of the latent variable *type of aircraft* by four sensors.

We similarly proceed for all the latent variables requiring the fusion of information from many sensors. This practically means that we add a subnetwork similar to the one reported in Figure 3 to each light gray

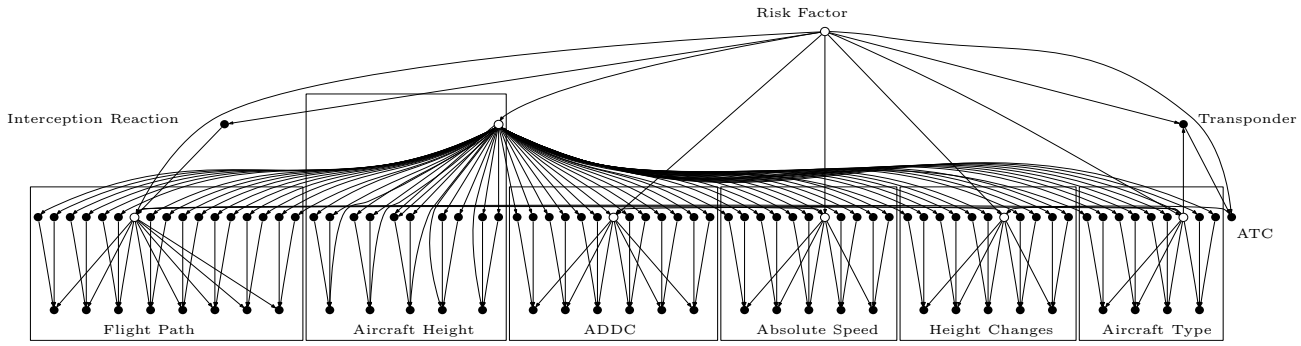


Figure 4: The complete structure of the credal network. Black nodes denote manifest variables, while latent variables correspond to white nodes. Boxes are used to highlight the different subnetworks modeling the observations of the latent variable as in Figure 3.

node of the core network in Figure 1. The resulting directed graph, which is still acyclic, is shown in Figure 4.

5 Quantitative Assessment of the Network

As outlined in Section 3, the specification of a credal network over the variables associated to the directed acyclic graph in Figure 4 requires the specification of a conditional credal set for each variable and each possible configuration of its parents.

For the core variables, these credal sets have been obtained by means of probability intervals explicitly provided by the Expert (Section 5.1), while, regarding observations, presence and reliability, a quantification procedure to automatically transform Expert's qualitative judgments in conditional credal sets specifications has been developed (Section 5.2).

5.1 Quantification of the Network Core

Because of the scarcity of historical cases, the quantification of the conditional credal sets for the core variables in Figure 1 is mainly based upon military and technical considerations. Together with the Expert we have isolated a number of principles, later translated into probability intervals and hence into conditional credal sets. We point the reader to [10] for a detailed description of this quantification task. Here, we cite as an example only some of the principles used to quantify this part of the network: *a renegade is not expected to use balloons or gliders; the light aircraft is the type of aircraft more probable to be used by a terrorist; erroneous intruders are usually light aircrafts and we do not expect a business jet or an airliner to be an erroneous intruder; balloons and gliders are subject to defects due to the meteorological conditions.*

In some situations, the Expert was also able to identify logical constraint among the variables. As an example, the fact that *balloons cannot maintain high levels of height* represents a constraint between the possible values of the variables *type of aircraft* and *height*. These kinds of constraints have been embedded in the structure of the network by means of zero probability assessments.

5.2 Observations, Presence and Reliability

To complete the quantification of our credal network, we should discuss, for each sensor, the quantification of the variables associated to the observation, the reliability and the presence.

We begin by explaining how presence and reliability are specified. Consider the network in Figure 2. The Expert should quantify, for each of the four possible values of the variable *height*, a credal set for the reliability and a credal set for the presence of the sensor. In practice, the Expert is simply required to suggest a value for the presence and a value for the reliability. To assess the value of the presence, he should take into consideration only the structure of the identification architecture; while to assess the value for the reliability level, also the actual meteorological and geographical situation should be considered.

For each specified level of presence or reliability, the Expert should also decide whether or not he is uncertain about this value. His judgments are then translated into coherent probability intervals, from which we can compute the corresponding credal sets reflecting his beliefs. To this purpose, we have defined, together with the Expert, a set of fixed credal sets that are used to model the different combinations of values and uncertainty values. This procedure sub-

stantially simplifies the quantification of the network, while maintaining a large flexibility in the specification of presence and reliability.

Regarding the observations, a conditional credal set for each possible value of the corresponding latent variable and for each possible level of reliability and presence has been assessed. The Expert has answered questions like, *what is the probability (interval) that the ground-based observers have medium reliability in observing the type of aircraft of an intruder that is flying at low height, if the meteorological condition is characterized by dense low clouds and we are in the plateau?*

Clearly, it can be extremely difficult to answer dozens of questions of this kind in a coherent and realistic way. It is much easier to answer questions like the following, *what is the reliability level that you expect from ground-based observers observing the type of aircraft of an intruder that is flying at low height, if the meteorological condition is characterized by dense low clouds and we are in the plateau?* The latter question is much simpler than the former, because one is required to specify something more qualitative than probabilities. This is exactly the type of question that we asked the Expert to quantify the necessary probabilities in our network. In the following we explain, in order, our quantification of presence and reliability of sensors, and the observation mechanism.

Let X be a latent variable, and O the manifest variable corresponding to the observation of X as returned by a given sensor. For each combination of *reliability* and *presence*, we should assess, for each $x \in \Omega_X$ and $o \in \Omega_O$, the bounds $\underline{P}(X = x|O = o)$ and $\overline{P}(X = x|O = o)$.

This quantification step can be simplified by defining a symmetric non-transitive relation of *similarity* among the elements of Ω_X. The *similarities* between the possible values of a latent variable according to a specific sensor can be naturally represented by an undirected graph as in the example of Figure 5. In general, given a latent variable X, for each possible outcome $x \in \Omega_X$, there are outcomes of X that are similar to x and outcomes that are not similar to x.

Having defined, for each latent variable and each corresponding sensor, the similarities between its possible outcomes, we can then divide the possible observations in four categories: (i) observing the correct value of X; (ii) confounding the real value of X with a similar one; (iii) confounding the true value of X with a value that is not similar; (iv) the observation is MISSING. The idea is to quantify, instead of a probability interval for $P(X = x|O = o)$ for each $x \in \Omega_X$ and each $o \in \Omega_O$, only four probability intervals, corresponding to the four categories of observations described above.

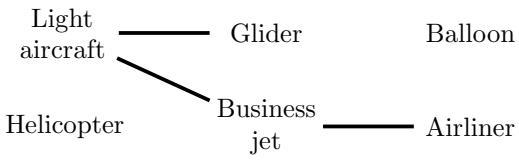


Figure 5: An undirected graph depicting *similarity* relations about the possible values of the variable *types of aircraft* according to the observation of a TV camera. Edges connect similar states. The sensor can mix up a light aircraft with a glider or a business jet, but not with a balloon or a helicopter.

Let us finally explain how the four probability intervals are quantified in our network for each combination of *reliability* and *presence* and for each sensor. The probability interval assigned to the case where the observation is missing depends uniquely on the *presence*. In particular, if the sensor is ABSENT, then the probability of having a MISSING observation is set equal to one and therefore the probability assigned to all the other cases are equal to zero. It follows that we have only seven combinations of *reliability* and *presence* to quantify. To this extent, we use constraints based on the concept of *interval dominance* to characterize the different combinations.[2] In order of accuracy of the observation, the combinations are the following:

1. HIGH, PRESENT: the correct observation dominates (clearly) the similar observations. The probability for not similar observations is zero and is therefore dominated by all the other categories.

2. HIGH,PARTIALLY PRESENT: the correct observation dominates the similar observations and dominates (clearly) the not similar observations. The similar observations dominates the not similar observations.

3. MEDIUM,PRESENT: the correct observation dominates the similar observations and dominates the not similar observations. The similar observations dominates the not similar observations.

4. MEDIUM, PARTIALLY PRESENT: the correct observation does not dominate the similar observations but dominates the not similar observations.

[2]Given a credal set $K(X)$ over a random variable X, and two possible values $x, x' \in \Omega_X$, we say that the x *dominates* x' if $P(X = x') < P(X = x)$ for each $P \in K(X)$. It is easy to show that that interval dominance, i.e., $\overline{P}(X = x') < \underline{P}(X = x)$, is a sufficient condition for dominance.

5. LOW,PRESENT: no dominance at all.

6. LOW,PARTIALLY PRESENT: no dominance at all.

7. ABSENT: the probability of a MISSING observation is equal to one, this value dominates all the other values.

6 Information Fusion by Imprecise Probabilities

The procedure described in Sections 4.2 and 5.2 to fuse the observations gathered by the sensors, can be regarded as a possible imprecise-probability approach to the general *information fusion* problem. In this section, we take a short detour from the military aspects to illustrate some key features of such an approach by a simple example.

Let us first formulate the general problem. Given a latent variable X, and the manifest variables $O_1, \ldots, O_n$ corresponding to the observations of X returned by n sensors, we want to update our beliefs about X, given the values $o_1, \ldots, o_n$ returned by the sensors.

The most common approach to this problem is to assess a (precise) probabilistic model over these variables and compute the conditional mass function $P(X|o_1, \ldots, o_n)$. That may be suited to model situations of *consensus* among the different sensors. The precise models tend to assign higher probabilities to the values of X returned by the majority of the sensors, which may be a suitable mathematical description of these scenarios.

The problem is more complex in case of *disagreement* among the different sensors. In these situations, precise models assign similar posterior probabilities to the different values of X. But a flat posterior probability mass function models *indifference*, while sensors disagreement seems to reflect instead a condition of *ignorance* about X.

Imprecise-probability models are more suited for these situations. Posterior ignorance about X can be represented by the impossibility of a precise specification of the conditional mass function $P(X|o_1, \ldots, o_n)$. The more disagreement we observe among the sensors, the wider we expect the posterior intervals to be, for the different values of X.

The case where the size of the posterior probability intervals results to be increased by conditioning is known in literature as *dilation* [11], and is relatively common with coherent imprecise probabilities.

The following simple example, despite its simplicity, is sufficient to outline how these particular features are obtained by our approach.

Example 1 *Consider a credal network over a latent variable X, and two manifest variables O_1 and O_2 denoting the observations of X returned by two identical sensors. Assume to be given the strong independencies coded by the graph in Figure 6. Let all the variables be Boolean. Assume $P(X)$ to be uniform and both $P(O_i = \mathrm{T}|X = \mathrm{T})$ and $P(O_i = \mathrm{F}|X = \mathrm{F})$ to take values in the interval $[1-\epsilon, 1]$, for each i=1,2, where $\epsilon > \frac{1}{2}$ models a (small) error in the observation mechanism. Since the network in Figure 6 can be regarded as a* naive credal classifier *[13], where the latent variable X plays the role of the class node and the observations correspond to the class attributes, we can exploit the algorithm presented in [13, Section 3.1] to compute the following posterior interval:*

$$P(X = \mathrm{T}|O_1 = \mathrm{T}, O_2 = \mathrm{T}) \in [\frac{(1-\epsilon)^2}{1-2\epsilon(1-\epsilon)}, 1].$$

It follows that, in case of consensus between the two sensors, the corresponding probability for the latent variable increases, given that the lower extreme is larger than $\frac{1}{2}$. In the case of disagreement, instead, we obtain that $P(X = \mathrm{T}|O_1 = \mathrm{F}, O_2 = \mathrm{T}) \in [0, 1]$, which means that our ignorance about X dilates, leading to a completely uninformative posterior interval.

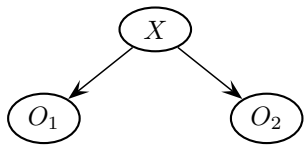


Figure 6: The credal network for Example 1.

Remarkably, assuming fixed levels of height, reliability and presence, Figure 3 reproduces the same structure of the prototypical example in Figure 6, with four sensors instead of two. The same holds for any sub-network modeling the relations between a latent variables and the relative manifest variables in our network.

7 Algorithmic Issues and Simulations

The discussion in Section 4 and Section 5 led us to the specification of a credal network, associated to the graph in Figure 4, over the whole set of random variables we consider, i.e., core variables, observations collected by the different sensors, reliability and presence levels.

At this point, we can evaluate the risk associated to an intrusion, by simply updating the probabilities for the four possible values of the *risk factor*, conditional on the values of the observations returned by the sensors

and on the levels of reliability and presence observed by the Expert.

As a preliminary test of the model, we have considered a simulated scenario of a single object in the Swiss Alps, like for example a dam, surveyed by an identification architecture that is characterized by the absence of interceptors and by a relatively good coverage of all the other sensors. We assumed as meteorological conditions discontinuous low clouds and daylight. The simulated scenario reproduces a situation where an agent provocateur is flying very low with a helicopter and without emitting any identification code. The decision maker is assumed to have uniform prior beliefs about the four classes of risk.

The size of the network suggests the opportunity of an approximate approach to this updating problem. In our approach, we have first reformulated our model as a *locally specified* credal network, according to the procedure developed in [1]. Then, we have transformed each non-binary variable of the credal network into a set of binary variables, according to the *binarization* algorithm, reported in [2]. The resulting credal net has been finally updated by the *loopy* version of the 2U algorithm (L2U) [7]. The overall procedure, which can be proved to be approximate only because of the L2U algorithm, can be implemented in polynomial time. In our case, the credal network has been updated in few seconds on a 2.8 GHz Pentium 4 machine, and convergence of L2U has been observed after seven iterations.

Figure 7.a depicts the posterior probability intervals for this simulated scenario. The upper probability for the outcome *renegade* is zero, and we can therefore exclude a terrorist attack. Similarly, the lower probability for the outcomes *agent provocateur* and *damaged intruder* are strictly greater than the upper probability for the state *erroneous*, and we can reject also this latter value because of interval dominance. Both these results are reasonable estimates for this simulated scenario.

Remarkably, the indecision between *agent provocateur* and *damaged intruder* disappears as we assume higher levels of reliability and presence for the sensors devoted to the observation of the *height*. The results, reported in Figure 7.b, state that the intruder is an *agent provocateur*, as we have assumed in the design of this simulation.

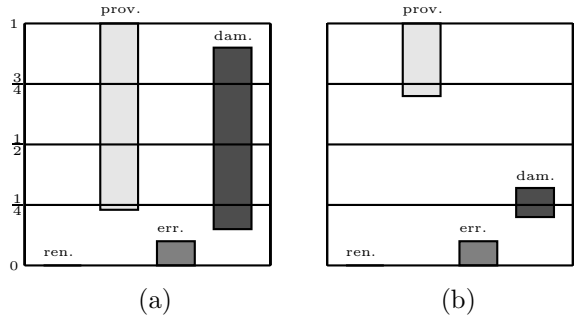


Figure 7: Posterior probability intervals for the risk factor, corresponding to a simulated scenario reproducing a helicopter entering the no-fly zone because of technical difficulties. The histogram bounds denote lower and upper probabilities. The sensors observing the *aircraft height* are assumed more reliable in (b) than in (a).

8 Conclusions and Future Work

A model for determining the risk of intrusion of a civil aircraft into no-fly zone has been presented. The model embeds in a single coherent mathematical framework human expertise expressed by imprecise-probability assessments, and a structure reproducing complex observation mechanisms and corresponding information fusion schemes.

The risk evaluation corresponds to the updating of the probabilities for the risk factor conditional on the observations of the sensors and the estimated levels of presence and reliability. Preliminary tests considered for a simulated scenario are consistent with the judgments of an expert domain for the same situation.

As future work we intend to test the model for other historical cases and simulated scenarios. The approximate updating procedure considered in the present work, as well as other algorithmic approaches will be considered, in order to determine the most suited for this specific problem.

In any case, it seems already possible to offer a practical support to the military experts in their evaluations. They can use the network to decide the risk level corresponding to a real scenario, but it is also possible to simulate situations and verify the effectiveness of the different sensors in order to design an optimal identification architecture.

Finally, we regard our approach to the fusion of the information collected by the different sensors as a sound and flexible approach to this kind of problems, able to work also in situations of contrasting observations between the sensors.

Acknowledgments

This research was partially supported by the Swiss NSF grants 200020-109295/1 and 200021-

113820/1. The Java implementation of the L2U algorithm included in the software tool *2UBayes* (http://www.pmr.poli.usp.br/ltd/) has been used to update the (binary) credal networks. The software tool *lrs* (http://cgm.cs.mcgill.ca/∼avis/C/lrs.htm) has been used to compute the extreme mass function of the conditional credal sets corresponding to the probability intervals provided by the Expert. The authors of this public software tool are gratefully acknowledged.

References

[1] A. Antonucci and M. Zaffalon. Locally specified credal networks. In *Proceedings of the third European Workshop on Probabilistic Graphical Models (PGM-2006)*, pages 25–34, Prague, 2006.

[2] A. Antonucci, M. Zaffalon, J.S. Ide, and F.G. Cozman. Binarization algorithms for approximate updating in credal nets. In IOS Press, editor, *STAIRS'06: Proceedings of the third European Starting AI Researcher Symposium*, pages 120–131, Amsterdam, 2006.

[3] L. Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.

[4] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[5] C. P. de Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1313–1318, Edinburgh, 2005.

[6] E. Demircioglu and L. Osadciw. A Bayesian network sensors manager for heterogeneous radar suites. In *IEEE Radar Conference*, Verona, NY, 2006.

[7] J. S. Ide and F. G. Cozman. IPE and L2U: Approximate algorithms for credal networks. In *Proceedings of the Second Starting AI Researcher Symposium*, pages 118–127, Amsterdam, 2004. IOS Press.

[8] J. S. Ide and F.G. Cozman. Approximate inference in credal networks by variational mean field methods. In F. G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 203–212. SIPTA, 2005.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.

[10] A. Piatti, M. Zaffalon, and A. Antonucci. Auswahl und provisorische spezifizierung der kredalen struktur. Technical report, Armasuisse, 2006.

[11] T. Seidenfeld and L. Wasserman. Dilation for sets of probability. *Annals of Statistics*, 21(3):1139–1154, 1993.

[12] B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7(3):95–120, 1992.

[13] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.

Constructing Predictive Belief Functions from Continuous Sample Data Using Confidence Bands

Astride Aregui
Suez Environnement and
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France

Thierry Denœux
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France

Abstract

We consider the problem of quantifying our belief in future values of a random variable X with unknown distribution P_X, based on the observation of a random sample from the same distribution. The adopted uncertainty representation framework is the Transferable Belief Model, a subjectivist interpretation of belief function theory. In a previous paper, the concept of predictive belief function at a given confidence level was introduced, and it was shown how to build such a function when X is discrete. This work is extended here to the case where X is a continuous random variable, based on step or continuous confidence bands.

Keywords. Dempster-Shafer Theory, Evidence Theory, Transferable Belief Model, p-box, distribution band.

1 Introduction

In the past few years, belief function theory has been developed as a tool for data fusion, but also for the management of uncertainty and various aspects of data mining or decision making. Different interpretations of this theory have been proposed [19]. In this paper, we shall adopt the Transferable Belief Model (TBM) interpretation [21], in which a belief function is considered as representing weighted opinions of an agent regarding some question of interest. This model provides a flexible framework even when the available information (data or expert knowledge) is poor. However, it is not always clear how to construct belief functions for a given problem.

In this paper, we consider the special case where the variable X of interest is defined from the result of a random experiment. It is thus a random variable, with unknown probability distribution P_X. The available information is assumed to consist in past observations collected from n independent repetitions of the same experiment, forming an independent random sample from P_X. Based on this information, we would like to express our beliefs regarding future values to be generated from P_X.

As the probability distribution of X is unknown, the available information is incomplete and the precision of the obtained belief function should depend on the number of observations. In [5], a formalization of this problem was suggested, using the concept of *predictive belief function* (PBF). A PBF was defined as a belief function less committed than P_X with some user-defined probability, and converging in probability towards P_X as the size of the sample tends to infinity. Practical methods for building belief functions were presented for the case where the domain $\mathcal{X}$ of X is discrete, based on multinomial confidence regions.

In this article, the above approach is extended to the case where X is a *continuous random variable*. The extension is based on confidence bands, which play a role similar to that of multinomial confidence regions in the discrete case. When a confidence band is defined by step upper and lower bounding functions, it is known to be equivalent to a belief function on the real line with a finite number of focal intervals. We first show that this belief function is a predictive belief function as defined in [5]. We then consider the generalization to continuous confidence band. In that case, the corresponding belief function is continuous, and we derive the expression of its basic belief density.

The paper is organized as follows. In Section 2, the reader is first reminded with the principles of belief functions theory and of the definition of predictive belief functions as introduced in [5]. The construction of a discrete predictive belief function from a step confidence band is then exposed in Section 3, and the construction of a continuous predictive belief function with a basic belief density from a continous confidence band is described in Section 4. Section 5 concludes the paper.

2 Background on Belief Functions

This section provides a short introduction to the main notions pertaining to the theory of belief functions that will be used throughout the paper, and in particular, its TBM interpretation. We first consider the case of belief functions defined on a finite domain [16], and then address the case of a continuous domain [20]. The concept of predictive belief function as introduced in [5] is then recalled.

2.1 Belief Functions on a Finite Frame

2.1.1 Definition of a Basic Belief Assignment

Let $\mathcal{X} = \{\xi_1, \ldots, \xi_K\}$ be a finite set, and let X be a variable taking values in $\mathcal{X}$. Given some evidential corpus, the knowledge held by a given agent at a given time over the actual value of variable X can be modeled by a so-called *basic belief assignment* (bba) m defined as a mapping from $2^{\mathcal{X}}$ into $[0, 1]$ such that:

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1. \tag{1}$$

Each mass $m(A)$ is interpreted as the part of the agent's belief allocated to the hypothesis that X takes some value in A [16, 21]. The mass $m(\mathcal{X})$ is often regarded as representing a degree of ignorance.

2.1.2 Belief Updating

A fundamental mechanism for belief updating in the TBM is the unnormalized *Dempster's rule of conditioning*, which is defined as follows [21]. Assume that the agent's beliefs about X are represented by a bba m, and the agent learns that the true value of X lies in $B \subseteq \mathcal{X}$. Then, m is transformed into the conditional bba $m[B]$ defined as:

$$m[B](A) = \sum_{C : C \cap B = A} m(C). \tag{2}$$

Upon learning that the truth lies in B, each mass of belief given to C is thus *transferred* to $C \cap B$, hence the term "*Transferable* Belief Model". Equivalent representations of a bba m include the belief, plausibility and commonality functions [16] defined as follows.

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \tag{3}$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{4}$$

and

$$q(A) = \sum_{B \supseteq A} m(B), \tag{5}$$

for all $A \subseteq \mathcal{X}$. In the TBM, $bel(A)$ represents the agent's total degree of belief in A. The plausibility $pl(A) = bel[A](A)$ may be interpreted as the maximal degree of belief that could be given to A after acquiring new information. Similarly, we observe that $q(A) = m[A](A)$. The commonality of A is thus the mass of belief that remains attached to A (i.e., the degree of ignorance) after conditioning by A.

2.1.3 Decision Making

The TBM is a two-level model in which belief representation and updating take place at a first level termed *credal level*, whereas decision making takes place at a second level called *pignistic level* [21]. To make decisions, any bba m such that $m(\emptyset) < 1$ is mapped into a pignistic probability function $Betp$ defined by

$$Betp(x) = \sum_{A \subseteq \mathcal{X}, A \neq \emptyset} \frac{m(A)}{1 - m(\emptyset)} \frac{1_A(x)}{|A|}, \quad \forall x \in \mathcal{X}, \tag{6}$$

where 1_A denotes the indicator function of A. A decision can then be made, based on $Betp$ and on a loss function, just as is done in Bayesian Probability Theory.

2.2 Belief Functions on Real Numbers

Let us now assume that variable X takes values in $\mathcal{X} = \mathbb{R}$. The above formalism can then be extended in at least two different ways.

2.2.1 Discrete Bba on $\mathbb{R}$

In the simplest approach, a bba is defined as above, with the constraint that the set $\mathcal{F}(m) = \{A_1, \ldots, A_n\}$ of focal elements is finite. This will be referred to as a *discrete* bba. Typically, focal elements are chosen among intervals or, more generally, Borel sets [23, 6, 24, 13]. Denoting $m_i = m(A_i)$, with $\sum_{i=1}^{n} m_i = 1$, and assuming $A_i \neq \emptyset$ for all i, Equations (3)-(5) become:

$$bel(A) = \sum_{A_i \subseteq A} m_i, \tag{7}$$

$$pl(A) = \sum_{A_i \cap A \neq \emptyset} m_i, \tag{8}$$

and

$$q(A) = \sum_{A_i \supseteq A} m_i, \tag{9}$$

for all $A \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel sigma-algebra on $\mathbb{R}$.

Equation (6) can be replaced by

$$Betp(x) = \sum_{i=1}^{n} m_i \frac{1_{A_i}(x)}{|A_i|}, \quad \forall x \in \mathbb{R}, \tag{10}$$

where $|A_i|$ now denotes the Lebesgue measure of A_i and we assume that $0 < |A_i| < \infty$ for all i. Equation (10) defines a probability density function [13]. In particular, if the A_is are bounded intervals, *Betp* is a finite mixture of continuous uniform distributions.

2.2.2 Basic Belief Density

A more complex generalization of the finite case is obtained by replacing the concept of bba by that of basic belief density (bbd) [4, 17, 20]. A normal bbd m is a function taking values from the set of closed real intervals into $[0, +\infty)$, such that

$$\iint_{x \leq y} m([x, y])\, dx\, dy = 1. \tag{11}$$

The belief, plausibility and commonality can be defined in the same way as in the finite case, replacing finite sums by integrals. The following definitions hold:

$$bel(A) = \iint_{[x,y] \subseteq A} m([x, y])\, dx\, dy, \tag{12}$$

$$pl(A) = \iint_{[x,y] \cap A \neq \emptyset} m([x, y])\, dx\, dy, \tag{13}$$

$$q(A) = \iint_{[x,y] \supseteq A} m([x, y])\, dx\, dy, \tag{14}$$

for all $A \in \mathcal{B}(\mathbb{R})$. In particular, when $A = [x, y]$,

$$bel([x, y]) = \int_x^y \int_u^y m([u, v]) dv du, \tag{15}$$

$$pl([x, y]) = \int_{-\infty}^{y} \int_{\max(x,u)}^{+\infty} m([u, v]) dv du, \tag{16}$$

$$q([x, y]) = \int_{-\infty}^{x} \int_{y}^{+\infty} m([u, v]) dv du, \tag{17}$$

for all $x \leq y$. The domains of these integrals may be represented as in Figure 1, where each point in the triangle corresponds to an interval with upper and lower bounds indicated on the horizontal and vertical axes, respectively.

Conversely, m may be recovered from *bel* or q as:

$$m([x, y]) = -\frac{\partial^2 bel([x, y])}{\partial x \partial y} = -\frac{\partial^2 q([x, y])}{\partial x \partial y}, \tag{18}$$

provided these derivatives exist.

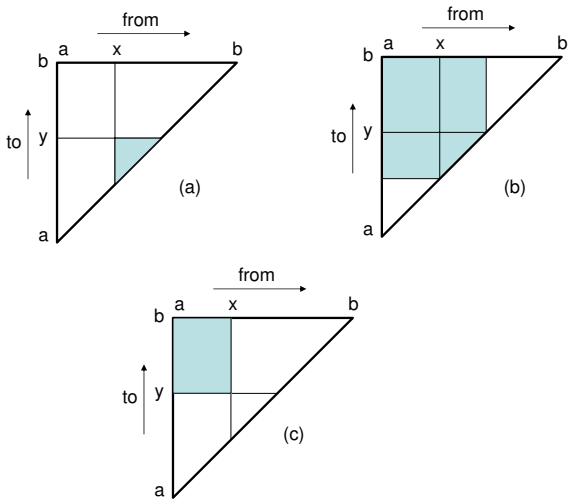


Figure 1: The belief, plausibility and commonality functions are defined as integrals of the bbd with support $[a, b]$ on the shaded area of triangles (a), (b) and (c), respectively.

The pignistic probability density becomes [20]:

$$Betp(x) = \lim_{\epsilon \to 0} \int_{-\infty}^{x} \int_{x+\epsilon}^{+\infty} \frac{m([u, v])}{v - u} dv du. \tag{19}$$

2.3 Predictive Belief Functions

In this section, we summarize the concept of predictive belief function introduced in [5]. Assume X is a random variable with unknown probability distribution P_X, and we have observed a realization $\mathbf{x} = (x_1, \ldots, x_n)$ of an independent and identically distributed (iid) random sample $\mathbf{X} = (X_1, \ldots, X_n)$ with parent distribution P_X. Based on this information, we would like to quantify our beliefs about the next value of X. As a toy example, consider the case where X denotes the color of a ball taken from an urn containing balls of different colors. Having observed the colors of n balls randomly taken from the urn with replacement, we would like to quantify our belief regarding the color of the next ball.

Let $bel(\cdot; \mathbf{X})$ denote a belief function on $\mathcal{X}$ constructed using $\mathbf{X}$. This is a function taking values from a sigma algebra $\mathcal{A}$ into $[0, 1]$. Typically, $\mathcal{A} = 2^{\mathcal{X}}$ if $\mathcal{X}$ is finite, and $\mathcal{A} = \mathcal{B}(\mathbb{R})$ if $\mathcal{X} = \mathbb{R}$ (only these two cases will be considered in this paper). In [5], we postulated that such a belief function should satisfy the following two requirements:

$$\forall A \in \mathcal{A}, \quad bel(A; \mathbf{X}) \xrightarrow{P} P_X(A), \text{ as } n \to \infty, \tag{20}$$

where $\xrightarrow{P}$ denotes convergence in probability, and

$$P\{bel(A; \mathbf{X}) \leq P_X(A), \forall A \in \mathcal{A}\} \geq 1 - \alpha, \tag{21}$$

where $\alpha \in (0,1)$.

Requirement (20) means that $bel(\cdot;\mathbf{X})$ should become closer to P_X as the sample size tends to infinity.

For finite n, $bel(\cdot;\mathbf{X})$ should be less informative than P_X, hence the condition $bel(\cdot;\mathbf{X}) \leq P_X$. However, this condition cannot be satisfied for all realizations of the random sample[1], hence requirement (21), which states that it should be satisfied asymptotically for at least a fraction $1-\alpha$ of the samples.

A belief function $bel(\cdot;\mathbf{X})$ satisfying requirements (20) and (21) is called a *predictive belief function at confidence level* $1-\alpha$. Methods for constructing such belief functions in the case where random variable X is discrete were described in [5], based on multinomial confidence regions.

The construction of predictive belief functions in the continuous case ($\mathcal{X} = \mathbb{R}$) is the main topic of this paper. It will be addressed in the following two sections.

3 Discrete Predictive Belief Functions on $\mathbb{R}$

In this section, the construction of a discrete predictive belief function on $\mathbb{R}$ from a step confidence band is addressed. Basic definitions related to confidence bands are first recalled in Section 3.1, and the construction of Kolmogorov confidence bands is exposed in Section 3.2. In Section 3.3, we show that the discrete belief function with interval focal sets equivalent to a Kolmogorov confidence band is a predictive belief function. The random set interpretation of a p-box is finally recalled in Section 3.4, as a way to introduce the continuous generalization presented in the next section.

3.1 Confidence Bands: Definitions

Let us assume that we have a random variable X with cumulative distribution function (cdf) F_X. In some cases, F_X is not precisely known, but we can specify a lower bounding function $\underline{F} : \mathbb{R} \to [0,1]$ and an upper bounding function $\overline{F} : \mathbb{R} \to [0,1]$ such that $\underline{F}(x) \leq F_X(x) \leq \overline{F}(x)$ for all $x \in \mathbb{R}$. The convex set of probabilities compatible with these constraints

$$\Gamma_X(\underline{F},\overline{F}) = \{P | \forall x \in \mathbb{R}, \underline{F}(x) \leq P((-\infty,x]) \leq \overline{F}(x)\}$$

is called a *distribution band* [11].

In the special case where $\underline{F}$ and $\overline{F}$ are step functions, then $\Gamma_X(\underline{F},\overline{F})$ is called a *probability box*[2], or p-box for short [6]. A continuous distribution bound can always be enclosed in a p-box. The smallest discrete approximation is always obtained by choosing the lower and upper bounding step functions to be right and left-continous, respectively [6]. From now on, only p-boxes possessing this property will be considered.

Suppose now that the available information about F_X takes the form of an iid random sample $\mathbf{X} = (X_1,\ldots,X_n)$ with parent distribution F_X. Let $\underline{F}(\cdot;\mathbf{X})$ and $\overline{F}(\cdot;\mathbf{X})$ be two functions computed from $\mathbf{X}$ and such that $\underline{F}(\cdot;\mathbf{X}) \leq \overline{F}(\cdot;\mathbf{X})$. The distribution band $\Gamma_X(\underline{F}(\cdot;\mathbf{X}),\overline{F}(\cdot;\mathbf{X}))$ is called a *confidence band at level* $\alpha \in (0,1)$ [12, page 334] iff

$$P\left\{\underline{F}(x;\mathbf{X}) \leq F_X(x) \leq \overline{F}(x;\mathbf{X}),\ \forall x \in \mathbb{R}\right\} = 1-\alpha,$$

or, equivalently:

$$P\left\{P_X \in \Gamma_X\left(\underline{F}(\cdot;\mathbf{X}),\overline{F}(\cdot;\mathbf{X})\right)\right\} = 1-\alpha.$$

Note that, in the above equalities, F_X and P_X are fixed unknown functions, whereas $\underline{F}(\cdot;\mathbf{X})$ and $\overline{F}(\cdot;\mathbf{X})$ depend on random sample $\mathbf{X}$.

3.2 Kolmogorov Confidence Bands

Let us assume that X is a continuous random variable. The simplest way to obtain a confidence band for F_X is to use Kolmogorov's statistic

$$D_n = \sup_x |S_n(x;\mathbf{X}) - F_X(x)|,$$

where $S_n(\cdot;\mathbf{X})$ is the sample distribution function defined by

$$S_n(x;\mathbf{X}) = \begin{cases} 0, & x < X_{(1)} \\ k/n, & X_{(k)} \leq x < X_{(k+1)} \\ 1, & X_{(n)} \leq x, \end{cases} \qquad (22)$$

for all $x \in \mathbb{R}$, where $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denote the observations sorted in increasing order.

The distribution of D_n does not depend on F_X. It was computed for fixed n by Kolmogorov [10], who also computed the asymptotic distribution of D_n. Let $d_{n,\alpha}$ denote the critical value of D_n defined as $P(D_n > d_{n,\alpha}) = \alpha$. Thus,

$$P\{\ S_n(x;\mathbf{X}) - d_{n,\alpha} \leq F_X(x) \leq S_n(x;\mathbf{X}) + d_{n,\alpha}, \forall x \in \mathbb{R}\ \} = 1-\alpha, \quad (23)$$

which implies that $S_n \pm d_{n,\alpha}$ defines a confidence bound at level $1-\alpha$ [9, page 481]. This band may

[1] Indeed, such a requirement would lead to the vacuous belief function.

[2] Ferson *et al.* [6] actually used the term "p-box" as a synonym to "distribution band". However, following Kriegler and Held [11], we prefer to reserve the term "p-box" for the important case where the bounding functions are step functions.

be narrowed by using the inequalities $0 \leq F_X(x) \leq 1$ for all x. Hence, we have:

$$\underline{F}(x;\mathbf{X}) = \max(0, S_n(x;\mathbf{X}) - d_{n,\alpha}), \quad (24)$$

$$\overline{F}(x;\mathbf{X}) = \min(1, S_n(x;\mathbf{X}) + d_{n,\alpha}). \quad (25)$$

If the support of X is bounded and known to be included in $[b, B]$, then the above bounds can be further narrowed.

Note that $S_n(\cdot|\mathbf{X})$ as defined by (22) and, consequently, both $\underline{F}(\cdot;\mathbf{X})$ and $\overline{F}(\cdot;\mathbf{X})$ are right-continuous step functions. However, $\overline{F}(\cdot;\mathbf{X})$ can be replaced by the left-continuous function $\overline{F}'(\cdot;\mathbf{X})$ taking the same values everywhere except at sample points, defined as $\overline{F}'(x;\mathbf{X}) = \lim_{h\to x^-} \overline{F}(h;\mathbf{X})$. The pair $(\underline{F}, \overline{F}')$ still defines a confidence band at level $1-\alpha$, i.e.,

$$P\left\{P_X \in \Gamma_X(\underline{F}, \overline{F}')\right\} = 1 - \alpha. \quad (26)$$

Example 1. *The data reported in [14] consists in the operational lives (in hours) of 20 bearings. These are 2398, 2812, 3113, 3212, 3523, 5236, 6215, 6278, 7725, 8604, 9003, 9350, 9460, 11584, 11825, 12628, 12888, 13431, 14266, 17809. Here, the variable of interest, denoted X (the lifetime of a bearing), has a lower bound $b = 0$ and no upper bound ($B = \infty$). Figure 2 shows the sample cdf of this data, together with the lower and upper bounding functions defining the Kolmogorov confidence band at level $1-\alpha = 0.95$.*

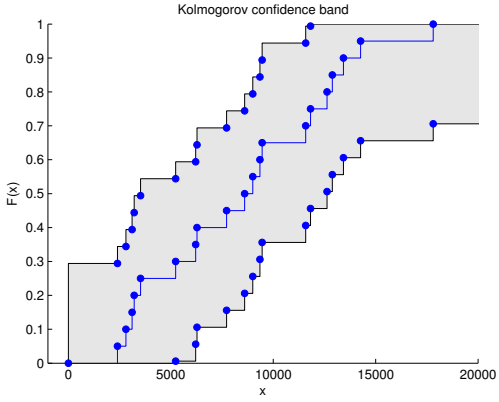


Figure 2: Sample cdf S_n and Kolmogorov confidence band at level $1-\alpha = 0.95$ for the bearings data.

3.3 Predictive Belief Function Induced by a Kolmogorov Confidence Band

The above method for constructing a confidence band yields a pair of lower and upper step functions, i.e., a p-box. The relationship between p-boxes and belief functions has been studied by several authors [23, 6, 22]. Recently, the exact correspondance between p-boxes with bounded support and discrete belief functions was proved by Kriegler and Held [11], who also proposed an algorithm for the rigorous construction of a discrete mass function m on $\mathbb{R}$ equivalent to a p-box.

The principle of this construction is illustrated in Figure 3. The lower and upper bounding functions are assumed to be right and left continuous, respectively. Each rectangle A_i in this figure corresponds to a focal interval $[a_i, b_i)$, with mass $m([a_i, b_i)) = d_i - c_i$.

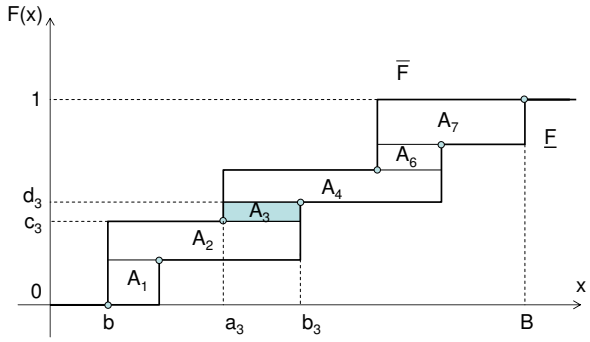


Figure 3: Principle of the construction of a basic belief assignment m from a p-box $(\underline{F}, \overline{F})$. Each rectangle A_i in the area between the lower and upper bounding functions corresponds a focal interval $[a_i, b_i)$ of m, with mass $d_i - c_i$.

Let $\Gamma_X(bel)$ denote the set of probability measures compatible with bel, the belief function induced by m, i.e.,

$$\Gamma_X(bel) = \{P | bel(A) \leq P(A), \forall A \in \mathcal{B}(\mathbb{R})\}.$$

Kriegler and Held [11] proved that $(\underline{F}, \overline{F})$ and bel are two equivalent representations of a unique family of probabilities, i.e.,

$$\Gamma_X(bel) = \Gamma_X(\underline{F}, \overline{F}). \quad (27)$$

If bel and pl denote the corresponding belief and plausibility functions, and if $\underline{P}$ and $\overline{P}$ denote the lower and upper envelopes of $\Gamma_X(\underline{F}, \overline{F})$, we have $bel = \underline{P}$ and $pl = \overline{P}$. In particular, $bel((-\infty, x]) = \underline{F}(x)$ and $pl((-\infty, x]) = \overline{F}(x)$ for all $x \in \mathbb{R}$.

Note that, although Kriegler and Held only considered the case of p-boxes with bounded support, their algorithm and result may be applied directly to the case of p-boxes with unbounded support.

Let us now consider the case where $\underline{F}$ and $\overline{F}$ are the lower and upper bounding functions of Kolmogorov confidence band at level $1-\alpha$, as defined by (24)-(25). Let $bel(\cdot;\mathbf{X})$ denote the belief function on $\mathbb{R}$ con-

structed from p-box $(\underline{F}, \overline{F}')$ using Kriegler and Held's algorithm. The following proposition holds.

Proposition 1. *$bel(\cdot;\mathbf{X})$ is a predictive belief function at level $1-\alpha$.*

Sketch of proof. First, requirement (21) is obviously satisfied as a direct consequence of (26) and (27): since $\Gamma_X(bel(\cdot;\mathbf{X})) = \Gamma_X(\underline{F}, \overline{F}')$, we have

$$P\{bel(A;\mathbf{X}) \leq P_X(A), \forall A \in \mathcal{A}\} = P\{P_X \in \Gamma_X(bel(\cdot;\mathbf{X}))\} = 1-\alpha.$$

Moreover, given that $\underline{F}(x) \xrightarrow{P} F_X(x)$ and $\overline{F}'(x) \xrightarrow{P} F_X(x)$ for all $x \in \mathbb{R}$, it can easily be shown that $bel(A;\mathbf{X}) \xrightarrow{P} P_X(A)$ for all interval A. Lastly, for any $B = \bigcup_{i\in I} A_i$ where $(A_i)_{i\in I}$ with $I \in \mathbb{N}$ is a countable family of intervals, we have

$$bel(B;\mathbf{X}) = \sum_{i\in I} bel(A_i;\mathbf{X}) \xrightarrow{P} \sum_{i\in I} P_X(A_i) = P_X(B),$$

which proves that requirement (20) is satisfied, and completes the proof. □

Example 2. *To illustrate the construction of a predictive belief function from a Kolmogorov confidence band, let us consider again the data of Example 1. Based on this data, we would like to express our beliefs regarding the lifetime X of a new bearing taken randomly from the same population. For commodity of representation, let us adopt the reasonable assumption that X has an upper bound, which will arbitrarily be set to 30000, so that the support of X is assumed to be $[0, 30000]$.*

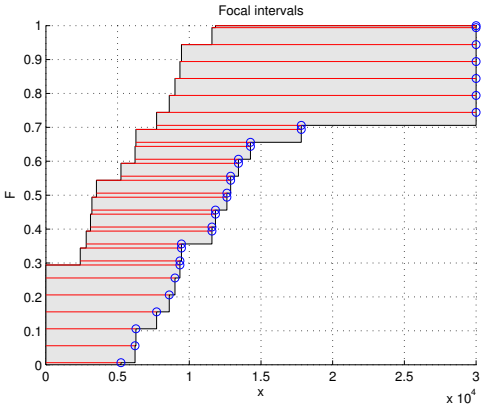


Figure 4: Focals intervals of the PBF constructed from the Kolmogorov confidence band at level $1-\alpha = 0.95$ (bearings data). The height of each segment representing a focal interval is equal to the cumulated mass.

The focal intervals of the corresponding PBF $bel(\cdot;\mathbf{X})$ are displayed in Figure 4. Figures 5 and 6 are examples of graphical displays that reveal different aspects of the information contained in the belief function $bel(\cdot;\mathbf{X})$. Figure 5 shows the plausibility profile function $x \to pl(\{x\};\mathbf{X})$ and the pignistic probability density function Betp computed from (6), which are two left-continuous real-valued step functions with simple interpretation. Figure 6 shows grey level representations of $bel([x,y];\mathbf{X})$, $pl([x,y];\mathbf{X})$ and $q([x,y];\mathbf{X})$ as two-dimensional functions of (x,y).

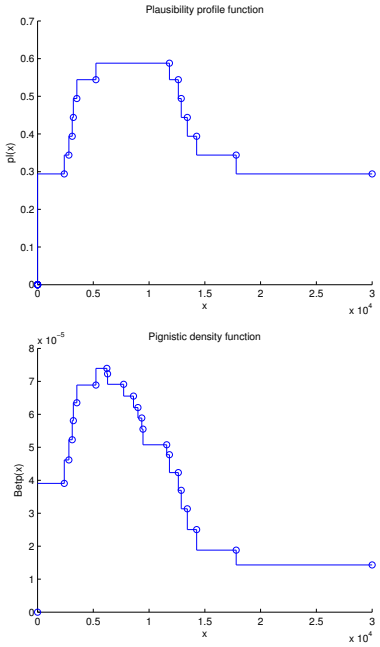


Figure 5: Plausibility profile function (up) and pignistic probability density function (down) of the discrete PBF constructed from the Kolmogorov confidence band (Bearings data).

3.4 Random Set Interpretation

The bba m associated to a p-box $(\underline{F}, \overline{F})$ may also be shown to correspond formally to a random set [1]. Let $\underline{F}^{-1}$ and $\overline{F}^{-1}$ be the pseudo-inverses of $\underline{F}$ and $\overline{F}$, defined, respectively, as:

$$\underline{F}^{-1}(\alpha) = \inf\{x \in \mathbb{R}, \underline{F}(x) \geq \alpha\},$$

$$\overline{F}^{-1}(\alpha) = \inf\{x \in \mathbb{R}, \overline{F}(x) \geq \alpha\},$$

for all $\alpha \in [0,1]$. Let us consider the mapping ρ from $[0,1]$ to the set of real intervals, such that $\rho(\alpha) = (\underline{F}^{-1}(\alpha), \overline{F}^{-1}(\alpha)]$, and let us consider the uniform probability distribution P_U on $[0,1]$. Then ρ is a random set, and it is formally equivalent to m. Let $\mathcal{F} = \{(\underline{F}^{-1}(\alpha), \overline{F}^{-1}(\alpha)], \alpha \in [0,1]\}$. For all $A \in \mathcal{F}$, we have

$$m(A) = P_U(\rho^{-1}(A)).$$

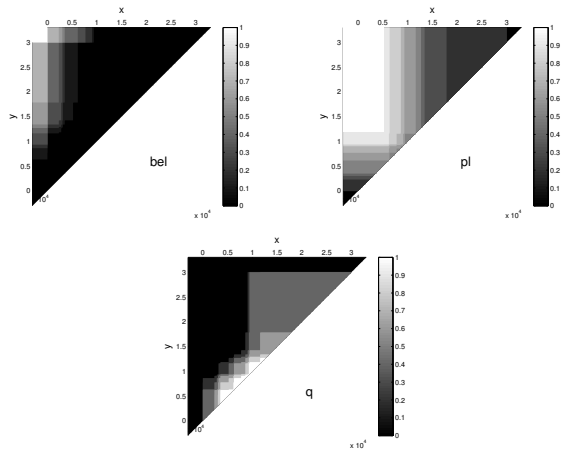


Figure 6: Contour plots of functions $bel[\mathbf{X}]([x,y])$, $pl[\mathbf{X}]([x,y])$ and $q[\mathbf{X}]([x,y])$ constructed from Kolomogorov's confidence band (Bearings data).

Note that the uniform probability distribution on $[0,1]$ and the mapping ρ are only considered here as mathematical constructs. In the TBM, only belief functions have an interpretation, and an underlying multi-valued mapping is not assumed. However, the random set point of view will guide us in the following section to propose a generalization of the above results in the case of continuous distribution bands.

4 Continuous Predictive Belief Functions on $\mathbb{R}$

Kolmogorov's confidence bands have the advantage of being exact and non parametric. However, they have a constant vertical width, which makes them unnecessarily broad in the tails. As a result, the equivalent belief functions may be excessively imprecise. Narrower confidence bands can be computed using parametric methods, but they are defined by continuous bounding functions. The usual approach to continuous distribution bands is to approximate them using a p-box [6]. Here, we show that this approximation can be avoided, and a continuous predictive belief function on $\mathbb{R}$ can be constructed from a continuous confidence band, thus providing an extension to the results presented in the previous section. Parametric confidence bands are first briefly reviewed in the following section.

4.1 Parametric Confidence Bands

Methods for the construction of continuous confidence bands as described above were proposed by several authors, including Kanofsky and Srinivasan [8] and Cheng and Iles [3]. In the sequel, Cheng and Iles' method, which will be used later to demonstrate the main findings of this paper, will briefly be recalled.

Let us assume that X is a continuous random variable with cdf $F_X(x,\theta)$, where θ is vector of r unknown parameters. Cheng and Iles' approach consists in determining lower and upper bounds of the cdf when θ varies in a confidence region R. This confidence region is built from the statistics

$$Q(\theta) = (\widehat{\theta} - \theta)^T I(\theta)(\widehat{\theta} - \theta),$$

where $\widehat{\theta}$ is the maximum likelihood estimate of θ, and $I(\theta)$ is the Fisher information matrix. It is known that $Q(\theta)$ is asymptotically a chi-squared variable with r degrees of freedom. In [3], Cheng and Iles apply their method in the case of a general location-scale parametric model of the form:

$$F_X(x) = G\left(\frac{x-\mu}{\sigma}\right),$$

where G is a fixed distribution function, and μ and σ are the unknown location and scale parameters. In that case the Fisher information matrix is of the form

$$I(\mu,\sigma) = \frac{n}{\sigma^2}\begin{pmatrix} k_0 & -k_1 \\ -k_1 & k_2 \end{pmatrix},$$

where k_0, k_1 and k_2 are constants independent of μ and σ. The bounds of the confidance band then have the following expressions:

$$\overline{F}(x) = G(\xi + h), \tag{28}$$
$$\underline{F}(x) = G(\xi - h), \tag{29}$$

where $\xi = (x - \widehat{\mu})/\widehat{\sigma}$, $\widehat{\mu}$ and $\widehat{\sigma}$ are the maximum likelihood estimates of μ and σ, and

$$h = \sqrt{\frac{\gamma}{n\,k_0}\left(1 + \frac{(k_0\xi + k_1)^2}{k_0 k_2 - k_1^2}\right)}. \tag{30}$$

Coefficient γ is the value for which $P(Q(\mu,\sigma) \leq \gamma) = 1-\alpha$. It can be approximated by the chi-squared quantile $\chi^2_2(\alpha)$. Cheng and Iles [3] demonstrate the application of these formula for the cases of the normal, lognormal, extreme-value (log-Weibull) and Weibull distributions. In the case of the normal distribution, $k_0 = 1$, $k_1 = 0$, and $k_2 = 2$.

4.2 PBF Induced by a Continuous Confidence Band

Let $(\underline{F}, \overline{F})$ be a continuous distribution band for some continuous random variable X, and assume that the lower and upper bounding functions $\underline{F}$ and $\overline{F}$ are strictly increasing. Consider the mapping ρ from $[0,1]$ to the set of real intervals, such that $\rho(\alpha) =$

$[\underline{F}^{-1}(\alpha), \overline{F}^{-1}(\alpha)]$, where $\underline{F}^{-1}$ and $\overline{F}^{-1}$ are the inverses of $\underline{F}$ and $\overline{F}$, respectively. If the $[0,1]$ interval is endowed with a uniform probability distribution, then mapping ρ defines a random set, which corresponds to a continuous belief function *bel* on $\mathbb{R}$ as described in Section 2.2.2.

This belief function is such that $bel([x,y]) = \underline{P}([x,y])$ for all $x \leq y$, $\underline{P}$ being the lower envelope of the distribution band. In particular, we have $bel((-\infty, x]) = \underline{F}(x)$ and $pl((-\infty, x]) = \overline{F}(x)$, for all $x \in \mathbb{R}$. As we are working within the TBM, this random set is for us a purely mathematical construct, and we would like to express *bel* directly through its bbd $m([x,y])$, $x \leq y$. This can be achieved using (18). The following proposition holds.

Proposition 2. *The bbd associated to a continuous distribution band $(\underline{F}, \overline{F})$ is defined by*

$$m([x,y]) = -\frac{\partial^2 bel([x,y])}{\partial x \partial y},$$

with:

$$\begin{aligned}
\frac{\partial^2 bel([x,y])}{\partial x \partial y} &= -\overline{f}(x)\underline{f}(y)\delta(\underline{F}(y) - \overline{F}(x)), && (31)\\
&= -\overline{f}(x)\delta(y - \underline{F}^{-1} \circ \overline{F}(x)), && (32)\\
&= -\underline{f}(y)\delta(x - \overline{F}^{-1} \circ \underline{F}(y)), && (33)
\end{aligned}$$

where $\underline{f}$ and $\overline{f}$ are the first derivatives of $\underline{F}$ and $\overline{F}$, respectively, and δ is the Dirac delta function.

Proof. We have

$$\begin{aligned}
bel([x,y]) &= \underline{P}([x,y]) && (34)\\
&= \max(0, \underline{F}(y) - \overline{F}(x)) && (35)\\
&= (\underline{F}(y) - \overline{F}(x))H(\underline{F}(y) - \overline{F}(x)) && (36)
\end{aligned}$$

where H is the Heaviside function. By differentiating with respect to x and y, we get:

$$\begin{aligned}
\frac{\partial^2 bel([x,y])}{\partial x \partial y} = -\overline{f}(x) \big(&\delta(\underline{F}(y) - \overline{F}(x))\underline{f}(y)\\
&+ \underline{f}(y)\delta(\underline{F}(y) - \overline{F}(x))\\
&+(\underline{F}(y) - \overline{F}(x))\delta'(\underline{F}(y) - \overline{F}(x))\underline{f}(y)\big). \quad (37)
\end{aligned}$$

Now, from the property of the delta function: $x\delta'(x) = -\delta(x)$, we have:

$$(\underline{F}(y) - \overline{F}(x))\delta'(\underline{F}(y) - \overline{F}(x)) = -\delta(\underline{F}(y) - \overline{F}(x)).$$

Hence, (37) is equivalent to (31).

In order to prove that (32) and (33) can be deduced from (31), the following property of the delta function can be used. For all function g,

$$\delta\left(g(x)\right) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|},$$

where the x_i are the roots of g. For fixed x, $\underline{F}(y) - \overline{F}(x)$ is a function of y with a unique root $\underline{F}^{-1} \circ \overline{F}(x)$. Hence,

$$\overline{f}(x)\underline{f}(y)\delta(\underline{F}(y) - \overline{F}(x)) = \overline{f}(x)\underline{f}(y)\frac{\delta(y - \underline{F}^{-1} \circ \overline{F}(x))}{\underline{f}(\underline{F}^{-1} \circ \overline{F}(x))} \quad (38)$$

The left-hand side of (38) is equal to 0 if $y \neq \underline{F}^{-1} \circ \overline{F}(x)$, and $\overline{f}(x)\delta(y - \underline{F}^{-1} \circ \overline{F}(x))$ otherwise. Consequently,

$$\overline{f}(x)\underline{f}(y)\delta(\underline{F}(y) - \overline{F}(x)) = \overline{f}(x)\delta(y - \underline{F}^{-1} \circ \overline{F}(x)).$$

Equation (33) can be deduced from (31) in a similar way, by fixing y and treating $\underline{F}(y) - \overline{F}(x)$ as a function of x. □

It can be checked that (35) may be recovered from $m([x,y])$ using (15). Similarly, the expressions of $pl([x,y])$ and $q([x,y])$ can be obtained from $m([x,y])$ using (16) and (17). The following proposition holds.

Proposition 3. *Let m be the bbd associated to a continuous distribution band $(\underline{F}, \overline{F})$. The plausibility and the commonality of any real interval $[x,y]$ are given by:*

$$pl([x,y]) = \overline{F}(y) - \underline{F}(x), \quad (39)$$

$$q([x,y]) = \max(0, \overline{F}(x) - \underline{F}(y)). \quad (40)$$

Proof. Let us prove (40). We have

$$\begin{aligned}
q([x,y]) &= \int_{-\infty}^{x}\int_{y}^{+\infty} m([u,v])dvdu\\
&= \int_{-\infty}^{x} \overline{f}(u)I(u)du,
\end{aligned}$$

with

$$I(u) = \int_y^{+\infty} \delta\left(v - \underline{F}^{-1} \circ \overline{F}(u)\right) dv.$$

Now, $I(u) = 1$ if $\underline{F}^{-1} \circ \overline{F}(u) \geq y$, i.e., if $u \geq \overline{F}^{-1} \circ \underline{F}(y)$, and 0 otherwise. Hence $q([x,y]) = 0$ if $\overline{F}^{-1} \circ \underline{F}(y) \geq x$, i.e., if $\underline{F}(y) \geq \overline{F}(x)$; otherwise,

$$q([x,y]) = \int_{\overline{F}^{-1} \circ \underline{F}(y)}^{x} \overline{f}(u)du = \overline{F}(x) - \underline{F}(y).$$

The proof of (39) is similar. □

Finally, the expression of the pignistic probability density associated to bbd m is given by the following proposition.

Proposition 4. *Let m be the bbd associated to a continuous distribution band $(\underline{F}, \overline{F})$. The associated pignistic probability density Betp is given by*

$$Betp(x) = \int_{\overline{F}^{-1} \circ \underline{F}(x)}^{x} \frac{\overline{f}(u)}{\underline{F}^{-1} \circ \overline{F}(u) - u}du.$$

Proof. From (19), we get

$$Betp(x) = \lim_{\epsilon \to 0} \int_{-\infty}^{x} J(u)du,$$

with

$$\begin{aligned} J(u) &= \overline{f}(u) \int_{x+\epsilon}^{+\infty} \frac{\delta\left(v - \underline{F}^{-1} \circ \overline{F}(u)\right)}{v-u} dv \\ &= \begin{cases} \frac{\overline{f}(u)}{\underline{F}^{-1} \circ \overline{F}(u) - u} & \text{if } \underline{F}^{-1} \circ \overline{F}(u) \geq x + \epsilon \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The condition $\underline{F}^{-1} \circ \overline{F}(u) \geq x + \epsilon$ can be expressed as $u \geq \overline{F}^{-1} \circ \underline{F}(x + \epsilon)$, hence

$$\begin{aligned} Betp(x) &= \lim_{\epsilon \to 0} \int_{\overline{F}^{-1} \circ \underline{F}(x+\epsilon)}^{x} \frac{\overline{f}(u)}{\underline{F}^{-1} \circ \overline{F}(u) - u} du \\ &= \int_{\overline{F}^{-1} \circ \underline{F}(x)}^{x} \frac{\overline{f}(u)}{\underline{F}^{-1} \circ \overline{F}(u) - u} du. \end{aligned}$$

□

The above results are valid for any continuous distribution band $(\underline{F}, \overline{F})$. When $(\underline{F}, \overline{F})$ is a confidence band at level $1 - \alpha$, then it is easy to see, using the same line of reasoning as in Section 3.3, that the corresponing belief function is a predictive belief function at level $1 - \alpha$.

Example 3. *This method for computing a continuous predictive belief function was applied to the bearings data of examples 1 and 2 As in [3], we assumed these data have a lognormal distribution. Figure 7 shows the 95 % confidence band and the estimated cdf. The plausibility profile function* $x \to pl(\{x\}; \mathbf{X})$ *is shown in Figure 8, and contour plots of* $bel([x, y]; \mathbf{X})$, $pl([x, y]; \mathbf{X})$ *and* $q([x, y]; \mathbf{X})$ *are shown in Figure 9. These figures should be compared to Figures 2, 5 and 6, respectively.*

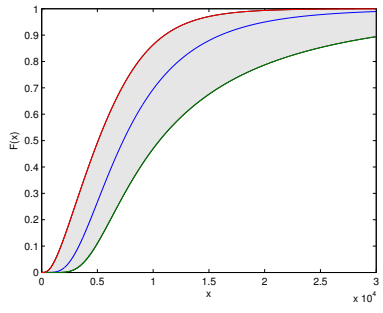


Figure 7: Continuous confidence band and cumulative density function estimated through Cheng and Iles' algorithm.

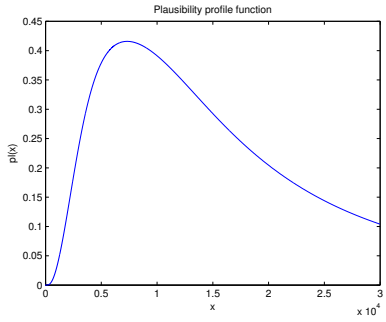


Figure 8: Plausibility profile function obtained from the continuous confidence band of Figure 7.

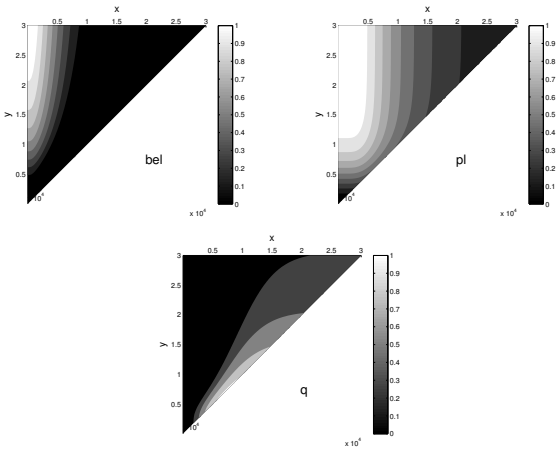


Figure 9: Contour plots of functions $bel([x, y]; \mathbf{X})$, $pl([x, y]; \mathbf{X})$ and $q([x, y]; \mathbf{X})$ constructed from Cheng and Iles' confidence band.

5 Conclusion

We have addressed the problem of constructing predictive belief functions as defined in [5], in the case where the random variable X is continuous. We have shown that such belief functions can be constructed from confidence bands, which play the same role as multinomial intervals in [5]. The methods yields a discrete BF with a finite number of interval focal sets when applied to a Kolomogov confidence band, and a basic belief density as studied in [20] when applied to a continuous parametric confidence band. These belief functions are interpreted as quantifying our belief in a future realization of X, based on a realization of a random sample from the same distribution.

An application of these results to novelty detection is described in [2]. Assume that we have defined a novelty measure T using, e.g., one-class support vector machines [15] or kernel principal component analysis [7]. Based on observations $T_1, \ldots, T_n$ of T for data recorded while the system under study was in the

normal state ω_0, we may compute a predictive belief function on T, *given that the system is in the normal state*. Using the General Bayesian Theorem [18] with some assumptions, it is then possible to build a belief function on $\Omega = \{\omega_0, \overline{\omega}_0\}$ (where $\overline{\omega}_0$ denotes the hypothesis that the system is not in the normal state), given T. This belief function may be combined with other information or used for decision making.

References

[1] D. A. Alvarez. On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43(3):241–267, 2006.

[2] A. Aregui and T. Denœux. Fusion of one-class classifiers in the belief function framework. In *Submitted to the 10th Int. Conf. on Information Fusion*, Quebec, Canada, July 2007.

[3] R. C. H. Cheng and T. C. Iles. Confidence bands for cumulative distribution functions of continuous random variables. *Technometrics*, 25(1):77–86, 1983.

[4] A. P. Dempster. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 39(3):957–966, 1968.

[5] T. Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.

[6] S. Ferson, V. Kreinovitch, L. Ginzburg, D. S. Myers, and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Technical Report SAND2002-4015, Sandia National Laboratories, Albuquerque, NM, 2003.

[7] Heiko Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40:863–874, March 2007.

[8] P. Kanofsky and R. Srinivasan. An approach to the construction of parametric confidence bands on cumulative distribution functions. *Biometrica*, 59(3):623–631, 1972.

[9] M. Kendall and A. Stuart. *The advanced theory of statistics*, volume 2. Charles Griffin and Co Ltd, London, fourth edition, 1979.

[10] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Istituto Italiano degli Attuari*, 4:83–91, 1933.

[11] E. Kriegler and H. Held. Utilizing belief functions for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39(2–3):185–209, 2005.

[12] E. L. Lehman. *Testing statistical hypotheses*. Springer-Verlag, New-York, 2nd edition, 1986.

[13] S. Petit-Renaud and T. Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004.

[14] R. E. Schafer and J. J. Angus. Estimation of Weibull quantiles with minimum error in the distribution function. *Technometrics*, 21:367–370, 1979.

[15] B. Schölkopf and A.J. Smola. *Learning with kernels*. MIT Press, 2002.

[16] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[17] Ph. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical*. PhD thesis, Université Libre de Bruxelles, Brussels, Belgium, 1978. (in French).

[18] Ph. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[19] Ph. Smets. What is Dempster-Shafer's model ? In R. R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 5–34. Wiley, New-York, 1994.

[20] Ph. Smets. Belief functions on real numbers. *International Journal of Approximate Reasoning*, 40(3):181–223, 2005.

[21] Ph. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[22] F. Tonon. Using random set theory to propagate epistemic uncertainty through a mechanical system. *Reliability Engineering & System Safety*, 85(1–3):169–181, 2004.

[23] R. R. Yager. Arithmetic and other operations on Dempster-Shafer structures. *Int. J. Man-Machines Studies*, 25:357–366, 1986.

[24] R. R. Yager. Cumulative distribution functions from Dempster-Shafer belief structures. *IEEE Trans. on Systems, Man and Cybernetics B*, 34(5):2080–2087, 2004.

Uncertainty analysis in food engineering involving imprecision and randomness

Cédric Baudrit cbaudrit@grignon.inra.fr
Arnaud Hélias arnaud.helias@grignon.inra.fr
Nathalie Perrot nathalie.perrot@grignon.inra.fr
UMR782 Génie et Microbiologie des Procédés Alimentaires
INRA, AgroParisTech
F-78850 Thiverval-Grignon, France

Abstract

During the cheese ripening, airflow pattern and climatic conditions inside cheese-ripening rooms are determinant for cheese weight losses. Due to the variation of air velocity inside ripening chambers, homogeneity in the distribution of climatic conditions is very hard to achieve at every single point of it. We are hence faced with imprecise and incomplete knowledge. In practice, it is common that some model parameters may be represented by single probability distributions, justified by substantial data, while others are more faithfully represented by possibility distributions due to the partial nature of available knowledge. This paper applies recent methods, designed for the joint propagation of variability and imprecision, to a cheese ripening mass loss model. Joint propagation methods provide lower & upper probability bounds of exceeding a certain value of cheese mass losses.

Keywords. Imprecise probabilities, p-boxes, belief functions, possibility, food processing, cheese ripening.

1 Introduction

In the food industry, end-products must achieve a compromise between several properties, including sensory, sanitary, technological properties. Among the latter, sensory and sanitary properties are essential because they influence consumer choice and preference. Nevertheless, managing these properties right from the fabrication stage with the aim of controlling them is no easy task ([23, 24]). One of the key reasons of this difficulty is the uncertainty that should be managed at different levels:

- Uncertainty (more specifically imprecision) on the measurements, especially the measurements of the sensory properties [15]. It is obvious and accepted that there is a lack of efficient sensors, and that existing sensors often provide incomplete information for taking action decisions on the process [17]. Moreover, when adequate sensors exist, the configurations of industrial processes do not often allow an efficient placement.
- Uncertainty on the phenomenon involved, even for control purposes. As a consequence the management of uncertainty on the parameters and also the structure of the models built are crucial [16].

Few contributions about this topic are available. Among them Davidson et al. [5] used a fuzzy arithmetic that estimates peanut eating time and browning to control peanut roasting. Perrot et al. [24] developed a decision help system to control the cheese ripening process, integrating the uncertainty of human measurements. Petermeier et al. [25] used a hybrid approach to develop a model of the fouling behavior of an arbitrary heat treatment device for milk. This is developed by combining deterministic differential equations with cognitive elements for the unknown parts of the knowledge model. These authors emphasize the relevance of this open field of research in the context of food processes and the interest of fuzzy symbolic representation of expert reasoning. Nevertheless, they call into question the optimality of the approaches developed on the basis of imperfect and incomplete expert knowledge.

The ripening process is one most important step for many cheese makers. Microbial activities, responsible for the organoleptic characteristics of cheeses, are influenced by climatic conditions (air temperature and relative humidity, gas concentration). So, controlling these climatic conditions inside cheese-ripening rooms is of paramount importance. Cheese mass loss dynamic is a key point in ripening process, with consequences on productivity and it introduces a risk that resulting product may be dropped in status (e.g., the Camembert-Normandie protected designation of origin requires a final weight of 0.25 kg).

Ventilation is used to evacuate heat and humidity generated by cheeses and the spatial distribution of climatic conditions inside cheese-ripening rooms is dependent on airflow (air velocity, air change rate). Nevertheless, only a few studies on interaction between climatic conditions and

airflow can be found in the literature due to confidentiality conditions. The distribution of climatic conditions is very hard to achieve at every single point of ripening chambers. In a industrial context, computational fluid dynamic model of ripening rooms [20] can not be carried out without an exhaustive room description. However, it is inconceivable to install sensors everywhere inside ripening chambers to pick up for instance temperature and relative humidity. Hence, we are faced with imprecise knowledge relative to the spatial variability of climatic conditions. Heat and mass transfert are a well studied phenomenon in cooking or drying process. However, little data have been published for the cheese ripening and transfert coefficients between cheese and atmosphere are not precisely described.

It is more and more acknowledged that uncertainty regarding model parameters has essentially two origins [12]. It may arise from randomness (often referred to as "stochastic uncertainty") due to natural variability of observations resulting from heterogeneity or the fluctuations of a quantity in time. Or it may be caused by imprecision (often referred to as "epistemic uncertainty") due to a lack of information. This lack of knowledge may stem from a partial lack of data, either this data is impossible to collect, or because only experts can provide some imprecise information. For example, it can be quite common for an expert to estimate numerical values of parameters in the form of confidence intervals according to his/her experience and intuition. The uncertainty pervading model parameters is not of a single nature, namely, randomness and incomplete knowledge may coexist, especially due to the presence of several, heterogeneous sources of knowledge, as for instance statistical data and expert opinions. The most general setting to recognize incompleteness as a feature distinct from randomness is the one of imprecise probabilities developed at length by Peter Walley [29]. In this theory, sets of probability distributions capture the notion of partial lack of probabilistic information. In practice, while information regarding variability is best conveyed using probability distributions, information regarding imprecision is more faithfully conveyed using families of probability distributions. Probability boxes [11] or possibility distributions (also called fuzzy intervals) [9] or yet belief function introduced by Dempster [7] (and elaborated further by Shafer [27] and Smets [28] in a different context) allow to encode such families. Most researchers typically use either one or the other of these modes of uncertainty representation [5, 23, 24, 25]. But to date, such combinations of these different modes of representation have never been applied to food processing.

In Section 2, we recall basic concepts of probability-boxes, numerical possibility theory and belief function in connection with imprecise probabilities. In Section 3, we present methods for propagating objective (variability) and subjective (imprecision) information through multivariate function. We also present post-processing to estimate confidence intervals and/or the probability of exceeding a threshold. In Section 4, we give an overview of a simplified cheese mass loss dynamic model with available knowledge associated to model inputs and their representation. Lastly, in Section 5, we process uncertainty on cheese mass loss model.

2 Concise representations of imprecise probability

Consider a probability space $(\Omega,\mathcal{A},P)$. Let $\mathcal{P}$ be a probability family on the referential Ω and X be a random variable associated with probability measure P. In the following, we consider three frameworks for representing special sets of probability functions, which are more convenient for a practical handling.

2.1 Probability boxes

Suppose $\overline{F}_X$ and $\underline{F}_X$ are nondecreasing functions from the real line $\mathbb{R}$ into $[0,1]$ such that $\underline{F}_X(x) \leq F_X(x) \leq \overline{F}_X(x)$, $\forall x \in \mathbb{R}$. The interval $[\underline{F}_X, \overline{F}_X]$ is called a "probability box" or "p-box" [11]. It encodes the class of probability measures whose cumulative distribution functions F_X are restricted by the bounding pair of cumulative distribution functions $\underline{F}_X$ and $\overline{F}_X$.
A p-box can be induced from the probability family $\mathcal{P}$ by $\forall x \in \mathbb{R}$:

$$\underline{F}(x) = \inf_{P\in\mathcal{P}} P((-\infty, x]); \overline{F}(x) = \sup_{P\in\mathcal{P}} P((-\infty, x]). \quad (1)$$

Let $\mathcal{P}(\underline{P} < \overline{P}) = \{P, \forall A \subseteq \Omega \text{ measurable}, \underline{P}(A) \leq P(A) \leq \overline{P}(A)\}$ be the probability family limited by upper $\overline{P}$ and lower $\underline{P}$ probabilities induced from $\mathcal{P}$. Clearly $\mathcal{P}$ is a proper subset of $\mathcal{P}(\underline{P} < \overline{P})$ generally. Let $\mathcal{P}(\underline{F}_X \leq \overline{F}_X)$ be the probability family containing $\mathcal{P}$ and defined by

$$\mathcal{P}(\underline{F}_X \leq \overline{F}_X) = \{P \in \mathcal{P}, \forall x \in \mathbb{R}, \underline{F}_X(x) \leq F(x) \leq \overline{F}_X(x)\}. \quad (2)$$

Generally $\mathcal{P}(\underline{F}_X \leq \overline{F}_X)$ strictly contains $\mathcal{P}(\underline{P} < \overline{P})$, hence also the set $\mathcal{P}$ it is built from. The probability box $[\underline{F}_X, \overline{F}_X]$ provides a bracketing of some ill-known cumulative distribution function and the gap between $\underline{F}_X$ and $\overline{F}_X$ reflects the incomplete nature of the knowledge, thus picturing the extent of what is ignored.

2.2 Numerical possibility theory

Possibility theory [9] is relevant to represent consonant imprecise knowledge. A possibility distribution can model imprecise information regarding a fixed unknown parameter and it can also serve as an approximate representation

of incomplete observation of a random variable. The basic notion is the possibility distribution, denoted π, describing the more or less plausible values of some uncertain variable X. Possibility theory provides two evaluations of the likelihood of an event: the possibility Π and the necessity N. The normalized measure of possibility Π (respectively necessity N) is defined from the possibility distribution $\pi : \mathbb{R} \to [0, 1]$ such that $\sup_{x\in\mathbb{R}} \pi(x) = 1$ as follows:

$$\Pi(A) = \sup_{x\in A} \pi(x),\ N(A) = 1 - \Pi(\overline{A}) = \inf_{x\notin A} (1 - \pi(x)). \quad (3)$$

Numerical possibility distribution may also be viewed as a nested set of confidence intervals, which are the α-cuts $[\underline{x}_\alpha, \overline{x}_\alpha] = \{x, \pi(x) \geq \alpha\}$ of π. The degree of certainty that $[\underline{x}_\alpha, \overline{x}_\alpha]$ contains X is $N([\underline{x}_\alpha, \overline{x}_\alpha])$ ($= 1 - \alpha$ if π is continuous). Conversely, given a nested set of intervals A_i with degrees of certainty λ_i that A_i contains X is equivalent to the possibility distribution

$$\pi(x) = \min_{i=1\ldots n} \{1 - \lambda_i, x \in A_i\}, \quad (4)$$

provided that λ_i is interpreted as a lower bound on $N(A_i)$, and π is chosen as the least specific possibility distribution satisfying these inequalities [10].
We can interpret any pair of dual functions necessity/possibility $[N, \Pi]$ as upper and lower probabilities induced from specific probability families.

- Let π be a possibility distribution inducing a pair of functions $[N, \Pi]$. We define the probability family $\mathcal{P}(\pi) = \{P, \forall A$ measurable, $N(A) \leq P(A)\} = \{P, \forall A$ measurable, $P(A) \leq \Pi(A)\}$. In this case, $\sup_{P\in\mathcal{P}(\pi)} P(A) = \Pi(A)$ and $\inf_{P\in\mathcal{P}(\pi)} P(A) = N(A)$ (see [6, 10]) hold. In other words, the family $\mathcal{P}(\pi)$ is entirely determined by the probability intervals it generates.
- Suppose pairs (interval A_i, necessity weight λ_i) supplied by an expert are interpreted as stating that the probability $P(A_i)$ is at least equal to λ_i where A_i is a measurable set. We define the probability family as follows: $\mathcal{P}(\pi) = \{P, \forall A_i, \lambda_i \leq P(A_i)\}$. We thus know that $\sup_{P\in\mathcal{P}(\pi)} P(A) = \Pi(A)$ and $\inf_{P\in\mathcal{P}(\pi)} P(A) = N(A)$ (see [10], and in the infinite case [6]).

We can define a particular p-box $[\underline{F}, \overline{F}]$ from the possibility distribution π such that $\underline{F}(x) = N((-\infty, x])$ and $\overline{F}(x) = \Pi((-\infty, x])\ \forall x \in \mathbb{R}$. But this p-box contains many more probability functions than $\mathcal{P}(\pi)$ (see [1] for more details about comparative expressivity of p-box and possibility distribution).

2.3 Belief function induced from random sets

The theory of imprecise probabilities introduced by Dempster [7] (and elaborated further by Shafer [27] and Smets [28] in a different context) allows imprecision and variability to be treated separately within a single framework. Indeed, it provides mathematical tools to process information which is at the same time of random and imprecise nature. A random set on Ω is defined by a mass assignment ν which is a probability distribution on the power set of Ω. We assume that ν assigns a positive mass only to a finite family of subsets of Ω called the set $\mathcal{F}$ of focal subsets. Generally $\nu(\emptyset) = 0$ and $\sum_{E\in\mathcal{F}} \nu(E) = 1$. A random set induces set functions called plausibility and belief measures respectively denoted by *Pl* and *Bel*, and defined by Shafer [27] as follows:

$$Bel(A) = \sum_{E, E\subseteq A} \nu(E),\ Pl(A) = \sum_{E, E\cap A\neq\emptyset} \nu(E). \quad (5)$$

$Bel(A)$ gathers the imprecise evidence that asserts A; $Pl(A)$ gathers the imprecise evidence that does not contradict A.

These set-functions can be interpreted as families of probability measures, even if this view does not match the original motivation of Shafer [27] and Smets [28] for belief functions. A mass distribution ν may encode the probability family $\mathcal{P}(\nu) = \{P \in \mathcal{P}/\forall A \subseteq \Omega, Bel(A) \leq P(A)\} = \{P \in \mathcal{P}/\forall A \subseteq \Omega, P(A) \leq Pl(A)\}$. This family generates lower and upper probability functions that coincide with the belief and plausibility functions, *i.e.*

$$Pl(A) = \sup_{P\in\mathcal{P}(\nu)} P(A), \quad Bel(A) = \inf_{P\in\mathcal{P}(\nu)} P(A) \quad (6)$$

Originally, Dempster [7] considered imprecise probabilities induced from a probability space via a set-valued mapping Γ from a probability space $(\Omega, \mathcal{A}, P)$ to S (yielding a random set). For simplicity assume $\forall\omega \in \Omega, \Gamma(\omega) \neq \emptyset$. Let $X : \Omega \to S$ be a random variable such that $\forall\omega \in \Omega$, $X(\omega) \in \Gamma(\omega)$ and P_X be its associated probability measure such that $P_X(A) = P(X^{-1}(A))$. Define upper and lower probabilities as follows:

$$\overline{P}(A) = \sup_{X\in S(\Gamma)} P_X(A) \quad \underline{P}(A) = \inf_{X\in S(\Gamma)} P_X(A) \quad (7)$$

where $S(\Gamma) = \{X : \Omega \to S | X(\omega) \in \Gamma(\omega), \forall\omega \in \Omega\}$. For all measurable subsets $A \subseteq \Omega$, we have $\underline{A} \subseteq A \subseteq \overline{A}$ where $\underline{A} = \{\omega \in \Omega/\Gamma(\omega) \subseteq A\}$ and $\overline{A} = \{\omega \in \Omega/\Gamma(\omega) \cap A \neq \emptyset\}$. By defining the mass distribution ν_Γ on Ω by $\nu(E) = P(\{\omega/\Gamma(\omega) = E\})$. We thus retrieve belief and plausibility functions as follows:

$$\underline{P}(A) = P(\underline{A}) = Bel_\Gamma(A) = \sum_{E\subseteq A} \nu_\Gamma(E) \quad (8)$$

$$\overline{P}(A) = P(\overline{A}) = Pl_\Gamma(A) = \sum_{E\cap A\neq\emptyset} \nu_\Gamma(E) \quad (9)$$

We may define an upper $\overline{F}$ and a lower $\underline{F}$ cumulative distribution function (a particular p-box) such that $\forall x \in \mathbb{R}$, $\underline{F}(x) \leq F(x) \leq \overline{F}(x)$ with :

$$\overline{F}(x) = Pl(X \in (-\infty, x]);\ \underline{F}(x) = Bel(X \in (-\infty, x]). \quad (10)$$

But this p-box contains many more probability functions than $\mathcal{P}(\nu)$.

2.4 Discretized encoding of probability, possibility and p-boxes a random sets

Belief functions [7, 27] encompass possibility, probability and probabiliy-boxes theories in the discrete case. Hence, we can encode probability distribution p, p-box $[\underline{F}_X, \overline{F}_X]$ and possibility distribution π by using mass distribution ν. In the continuous case, the representation will be approximate in a discrete framework for being able to do computations.

1. Probability → Belief function.
   Let X be a real random variable. In the discrete case, focal elements are singletons $(\{x_i\})_i$ and the mass distribution ν is defined by $\nu(\{x_i\}) = P(X = x_i)$. In the continuous case, we define focal intervals $((x_i, x_{i+1}])_i$ by discretizing probability density into m intervals and a mass distribution ν is defined by $\nu((x_i, x_{i+1}]) = P(X \in (x_i, x_{i+1}]), \forall\, i = 1 \dots m$.

2. Possibility → Belief function.
   Let X be a ill-known random variable described by a possibility distribution π. Focal sets correspond to the α-cuts

   $$\pi_{\alpha_j} = \{x | \pi(x) \geq \alpha_j\},\ \forall\ j = 1...q$$

   of possibility distribution π associated with X such that $\alpha_1 = 1 \geq \alpha_j \geq \alpha_{j+1} \geq \alpha_q > 0$ and $\pi_{\alpha_j} \subseteq \pi_{\alpha_{j+1}}$. Mass distribution ν is defined by $\nu(\pi_{\alpha_j}) = \alpha_j - \alpha_{j+1}\ \forall$ $j = 1 \dots q$ where $\alpha_{q+1} = 0$.

3. P-box → Belief function.
   Let X be a ill-defined random variable represented by a p-box $[\underline{F}_X, \overline{F}_X]$. By putting

   $$\underline{F}_X^{-1}(p) = \min\{x | \underline{F}_X(x) \geq p\},\ \forall\ p \in [0, 1] \qquad (11)$$

   $$\overline{F}_X^{-1}(p) = \min\{x | \overline{F}_X(x) \geq p\},\ \forall\ p \in [0, 1] \qquad (12)$$

   we can choose focal sets of the form $([\overline{F}_X^{-1}(p_i), \underline{F}_X^{-1}(p_i)])_i$ and the mass distribution ν such that $\nu([\overline{F}_X^{-1}(p_i), \underline{F}_X^{-1}(p_i)]) = p_i - p_{i-1}$ where $1 \geq p_i > p_{i-1} > 0$. In this case, Kriegler et al. [18] have showed that we have $\mathcal{P}(\underline{F}_X \leq \overline{F}_X) = \mathcal{P}(\nu)$.

3 Propagating general heterogeneous information

This section is dedicated to the combination and the propagation of three kinds of information: pure random variables, imprecisely known fixed quantities, and imprecise random variables (see [2, 3] for more details about the joint propagation methods of variability and imprecision). $\overrightarrow{X} : \Omega \to \mathbb{R}^k$ is a random vector that is observed with total precision; $\overrightarrow{Y} = (y_1, \dots, y_l)$ is a deterministic vector and we have partial information about it. Finally, $\overrightarrow{Z} : \Omega \to \mathbb{R}^n$ is a random vector observed with imprecision. In our model we suppose that there exists an unidimensional random variable, $T : \Omega \to \mathbb{R}$, that can be expressed of the form $T = f(\overrightarrow{X}, \overrightarrow{Y}, \overrightarrow{Z})$, where the mathematical model described by the function $f : \mathbb{R}^{k+l+n} \to \mathbb{R}$ is totally well-known. We will try to represent the information about the probability distribution of T based on the information available, about $\overrightarrow{X}$, $\overrightarrow{Y}$ and $\overrightarrow{Z}$, respectively.

First, as $\overrightarrow{X}$ is a random vector, it can be considered as a particular case of multidimensional random set (a singleton in $\mathbb{R}^k$). Thus, in our model, we can assume it as part of vector $\overrightarrow{Z}$.

To simplify the notation, suppose $\overrightarrow{Z} = (Z_1, Z_2)$ and $\overrightarrow{Y} = (y_1, y_2)$. The imprecise knowledge about y_1 (resp. y_2) is modeled by a possibility distribution π^1 (resp. π^2). Thus, with a confidence level $1 - \alpha$, the parameter y_1 (resp. y_2) belongs to α-cut $\pi^1_\alpha = \{x \in \mathbb{R} \mid \pi^1(x) \geq \alpha\}$ (resp. $\pi^2_\alpha = \{x \in \mathbb{R} \mid \pi^2(x) \geq \alpha\}$). Let us encode π^1 as belief function by their focal sets:

$$\pi^1_{\alpha_i} = \{x \in \mathbb{R} \mid \pi^1(x) \geq \alpha_i\}, \forall i = 1 \dots q \qquad (13)$$

such that $\pi^1_{\alpha_i} \subseteq \pi^1_{\alpha_{i+1}}$ with respective masses $\nu^1_i = \nu(\pi^1_{\alpha_i}) = \alpha_i - \alpha_{i+1},\ \forall i = 1 \dots q$ where $\alpha_1 = 1 \geq \alpha_i \geq \alpha_{i+1} \geq \alpha_q > 0$ and $\alpha_{q+1} = 0$. We proceed in the same way for π^2. Let $(C^1_j, m^1_j)_{j=1\dots r}$ (resp. $(C^2_l, m^2_l)_{l=1\dots r}$) be the focal sets and the mass distribution associated to Z_1 (resp. Z_2).

Now, we need to represent the available information about the probability measure P_T induced by T. The probability measure of T is imprecisely determined by means of the basic assignment (denoted ν^T_{ijkl}), associated with the focal sets

$$T_{ijkl} = f(\pi^1_{\alpha_i}, \pi^2_{\alpha_j}, C^1_k, C^2_l)$$

of T by $\forall i, j, k, l$:

$$\nu^T_{ijkl} = P(Y_1 = \pi^1_{\alpha_i}, Y_2 = \pi^2_{\alpha_j}, Z_1 = C^1_k, Z_2 = C^2_l)$$

In practice, only the marginals of the joint mass assignment are known, because no assumption is made about the relationship between the observation processes. If, in particular, independence between focal sets is assumed, the mass distribution becomes:

$$\forall i, j, k, l \quad \nu^T_{ijkl} = \nu^1_i \times \nu^2_j \times m^1_k \times m^2_l$$

Hence, if we want to estimate $Pl^T(A)$ for all measurable

set A, using the definition of v^T_{ijkl}, we have:

$$Pl^T(A) = \sum_{(i,j,k,l):A\cap T_{ijkl}\neq\emptyset} v^T_{ijkl},\ Bel^T(A) = \sum_{(i,j,k,l):T_{ijkl}\subseteq A} v^T_{ijkl}$$

It corresponds to applying a Monte-Carlo method to all variables. For each possibility distribution, an α-cut is independently selected. This approach is a conservative counterpart to the calculus of probabilistic variables under stochastic independence [4].

Suppose now the same value of α is selected in the Monte-Carlo simulation for y_1 and y_2. Then, $\forall i, j, k, l$:

$$\begin{array}{ll} \alpha_i = \alpha_j & v^T_{ijkl} = v^{y_1,y_2}_{\alpha_i} m^1_k m^2_l \\ \alpha_i \neq \alpha_j & v^T_{ijkl} = 0 \end{array}$$

The joint possibility distribution π associated to (y_1, y_2) is characterized by $min(\pi^1, \pi^2)$ which corresponds to the nested cartesian products of α-cuts and v^{y_1,y_2}_i is the mass associated to the Cartesian product $\pi^1_{\alpha_i} \times \pi^2_{\alpha_i}$. The use of "minimum" assumes the non-interaction of y_1, y_2, which expresses a lack of knowledge about the links between the values of y_1, y_2 and a lack of commitment as to whether y_1, y_2 are linked or not. We thus assume a total dependence between focal elements associated to possibilistic variables. This suggests that, if the source informing on y_1 is rather precise, then the one informing on y_2 is also precise (for instance it is the same source). However, this form of dependence does not presuppose any genuine functional (objective) dependence between possibilistic variables inside the domain $\pi^1_\alpha \times \pi^2_\alpha$ (observed phenomenons). Hence, if we want to estimate $Pl^T(A)$ and $Bel^T(A)$ using the last definition of v^T_{ijkl}, we deduce:

$$Pl^T(A) = \sum_{jk} \Pi^T_{jk}(A) \times m^1_k \times m^2_l,\ Bel^T(A) = \sum_{jk} N^T_{jk}(A) \times m^1_k \times m^2_l$$

where Π^T_{jk} are the possibility measures associated with the joint non-interactive possibility distribution π^T_{jk} obtained by means of the extension principle [8]:

$$\pi^T_{jk}(t) = \sup_{\substack{(y_1,y_2)\in\mathbb{R}^2,\\ (z_1,z_2)\in C^1_j\times C^2_k,\\ f(y_1,y_2,z_1,z_2)=t}} \min(\pi^1(y_1), \pi^2(y_2))$$

This technique thus computes the eventwise weighted average of the possibility measures associated with each output fuzzy interval, and applies to any event. It is easy to extend this propagation method for more than four variables.

4 Case description

Our example is concerned with the ripening of a soft mould cheese (camembert type). A model has been built [14] to estimate the mass loss of a cheese during ripening according to the close atmosphere. Our aim is to estimate confidence intervals or probability that cheese weight exceeds a threshold during ripening by taking into account uncertainty relative to measures and model parameters.

4.1 The ripening chamber

Soft cheeses (Camembert type) were manufactured in a sterile environment as previously described [22]. After drainage, 45 cheeses were aseptically transferred to a sterile pilot ripening chamber (see Figure 1). The average weigh of cheese was 0.333 kg with a standard deviation of 0.023 kg.

The ripening chamber (0.91m^3) was placed into a refrigerated room to allow the temperature regulation (see Figure 1). A cheese was continuously weighted with an electronic balance. Two combined sensors measured atmospheric temperature and relative humidity: 6 cm above the weighted cheese (see Figure 1, position a) and in the center of ripening chamber (see Figure 1, position b). Atmospheric changes were also characterized with CO_2 and O_2 sensors [26]. When the ripening chamber was used without input airflow, variations of these gas concentrations were depending only of cheese respiratory activity (CO_2 production and O_2 consumption). The ripening was performed with a periodically renewed atmosphere: if necessary, the CO_2 concentration was decreased to 2% by daily air injection with 6 m^3/h flow rate. In practical, the atmosphere was not renewed except 30 min per day. The ripening duration was 15 days, cheese were turned over on day 5. All online data were carried out with a 6 min acquisition period.

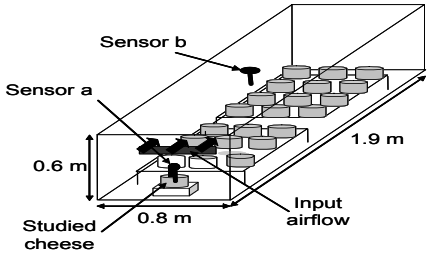


Figure 1: cheese-ripening room

4.2 Model of Cheese mass loss

Cheese mass loss dynamic results from exchange between product (cheese) and close atmosphere. A schematic view of system is illustrated by Figure 2. Biological activities induce a matter flux between the cheeses and atmosphere of the ripening chamber: oxygen consumption and carbon dioxide release. r_{o_2}, the O_2 consumption, and r_{co_2}, the CO_2 production rates (mol.m^{-2}.s^{-1}) are obtained by deriving CO_2 and O_2 atmospheric concentrations. The respiration

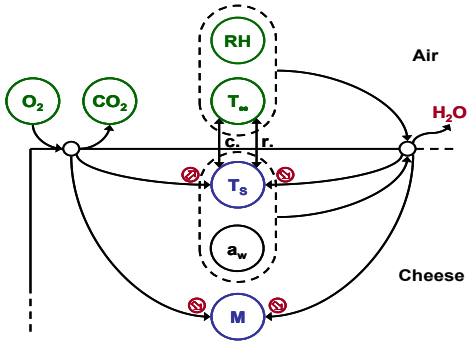


Figure 2: Schematic view of mass loss phenomenon.

matter flux ϕ_r (kg.m^{-2}.s^{-1}), is obtained by the difference between these two rates balanced by the molar masses

$$\phi_r = w_{o_2} r_{o_2} - w_{co_2} r_{co_2} \tag{14}$$

with w_{o_2} and w_{co_2} the respective molar masses (kg.mol^{-1}). Because the O_2 consumption and CO_2 production rates have the same dynamic, the following simplification is used:

$$\phi_r \simeq (w_{o_2} - w_{co_2})\, r = w_c r \tag{15}$$

with

$$r = \left(\frac{r_{o_2} + r_{co_2}}{2}\right) \tag{16}$$

The two rates are merged in r, corresponding to the respiratory activity. This simplification can be easily done because the carbon loss represents only 3% of the total mass loss.

The difference between water vapor pressure in the atmosphere and at the cheese surface causes an evaporative flux ϕ_w classically represented as following:

$$\phi_w = k\,(a_{ws} P_{sv}(T_s) - rh P_{sv}(T_\infty)) \tag{17}$$

with a_{ws} the cheese surface water activity, T_s and T_∞ the average surface and atmospheric temperatures respectively (K), rh the relative humidity (expressed between 0 and 1), $P_{sv}(T_\star)$ (Pa) the saturation vapor pressure at the temperature $T_\star$, and k the average water transfer coefficient (kg.m^{-2}.Pa^{-1}.s^{-1}).

The saturation vapor pressures are classically calculated with empirical relations as the Goff-Gratch equation [30]. However, the ripening temperature is usually between 12 °C and 14 °C. For this low range of temperature, an approximation can be done for saturation vapor pressure values, using a linear regression on the Goff-Gratch equation. The following relation is used:

$$P_{sv}(T_\star) = \beta_1 T_\star + \beta_2 \tag{18}$$

where $\beta_1 = 102$ Pa.K^{-1} and $\beta_2 = -27643$ Pa. The relative error (residual standard deviation over value range) is equal to 0.48%.

Direct heat exchange between the cheese and the atmosphere result from convective and radiative fluxes (see Figure 2)

$$\psi_{cr} = h\,(T_s - T_\infty) + \epsilon\sigma\left(T_s^4 - T_\infty^4\right) \tag{19}$$

with h the average convective heat transfer coefficient (W.m^{-2}.K^{-1}), ϵ the product emissivity (dimensionless) and σ the Stefan-Boltzmann constant (W.m^{-2}.K^{-4}). The radiative heat flux relation causes a strong nonlinearity; it can be approximated as following:

$$\epsilon\sigma\left(T_s^4 - T_\infty^4\right) \simeq 4\epsilon\sigma\overline{T}_\infty^3\,(T_s - T_\infty) \tag{20}$$

where $\overline{T}_\infty$ is the atmospheric temperature mean value. It is then possible to define a global heat transfer coefficient $h^\star = h + 4\epsilon\sigma\overline{T}_\infty^3$. From [21], we define an empirical relation between h and k for product with cheese shape

$$k = 0.75 \times 10^{-8} h \tag{21}$$

In addition, the moisture loss induces an heat consumption flux $\psi_w = \lambda\phi_w$ for the evaporation, with λ the water latent vaporization heat (J.kg^{-1}).

High biological activity is observed during the ripening for Camembert-type cheeses with an important mycelial development on the rind. This phenomenon induces a respirative heat production. The generic glucose aerobic respiration equation is

$$\begin{aligned} C_6H_{12}O_6 + \quad & 6O_2 \longrightarrow \\ & 6H_2O + 6CO_2 + 2.816 \times 10^6 \text{J.mol}^{-1} \end{aligned} \tag{22}$$

We have with this equation an equimolarity between O_2 and CO_2. During ripening, many substrates are oxidized (lactose, lactate, lipids and proteins), which can induce small differences between O_2 consumption and CO_2 production. This variability is then represented by the average of the gases rates r.

The cheese temperature dynamical model is

$$\frac{dT_s}{dt} = \frac{s}{mC}(-\psi_{cr} - \lambda\phi_w + \alpha\, r) \tag{23}$$

with m the mass of a cheese, s (m^2) the surface exchange of the cheese, C the specific heat (J.kg^{-1}.K^{-1}) and α the respiration heat (J.mol^{-1}) determined according to (22).

The mass loss dynamic is very slow compared to temperature dynamic, what allows to take T_s at the steady-state. We can thus write

$$T_s = \frac{h^\star T_\infty - \lambda k\,(a_{ws}\beta_2 - rh\,(\beta_1 T_\infty + \beta_2)) + \alpha r}{h^\star + \lambda k a_{ws}\beta_1} \tag{24}$$

and the mass loss rate q_m is defined by

$$\begin{aligned} q_m = \quad & \gamma h^\star (a_{ws} - rh)(\beta_1 T_\infty + \beta_2) \\ & + (\gamma a_{ws}\beta_1\alpha + w_c)\, r \end{aligned} \tag{25}$$

with

$$\gamma = \frac{k}{h^\star + \lambda k a_{ws}\beta_1}$$

4.3 Information representation

In this Section, we try to represent the available information faithfully relative to input variables and model parameters.

4.3.1 model parameters

Knowledge about heat respiration α, water latent vaporization heat λ, product emissivity ϵ, Stefan-Boltzmann constant σ and molar mass w_c come from literature (see Table 1).

Surface water activity (a_{ws}) is a key parameter for relation (17). Experimental measurements allows us to assume a_{ws} as constant equal to 0.976.

Due to low airflow velocity inside ripening chamber, available knowledge about the convective heat transfert coefficient h is imprecise and incomplete. Experts consider that heat transfert coefficient is most likely to lie between 3 and 3.2 $W.m^{-2}.K^{-4}$ but they do not exclude values as low as 2.5 and as high as 3.5 $W.m^{-2}.K^{-4}$. Hence, the knowledge of convective heat transfert coefficient h is represented by means of a trapezoidal possibility distribution of core [3,3.2] and support [2.5,3.5] (see Figure 3). According to

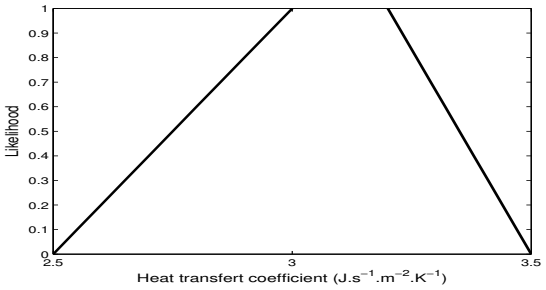


Figure 3: Trapezoidal possibility distribution representing convective heat transfert coefficient h

equation (21), knowledge about the average water transfert coefficient k is also represented by a trapezoidal possibility distribution.

4.3.2 input variables

The mass loss rate (25) is a function of 3 input variables which describe the gas exchanges (r) between cheese and atmosphere and the climatic condition (T_∞, rh). Measurements have not shown significant spatial gradient for O_2 and CO_2 concentrations inside the ripening room. Consequently, the measurements carried out at the position b (see Figure 1) are assumed as representative of gas concentrations close to the cheese. In ripening rooms, as well as cold chambers, due to air condition control, spatial variations of humidity rh and temperature T_∞ are always observed. These gradients are determined by the shape of the

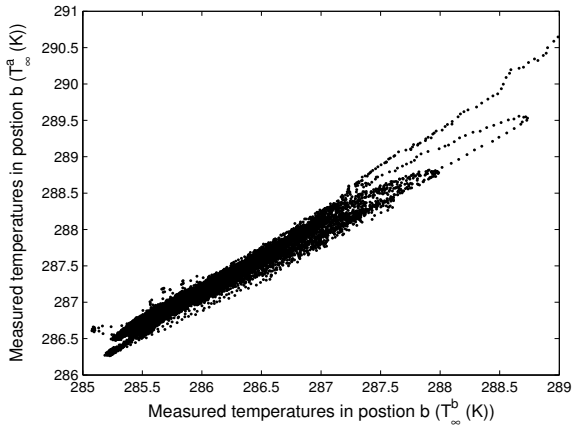


Figure 4: Temperature acquisition in postion a (T^a_∞) vs temperature acquisition in position b (T^b_∞) (see Figure 1).

room and the air regulation device. They are apprehended with detailed measurements and computational fluid dynamic models (*e.g.* see [19]) but these approaches can not be easily performed. We recall that it is inconceivable to install control sensors everywhere inside ripening rooms in order to model the behavior of humidity rh and temperature T_∞ inside chamber. In the present work, the aim is thus to estimate climatic conditions (T_∞, rh) close to cheeses (see Figure 1) from on-line temperature and relative humidity measurements in position b (denoted T^b_∞, rh^b), which is realistic in an industrial context. It is observed a linear relationship (see Figure 4 for temperatures) between climatic conditions (T^a_∞, rh^a) measured by sensors in position a and (T^b_∞, rh^b) measured in position b (see Figure 1, position a & position b). From a linear regression analysis, we obtain:

$$T^a_\infty = 0.91T^b_\infty + 2.31 \text{ and } rh^a = 1.029rh^b - 0.064 \quad (26)$$

with residual standard deviations $\sigma_{T_\infty} = 6.6\%$, $\sigma_{rh} = 0.5\%$. Due to low airflow velocity, experts assume that linear relationships remain valid between measured climatic conditions (T^b_∞, rh^b) and those close to by cheeses everywhere inside ripening chamber. However, due to the ill-known spatial variations of humidity and temperature inside ripening room, the linear relationships are tainted with imprecision. According to expert opinions, the imprecision about linear relationships is characterized by the imprecise bias of linear models. That means that temperature T_∞ and relative humidity rh perceived by each cheese can be encoded by:

$$T_\infty = a_{T_\infty} \times T^b_\infty + b_{T_\infty} \text{ and } rh = a_{rh} \times rh^b + b_{rh}$$

where $b_{T_\infty} \in [\underline{b}_{T_\infty}, \overline{b}_{T_\infty}]$ and $b_{rh} \in [\underline{b}_{rh}, \overline{b}_{rh}]$. Finally, by using linear regressions (26) and the empirical knowledge of system by experts, we decided, in the present work, to represent T_∞ (resp. rh) by an imprecise normal distribution

Symbol	Mode of representation
a_{ws} (unit less)	0.976
h (W.m^{-2}.K^{-1})	Trapezoidal possibility distribution support=[2.5,3.5], core=[3,3.2]
k (kg.m^{-2}.Pa^{-1}.s^{-1})	0.75×10^{-8}h
r (unit less)	measures
T_∞ (K)	Imprecise normal distribution $\mathcal{N}(a_{T_\infty}T_\infty^b + b_{T_\infty}, \sigma_{T_\infty})$ $a_{T_\infty} = 0.91$ $b_{T_\infty} \in [2.26, 2.36]$ $\sigma_{T_\infty} = 0.075$
rh (unit less)	Imprecise normal distribution $\mathcal{N}(a_{rh}rh + b_{rh}, \sigma_{rh})$ $a_{rh} = 1.029$ $b_{rh} \in [-0.066, -0.062]$ $\sigma_{rh} = 0.005$
α (J.mol^{-1})	4.693×10^5
λ (J.kg^{-1})	2.47×10^6
ϵ (unit less)	0.91
σ(W.m^{-2}.K^{-4})	5.67×10^{-8}

Table 1: Representation of model parameters & input variables.

$\mathcal{N}(a_{T_\infty}T_\infty^b + b_{T_\infty}, \sigma_{T_\infty})$ (resp. $\mathcal{N}(a_{rh}rh^b + b_{rh}, \sigma_{rh})$) where $(a_{T_\infty}, b_{T_\infty}, \sigma_{T_\infty}) \in \{0.91\}\times[2.31-0.5, 2.31+0.5]\times\{0.075\}$ (resp. $(a_{rh}, b_{rh}, \sigma_{rh}) \in \{1.029\}\times[-0.064-0.02, -0.064+0.02]\times\{0.005\}$. Table 1 summarizes modes of representation selected for the different model parameters and input variables.

5 Uncertainty processing

In this Section, we acknowledge the imprecise nature of available information regarding certain model parameters & input variables (see Table 1) and attempt to jointly propagate variability and imprecision in the estimation of cheese mass loss through ripening process. We assume stochastic independence between the group of random sets (T_∞, rh) and the group of possibilistic variables (h, k). Lastly, we assume independence between information sources pertaining to (T_∞, rh). According to the propagation method described in Section 3, the sketch for estimating the probability measure of cheese mass loss through ripening process is the following:

1. For time $t = t_0$.
2. Select a size L of the input sample.
3. We perform a random selection among focal elements by taking into account dependencies described previously:

$$\begin{pmatrix} a_{ws} & \pi^h_{\alpha_1} & \pi^k_{\alpha_1} & T^1_\infty(t) & rh^1(t) & r(t) & \alpha & \lambda & \epsilon & \sigma \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ a_{ws} & \pi^h_{\alpha_L} & \pi^k_{\alpha_L} & T^L_\infty(t) & rh^L(t) & r(t) & \alpha & \lambda & \epsilon & \sigma \end{pmatrix}$$

where, $\forall i = 1, \ldots, L$,

$$T^i_\infty(t) = [a_{T_\infty}T^b_\infty(t) + \underline{b}_{T_\infty} + \sigma_{T_\infty}u_i, a_{T_\infty}T^b_\infty(t) + \overline{b}_{T_\infty} + \sigma_{T_\infty}u_i]$$

$$rh^i(t) = [a_{rh}rh^b(t) + \underline{b}_{rh} + \sigma_{rh}v_i, a_{rh}rh^b(t) + \overline{b}_{rh} + \sigma_{rh}v_i]$$

and $(u_1, \ldots, u_L)$, $(v_1, \ldots, v_L)$ are random sampling from $\mathcal{N}(0, 1)$.

4. Propagate the sample through the model $q_m(t)$, we obtain a random set with focal elements $([\underline{q}^i_m(t), \overline{q}^i_m(t)])_{i=1,\ldots,L}$ defined by:

$$\underline{q}^i_m(t) = \inf_{(h,k,T_\infty,rh)\in\pi^h_{\alpha_i}\times\pi^k_{\alpha_i}\times T^i_\infty(t)\times rh^i(t)}$$

and

$$\overline{q}^i_m(t) = \sup_{(h,k,T_\infty,rh)\in\pi^h_{\alpha_i}\times\pi^k_{\alpha_i}\times T^i_\infty(t)\times rh^i(t)}$$

5. Hence, we can estimate $Bel(q_m(t) \in A)$ by:

$$Bel(q_m(t) \in A) = \frac{1}{L}Card\{i | [\underline{q}^i_m(t), \overline{q}^i_m(t)] \subseteq A\}$$

6. $t = t + \delta t$, return to step 1.

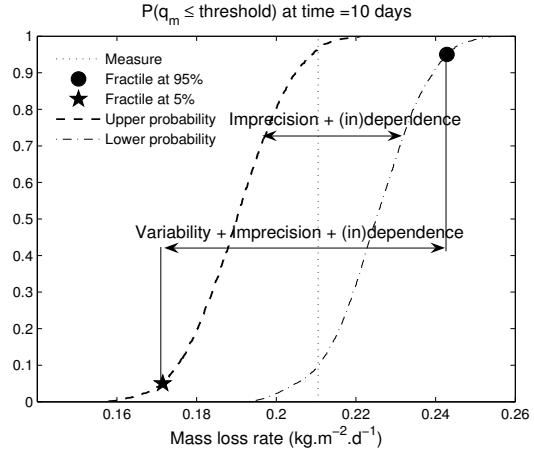


Figure 5: Upper & lower cumulative probabilities of mass loss rate for the tenth day.

In order to illustrate the impacts of imprecision and variability on mass loss rate, We decided to show, through Figure 5, the upper ($Pl(q_m(10) \leq .)$) & lower ($Bel(q_m(10) \leq .)$) cumulative distribution functions of it for the tenth day. It also illustrates a comparison with the mass loss rate obtained from online acquisition in position a. The gap between these two distributions is primarily a consequence

of the imprecise nature of available information and, to a lesser extent, of the choice of the dependence in propagation method. According to Figure 5, there is a 5% (resp. 95%) of plausibility (resp. belief) of being lower than 0.172 kg.m^{-2}.d^{-1} (resp. 0.243 kg.m^{-2}.d^{-1}). We can

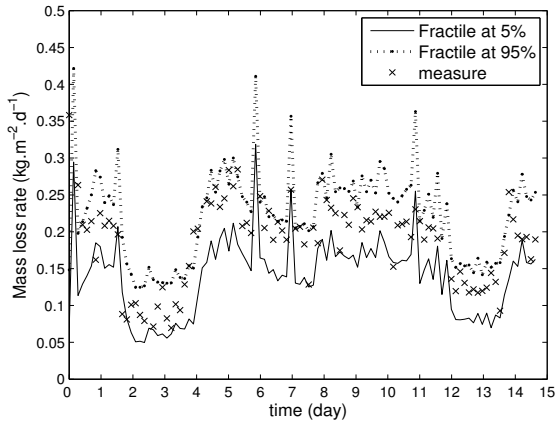


Figure 6: Uncertainty margins of 5% & 95% percentiles pertaining to the mass loss rate through ripening process.

then summarize the uncertainty on mass loss rate for the

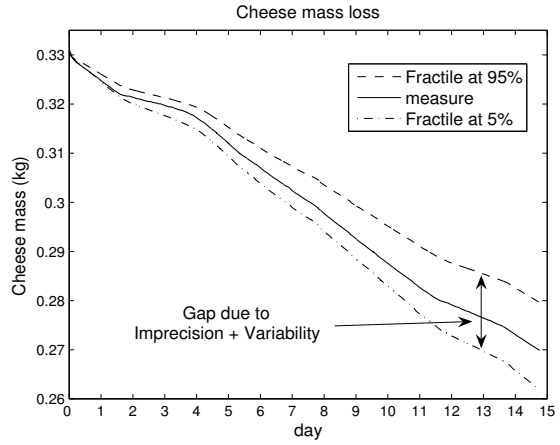


Figure 7: Uncertainty margins of 5% & 95% percentiles pertaining to the cheese mass loss through ripening process.

tenth day by means of interval [0.172 0.243] which can be seen as "confidence interval" containing imprecision. That means, for instance, we are sure at 95% that mass loss rate exceeds 0.172 kg.m^{-2}.d^{-1} for the tenth day. Figure 6 presents uncertainty margins of 5% and 95% percentiles pertaining to the mass loss rate for each time step through ripening process.

After integrate mass loss rate, Figure 7 presents uncertainty margins of 5% and 95% percentiles pertaining to the cheese mass loss for each time step through ripening process. That means $[e_1^t, e_2^t]$ such that

$$Pl(m_{init} - \int_0^t q_m(u)du \leq e_1^t) = 5\%$$

$$Bel(m_{init} - \int_0^t q_m(u)du \leq e_2^t) = 95\%$$

where $m_{init} = 0.333kg$. Hence, the probability, at the fifteenth day, of being lower than 0.263 kg is inferior to 5% and the probability of being lower than 0.278 kg is superior to 95%. That means that mass cheese is upper than 0.263 kg and lower than 0.278 kg with a confidence level superior to 90% at the fifteenth day of ripening process. On the one hand, we are sure at 95% that the mass loss of cheese does not exceed 67 g through 15 days of ripening; on the other hand, we are sure at 95% that cheese losses at least 52 g during 15 days.

6 Conclusion

During cheese ripening, a mass loss occurs resulting from heat and mass transfers from cheese to atmosphere. This phenomenon is based on physical laws and biological activity. The state of knowledge to model the process induces uncertainty on some phenomenon and as a consequence on some parameters of it. In this paper, we have quantified uncertainty on the model of cheese ripening mass loss by treating imprecision and variability.

Propagating imprecision on the basis of the results shown, shall help us to improve the control process. It is interesting to notice that a strategy to complete this knowledge can be elaborated as to be able to give a better estimation of the mass loss at the end of the ripening process. For example, considering the large gap between the upper and lower bounds on probability in Figure 5, it is clear that further studies on heat transfert coefficient and climatic conditions would be needed in order to reduce the subjective uncertainty regarding these quantities.

Such a result shows that it is possible to integrate and process mathematically the uncertainty on a complex process such as cheese ripening. Further studies will focus on this last point and moreover on the way to process uncertainty on a more general frame of knowledge integration and dynamic reconstruction.

References

[1] Baudrit, C., Dubois, D. Practical Representation of Incomplete Probabilistic Knowledge. *Comput. Stat. Data Anal.*, 51(1), 86-108, 2006.

[2] Baudrit, C., Couso, I., Dubois, D. Joint Propagation of Probability and Possibility in Risk Analysis: toward a formal framework. *Int. J. of Appro. Reason.*, 45(1), 82-105,

2007.

[3] Baudrit, C., Dubois, D., Guyonnet, D. Joint Propagation and Exploitation of Probabilistic and Possibilistic Information in Risk Assessment Models. *IEEE Trans. on Fuzzy Syst.*, 14(5), 593-608, 2006.

[4] Couso, I., Moral, S., Walley, P. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5, 165-180, 2000.

[5] Davidson, V.J., Brown, R.B., Landman, J.J. Fuzzy control system for peanut roasting. *J. Food Eng.*, 41, 141-146, 1999.

[6] De Cooman, G., Aeyels, D. Supremum-preserving upper probabilities. *Information Sciences*, 118, 173-212, 1999.

[7] Dempster, A.P. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Ann. Math. Stat.*, 38, 325-339, 1967.

[8] Dubois, D., Kerre, E., Mesiar, R, and Prade, H. Fuzzy interval analysis.*Fundamentals of Fuzzy Sets*, D. Dubois, H. Prade, Eds. Boston MA: Kluwer, 483-581, 2000.

[9] Dubois, D., Nguyen, H.T., Prade, H. Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. *Fundamentals of Fuzzy Sets*, Dubois,D. Prade,H., Eds: Kluwer , Boston, Mass, 343-438 , 2000.

[10] Dubois, D., Prade, H. When upper probabilities are possibility measures. *Fuzzy Sets Syst.*, 49, 65-74, 1992.

[11] Ferson, S., Ginzburg, L., Kreinovich, V., Myers, D.M., Sentz, K.Construction Probability Boxes and Dempster-Shafer structures. *Sandia National Laboratories, Technical report SANDD2002-4015*, 2003. URL:www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf.

[12] Ferson, S., Ginzburg, L.R. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety*, 54, 133-144, 1996.

[13] Goodman, I.R., Nguyen, H.T. *Uncertainty Models for Knowledge-Based Systems; A Unified Approach to the Measurement of Uncertainty*, Elsevier Science Inc., New York, NY, 1985.

[14] Helias, A., Mirade, P.S., Corrieu, C. Sensitivity analysis of a simplified cheese ripening mass loss model. Accepted to *10th Computer Application in Biotechnology*, Cancún, México, 2007.

[15] Ioannou, I., Mauris, G., Trystram, G., Perrot, N. Backpropagation of imprecision in a cheese ripening fuzzy model based on human sensory evaluations. *Fuzzy Sets Syst.*, 157(9), 1179-1187, 2006.

[16] Ioannou, I., Perrot, N., Mauris G., Trystram, G. Building of a control system using fuzzy set theory applied to a browning process, parts 1 and 2, *J. Food Eng.*, 64, 497-514, 2004.

[17] Ioannou, I., Perrot, P., Mauris G., Trystram G. Experimental analysis of sensory measurement imperfection impact for a cheese ripening fuzzy model. *Fuzzy Sets Syst. IFSA 2003*, eds T. Bilgic, B. De Baets, O. Kaynak, Springer, 595-602, 2003.

[18] Kriegler, E., Hermann, H. Utilizing belief functions for the estimation of future climate change. *Int. J. Appro. Reason.*, 39, 185-209, 2005.

[19] Mirade P.-S., Rougier T., Daudin J.-D., Picque D. and Corrieu G. Effect of design of blowing duct on ventilation homogeneity around cheeses in a ripening chamber. *Journal of Food Engineering*, 75(1), 59-70.

[20] Mirade, P.S., Daudin, J.D. Computational fluid dynamics prediction and validation of gas circulation in a cheese-ripening room. *I. Dairy Journal*, 16(8), 920-930, 2006.

[21] Mirade, P.S. and T. Rougier and A. Kondjoyan and J.D. Daudin and D. Picque & G. Corrieu. Caractérisation expérimentale de l'aéraulique d'un hâloir de fromagerie et des changes air-produit. *Lait*, 84, 483-500, 2004.

[22] Leclercq-Perlat, M.N., and F. Buono and D. Lambert and E. Latrille and H.-E. Spinnler and G. Corrieu. Controlled production of Camembert-type cheeses. Part 1: Microbiological and physicochemical evolutions. *J. Dairy Res.*, 35, 346-354, 2004.

[23] Perrot, N., Ioannou, I., Allais, I., Curt, C., Hossenlopp J. & Trystram, G. Fuzzy concepts applied to food product quality control: A review. *Fuzzy Sets Syst.*, 157(9), 1145-1154, 2006.

[24] Perrot, N., Agioux, L., Ioannou, I., Trystram, G., Mauris G. & Corrieu, G. Decision support system design using the operator skill to control cheese ripening - Application of the fuzzy symbolic approach. *J. Food Eng.*, 64, 321-333, 2004.

[25] Petermeier, H., Benning, R., Delgado, A;, Kulozik, U., Hinrichs, J., Becker, T. Hybrid model of the fouling process in tubular heat exchangers for the dairy industry. *J.Food Eng.*, 55, 9-17, 2002.

[26] Picque, D., and M.-N. Leclercq-Perlat and G. Corrieu. Effects of Atmospheric Composition on Respiratory Behavior, Weight Loss, and Appearance of Camembert-Type Cheeses During Chamber Ripening. *J.Dairy Sci.*, 89, 3250-3259, 2006.

[27] Shafer, G. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[28] Smets P. and Kennes R. (1994). The transferable belief model. *Artificial Intelligence*, 66, 191-234.

[29] Walley, *P. Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, 1991.

[30] WMO. General meteorological standards and recommended practices, Appendix A, corrigendum. *World Meteorological Organization Technical Regulations*, Geneva, 49, 2000.

Some results on imprecise conditional prevision assessments

Veronica Biazzo[1] & **Angelo Gilio**[2]

[1] Dipartimento di Matematica e Informatica
Viale A. Doria, 6 - 95125 Catania (Italy)
[2] Dipartimento di Metodi e Modelli Matematici
Via A. Scarpa, 16 - 00161 Roma (Italy)

Abstract

In this paper we consider conditional prevision assessments on random quantities with finite set of possible values. After some preliminaries, we give the notions of generalized coherence and total coherence for imprecise conditional prevision assessments on finite families of conditional random quantities. Then, we examine some results on total coherence of such conditional previsions under different assumptions for the conditioning events. We first consider the case of logically incompatible conditioning events; then, we examine the case of logical independence. Finally, we examine the general case in which there may be some logical dependencies among the conditioning events. We show that in this case the property of total coherence is generally lost, while it is always valid a connection property. By exploiting such property, we obtain suitable totally coherent sets of conditional prevision assessments. We also give a necessary and sufficient condition of total coherence for interval-valued conditional prevision assessments.

Keywords. conditional random quantities, imprecise conditional prevision assessments, generalized coherence, total coherence, connection property.

1 Introduction

The probabilistic treatment of uncertainty plays a relevant role in many applications of Artificial Intelligence, e.g. reasoning under uncertainty with a vague and partial information. In these applications typically the set of conditional events and/or random quantities at hand doesn't have any particular algebraic structure. Then, to obtain a flexible and consistent probabilistic approach we can use imprecise conditional probability and/or conditional prevision assessments, by exploiting suitable generalization of the coherence principle of de Finetti, or similar principles like that ones adopted for lower/upper probabilities and/or previsions (see, e.g., [1], [2], [5], [6], [7], [10], [14], [15], [17], [18], [19]).

In this paper we examine interval-valued conditional prevision assessments on finite families of conditional random quantities having a finite set of possible values. Even if this is not the more general case from a theoretical viewpoint, notwithstanding it is surely important in many applications. We use a notion of generalized coherence (g-coherence) which is equivalent to avoiding uniform loss property (AUL) introduced by Walley for lower previsions. We first recall some results on precise and imprecise conditional probability assessments on finite families of conditional events ([3], [4]). Then, we obtain some results concerning the more general case of conditional prevision assessments on finite families of conditional random quantities; in particular, we illustrate a connection property of the set Π_n of coherent conditional prevision assessments on a family $\mathcal{F}_n$ of n conditional random quantities. Such a property may be important when we want to determine conditional prevision assessments which are intermediate between other assessments which are judged too extreme, or not reasonable in some sense. For instance, we can imagine that we have two different assessments $\mathcal{M}', \mathcal{M}''$, given by two experts, on the same family $\mathcal{F}_n$, but we want determine some assessment $\mathcal{M}$ which is intermediate between $\mathcal{M}'$ and $\mathcal{M}''$. Then, the connection property assures us that we can choose $\mathcal{M}$ on a suitable curve $\mathcal{C}$, each point of which is a generalized convex combination of the extreme points $\mathcal{M}', \mathcal{M}''$; we observe that in general $\mathcal{C}$ could be constructed in an infinite number of ways. It would be interesting to investigate possible applications of the connection property in decisional problems where there are several probability assessors; but, the deepening of this aspect and a comparison with other approaches to imprecise probabilities is out of the scope of this paper. By exploiting the connection property, we obtain theoretical results on totally coherent sets of conditional prevision assessments. We observe that, given a family of n conditional random quantities $\mathcal{F}_n$, the total coherence of

a set $\mathcal{S} \subseteq \boldsymbol{R}^n$ means that, for every $\mathcal{M} \in \mathcal{S}$, the point $\mathcal{M}$ is a coherent conditional prevision assessment on $\mathcal{F}_n$. In particular, we obtain a necessary and sufficient condition of total coherence for interval-valued conditional prevision assessments. This property assures that, considering the interval I associated with an interval-valued conditional prevision assessment on a family of random quantities, if each vertex of I is a coherent (precise) conditional prevision assessment, then every point $\mathcal{M} \in I$ is coherent too. This allows to choose, if needed, in a very flexible way a precise conditional prevision assessment $\mathcal{M} \in I$, being sure that $\mathcal{M}$ is coherent.

We recall that an extension of a totally coherent *interval-valued* conditional probability assessment doesn't always exist ([12]); however, while the "least-committal" coherent interval-valued assessment "approximates" and contains the set Π_n of the coherent precise assessments on $\mathcal{F}_n$, in a dual way we could use (when possible) a suitable union of totally coherent interval-valued assessments, with the aim of approximating Π_n by a subset of it.

The paper is organized as follows. In Section 2 we recall some preliminary notions on precise and imprecise conditional probability and/or prevision assessments. Then, we give the notion of g-coherence for interval-valued prevision assessments, by remarking its equivalence with the notion of AUL lower previsions. In Section 3, after some preliminary aspects, we define the notions of g-coherence and of total coherence for a set of conditional prevision assessments. In Section 4 we give a result on totally coherent sets of conditional prevision assessments when the conditioning events are logically incompatible. In Section 5, after an introductory example, we give a result on coherent conditional probability assessments under suitable hypotheses of logical independence; then, we obtain a result on total coherence of conditional prevision assessments. In Section 6 we examine the general case in which among the conditioning events there exist some (possibly partial) logical dependencies. We show by an example that the property of total coherence is lost. Then, we give a theoretical result concerning a connection property which assures that, given two coherent conditional prevision assessments $\mathcal{M}', \mathcal{M}''$, we can construct (in general, in an infinite number of ways) a curve $\mathcal{C}$ each point of which is a coherent intermediate assessment between $\mathcal{M}', \mathcal{M}''$. In Section 7, exploiting the connection property, we give some further results on total coherence; in particular, we obtain a necessary and sufficient condition of total coherence for interval-valued conditional prevision assessments. Finally, in Section 8 we give some conclusions and comments on possible further developments of the work.

2 Some preliminary notions

We give some preliminary notions on coherence and generalized coherence of precise and imprecise conditional prevision assessments on finite families of conditional random quantities. We assume that each random quantity has a finite set of possible values. We denote by A^c the negation of A and by $A \vee B$ (resp., AB) the logical union (resp., intersection) of A and B. We use the same symbol to denote an event and its indicator. For each integer n, we set $J_n = \{1, 2, \ldots, n\}$.

2.1 Precise conditional prevision assessments

Given a real function $\mathbb{P}$ defined on an arbitrary family of conditional random quantities $\mathcal{K}$, let $\mathcal{F}_n = \{X_i|H_i,\, i \in J_n\}$ be a finite subfamily of $\mathcal{K}$ and $\mathcal{M}_n$ the vector $(\mu_i,\, i \in J_n)$, where $\mu_i = \mathbb{P}(X_i|H_i)$. With the pair $(\mathcal{F}_n, \mathcal{M}_n)$ we associate the random gain $\mathcal{G}_n = \sum_{i \in J_n} s_i H_i (X_i - \mu_i)$, where $s_1, \ldots, s_n$ are arbitrary real numbers and $H_1, \ldots, H_n$ denote the indicators of the corresponding events. We set $\mathcal{H}_n = H_1 \vee \cdots \vee H_n$; moreover, we denote by $\mathcal{G}_n|\mathcal{H}_n$ the restriction of $\mathcal{G}_n$ to $\mathcal{H}_n$. Then, using the *betting scheme* of de Finetti (see, e.g., [13]), we have

Definition 1. The function $\mathbb{P}$ is coherent if and only if, $\forall\, n \geq 1, \forall\, \mathcal{F}_n \subseteq \mathcal{K}, \forall\, s_1, \ldots, s_n \in \mathbb{R}$, it is $sup\ \mathcal{G}_n|\mathcal{H}_n \geq 0$.

We denote by Π_n the set of coherent conditional prevision assessments on $\mathcal{F}_n$. Given two points

$$\mathcal{M}' = (\mu_i',\, i \in J_n) \in \Pi_n\,, \quad \mathcal{M}'' = (\mu_i'',\, i \in J_n) \in \Pi_n\,,$$

we set

$$\begin{array}{l} \mu_i^m = \min\{\mu_i', \mu_i''\}\,, \;\; \mu_i^M = \max\{\mu_i', \mu_i''\}\,, \\ \mathcal{M}^m = \mathcal{M}' \wedge \mathcal{M}'' = (\mu_i^m,\; i \in J_n)\,, \\ \mathcal{M}^M = \mathcal{M}' \vee \mathcal{M}'' = (\mu_i^M,\; i \in J_n)\,. \end{array} \tag{1}$$

Moreover, given any pair of points

$$\mathbf{x} = (x_i,\, i \in J_n), \quad \mathbf{y} = (y_i,\, i \in J_n)\,,$$

we set $\mathbf{x} \leq \mathbf{y}$ if and only if $x_i \leq y_i$, $\forall\, i \in J_n$.
Then, $\mathcal{M}^m \leq \mathcal{M}^M$, for every $\mathcal{M}', \mathcal{M}''$.
In particular, given two probability assessments

$$\mathcal{P}' = (p_i',\; i \in J_n)\,, \;\; \mathcal{P}'' = (p_i'',\; i \in J_n)$$

on n conditional events $E_1|H_1, \ldots, E_n|H_n$, as in (1) we set

$$\mathcal{P}^m = \mathcal{P}' \wedge \mathcal{P}''\,, \;\; \mathcal{P}^M = \mathcal{P}' \vee \mathcal{P}''\,.$$

We remark that, given any point $\mathcal{P} = (p_i,\, i \in J_n)$, we have $\mathcal{P}^m \leq \mathcal{P} \leq \mathcal{P}^M$ if and only if there exists a vector $\Delta = (\delta_i,\, i \in J_n) \in [0,1]^n$ such that

$$p_i = (1 - \delta_i) p_i' + \delta_i p_i''\,, \;\; i \in J_n\,.$$

In this case we say that $\mathcal{P}$ is a *generalized convex combination* of $\mathcal{P}', \mathcal{P}''$. Below, we recall (in a slightly modified version) a result given in [3] which concerns conditional events.

Theorem 1. Let $\mathcal{P}' = (p'_i,\ i \in J_n)$, $\mathcal{P}'' = (p''_i,\ i \in J_n)$ be two coherent probability assessments defined on $\mathcal{F}_n = \{E_i|H_i, i \in J_n\}$. There exists a continuous curve Γ with extreme points $\mathcal{P}', \mathcal{P}''$ such that:
(i) each $\mathcal{P} \in \Gamma$ is a generalized convex combination of $\mathcal{P}', \mathcal{P}''$, i.e. $\mathcal{P}^m \leq \mathcal{P} \leq \mathcal{P}^M$; (ii) $\Gamma \subseteq \Pi_n$.

Theorem 1 assures that, for every pair of coherent assessments $\mathcal{P}', \mathcal{P}''$ on $\mathcal{F}_n$, we can construct (at least) a continuous curve $\Gamma \subseteq \Pi_n$ (from $\mathcal{P}'$ to $\mathcal{P}''$) whose points are intermediate coherent assessments between $\mathcal{P}'$ and $\mathcal{P}''$. We remark that in general the number of such curves is infinite.
Theorem 1 will be generalized to the case of conditional random quantities by Theorem 4.
By Theorem 1, we obtain

Corollary 1. Given any quantities $p_1, \ldots, p_{i-1}$, $l_i \leq u_i$, $p_{i+1}, \ldots, p_n$, let us define

$$\mathcal{P}' = (p_1, \ldots, p_{i-1}, l_i, p_{i+1}, \ldots, p_n)\,,$$
$$\mathcal{P}'' = (p_1, \ldots, p_{i-1}, u_i, p_{i+1}, \ldots, p_n)\,.$$

Moreover, let $\mathcal{I} = \mathcal{P}'\mathcal{P}''$ be the segment $\{(p_1, \ldots, p_i, \ldots, p_n) : l_i \leq p_i \leq u_i\}$, with set of vertices $\mathcal{V} = \{\mathcal{P}', \mathcal{P}''\}$. Then: $\mathcal{I} \subseteq \Pi_n \iff \mathcal{V} \subset \Pi_n$.

We remark that Corollary 1 is also an immediate consequence of the extension theorem for coherent conditional probabilities. Conversely, as shown in [3], the extension theorem can be obtained by Corollary 1 and the closure property of coherent conditional probability assessments.

2.2 Interval-valued conditional prevision assessments

Let $\mathcal{A}_n = ([l_i, u_i],\ i \in J_n)$ be any interval-valued conditional prevision assessment on a family $\mathcal{F}_n = \{X_i|H_i,\ i \in J_n\}$. We give below a notion of generalized coherence (g-coherence), already used in [1] for the case of conditional events (and simply named 'coherence' in [9]).

Definition 2. An interval-valued conditional prevision assessment $\mathcal{A}_n = ([l_i, u_i], i \in J_n)$, defined on a family of n conditional random quantities $\mathcal{F}_n = \{X_i|H_i, i \in J_n\}$, is g-coherent if there exists a coherent precise conditional prevision assessment $\mathcal{M}_n = (\mu_i, i \in J_n)$ on $\mathcal{F}_n$, with $\mu_i = \mathbb{P}(X_i|H_i)$, which is consistent with $\mathcal{A}_n$, that is such that $l_i \leq \mu_i \leq u_i$ for each $i \in J_n$.

Remark 1. Notice that, as $\mathbb{P}(X_i|H_i) \leq u_i$ amounts to $\mathbb{P}(-X_i|H_i) \geq -u_i$, g-coherence can be expressed by using only lower bounds. Then, g-coherence means that there exists a dominating coherent precise prevision and hence it is equivalent to *avoiding uniform loss* property of lower previsions given in [17]. Below we briefly comment on such equivalence. We recall that a lower prevision $\underline{P}$ on a family of conditional random quantities $\mathcal{K}$ *avoids uniform loss* (AUL) if, for every

$$\mathcal{F}_n = \{X_1|H_1, \ldots, X_n|H_n\} \subseteq \mathcal{K}\,,$$

defining $\underline{P}(X_i|H_i) = l_i\,,\ i \in J_n$ and

$$\mathcal{G}_n = \sum_{i=1}^{n} s_i H_i (X_i - l_i)\,,\ \ \mathcal{H}_n = H_1 \vee \cdots \vee H_n\,,$$

the inequality $sup\ \mathcal{G}_n|\mathcal{H}_n \geq 0$ is satisfied for every $s_1 \geq 0, \ldots, s_n \geq 0$. By exploiting the conjugacy condition $\overline{P}(X|H) = -\underline{P}(-X|H)$, we can express upper previsions in terms of lower previsions. As is well known, every AUL conditional prevision assessment admits the natural extension (see, e.g., [18]) which, being coherent, is a lower envelope of a set of coherent precise previsions (see [19], and for a review of this basic paper see [16]) which dominate the natural extension and hence the AUL assessment too. Conversely, as AUL property is given in terms of gains, it can be verified that every assessment dominated by a precise prevision is AUL.
A different method to show the equivalence between g-coherence and AUL property of a lower prevision assessment on a *finite* family $\mathcal{K}$ of conditional random quantities, is based on the following two steps:
(i) for each $\mathcal{F} \subseteq \mathcal{K}$, let $\mathcal{G}$ and $\mathcal{H}$ be respectively the random gain and the union of conditioning events associated with $\mathcal{F}$. Then, by an alternative theorem ([8], Th. 2.10) it can be verified that the condition $sup\ \mathcal{G}|\mathcal{H} \geq 0$ is equivalent to solvability of a suitable linear system Σ associated with $\mathcal{F}$;
(ii) it can be shown that the given lower prevision assessment is g-coherent if and only if, for each $\mathcal{F} \subseteq \mathcal{K}$, the associated system Σ is solvable.
This alternative method may be useful in real applications as, using a finite number of linear systems, we may construct, for the conditional random quantities in $\mathcal{K}$, a probability distribution assessment consistent with the given lower prevision assessment.

We denote by $\Im_n$ the set of g-coherent interval-valued conditional prevision assessments on $\mathcal{F}_n$. We recall below (in a slightly modified version) a result (see [4], Theorem 12) which generalizes Theorem 1 to the case of interval-valued conditional probability assessments, by showing how to construct an infinite class

of interval-valued assessments $\mathcal{A}_n = ([l_i, u_i],\, i \in J_n)$ which are intermediate between two given interval-valued assessments

$$A'_n = ([l'_i, u'_i]\,,\, i \in J_n)\,, \quad A''_n = ([l''_i, u''_i],\, i \in J_n)\,;$$

this means that there exists a vector $\Delta = (\delta_i,\, i \in J_n) \in [0,1]^n$ such that

$$l_i = (1-\delta_i)l'_i + \delta_i l''_i\,, \quad u_i = (1-\delta_i)u'_i + \delta_i u''_i\,, \quad i \in J_n\,.$$

As already made in the case of precise probability assessments, we say that $\mathcal{A}_n$ is a generalized convex combination of $\mathcal{A}'_n, \mathcal{A}''_n$, also denoted by $\mathcal{A}_\Delta$.

Theorem 2. Let be given two g-coherent interval-valued assessments $\mathcal{A}'_n = ([l'_i, u'_i],\, i \in J_n)$, $\mathcal{A}''_n = ([l''_i, u''_i],\, i \in J_n)$, on a family of n conditional events $\mathcal{F}_n = \{E_i|H_i, i \in J_n\}$. Then, we can construct an infinite class Υ of interval-valued probability assessments on $\mathcal{F}_n$ such that: (i) each $\mathcal{A}_n \in \Upsilon$ is a generalized convex combination between $\mathcal{A}'_n, \mathcal{A}''_n$; i.e., $\mathcal{A}_n = \mathcal{A}_\Delta$ for some $\Delta = (\delta_i,\, i \in J_n) \in [0,1]^n$; (ii) $\mathcal{A}_n \subseteq \Im_n$.

By Theorem 2, we can move in a continuous way from $\mathcal{A}'_n$ to $\mathcal{A}''_n$; then, by analogy with Theorem 1, we can say that $\mathcal{A}'_n, \mathcal{A}''_n$ are *connected* by the interval-valued probability assessments contained in Υ.

3 Some preliminary aspects

We recall that we consider random quantities with finite sets of possible values. Let X be a random quantity, with $X \in \mathcal{X} = \{x_1, \ldots, x_n\}$. We denote by E_i, the event $(X = x_i)$, $i \in J_n$. Moreover, given any event $H \neq \emptyset$, for each i we set $p_i = P(E_i|H)$; then, for the prevision of $X|H$ we have $\mathbb{P}(X|H) = \sum_i p_i x_i$. Of course, the coherence of a given assessment $\mathbb{P}(X|H) = \mu$ amounts to the existence of a nonnegative vector $(p_1, \ldots, p_n)$, with $\sum_i p_i = 1$, such that $\sum_i p_i x_i = \mu$. In equivalent terms, observing that $E_i H = \emptyset$ implies $p_i = 0$ and denoting by $\mathcal{X}_H \subseteq \{x_1, \ldots, x_n\}$ the set of possible values of X compatible with H, μ is coherent if and only if the following condition is satisfied

$$min_{x_i \in \mathcal{X}_H}\; x_i \leq \mu \leq max_{\;x_i \in \mathcal{X}_H}\; x_i\,. \qquad (2)$$

We denote by I_H the interval with vertices having the values $min_{x_i \in \mathcal{X}_H}$, $max_{x_i \in \mathcal{X}_H}$; i.e. we set

$$I_H = [min_{x_i \in \mathcal{X}_H}\; x_i\,,\; max_{x_i \in \mathcal{X}_H}\; x_i]\,. \qquad (3)$$

Of course, given two coherent assessments $\mathbb{P}(X|H) = \mu', \mathbb{P}(X|H) = \mu''$, it is $[\mu', \mu''] \subseteq I_H$; hence, the assessment $\mathbb{P}(X|H) = \mu$ is coherent, $\forall\, \mu \in (\mu', \mu'')$.
Given any pair of events H, K, we set $\mathbb{P}(X|H) = \mu_H, \mathbb{P}(X|K) = \mu_K$. As noted above, the coherence of μ_H (resp. μ_K) amounts to $\mu_H \in I_H$ (resp. $\mu_K \in I_K$). We set $I_{HK} = I_H \times I_K$. Of course, given an assessment $\mathcal{M} = (\mu_H, \mu_K)$ on $\{X|H, X|K\}$, the coherence of $\mathcal{M}$ amounts to the existence of two nonnegative vectors $(p_1, \ldots, p_n), (\pi_1, \ldots, \pi_n)$, with

$$\sum_i p_i x_i = \mu_H\,, \;\; \sum_i \pi_i x_i = \mu_K\,, \;\; \sum_i p_i = \sum_i \pi_i = 1\,,$$

such that the assessment $(p_1, \ldots, p_n, \pi_1, \ldots, \pi_n)$ on $\{E_1|H, \ldots, E_n|H, E_1|K, \ldots, E_n|K\}$ is coherent.
We recall that, if $E_i H = \emptyset$ (resp. $E_i K = \emptyset$), then $p_i = 0$ (resp. $\pi_i = 0$).
More generally, given n events $H_1, \ldots, H_n$ and n random quantities $X_1, \ldots, X_n$, we denote by $\mathcal{X}_{H_r} = \{x_{r1}, \ldots, x_{rk_r}\}$ the set of values of X_r compatible with H_r; then, for each $r \in J_n$, we set

$$I_r = [min_{x_{rj} \in \mathcal{X}_{H_r}}\; x_{rj}\,,\; max_{x_{rj} \in \mathcal{X}_{H_r}}\; x_{rj}] \qquad (4)$$

and $I_{1 \cdots n} = I_1 \times \cdots \times I_n$. Then, based on Definition 2 we give the following

Definition 3. Let $\mathcal{S}$ be a subset of the interval $I_{1 \cdots n}$. We say that $\mathcal{S}$ is g-coherent if there exists $\mathcal{M} = (\mu_1, \ldots, \mu_n) \in \mathcal{S}$ such that $\mathcal{M}$ is a coherent conditional prevision assessment on $\{X_1|H_1, \ldots, X_n|H_n\}$; in this case we simply say that $\mathcal{M}$ is coherent. We say that $\mathcal{S}$ is totally coherent if, for every $\mathcal{M} \in \mathcal{S}$, $\mathcal{M}$ is coherent.

We remark that in general the checking for total coherence of an (arbitrary) set S may be intractable, while the situation is different for the case of interval-valued assessments. In particular, considering the case of conditional events, let be given an interval-valued assessment $\mathcal{A}_n = ([l_1, u_1], \ldots, [l_n, u_n])$ on a family of n conditional events $\mathcal{F}_n$ and the associated interval and set of vertices

$$\mathcal{I} = [l_1, u_1] \times \cdots \times [l_n, u_n]\,, \quad \mathcal{V} = \{l_1, u_1\} \times \cdots \times \{l_n, u_n\}\,.$$

Then, a necessary and sufficient condition of total coherence for $\mathcal{I}$, obtained in [11], is given below.

Theorem 3. Given an interval-valued probability assessment $\mathcal{A}_n = ([l_1, u_1], \ldots, [l_n, u_n])$ on $\mathcal{F}_n$, one has $\mathcal{I} \subseteq \Pi_n$ if and only if $\mathcal{V} \subseteq \Pi_n$.

This necessary and sufficient condition says that total coherence of the interval $\mathcal{I}$ is equivalent to coherence of each of its vertices.

4 Logically incompatible conditioning events

In this section we give a result on totally coherent conditional prevision assessments when the conditioning events are logically incompatible. We have

Proposition 1. Given the conditional random quantities $X_1|H_1, \ldots, X_n|H_n$, let I_j be the interval associated with the set of possible values of X_j compatible with H_j, $j \in J_n$. Moreover, let $I_{1\cdots n}$ denote the interval $I_1 \times \cdots \times I_n$. If $H_i H_j = \emptyset$ for every $i \neq j$, then $I_{1\cdots n}$ is totally coherent.

Proof. Given any $\mathcal{M} = (\mu_1, \ldots, \mu_n) \in I_{1\cdots n}$, we have $\mu_j \in I_j$, $j \in J_n$; hence $\mu_1, \ldots, \mu_n$ are (separately) coherent. Then, there exist n nonnegative vectors

$$(p_{i1}, \ldots, p_{ik_i})\,,\ i \in J_n\,,$$

such that

$$\sum_{j=1}^{k_i} p_{ij} = 1\,, \quad \sum_{j=1}^{k_i} p_{ij} x_{ij} = \mu_i\,,\ \ i \in J_n\,.$$

Based on well known results, it follows that the probability assessment

$$(p_{11}, \ldots, p_{1k_1}, \ldots, p_{n1}, \ldots, p_{nk_n})$$

on the family of conditional events

$$\{A_{11}|H_1, \ldots, A_{1k_1}|H_1, \ldots, A_{n1}|H_n, \ldots, A_{nk_n}|H_n\}$$

is coherent; hence $\mathcal{M}$ is coherent. Therefore, $I_{1\cdots n}$ is totally coherent. □

We remark that the previous result can be related to the notion of *separate coherence* given in ([17], 6.2.2) for the case of conditioning events belonging to a finite partition of the sure event.
By our result we have that, when the conditioning events are logically incompatible, separate coherence implies total coherence.

5 Logically independent conditioning events

In this section we relax the assumption of logical incompatibility among conditioning events, by assuming some suitable hypotheses of logical independence. We recall that n events $E_1, \ldots, E_n$ are defined logically independent if and only if the number of constituents is maximum, that is 2^n. We first give an introductory example.

Example 1. Let be given four events A_1, A_2, H_1, H_2 satisfying the following logical conditions:
(i) A_1 and A_2 are logically incompatible;
(ii) A_1, H_1, H_2 are logically independent;
(iii) A_2, H_1, H_2 are logically independent.
It could be shown that every non negative vector (p_1, p_2, π_1, π_2) such that $p_1 + p_2 \leq 1$, with

$$p_1 + p_2 = 1 \quad \text{if} \quad A_1^c A_2^c H_1 = \emptyset\,,$$

and $\pi_1 + \pi_2 \leq 1$, with

$$\pi_1 + \pi_2 = 1 \quad \text{if} \quad A_1^c A_2^c H_2 = \emptyset\,,$$

is a coherent probability assessment on the family of conditional events $\{A_1|H_1, A_2|H_1, A_1|H_2, A_2|H_2\}$.

More in general, we have

Lemma 1. Let be given $k + n$ events $A_1, \ldots, A_k$, $H_1, \ldots, H_n$ satisfying the following logical conditions:
(i) $A_1, \ldots, A_k$ are logically incompatible;
(ii) for each index $i \in J_k$ the events $A_i, H_1, \ldots, H_n$ are logically independent.
Then, given any n nonnegative vectors

$$(p_1^{(1)}, \ldots, p_k^{(1)})\,, \ldots\,, (p_1^{(n)}, \ldots, p_k^{(n)})\,,$$

such that $\sum_i p_i^{(r)} \leq 1$, with $\sum_i p_i^{(r)} = 1$ if $A_1^c \cdots A_k^c H_r = \emptyset$, $r \in J_n$, the probability assessment

$$\mathcal{P} = (p_1^{(1)}\,, \ldots\,, p_k^{(1)}\,, \ldots\,, p_1^{(n)}, \ldots, p_k^{(n)})$$

on

$$\mathcal{F} = \{A_1|H_1\,, \ldots\,, A_k|H_1\,, \ldots\,, A_1|H_n\,, \ldots\,, A_k|H_n\}$$

is coherent.

Proof. Given any sub-family $\mathcal{F}' \subseteq \mathcal{F}$, we denote by $\mathcal{P}'$ the associated sub-assessment of $\mathcal{P}$ and by $\mathcal{G}'$ the random gain associated with the pair $(\mathcal{F}', \mathcal{P}')$. Moreover, we denote by $\mathcal{H}'$ the union of those conditioning events H_j's such that $A_i|H_j \in \mathcal{F}'$ for some index i; in particular, we set $\mathcal{H} = H_1 \vee \cdots \vee H_n$. We will verify the coherence condition

$$sup\ \mathcal{G}'|\mathcal{H}' \geq 0\,,\ \ \forall\, \mathcal{F}' \subseteq \mathcal{F}\,,$$

by the following steps:
1. We preliminarily observe that each nonnegative vector $P_r = (p_1^{(r)}, \ldots, p_k^{(r)})$ such that $\sum_i p_i^{(r)} \leq 1$, with $\sum_i p_i^{(r)} = 1$ if $A_1^c \cdots A_k^c H_r = \emptyset$, is a coherent assessment on the sub-family $F_r = \{A_1|H_r, \ldots, A_k|H_r\}$; so that, denoting by G_r the random gain associated with the pair (F_r, P_r), it is

$$sup\ G_r|H_r \ \geq\ 0\,,\ \ \forall\, r \in J_n\,.$$

For each $h \in J_k$ we denote by $g_h^{(r)}$ the value of $G_r|H_r$ associated with the constituent

$$H_r A_1^c \cdots A_{h-1}^c A_h A_{h+1}^c \cdots A_k^c\,;$$

moreover, if $H_r A_1^c \cdots A_k^c \neq \emptyset$, we denote by $g_{k+1}^{(r)}$ the corresponding value of $G_r|H_r$. Hence

$$sup\ G_r|H_r = sup_h\ g_h^{(r)}\ \geq\ 0\,.$$

2. By the logical assumptions, the set of constituents associated with the pair $(\mathcal{F}, \mathcal{P})$ contains, for each $r \in J_n$, the following ones

$$(\bigwedge_{j \neq r} H_j^c) H_r A_1^c \cdots A_{h-1}^c A_h A_{h+1}^c \cdots A_k^c \,,\ h \in J_k \,,$$

denoted $C_1^{(r)}, \ldots, C_k^{(r)}$, and, if not impossible, the further constituent

$$C_{k+1}^{(r)} = (\bigwedge_{j \neq r} H_j^c) H_r A_1^c \cdots A_k^c \,.$$

We make two remarks:
a) the gains associated with the constituents above are

$$s_1^{(r)} - \sum_{i=1}^{k} p_i^{(r)} s_i^{(r)} \,, \ \ldots \,, s_k^{(r)} - \sum_{i=1}^{k} p_i^{(r)} s_i^{(r)} \,,$$

(and possibly $-\sum_{h=1}^{n} p_h^{(r)} s_h^{(r)}$);
b) these gains coincide respectively with the values $g_1^{(r)}, \ldots, g_k^{(r)}$ (and possibly $g_{k+1}^{(r)}$) of $G_r | H_r$.
Then, denoting by $\mathcal{G}$ the random gain associated with the pair $(\mathcal{F}, \mathcal{P})$, as

$$sup\, \mathcal{G}|\mathcal{H} \ \geq \ s_h^{(r)} - \sum_{i=1}^{k} p_i^{(r)} s_i^{(r)} \,, \ \ \forall\, h \in J_k \,,$$

and (from coherence of the assessment P_r on F_r) $sup_h\, g_h^{(r)} \geq 0$, it follows $sup\, \mathcal{G}|\mathcal{H} \geq 0$.
3. Now, given any pair $(\mathcal{F}', \mathcal{P}')$, where $\mathcal{F}'$ is a sub-family of $\mathcal{F}$ and $\mathcal{P}'$ is the corresponding sub-vector of $\mathcal{P}$, we observe that the structure of $(\mathcal{F}', \mathcal{P}')$ is similar to that of $(\mathcal{F}, \mathcal{P})$; in particular, the hypotheses (i) and (ii), of logical incompatibility and of logical independence, still hold for the sub-family of events $\{A_i, H_r : A_i | H_r \in \mathcal{F}'\}$. Then, by the same reasoning, we can verify that the (necessary) coherence condition associated with $(\mathcal{F}', \mathcal{P}')$, i.e. $sup\, \mathcal{G}'|\mathcal{H}' \geq 0$, is satisfied, $\forall\, (\mathcal{F}', \mathcal{P}')$. Thus, the probability assessment $\mathcal{P}$ on the family $\mathcal{F}$ is coherent. □

Now, we will consider the events $E_j = (X = x_j)$, $j \in J_k$, which are a partition of the sure event Ω, denoting by I the interval associated with the set of possible values of X. By Lemma 1, we have

Proposition 2. Given the conditional random quantities $X|H_1, \ldots, X|H_n$, let I_j be the interval associated with the set of possible values of X compatible with H_j, $j \in J_n$. Moreover, let be $I_{1 \cdots n} = I_1 \times \cdots \times I_n$. If, for each $j \in J_n$, the events $E_j, H_1, \ldots, H_n$ are logically independent, then $I_j = I$, $\forall\, j \in J_n$, and $I_{1 \cdots n}$ is totally coherent.

Proof. By the hypotheses of logical independence it immediately follows $I_1 = \cdots = I_n = I$. Given any $\mathcal{M} = (\mu_1, \ldots, \mu_n) \in I_{1 \cdots n}$, we have $\mu_j \in I, j \in J_n$; hence, for each j, μ_j is (separately) coherent. Then, there exist n nonnegative vectors $(p_1^{(r)}, \ldots, p_k^{(r)})$, with $\sum_j p_j^{(r)} = 1$, $r \in J_n$, where $p_j^{(r)} = P(E_j | H_r)$, such that $\sum_j p_j^{(r)} x_j = \mu_r$, $r \in J_n$. By Lemma 1, the probability assessment $(p_1^{(1)}, \ldots, p_k^{(1)}, \ldots, p_1^{(n)}, \ldots, p_k^{(n)})$ on $\{E_1|H_1, \ldots, E_k|H_1, \ldots, E_1|H_n, \ldots, E_k|H_n\}$ is coherent. Hence $\mathcal{M}$ is coherent too; thus $I_{1 \cdots n}$ is totally coherent. □

A comparison with other approaches to precise and/or imprecise probabilities is out of the scope of this paper; however, it is presumable that the results of the sections 4 and 5 could be obtained by similar methods proposed by other authors (see, e.g., [6], [18]).

6 Logically dependent conditioning events

In this section we will give some results in the general case in which among the conditioning events there exist some (possibly partial) logical dependencies. In this case generally the property of total coherence is lost. We will illustrate this aspect in the following

Example 2. Given a random quantity $X \in \{x_1, \ldots, x_n\}$ and two events H, K, let us consider the conditional random quantities $K|H, X|HK, XK|H$. Then, let $\mathcal{M}_1 = (m_1, m_2, m_3), \mathcal{M}_2 = (\mu_1, \mu_2, \mu_3)$ be two conditional prevision assessments on the family $\mathcal{F}_3 = \{K|H, X|HK, XK|H\}$. As is well known, if $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) is coherent, then $m_3 = m_2 m_1$ (resp. $\mu_3 = \mu_2 \mu_1$). Then, denoting respectively by I_1, I_2, I_3 the intervals associated with the set of possible values of $K|H, X|HK, XK|H$, let be $I = I_1 \times I_2 \times I_3$. We observe that, even assuming $I_1 \times I_2$ totally coherent, the interval I is not totally coherent; that is, given any $\mathcal{M} = (x, y, z) \in I$, if $z \neq xy$, then $\mathcal{M}$ is not coherent. In particular, we observe that if $\mathcal{M}$ is a point of the segment $\mathcal{M}_1 \mathcal{M}_2$, generally $\mathcal{M}$ is not coherent. Hence, the set Π_3 of coherent conditional prevision assessments on $\mathcal{F}_3$ is a strict non convex subset of I. However, if we are searching for a (coherent) assessment $\mathcal{M} = (x, y, xy)$ which is "intermediate" between $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. such that

$$min\,\{x_1, x_2\} \leq x \leq max\,\{x_1, x_2\} \,,$$

$$min\,\{y_1, y_2\} \leq y \leq max\,\{y_1, y_2\} \,,$$

$$min\,\{x_1 y_1, x_2 y_2\} \leq xy \leq max\,\{x_1 y_1, x_2 y_2\} \,,$$

generally we can choose it in an infinite number of ways. For instance, assuming

$$x_1 < x_2 \,, \quad y_1 > y_2 \,, \quad x_1 y_1 < x_2 y_2 \,,$$

any coherent assessment $\mathcal{M} = (x, y, xy)$, such that

$$x_1 \leq x \leq x_2 \, \frac{y_2}{y_1} \,, \quad max\,\{y_1 \, \frac{x_1}{x}, \, y_2\} \leq y \leq y_1 \,,$$

satisfies the inequalities

$$x_1 \leq x \leq x_2 \,, \quad y_2 \leq y \leq y_1 \,, \quad x_1 y_1 \leq xy \leq x_2 y_2 \,;$$

hence, $\mathcal{M}$ is intermediate between $\mathcal{M}_1$ and $\mathcal{M}_2$.
In general, we can construct an infinite number of continuous curves $\mathcal{C}$ connecting $\mathcal{M}_1$ and $\mathcal{M}_2$, with $\mathcal{C} \subseteq \Pi_3$, as is shown by the following examples, where $I_1 \times I_2$ is assumed totally coherent:
(i) defining $\mathcal{M} = (x_2, y_1, x_2 y_1)$, the two segments

$$\mathcal{M}_1\mathcal{M} = \{(x, y_1, xy_1) : x = x_1 + t(x_2 - x_1),\, 0 \leq t \leq 1\} \,,$$

$$\mathcal{M}\mathcal{M}_2 = \{(x_2, y, x_2 y) : y = y_1 + t(y_2 - y_1),\, 0 \leq t \leq 1\} \,,$$

belong to Π_3. Then, the polygonal $\mathcal{C} = \mathcal{M}_1\mathcal{M} \cup \mathcal{M}\mathcal{M}_2$ is contained in Π_3 and connects $\mathcal{M}_1, \mathcal{M}_2$.
(ii) defining $\mathcal{M} = (x_1, y_2, x_1 y_2)$, the polygonal $\mathcal{C} = \mathcal{M}_1\mathcal{M} \cup \mathcal{M}\mathcal{M}_2$ is contained in Π_3 and connects $\mathcal{M}_1, \mathcal{M}_2$.
(iii) given suitable values a, b, c, let Γ be the arc of parabola defined as

$$\Gamma = \{(x, y) \in I_1 \times I_2 \,:\, y = ax^2 + bx + c\} \,.$$

Then the curve

$$\mathcal{C} = \{(x, y, z) : (x, y) \in \Gamma,\, z = xy = ax^3 + bx^2 + cx\}$$

is contained in Π_3 and connects $\mathcal{M}_1, \mathcal{M}_2$.
(iv) more in general, given a suitable interval $[t_1, t_2]$ and a continuous parameter $t \in [t_1, t_2]$, let Γ be a continuous curve contained in $I_1 \times I_2$, with parametric equations $x = x(t)\,,\, y = y(t)\,,\, t \in [t_1, t_2]$. Then, the continuous curve $\mathcal{C}$, with parametric equations

$$x = x(t)\,,\, y = y(t)\,,\, z(t) = x(t)y(t)\,,\, t \in [t_1, t_2]\,,$$

is contained in Π_3 and connects $\mathcal{M}_1, \mathcal{M}_2$.

As shown by Example 2, when there exist logical dependencies, the property of total coherence is generally lost; however, the possibility of searching for "intermediate" assessments is preserved. By generalizing Theorem 1, we will show that given any pair of coherent conditional prevision assessments $\mathcal{M}', \mathcal{M}''$, we can construct (in general, in an infinite number of ways) a continuous curve $\mathcal{C}$ connecting $\mathcal{M}', \mathcal{M}''$, such that, for every $\mathcal{M} \in \mathcal{C}$, $\mathcal{M}$ is coherent. We will see that each point $\mathcal{M}$ of $\mathcal{C}$ is an intermediate conditional prevision assessment between $\mathcal{M}'$ and $\mathcal{M}''$. We have

Theorem 4. Given n events $H_1, \ldots, H_n$ and n random quantities $X_1, \ldots, X_n$, for each $r \in J_n$ denote by $\mathcal{X}_{H_r}$ the set $\{x_{r1}, \ldots, x_{rk_r}\}$ of possible values of X_r compatible with H_r and by I_r the interval associated with $\mathcal{X}_{H_r}$, as defined by (4). Moreover, let $\mathcal{M}' = (\mu'_1, \ldots, \mu'_n)$, $\mathcal{M}'' = (\mu''_1, \ldots, \mu''_n)$ be two coherent conditional prevision assessments on the family $\mathcal{F}_n = \{X_1|H_1, \ldots, X_n|H_n\}$. Then, there exists (at least) a continuous curve $\mathcal{C}$ contained in the interval $I_{1\cdots n} = I_1 \times \cdots \times I_n$ such that for every $\mathcal{M} = (\mu_1, \ldots, \mu_n) \in \mathcal{C}$, we have:
(i) $\mathcal{M}$ is a coherent conditional prevision assessment on $\mathcal{F}_n$;
(ii) each $\mathcal{M} \in \mathcal{C}$ is a generalized convex combination of $\mathcal{M}', \mathcal{M}''$; i.e. $\min\,\{\mu'_i, \mu''_i\} \leq \mu_i \leq \max\,\{\mu'_i, \mu''_i\}$, $\forall\, i \in J_n$.

Proof. From coherence of $\mathcal{M}'$ and $\mathcal{M}''$, there exist two suitable nonnegative vectors

$$\mathcal{P}_1 = (p^{(1)}_{11}, \ldots, p^{(1)}_{1k_1}, \ldots, p^{(1)}_{n1}, \ldots, p^{(1)}_{nk_n})$$

$$\mathcal{P}_2 = (p^{(2)}_{11}, \ldots, p^{(2)}_{1k_1}, \ldots, p^{(2)}_{n1}, \ldots, p^{(2)}_{nk_n}) \,,$$

with

$$\sum_{j=1}^{k_1} p^{(1)}_{1j} = \cdots = \sum_{j=1}^{k_n} p^{(1)}_{nj} = \sum_{j=1}^{k_1} p^{(2)}_{1j} = \cdots = \sum_{j=1}^{k_n} p^{(2)}_{nj} = 1 \,,$$

which represent coherent assessments on the family

$$\{A_{i1}|H_i \,,\, \ldots \,,\, A_{ik_i}|H_i \,,\, i \in J_n\} \,;$$

that is, under the assessment $\mathcal{P}_1$ it is

$$P(A_{i1}|H_i) = p^{(1)}_{i1}, \ \ldots, \ P(A_{ik_i}|H_i) = p^{(1)}_{ik_i} \,,\ i \in J_n \,,$$

while under the assessment $\mathcal{P}_2$ it is

$$P(A_{i1}|H_i) = p^{(2)}_{i1}, \ \ldots, \ P(A_{ik_i}|H_i) = p^{(2)}_{ik_i} \,,\ i \in J_n \,;$$

moreover, $\mathcal{P}_1$ and $\mathcal{P}_2$ are such that

$$\sum_{j=1}^{k_1} p^{(1)}_{1j} x_{1j} = \mu'_1 \,, \ \ldots \,, \ \sum_{j=1}^{k_n} p^{(1)}_{nj} x_{nj} = \mu'_n \,,$$

$$\sum_{j=1}^{k_1} p^{(2)}_{1j} x_{1j} = \mu''_1 \,, \ \ldots \,, \ \sum_{j=1}^{k_n} p^{(2)}_{nj} x_{nj} = \mu''_n \,.$$

By Theorem 1, there exists a continuous curve Γ connecting $\mathcal{P}_1, \mathcal{P}_2$, with

$$\mathcal{P}^m = \mathcal{P}_1 \wedge \mathcal{P}_2 \leq \mathcal{P} \leq \mathcal{P}_1 \vee \mathcal{P}_2 = \mathcal{P}^M \,, \ \ \forall\, \mathcal{P} \in \Gamma \,.$$

Moreover, each component p_{ij} of $\mathcal{P}$ is a convex combination of the corresponding components $p^{(1)}_{ij}$, $p^{(2)}_{ij}$ of $\mathcal{P}_1, \mathcal{P}_2$, say $p_{ij} = (1 - t_{ij})p^{(1)}_{ij} + t_{ij}p^{(2)}_{ij}$, with $t_{ij} \in [0, 1]$.

Then, from coherence of $\mathcal{P}$ it follows that the conditional prevision assessment $\mathcal{M} = (\mu_1, \ldots, \mu_n) \in \mathcal{C}$ on $\mathcal{F}_n = \{X_1|H_1, \ldots, X_n|H_n\}$, where

$$\mu_i = \mathbb{P}(X_i|H_i) = \sum_{j=1}^{k_i} p_{ij} x_{ij}\,,\ i \in J_n\,,$$

is coherent too. Moreover, it is

$$\sum_{j=1}^{k_i} p_{ij} x_{ij} = (1 - t_{ij}) \sum_{j=1}^{k_i} p_{ij}^{(1)} x_{ij} + t_{ij} \sum_{j=1}^{k_i} p_{ij}^{(2)} x_{ij} =$$

$$= (1 - t_{ij})\mu_i' + t_{ij}\mu_i''\,;$$

or, equivalently,

$$\min\{\mu_i', \mu_i''\} \le \mu_i \le \max\{\mu_i', \mu_i''\}\,,\ \ i \in J_n\,.$$

Hence, $\mathcal{M}$ is a generalized convex combination of $\mathcal{M}', \mathcal{M}''$; of course $\mathcal{M} \in I_{1\cdots n}$. Finally, by moving the point $\mathcal{P}$ on the curve Γ from $\mathcal{P}_1$ to $\mathcal{P}_2$, we construct a continuous curve $\mathcal{C}$, contained in the interval $I_{1\cdots n}$, which connects $\mathcal{M}', \mathcal{M}''$. □

By Theorem 4, it follows

Corollary 2. Given n conditional random quantities $X_1|H_1, \ldots, X_n|H_n$ and any quantities $\mu_1, \ldots, \mu_{i-1}$, $l_i \le u_i$, $\mu_{i+1}\,, \ldots\,, \mu_n$, let

$$\mathcal{M}' = (\mu_1, \ldots, \mu_{i-1}, l_i, \mu_{i+1}, \ldots, \mu_n)\,,$$
$$\mathcal{M}'' = (\mu_1, \ldots, \mu_{i-1}, u_i, \mu_{i+1}, \ldots, \mu_n)\,,$$

be two conditional prevision assessments on $\{X_1|H_1, \ldots, X_n|H_n\}$. Moreover, let $I = \mathcal{M}'\mathcal{M}''$ be the segment $\{(\mu_1, \ldots, \mu_i, \ldots, \mu_n) : l_i \le \mu_i \le u_i\}$, with vertices $\mathcal{M}', \mathcal{M}''$. Then, the segment I is totally coherent if and only if $\mathcal{M}'$ and $\mathcal{M}''$ are both coherent.

Proof. The proof immediately follows by observing that in our case the interval $I_{1\cdots n}$ coincides with the segment I; therefore, the unique curve connecting $\mathcal{M}', \mathcal{M}''$ is the segment I. □

We observe that Corollary 2, which generalizes Corollary 1 to the case of conditional prevision assessments, is also an immediate consequence of the extension theorem for coherent conditional previsions.

7 Further results on total coherence

In this section we exploit the results of Section 6 to obtain some related results on total coherence of suitable sets of conditional prevision assessments. We have

Theorem 5. Given two conditional random quantities $X|H, Y|K$, let $\mathcal{M}_1 = (m_1, \mu_1)$, $\mathcal{M}_2 = (m_1, \mu_2)$, $\mathcal{M}_3 = (m_2, \mu_3)$, $\mathcal{M}_4 = (m_2, \mu_4)$ be four coherent conditional prevision assessments on $\{X|H, Y|K\}$. Moreover, let $\mathcal{C}_1, \mathcal{C}_2$ be two curves connecting, respectively, $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3, \mathcal{M}_4$, such that for every $\mathcal{M}' \in \mathcal{C}_1, \mathcal{M}'' \in \mathcal{C}_2$, both $\mathcal{M}'$ and $\mathcal{M}''$ are coherent conditional prevision assessments on $\{X|H, Y|K\}$. Then, the closed set $\mathcal{S}$, delimited by the curves $\mathcal{C}_1, \mathcal{C}_2$ and by the vertical segments $\mathcal{M}_1\mathcal{M}_2$ and $\mathcal{M}_3\mathcal{M}_4$, is totally coherent.

Proof. We need to show that, for every $\mathcal{M} \in \mathcal{S}$, $\mathcal{M}$ is a coherent conditional prevision assessment on $\{X|H, Y|K\}$. Without loss of generality we can assume: (i) $m_1 \le m_2$; (ii) for every $\mathcal{M}' = (m', \mu') \in \mathcal{C}_1$, $\mathcal{M}'' = (m'', \mu'') \in \mathcal{C}_2$, if $m' = m''$, then $\mu' \le \mu''$. For each $m \in [m_1, m_2]$ we denote by I_m the segment with vertices the points $\mathcal{M}' = (m, \mu') \in \mathcal{C}_1$, $\mathcal{M}'' = (m, \mu'') \in \mathcal{C}_2$. Then, by Corollary 2, the coherence of $\mathcal{M}', \mathcal{M}''$ implies the total coherence of I_m, for every $m \in [m_1, m_2]$. Finally, as $\mathcal{S} = \bigcup_{m \in [m_1, m_2]} I_m$, $\mathcal{S}$ is totally coherent. □

Remark 2. A particular interesting case of Theorem 5 is obtained when $\mu_3 = \mu_1\,,\ \mu_4 = \mu_2$. In this case the interval $\mathcal{I}_2 = [m_1, m_2] \times [\mu_1, \mu_2]$ is totally coherent if and only if the conditional prevision assessments $\mathcal{M}_1 = (m_1, \mu_1)$, $\mathcal{M}_2 = (m_1, \mu_2)$, $\mathcal{M}_3 = (m_2, \mu_1)$, $\mathcal{M}_4 = (m_2, \mu_2)$ are all coherent. Of course, the reasoning is the same as in the proof of Theorem 5.

More in general, we have

Theorem 6. Given a family of n conditional random quantities $\mathcal{F}_n = \{X_1|H_1, \ldots, X_n|H_n\}$, let us consider the interval $\mathcal{I}_n = [m_1, \mu_1] \times \cdots \times [m_n, \mu_n]$ associated with the imprecise conditional prevision assessment $\mathcal{A}_n$ on $\mathcal{F}_n$, defined by

$$m_i \le \mathbb{P}(X_i|H_i) \le \mu_i\,,\ \ i = 1, \ldots, n\,. \qquad (5)$$

Then, defining $\mathcal{V} = \{m_1, \mu_1\} \times \cdots \times \{m_n, \mu_n\}$, the interval $\mathcal{I}_n$ is totally coherent if and only if each vertex $V \in \mathcal{V}$ is coherent.

Proof. We set

$$\mathcal{V}' = \{m_1, \mu_1\} \times \cdots \times \{m_{n-1}, \mu_{n-1}\} \times \{m_n\}\,,$$

$$\mathcal{V}'' = \{m_1, \mu_1\} \times \cdots \times \{m_{n-1}, \mu_{n-1}\} \times \{\mu_n\}\,.$$

We observe that $\mathcal{V} = \mathcal{V}' \cup \mathcal{V}''$; moreover, $\mathcal{V}'$ and $\mathcal{V}''$ are, respectively, the sets of vertices of the intervals

$$\mathcal{I}' = [m_1, \mu_1] \times \cdots \times [m_{n-1}, \mu_{n-1}] \times [m_n, m_n]\,,$$

$$\mathcal{I}'' = [m_1, \mu_1] \times \cdots \times [m_{n-1}, \mu_{n-1}] \times [\mu_n, \mu_n]\,.$$

Of course, the total coherence of $\mathcal{I}_n$ implies the coherence of V, for every $V \in \mathcal{V}$.
Conversely, assume that V is coherent, $\forall\, V \in \mathcal{V}$. We proceed by the following steps:
1) m_1 and μ_1 are coherent, hence the interval $\mathcal{I}_1 = [m_1, \mu_1]$ is totally coherent;
2) from the coherence of $(m_1, m_2), (\mu_1, m_2)$ (resp. $(m_1, \mu_2), (\mu_1, \mu_2)$) we obtain the total coherence of the interval $[m_1, \mu_1] \times [m_2, m_2]$ (resp. $[m_1, \mu_1] \times [\mu_2, \mu_2]$); then, by reasoning as in Theorem 5, we obtain the total coherence of $\mathcal{I}_2$;

.....

n) by induction, assume that by iterating the reasoning we have obtained the total coherence of the interval $\mathcal{I}_{n-1} = [m_1, \mu_1] \times \cdots \times [m_{n-1}, \mu_{n-1}]$. The total coherence of the sets of vertices $\mathcal{V}', \mathcal{V}''$ imply the total coherence of the intervals $\mathcal{I}', \mathcal{I}''$; then, for each given point $(\pi_1, \ldots, \pi_{n-1}) \in \mathcal{I}_{n-1}$, the assessments

$$(\pi_1, \ldots, \pi_{n-1}, m_n)\,, \quad (\pi_1, \ldots, \pi_{n-1}, \mu_n)$$

are coherent. Hence, the segment

$$I_{\pi_n} = \{(\pi_1, \ldots, \pi_{n-1}, \pi_n) \,:\, m_n \leq \pi_n \leq \mu_n\}$$

is totally coherent. Finally, as

$$\mathcal{I}_n = \bigcup_{m_n \leq \pi_n \leq \mu_n} I_{\pi_n}\,,$$

we conclude that $\mathcal{I}_n$ is totally coherent. □

8 Conclusions

In the paper we have considered conditional prevision assessments on random quantities with finite sets of possible values. We have suitably extended the notions of g-coherence and total coherence, introduced in previous papers for the case of conditional probability assessments. We have remarked that the notion of g-coherence is equivalent to the avoiding uniform loss property of lower previsions introduced by Walley. We have obtained some results on total coherence of conditional prevision assessments under different assumptions for the conditioning events, by first considering the case of logical incompatibility. Then, we have examined the case of logical independence and the general case in which there exist logical dependencies among the conditioning events. We have shown that, while the property of total coherence is generally lost, the connection property is always valid. Such a property assures that, given a pair of coherent conditional prevision assessments $\mathcal{M}', \mathcal{M}''$ (representing for instance the probabilistic judgements of two different experts), we can construct (in general, in an infinite number of ways) a curve $\mathcal{C}$ whose points are intermediate coherent assessments between $\mathcal{M}', \mathcal{M}''$. Then, if the assessments $\mathcal{M}', \mathcal{M}''$ are judged "too extreme", we could use (for the decisional problem at hand) a suitable assessment $\mathcal{M} \in \mathcal{C}$. By exploiting the connection property we have obtained some theoretical results on total coherence of suitable sets of conditional prevision assessments. We have also obtained a necessary and sufficient condition of total coherence for interval-valued conditional prevision assessments. Interesting developments of the research, which were out of the scope of this paper, could be: (i) an investigation of possible applications where there are several probability assessors; (ii) a comparison with other approaches to imprecise probabilities.
Further work should also deepen the study of imprecise conditional prevision assessments by extending the results to more general random quantities.

Acknowledgements
We are grateful to the anonymous referees for their very useful criticisms and suggestions.

References

[1] Biazzo V., and Gilio A., A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments, *International Journal of Approximate Reasoning* 24, 251-272, 2000.

[2] Biazzo V., and Gilio A., On the linear structure of betting criterion and the checking of coherence, *Annals of Mathematics and Artificial Intelligence* 35, 83-106, 2002.

[3] Biazzo V., and Gilio A., Some theoretical properties of conditional probability assessments, Proc. ECSQARU'05, Barcelona, Spain, July 6-8, 2005, 775-787.

[4] Biazzo V., and Gilio A., Some theoretical properties of interval-valued conditional probability assessments, Proc. of the Fourth International Symposium on Imprecise Probabilities and their Applications (ISIPTA '05), Pittsburgh, PA, USA, July 20-23, 2005, 58-67.

[5] Biazzo V., Gilio A., and Sanfilippo G., Coherence Checking and Propagation of Lower Probability Bounds, *Soft Computing* 7, 310-320, 2003.

[6] Capotorti A., and Vantaggi B., Locally strong coherence in inference processes, *Annals of Mathematics and Artificial Intelligence* 35, 125-149, 2002.

[7] Coletti G., Scozzafava R., Probabilistic logic in a coherent setting, Kluwer Academic Publishers, 2002.

[8] Gale D., The theory of linear economic models, McGraw-Hill, New York, 1960.

[9] Gilio A., Probabilistic consistency of knowledge bases in inference systems, *Lecture Notes in Computer Science* 747 (M. Clarke, R. Kruse, and S. Moral, Eds.), Springer-Verlag, 160-167, 1993.

[10] Gilio A., Algorithms for precise and imprecise conditional probability assessments, in *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence* (Coletti, G.; Dubois, D.; and Scozzafava, R. eds.), New York: Plenum Press, 231-254, 1995.

[11] Gilio A., Ingrassia S., Totally coherent set-valued probability assessments, *Kybernetika* 34 (1), 3-15, 1998.

[12] Gilio A., Ingrassia S., Extension of totally coherent interval-valued probability assessments, Proc. IPMU'98, Paris, July 6-10, 1708-1715, 1998.

[13] Holzer S., On coherence and conditional prevision, Boll. U.M.I., Serie VI, Vol. IV-C, N. 1, 441-460, 1985.

[14] Pelessoni R., and Vicig P., A consistency problem for imprecise conditional probability assessments, in *Proc. of the Seventh Int. Conf. on "Information Processing and Management of Uncertainty in Knowledge-Based Systems" (IPMU '98)*, Paris, France, 1478-1485, 1998.

[15] Vicig P., An algorithm for imprecise conditional probability assessments in expert systems, in *Proc. of the Sixth Int. Conf. on "Information Processing and Management of Uncertainty in Knowledge-Based Systems" (IPMU '96)*, Granada, Spain, 61-66, 1996.

[16] Vicig P., Zaffalon M., Cozman F.G., Notes on "Notes on conditional previsions", Internat. J. Approx. Reason. 44, 358-365, 2007.

[17] Walley P., Statistical reasoning with imprecise probabilities, Chapman and Hall, London, 1991.

[18] Walley P., Pelessoni R., and Vicig P., Direct Algorithms for Checking Coherence and Making Inferences from Conditional Probability Assessments, *Journal of Statistical Planning and Inference*, 126(1), 119-151, 2004.

[19] Williams P.M., Notes on conditional previsions, Internat. J. Approx. Reason. 44, 366-383, 2007 (*the first version of this paper appeared as a Research Report, School of Math. and Phys. Sci., University of Sussex, U.K., 1975*).

Predicting the Next Pandemic: An Exercise in Imprecise Hazards

Miķelis Bickis
University of Saskatchewan
bickis@snoopy.usask.ca

Uğis Bickis
Phoenix OHC
bickis@sympatico.ca

Abstract

Influenza pandemics have swept the world numerous times during the last few centuries. Cases of bird flu infecting humans have prompted predictions that we are due for another pandemic soon, but skeptics dismiss such prognostications as panic caused by a misunderstanding of probability. The issue can be reduced mathematically to the question of whether the pandemic process has an increasing, constant, or decreasing hazard function. Historical data on past pandemics can be used to estimate the hazard function using imprecise probabilities, giving upper and lower predictive probabilities of an imminent pandemic, given past waiting times. In order to achieve smoother estimates of the imprecise hazard function, an autocorrelated imprecise Normal prior is proposed.

Keywords. Survival analysis, hazard function, autocorrelated prior.

1 Introduction

Observations of human cases of H5N1 avian influenza in recent years have sparked much discussion in both scientific literature and popular media about the prospects of another flu pandemic. Such pandemics have occurred several times in recent history and concern has been expressed about the prospects of another one. The most devastating occurrence was the Spanish flu of 1918, but lesser pandemics have occured since then, the most recent being an H1N1 strain in 1977 [7]. Experts disagree on the probability of an imminent pandemic, and in an attempt to elicit probabilities, the University of Iowa has even created an online market in avian influenza futures! [1].

One question that arises is whether the probability of an imminent pandemic increases the longer it has been since the last one. On the one hand, it has been argued that pandemics have historically occurred at 20 to 30 year intervals, and given that it has been 30 years since the last one, we are "due" for one. Countering that is the argument [3] that a long waiting time actually makes an imminent pandemic *less* likely since it indicates that the evolutionary course of prospective pathogens has wandered away from genotypes adapted to human transmission. Dismissing both these arguments are individuals with a little learning in probability who proclaim that the probability of a pandemic is unaffected by a long waiting time, since a pandemic is a random event.

A more sophisticated probabilistic view, of course, will acknowledge that any of the three scenarios are logically possible. Suppose that t represents the year of the last pandemic, and let $T + t$ be the year of the next one. If $t+s$ is the current year, then the relevant quantity is the discrete hazard rate

$$\begin{aligned} h(t) &= \Pr\{T + t = t + s + 1 | T + t > t + s\} \\ &= \frac{\Pr\{T = s + 1\}}{\Pr\{T > s\}}, \end{aligned} \tag{1}$$

the conditional probability that it will occur in the next year given that it has not happened yet. Equivalently, one can work with the *instantaneous hazard rate* defined as

$$\lambda(s) = \lim_{\delta \downarrow 0} \frac{\Pr\{T \le s + \delta\}}{\delta \Pr\{T > s\}}, \tag{2}$$

the two concepts being related by

$$S(t) = \Pr\{T > s\} = \exp\left(-\int_0^s \lambda(u)\, du\right) = e^{-\Lambda(s)} \tag{3}$$

and

$$h(s) = 1 - \exp\Big[\Lambda(s) - \Lambda(s+1)\Big] \approx \lambda(s). \tag{4}$$

S is called the survivor function, and Λ is the integrated hazard.

The contrary opinions expressed in the previous paragraph can now be described as believing that the haz-

ard rate is respectively increasing, decreasing, or constant. It is possible to construct probabilistic models that are consistent with any of these viewpoints. While the dynamics of viral evolution is too complex to describe by a simple model, even simplistic models exhibit increasing or decreasing or constant hazards. If the occurrence of a pandemic happens as a result of a number of steps with a strong selective drift, then the hazard will be increasing, since while we are waiting, the virus is getting closer to a pandemic state. On the other hand, if viral evolution is envisaged as a random walk in a space of genotypes then a decreasing hazard would be typical of hitting times in such processes. But if pandemics truly are like a Poisson process, then a constant hazard would be expected.

The purpose of this paper is not to delve into realistic models of viral evolution, nor to propose definitive predictions of an influenza pandemic. Rather, we will examine to what extent one can determine the nature of the hazard function for the pandemic process, based solely on the historical record of past occurrences, and show how principles of imprecise probability cast light on the uncertainty present in such estimates. We will also contrast these methods with classical statistical approaches.

2 Mathematical models and data

According to Patterson [7], influenza pandemics occurred in the following years: 1729, 1732, 1781, 1788, 1830, 1833, 1836, 1889, 1899, 1918, 1957, 1968, and 1977. Some pandemics may have lasted more than a year. We use the first year reported as indicating the beginning of the pandemic.

We consider the pandemics to be a renewal process, in which the time between occurrences are i.i.d. random variables. Thus we are assuming that after each pandemic the virulent strain dies out because of immunity and deaths of hosts, and the evolutionary process to a new strain of pandemic virulence begins anew. We are also assuming no secular trend in the intensity of the process. These assumptions are admittedly simplistic, and may be challenged. Variation of these assumptions would increase the imprecision in the estimates.

Patterson's record gives inter-pandemic periods of 3, 49, 7, 42, 3, 3, 53, 10, 19, 39, 11, and 9 years. To this data we can add the 30 pandemic-free years to the present, which becomes a censored observation.

3 Frequentist analysis

A classical approach to fitting the data is to use the Kaplan-Meier estimator [6]. This allows one to estimate the survivor function S allowing for censoring, but does not directly address the question of increasing or decreasing hazard. We can, however, fit a parametric model

$$\log \lambda(t) = \theta_1 + \theta_2 \log t, \tag{5}$$

which assumes that the interpandemic times have a Weibull distribution. Increasing, constant, and decreasing hazards correspond to positive, zero, and negative values, respectively, of the parameter θ_2.

Estimating the parameters by maximum likelihood gives the estimates $\hat{\theta}_1 = -3.329$, $\hat{\theta}_2 = 0.075$, suggesting a slightly increasing hazard. However, a likelihood ratio test finds that $\hat{\theta}_2$ does not differ significantly from zero, which some people (mistakenly) might interpret as evidence that the hazard is constant. Indeed the 95% confidence set on the parameters establishes only that $-0.44 < \theta_2 < 0.80$, indicating that the data are consistent (under this model) with decreasing, increasing, or a constant hazard. Figure 1 shows the estimated survivor functions for both the Kaplan-Meier and maximum likelihood estimates.

The hazard is more relevant than the survivor function for the predictive probability of an imminent pandemic after a waiting period. As shown in Figure 2, the estimated hazard is nearly constant at about 5%. However, the estimate has considerable uncertainty, which in classical terms is indicated by a confidence set.

Figure 2 also shows a set of hazard functions corresponding to the boundary of the 95% confidence set on the parameters. This envelope well displays the uncertainty, showing both increasing and decreasing hazards that are consistent with the data.

4 Imprecise probability models

While a confidence band on a predictive curve indicates the imprecision, it is not possible to interpret these bounds as predictive probabilities. It is difficult to explain the meaning of the upper envelope in Figure 2 in a way that is both understandable and mathematically correct. Imprecise probability bounds, on the other hand, can honestly be described as upper and lower predictive probabilities.

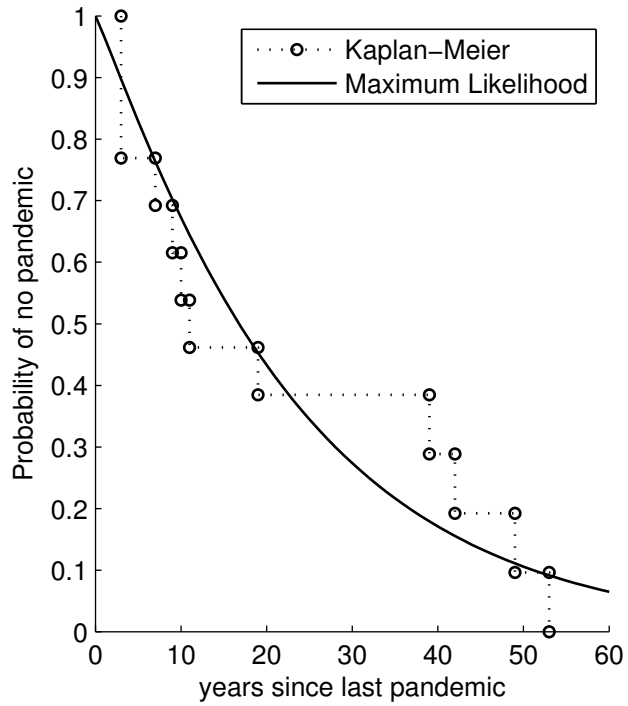


Figure 1: Kaplan-Meier and maximum likelihood estimates of survivor function for pandemic-free periods.

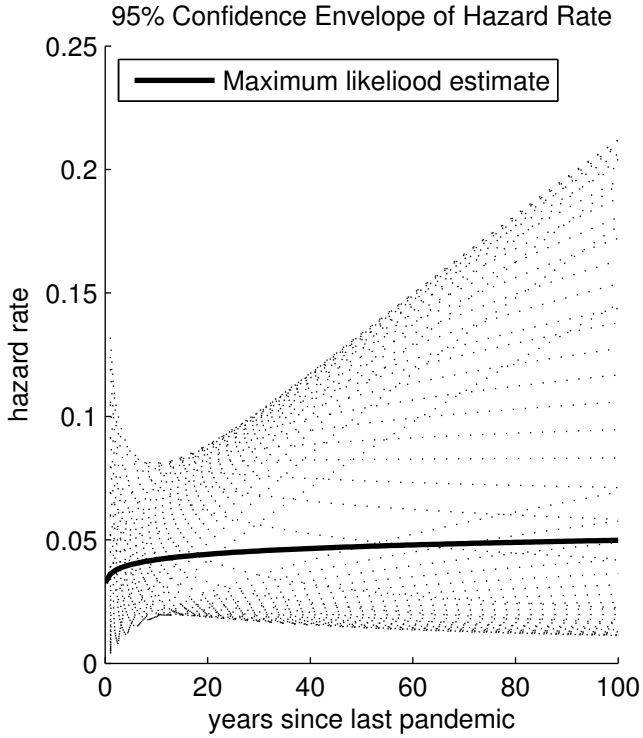


Figure 2: Maximum likelihood estimate of hazard function along with 95% confidence envelope.

4.1 Imprecise Dirichlet and product Beta models

Nonparametric estimates of survivor functions using Walley's [8] imprecise Dirichlet model were discussed by Coolen [4]. However, this model does not give useful descriptions of the hazard function, since upper and lower bounds on the survivor function do not translate into upper and lower bounds on the hazard. One can, however, estimate the hazard function directly.

Suppose that we have k time intervals (which we assume to be equally spaced). Suppose that θ_i represents the conditional probability of a failure (i.e., pandemic) by the end of the ith interval, given that there have been no failures in the preceding $i-1$ intervals. If we have data on m_{i-1} cases in which no failures have occurred, and n_i of them do fail, then this random variable will have a binomial distribution with success probability θ_i. Moreover, since failures in different intervals will be conditionally independent (given survival), the likelihood of a sample will be proportional to

$$\prod_{i=1}^{k}(1-\theta_i)^{m_i}\theta_i^{n_i}. \tag{6}$$

Conjugate to this likelihood would be a product Beta distribution. We can then use, for each interval, an imprecise Beta prior with hyperparameters $\alpha_i\nu$ and $(1-\alpha_i)\nu$ (using the notation of Bernard [2]) where α_i covers the interval $(0,1)$ to give the range of imprecise probabilities. We use the same ν for all intervals, although an argument could be made for varying it. The upper and lower predictive hazards, (i.e., the upper and lower posterior expectations of θ_i) then become $(n_i+\nu)/(n_i+m_i+\nu)$ and $n_i/(n_i+m_i+\nu)$, respectively. The upper and lower survivor functions can then be computed as

$$\overline{\hat{S}_i} = \prod_{j=1}^{i}\left(1-\frac{n_j+\nu}{n_j+m_j+\nu}\right) \tag{7}$$

$$\text{and} \quad \underline{\hat{S}_i} = \prod_{j=1}^{i}\left(1-\frac{n_j}{n_j+m_j+\nu}\right). \tag{8}$$

In the absence of censoring, $m_i = n_i + m_{i+1}$, and as $\nu \to 0$, $\overline{\hat{S}_i}$ becomes the Kaplan-Meier estimator.

Figure 3 shows the upper and lower probabilities of the survival function for both the imprecise Dirichlet model and the product Beta model, as well as the Kaplan-Meier estimator for comparison. Following the suggestion of Walley, we used a value of $\nu = 1$ as the imprecision parameter. It turns out that the upper probabilities of the Dirichlet and product Beta models are identical, whereas the Beta model gives a

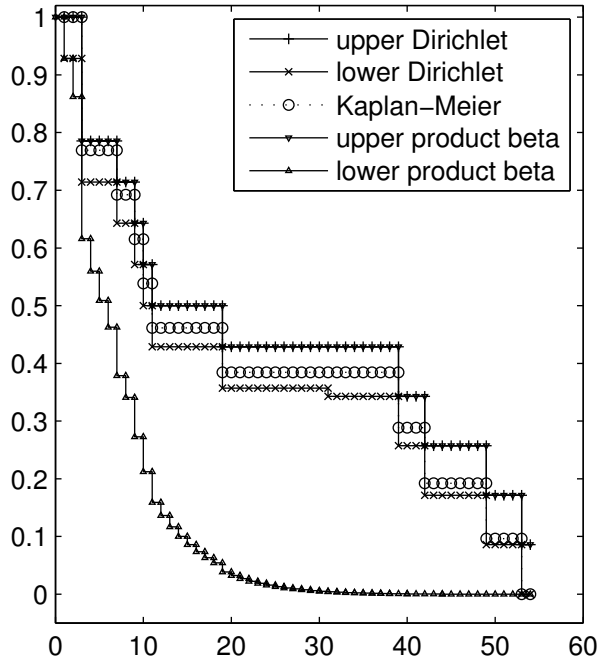


Figure 3: Upper and lower imprecise survivor function, showing both imprecise Dirichlet estimates, and product Betas estimates

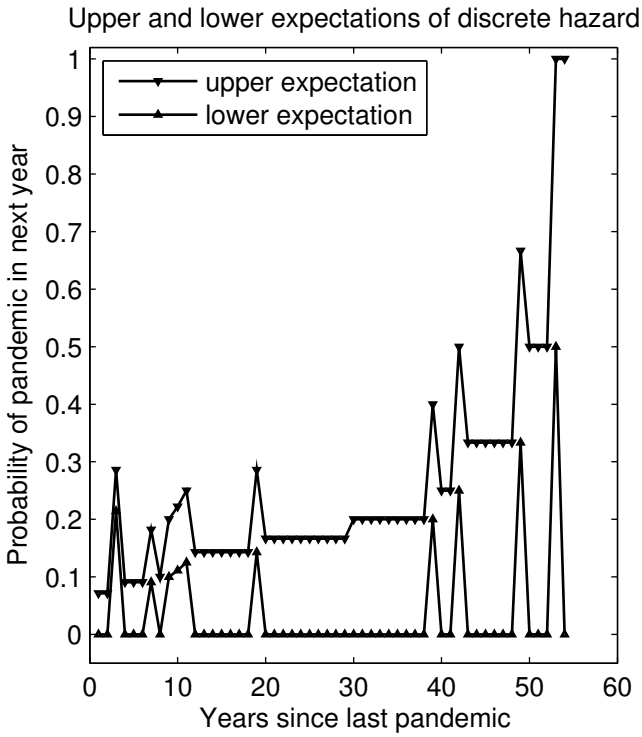


Figure 4: Upper and lower imprecise hazard function.

substantially smaller lower probability. The Kaplan-Meier estimator lies between the upper and lower probabilities, as was pointed out by Coolen [4].

Figure 4 shows the upper and lower probabilities for the hazard function. Note that when an interval has no occurrences, the lower probability is necessarily zero, while the upper probability can be quite high if the remaining sample numbers are low. The rather jagged shape of the curve can be explained by the fact that if the parameters are independent *a priori* then the form of the likelihood (6) makes them independent *a posteriori* as well.

4.2 Correlated imprecise Normal model

It would be preferable to make use of the prior information that the hazard function would be continuous and fairly smooth. We would not expect drastic changes in the probability of recurrence in a year. Thus, in place of the product Beta model, we are proposing that the prior distribution of the θ's be an autoregressive process. To make this tractable, we use a Gaussian prior on the log-odds.

Specifically, we assume that the $\omega_i = \log\big(\theta_i/(1-\theta_i)\big)$ has *a priori* a Normal distribution with mean μ and variance σ^2. Moreover, we assume that the sequence of ω_i's follow a stationary AR(1) process with autocorrelation ρ.

Using a Beta prior, the distribution of ω_i would be Fisher's-Z [5], which has lighter tails than the Normal. Thus the Normal prior tends to give somewhat less weight to extreme probabilities (which could be viewed as an advantage). Another difficulty is that the posterior distribution is harder to evaluate. Given binomial data of y successes out of n trials, the posterior distribution of ω has density

$$K(\mu,\sigma,n,y)\frac{\exp\left[-\Big(\omega-(\mu+\sigma^2 y)\Big)/(2\sigma^2)\right]}{(1+e^{\omega})^n} \qquad (9)$$

where K is a constant of integration. The posterior mean, (i.e., the predictive probability) appears not to be tractable, but can be computed numerically as

$$K\int_0^1 \exp\left[-\frac{\Big(\log\Big(\frac{\theta}{1-\theta}\Big)-(\mu+\sigma^2 y)\Big)^2}{2\sigma^2}\right]\frac{(1-\theta)^{n-1}}{\theta}\,d\theta. \qquad (10)$$

The imprecise Dirichlet model has the property that the prior probabilities are vacuous, but the posterior ones may have some precision. To achieve the same goal with the Normal model requires care. We use the

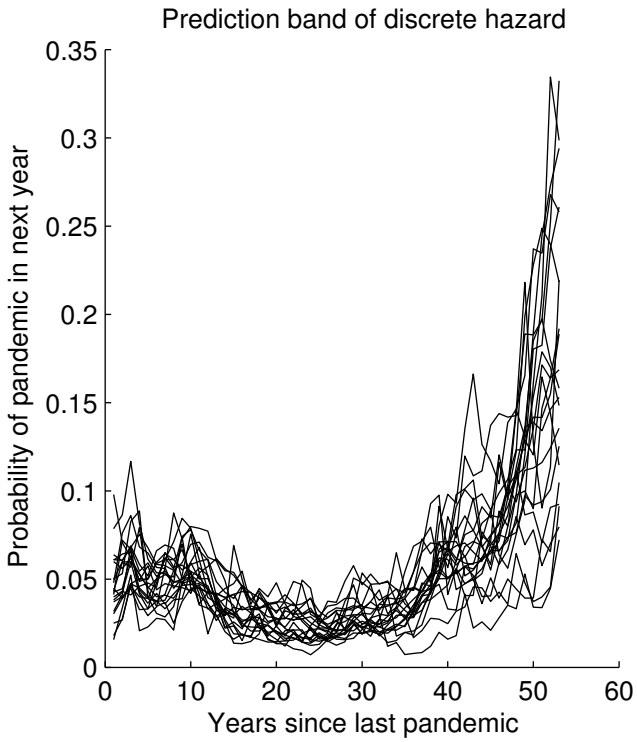


Figure 5: Sampled hazard functions from autocorrelated imprecise posterior

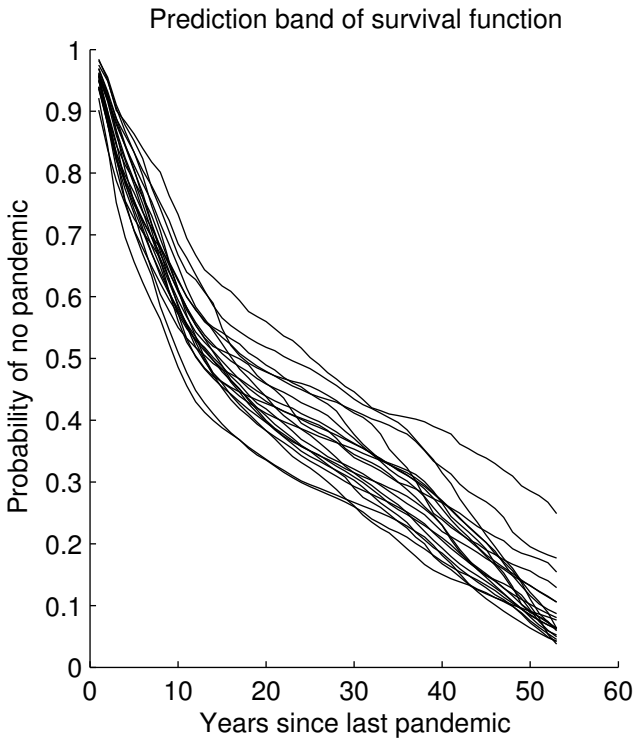


Figure 6: Sampled survival functions from autocorrelated imprecise posterior.

family of Normal distributions where

$$\sigma = \sigma_0 + \tau|\mu|^\gamma \tag{11}$$

where σ_0, τ and γ are viewed as tuning parameters. We use $\sigma_0 = 8/3$ as a value that (with $\mu = 0$) approximates the Beta$(1/2, 1/2)$ density for θ. Thus this symmetric prior distribution represents about the same level of uncertainty as a symmetric Beta distribution with $\nu = 1$. Putting $\tau = \gamma = 0.5$ and letting μ vary from $-\infty$ to ∞ appears to achieve our goal of providing upper and lower probabilities (although more work could be done here).

Extending the integral (9) to the multivariate case seemed intractable, so we estimated the smoothed hazard function using importance sampling. Letting μ vary from -8 to 2, 1000 samples were taken from a Gaussian AR(1) process with mean μ, $\rho = 0.99$ and variance given by (11). For each sample, the likelihood of the observed data was computed. These likelihoods were then used as weights in computing the predictive probabilities of both the hazard and survival functions. The results are shown in Figures 5 and 6. The bundle of curves displays the imprecision in the predictive probabilities.

5 Conclusion

From these displays we can see that although a constant hazard can (barely) fit inside this band, there is a rather strong suggestion of an increasing hazard after about 25 years. While this exercise cannot pretend to be the last word on predicting pandemics, it does show how ideas of imprecise probability can focus on realistic understanding of future risks. We hope that imprecise probability methods will be useful in other situations of estimating risks after waiting time. As extension of this work, we intend to examine how the hyperparameters of the stationary Gaussian process affect the performance of the estimates.

References

[1] Anonymous. Avian Influenza – the Iowa Health Prediction Market, web page. `http://fluprediction.uiowa.edu /fluhome/Market_AvianInfluenza.html`

[2] Jean-Marc Bernard. An Introduction to the Imprecise Dirichlet Model for Multinomial Data. *International Journal of Approximate Reasoning*, 39:123–150, 2005.

[3] Canadian Broadcasting Corporation. Scientific jury still out on prospects of avian flu pandemic, web page. `http://www.cbc.ca/health/story/2006/03/20/avian-flu060320.html`

[4] F. P. A. Coolen. An Imprecise Dirichlet Model for Bayesian Analysis of Failure Data Including Right-Censored Observations. *Reliability Engineering and System Safety*, 56:61–68, 1997.

[5] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[6] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[7] K. David Patterson. *Pandemic influenza, 1700-1900 : a study in historical epidemiology* Rowan & Littelfield, 2005.

[8] Peter Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society*, 58B:3–57, 1996.

Measuring Uncertainty with Imprecision Indices

Andrey Bronevich
Technological Institute of Southern Federal University, Taganrog, RUSSIA
brone@mail.ru

Alexander Lepskiy
Technological Institute of Southern Federal University, Taganrog, RUSSIA
lepskiy@mail.ru

Abstract

The paper is devoted to the investigation of imprecision indices, introduced in [8]. They are used for evaluating uncertainty (namely imprecision), which is contained in information, described by fuzzy (non-additive) measures, in particular, by lower and upper probabilities. We argue that there exist various types of uncertainty, for example, randomness, investigated in probability theory, imprecision, described by interval calculi, inconsistency, incompleteness, fuzziness and so on. In general these types of uncertainty have very complex behavior, caused by their interaction. Therefore, the choice of uncertainty measures is not unique, and depends on the problems addressed. The classical uncertainty measures are Shannon's entropy and Hartley's measure. In the paper imprecision indices and also linear ones are introduced axiomatically. The system of axioms allows us to define various imprecision indices. So we investigate the algebraic structure of all imprecision indices and investigate their families with best properties.

Keywords. Imprecision indices, lower and upper probabilities, uncertainty-based information.

1 Introduction

Measuring uncertainty plays a major role in uncertainty theories, in particular, probability theory, information theory, fuzzy sets theory and so on. There are some ways how to define such measures in the theory of evidence, in the theory of fuzzy (non-additive) measures and in the theory of imprecise probabilities. However, one can see that in such general theories the uncertainty measure with the best properties has not been found as yet. This situation is explained by the very complex interaction among various types of uncertainty, including randomness, inconsistence, imprecision, incompleteness of the analyzed information. We recall classical uncertainty measures, used in information theory and probability theory. Let X be a finite set of alternatives. Assigning to each alternative $x \in X$ some probability $P(\{x\})$, we have information, which is described by probability measure P, and in this case Shannon's entropy $S(P) = -\sum_{x \in X} P(\{x\}) \log_2 P(\{x\})$ can be used. Let we know only that the "true" alternative is in a nonempty set $B \subseteq X$. This situation can be described by the non-additive measure $\eta_{\langle B \rangle}(A) = \begin{cases} 1, B \subseteq A \\ 0, B \not\subseteq A \end{cases}$, $A \subseteq X$, which gives the lower probability of an event A, and Hartley's measure $H(\eta_{\langle B \rangle}) = \log_2 |B|$ can be justified. It is easily seen that in the first case uncertainty has a type that one call randomness, and the second case is more connected with imprecision of the information. The generalization of these two cases consists in the following. Consider a pair $(\underline{g}, \overline{g})$ of set functions $\underline{g} : 2^X \to [0,1]$, $\overline{g} : 2^X \to [0,1]$ defined on the powerset 2^X. We suggest that $\underline{g}(A) \le \overline{g}(A)$ for all $A \in 2^X$, $\underline{g}(\varnothing) = \overline{g}(\varnothing) = 0$, and there is a "true" probability measure P on 2^X with $\underline{g}(A) \le P(A) \le \overline{g}(A)$ for all $A \in 2^X$. In other words, set functions $\underline{g}, \overline{g}$ give us upper and lower bounds of probabilities, and for any event $A \in 2^X$ we have only the interval $[\underline{g}(A), \overline{g}(A)]$ of possible values of a "true" probability $P(A)$. In practical issues it is sufficient to define the lower probability $\underline{g}$, the upper probability can be calculated by $\overline{g}(A) = 1 - \underline{g}(\overline{A})$, where $A \in 2^X$ and $\overline{A}$ is the complement of A. Due to works of Abellan, Klir, Higashi, Harmanec and others (see [1,5,6,7]), there are two important uncertainty measures, which show the best properties in a sense of obeying axioms, which are similar to the axioms of Shannon's entropy. They are generalized Hartley's measure, and aggregate measure of

uncertainty. Let $\underline{g}$ be a belief function, i.e. it can be represented by $\underline{g} = \sum_{B \in 2^X} m(B)\eta_{\langle B \rangle}$, where $m(\varnothing) = 0$, $m(B) \geq 0$ for all $B \in 2^X$, and $\sum_{B \in 2^X} m(B) = 1$. Then generalized Hartley's measure is defined by

$$GH\left(\underline{g}\right) = \sum_{B \in 2^X \setminus \{\varnothing\}} m(B) \log_2 \left|B\right|.$$

The aggregate measure of uncertainty is calculated by

$$Au\left(\underline{g}\right) = \sup_{P \geq \underline{g}} S(P),$$

where sup is taken over all probability measures on 2^X, which are consistent with $\underline{g}$, i.e. $P(A) \geq \underline{g}(A)$ for all $A \in 2^X$. It is worth to mention that generalized Hartley's measure can be used for measuring imprecision and aggregate measure of uncertainty for total uncertainty. It is easy to check that aggregate measure of uncertainty coincides with Shannon's entropy for probability measures and with Hartley's measure for $\underline{g} = \eta_{\langle B \rangle}$, $B \neq \varnothing$.

The paper has the following structure. We remind first some definitions and results from the theory of non-additive measures and axiomatic of imprecision indices, formulated in [8]. Then we analyze so called linear imprecision indices on the set of upper and lower probabilities, giving their detailed description, and introducing their important families with symmetrical properties. We finish the paper with generalizing imprecision indices for the set of all monotone measures introducing in addition indices of inconsistency.

2 Basic definitions and problem statement

Let X be a finite set. In the sequel we will use the following notations:

1. M is the set of all real-valued set functions on the powerset 2^X;

2. $M_0 = \{g \in M \mid g(\varnothing) = 0\}$;

3. We write $g_1 \leq g_2$ for $g_1, g_2 \in M$ if $g_1(A) \leq g_2(A)$ for all $A \in 2^X$.

4. $M_{mon} \subset M_0$ is the set of all normalized monotone set functions on 2^X. It means that $g \in M_{mon}$ implies $g(\varnothing) = 0$, $g(X) = 1$, and $g(A) \leq g(B)$ if $A \subseteq B$.

5. $M_{\Pr}$ is the set of all probability measures on 2^X;

6. $M_{low} = \{g \in M_0 \mid \exists P \in M_{\Pr} : g \leq P\}$ is the set of all lower probabilities on 2^X.

6. $M_{up} = \{g \in M_0 \mid \exists P \in M_{\Pr} : g \geq P\}$ is the set of all upper probabilities on 2^X.

7. Let $g \in M$ then the dual of g is denoted by $\overline{g}$ and by definition: $\overline{g}(A) = g(X) - g\left(\overline{A}\right)$, $A \in 2^X$.

8. M_{bel} is the set of all belief functions on 2^X. Any $g \in M_{bel}$ has the following unique representation: $g = \sum_{B \in 2^X} m(B)\eta_{\langle B \rangle}$, where $m(B) \geq 0$ for all $B \in 2^X$, $m(\varnothing) = 0$, and $\sum_{B \in 2^X} m(B) = 1$.

9. M_{pl} is the set of all plausibility functions on 2^X. Any $g \in M_{pl}$ is represented uniquely by $g = \sum_{B \in 2^X} m(B)\overline{\eta}_{\langle B \rangle}$, where $m(B) \geq 0$ for all $B \in 2^X$, $m(\varnothing) = 0$, and $\sum_{B \in 2^X} m(B) = 1$.

We can consider the set M (or M_0) as a linear space w.r.t. to usual sum of set functions and usual product of set functions and real numbers. In non-additive measure theory, the basis, consisting of functions $\eta_{\langle B \rangle}$, $B \in 2^X$, is of interest. Let $g \in M$ and $g = \sum_{B \in 2^X} m_g(B)\eta_{\langle B \rangle}$ then the set function m_g is called Möbius transform of g. The function m_g is expressed by $m_g(B) = \sum_{A:A \subseteq B} (-1)^{|B \setminus A|} g(A)$. We will also use so-called dual Möbius transform of g. This transform is connected with the basis, consisting of set functions $\eta^{\langle B \rangle}$, $B \in 2^X$, defined by $\eta^{\langle B \rangle}(A) = \eta_{\langle \overline{B} \rangle}(\overline{A})$. Let $g = \sum_{B \in 2^X} m^g(B)\eta^{\langle B \rangle}$ then the set function m^g is called dual Möbius transform of g. It is calculated by $m^g(B) = \sum_{A:B \subseteq A} (-1)^{|A \setminus B|} g(A)$.

We remind now some definitions, introduced in [8].

Definition 1. A functional $f : M_{low} \to [0,1]$ is called *imprecision index* if the following conditions are fulfilled: 1) $g \in M_{\Pr}$ implies $f(g) = 0$; 2) $f(g_1) \geq f(g_2)$ for all $g_1, g_2 \in M_{low}$ such that $g_1 \leq g_2$; 3) $f\left(\eta_{\langle X \rangle}\right) = 1$.

Remark 1. We write $g_1 < g_2$ for $g_1, g_2 \in M$ if $g_1 \leq g_2$ and $g_1 \neq g_2$. Then *sensitive imprecision indices* have to obey: $f(g_1) > f(g_2)$ if $g_1, g_2 \in M_{low}$ and $g_1 < g_2$. In some works (e.g. [5,7]) there is an argumentation that uncertainty measures have to obey also subaddivity

property. Here we do not discuss this problem, because, in our opinion, this property is related to another kind of uncertainty, which can be called incompleteness of the information. However, adding the subadditivity property to the list of axioms for imprecision indices on M_{low} leads to the fact that there is no sensitive imprecision index with subadditivity property (for checking this statement you can use Example 1 in [1]).It is clear that there are many ways for defining imprecision indices. One class of them consisting of linear imprecision indices is described in the following definition.

Definition 2. An imprecision index f on M_{low} is called *linear* if for any linear combination $\sum_{j=1}^{k}\alpha_j g_j \in M_{low}$, $\alpha_j \in \mathbb{R}$, $g_j \in M_{low}$, $j=1,...,k$, we have $f\left(\sum_{j=1}^{k}\alpha_j g_j\right)=\sum_{j=1}^{k}\alpha_j f\left(g_j\right)$.

3 The investigation of linear imprecision indices

We notice first that any linear functional f on M is defined uniquely by its values on a chosen basis of M. This enables to define f by the set function $\mu_f : 2^X \to \mathbb{R}$ with the following property $\mu_f(B)=f\left(\eta_{\langle B\rangle}\right)$, $B\in 2^X$. Since any $g\in M_{low}$ is represented as a linear combination of $\left\{\eta_{\langle B\rangle}\right\}_{B\in 2^X\setminus\{\varnothing\}}$, we take by definition that $\mu_f(\varnothing)=0$ (or $f\left(\eta_{\langle\varnothing\rangle}\right)=0$) for any linear imprecision index f.

Proposition 1 [8]. *Let f be a linear imprecision index on M_{low} then $\mu_f \in M_{mon}$ with $\mu_f(\{x\})=0$ for any $x\in X$.*

The following proposition gives us the expression of any linear functional through the values of the transformed set function.

Proposition 2. *Let f be a linear functional on M then $f(g)=\sum_{B\in 2^X} m^{\mu_f}(B)g(B)$ for any $g\in M$.*

Proof. By definition $\mu_f=\sum_{B\in 2^X} m^{\mu_f}(B)\eta^{\langle B\rangle}$ and $g=\sum_{C\in 2^X} m_g(C)\eta_{\langle C\rangle}$, therefore,

$$f(g)=\sum_{C\in 2^X} m_g(C)\mu_f(C)$$
$$=\sum_{C\in 2^X} m_g(C)\sum_{B\in 2^X} m^{\mu_f}(B)\eta^{\langle B\rangle}(C)$$
$$=\sum_{B\in 2^X} m^{\mu_f}(B)\sum_{C\in 2^X} m_g(C)\eta_{\langle C\rangle}(B)$$
$$=\sum_{B\in 2^X} m^{\mu_f}(B)g(B).\blacksquare$$

The following theorem gives necessary and sufficient conditions on a linear functional to be an imprecision index through the dual Möbius transform of μ_f.

Theorem 1. *Let f be a linear functional on M then it is an imprecision index on M_{low} iff*

a) $m^{\mu_f}(X)=1$; $\sum_{D\in 2^X} m^{\mu_f}(D)=0$;

b) $\sum_{D:\, x\in D} m^{\mu_f}(D)=0$ *for all* $x\in X$;

c) $m^{\mu_f}(D)\le 0$ *for all* $D\in 2^X\setminus\{\varnothing, X\}$.

Proof. It is clear that the condition a) guarantees that $f\left(\eta_{\langle X\rangle}\right)=1$ and $f\left(\eta_{\langle\varnothing\rangle}\right)=0$. It is easy to show that b) is the necessary and sufficient condition that $f(g)=0$ for any $g\in M_{\Pr}$. Indeed, since $\eta_{\langle\{x\}\rangle}\in M_{\Pr}$ then $\mu_f(\{x\})=f\left(\eta_{\langle\{x\}\rangle}\right)=0$, $\mu_f(\{x\})=\sum_{D:\, x\in D} m^{\mu_f}(D)=0$. On the other hand, any $g\in M_{\Pr}$ can be represented as a convex sum of $\eta_{\langle\{x\}\rangle}$, i.e. $g=\sum_{x\in X} m_g(\{x\})\eta_{\langle\{x\}\rangle}$, hence,

$$f(g)=\sum_{x\in X} m_g(\{x\})f\left(\eta_{\langle\{x\}\rangle}\right)$$
$$=\sum_{x\in X} m_g(\{x\})\mu_f(\{x\})=0.$$

So b) is proved. c) is the sufficient and necessary condition of antimonotonicity of f on M_{low}. Let c) be fulfilled and $g_1\le g_2$ for $g_1, g_2\in M_{low}$ then by Proposition 2

$$f(g_1)-f(g_2)=\sum_{B\in 2^X} m^{\mu_f}(B)\left(g_1(B)-g_2(B)\right)$$
$$=\sum_{B\in 2^X\setminus\{\varnothing,X\}} m^{\mu_f}(B)\left(g_1(B)-g_2(B)\right).$$

Since $g_1(B)-g_2(B)\le 0$ for any $B\in 2^X$ and $m^{\mu_f}(B)\le 0$ for any $B\in 2^X\setminus\{\varnothing,X\}$, we get $f(g_1)\ge f(g_2)$, i.e. c) implies antimonotonicity of f. Vice versa, let f be antimonotone on M_{low} then for any $D\in 2^X\setminus\{\varnothing,X\}$ we can always find such $g_1, g_2\in M_{low}$ with $g_1(B)=g_2(B)$ for all $B\ne D$, and $g_1(D)<g_2(D)$. According to Proposition 2 $0\le f(g_1)-f(g_2)=m^{\mu_f}(D)(g_1(D)-g_2(D))$, i.e. $m^{\mu_f}(D)\le 0$. ■

Conditions of Theorem 1 can be transformed to the form, which is very close to the condition "avoiding sure loss" from the theory of imprecise probabilities [10]. It enables to get the implicit expression for an arbitrary linear imprecision index. We will further use the functions 1_B, $B\subseteq X$, on X defined by $1_B(x)=1$ if $x\in B$, and $1_B(x)=0$ otherwise.

Theorem 2. *Any linear imprecision index f on M_{low} can be uniquely represented by*

$$f(g) = 1 - \sum_{B \in 2^X} m(B) g(B),$$

where the set function m obeys the following conditions:
1) $m(\varnothing) = 0$, $m(X) = 0$, $m(B) \geq 0$ for all $B \in 2^X$;
2) $\sum_{B \in 2^X} m(B) 1_B = 1_X$.

Remark 2. The condition of "avoiding sure loss" from the theory of imprecise probabilities can be formulated with the help of the set function m from the Theorem 2 as follows: let $g \in M_0$ then $g \in M_{low}$ iff for any set function m obeying 1), 2) from Theorem 2, we have $\sum_{B \in 2^X} m(B) g(B) \leq 1$.

Theorem 3. *Let f be a linear functional on M then it is an imprecision index on M_{low} iff $\mu_f = a\mu - b\overline{\eta}_{\langle X \rangle}$, where $b > 0$, $a = 1 + b$, and $\mu \in M_{pl}$ with $\mu(\{x\}) = b/a$ for all $x \in X$.*

Proof. *Necessity.* Let f be a linear imprecision index on M_{low} then

$$\mu_f(B) = \sum_{A \in 2^X \setminus \{X, \varnothing\}} m^{\mu_f}(A) \eta_{\langle \overline{A} \rangle}(\overline{B}) + m^{\mu_f}(X) \eta_{\langle \varnothing \rangle}(\overline{B}) + m^{\mu_f}(\varnothing) \eta_{\langle X \rangle}(\overline{B}),$$

where $m^{\mu_f}(A) \leq 0$ for any $A \in 2^X \setminus \{X, \varnothing\}$ and $\eta_{\langle \varnothing \rangle} \equiv 1$, $m^{\mu_f}(X) = 1$. Let $a = -\sum_{A \in 2^X \setminus \{X, \varnothing\}} m^{\mu_f}(A)$ then taking $q(A) = -\frac{1}{a} m^{\mu_f}(\overline{A})$ for $A \in 2^X \setminus \{X, \varnothing\}$ and $m(A) = 0$ for $A \in \{X, \varnothing\}$, we get

$$\begin{aligned} \mu_f(B) &= -a \sum_{A \in 2^X} q(\overline{A}) \eta_{\langle \overline{A} \rangle}(\overline{B}) + 1 + m^{\mu_f}(\varnothing) \eta_{\langle X \rangle}(\overline{B}) \\ &= a \sum_{A \in 2^X} q(A) \left(1 - \eta_{\langle A \rangle}(\overline{B})\right) \\ &\quad - m^{\mu_f}(\varnothing) \left(1 - \eta_{\langle X \rangle}(\overline{B})\right) + m^{\mu_f}(\varnothing) + 1 - a. \end{aligned}$$

It is clear $m^{\mu_f}(\varnothing) + 1 - a = \sum_{A \in 2^X} m^{\mu_f}(A) = \mu_f(\varnothing) = 0$, hence, we get the representation required

$$\mu_f(B) = a \sum_{A \in 2^X} q(A) \overline{\eta}_{\langle A \rangle}(B) - b \overline{\eta}_{\langle X \rangle}(B),$$

where $\mu = \sum_{A \in 2^X} q(A) \overline{\eta}_{\langle A \rangle}$, $b = m^{\mu_f}(\varnothing)$, $a = 1 + b$.

It is easy to show that $\mu(\{x\}) = b/a$, $x \in X$, and $b > 0$. Actually, by Proposition 1 $\mu_f(\{x\}) = 0$ for all $x \in X$, i.e. $\mu(\{x\}) = b/a$ for all $x \in X$. On the other hand,

$$\mu_f(\{x\}) = a \sum_{A : x \in A} q(A) - b = 0,$$

i.e. $b \geq 0$ and if $b = 0$ then $q \equiv 0$ and this contradicts to the definition of imprecision index.

Sufficiency. Assume that we have the representation of μ_f from the theorem. We prove sufficiency if we check all conditions from Theorem 1. We see that $\mu_f(\varnothing) = 0$, $\mu_f(X) = 1$, and $\mu_f(\{x\}) = 0$ for all $x \in X$, i.e. conditions a), b) are true. We will further prove that $m^{\mu_f}(A) \leq 0$ for all $A \in 2^X \setminus \{\varnothing, X\}$. Since μ is a plausibility function, it is represented by $\mu = \sum_{A \in 2^X} q(A) \overline{\eta}_{\langle A \rangle}$, where $q(A) \geq 0$ for all $A \in 2^X$, $q(\varnothing) = 0$, and $\sum_{A \in 2^X} q(A) = 1$. We can write

$$\begin{aligned} \mu_f(B) &= a \sum_{A \in 2^X} q(A) \overline{\eta}_{\langle A \rangle}(B) - b \overline{\eta}_{\langle X \rangle}(B) \\ &= a \sum_{A \in 2^X} q(A) \left(1 - \eta_{\langle A \rangle}(\overline{B})\right) - b \left(1 - \eta_{\langle X \rangle}(\overline{B})\right) \\ &= a \sum_{A \in 2^X} q(\overline{A}) \left(1 - \eta_{\langle \overline{A} \rangle}(\overline{B})\right) - b \left(1 - \eta_{\langle X \rangle}(\overline{B})\right). \end{aligned}$$

The last expression implies $m^{\mu_f}(A) = -aq(\overline{A}) \leq 0$ for all $A \in 2^X \setminus \{\varnothing, X\}$, i.e. c) is also true. ■

From the proof of Theorem 3, we see that we can use the basis $\left\{\overline{\eta}_{\langle B \rangle}\right\}_{B \in 2^X \setminus \{\varnothing\}}$ of M_0 for defining other sufficient and necessary conditions on linear imprecision index. We formulate them in

Corollary 1. *Let f be a linear functional on M and $\mu_f = \sum_{A \in 2^X \setminus \{\varnothing\}} m(A) \overline{\eta}_{\langle A \rangle}$ then f is an imprecision index iff 1) $\mu_f \in M_0(X)$; 2) $\mu_f(\{x\}) = 0$ for all $x \in X$; 3) $m(A) \geq 0$ for all $A \in 2^X \setminus \{\varnothing, X\}$.*

The next theorem follows from Theorem 3.

Theorem 4. *Let f be a linear functional on M then it is an imprecision index on M_{low} iff 1) $\mu_f \in M_0$; 2) $\mu_f(\{x\}) = 0$ for all $x \in X$; 3) the set function $\mu_f^{\{x\}}$, defined by $\mu_f^{\{x\}}(B) = \mu_f(B \cup \{x\})$, $B \in 2^X$, is in M_{pl} for any $x \in X$.*

It seems to be logical in some problems that the quantity of imprecision in the situation, where we know only that the true alternative belongs to the set B, depends on $|B|$ and does not depend on other factors. In this case we assume that $f(\eta_{\langle B \rangle}) = f(\eta_{\langle C \rangle})$ or $\mu_f(B) = \mu_f(C)$ if $|B| = |C|$, and we call such linear imprecision indices symmetrical. In the sequel we will use the fact that such symmetrical monotone set functions can be viewed as distorted probabilities [3]. Let P be a probability measure on $X = \{x_1, ..., x_N\}$; let $\lambda : [0,1] \to [0,1]$ be non-decreasing function with $\lambda(0) = 0$, $\lambda(1) = 1$, then the set function $g = \lambda \circ P$ ($g(A) = \lambda(P(A))$, $A \in 2^X$) is called

distorted probability. We are interested in the case, where $P(\{x_i\}) = 1/N$, $i = 1,...,N$. Further we will use the following sufficient condition of total monotonicity [2]: let $g = \lambda \circ P$, then it is a belief function if λ is infinitely differentiable on $[0,1)$ and $d^n\lambda(t)/dt^n \geq 0$, $n = 1,2,...$, for any $t \in [0,1)$.

Theorem 5. *Let f be a linear functional on M and $\mu_f = \lambda \circ P$, i.e. μ_f is a distorted probability, mentioned above, and $P(\{x_i\}) = 1/N$, $i = 1,...,N$. Then f is an imprecision index if:* 1) $\lambda(1/N) = 0$; 2) λ *is infinitely differentiable on* $\left[\frac{1}{N},1\right)$ *and* $(-1)^{n-1} d^n\lambda(t)/dt^n \geq 0$, $n = 1,2,...$, *for any* $t \in \left[\frac{1}{N},1\right)$.

Proof. We will check that the all conditions from Theorem 4 are true. It is clear that $\mu_f \in M_0$ and $\mu_f(\{x\}) = 0$ for all $x \in X$. Now we prove that 3) is also true. In this case $\mu_f^{\{x\}}(B) = \lambda(P(B \cup \{x\}))$, $B \in 2^{X\setminus\{x\}}$, $\mu_f^{\{x\}}$ can be considered as a distorted probability on $2^{X\setminus\{x\}}$, and $\mu_f^{\{x\}} = \lambda_1 \circ P_1$, where $\lambda_1(t) = \lambda\left(\frac{1+t(N-1)}{N}\right)$, $t \in [0,1]$, and $P_1(\{y\}) = 1/(N-1)$, $y \in X \setminus \{x\}$. We find that $\overline{\mu_f^{\{x\}}}(A) = 1 - \lambda_1(P_1(\bar{A})) = 1 - \lambda_1(1 - P_1(A))$, i.e. $\overline{\mu_f^{\{x\}}} = \lambda_2 \circ P_1$ is a distorted probability and $\lambda_2(t) = 1 - \lambda_1(1-t) = 1 - \lambda\left(1 - \frac{t(N-1)}{N}\right)$. It is clear $\mu_f^{\{x\}} \in M_{pl}$ iff $\overline{\mu_f^{\{x\}}} \in M_{bel}$. Then we argue that $\mu_f^{\{x\}}$ is a plausibility function if $d^n\lambda_2(t)/dt^n \geq 0$, $n = 1,2,...$, for any $t \in [0,1)$, or $(-1)^{n-1} d^n\lambda(t)/dt^n \geq 0$, $n = 1,2,...$, for any $t \in \left[\frac{1}{N},1\right)$.■

In some cases it is suitable to define symmetrical μ_f by a non-decreasing function $\varphi : [1,+\infty) \to [0,+\infty)$ with $\varphi(1) = 0$ assuming that $\mu_f(A) = \varphi(|A|)/\varphi(|X|)$ for $A \neq \varnothing$. Then $\lambda(t) = \varphi(tN)/\varphi(N)$ for $t \in \left[\frac{1}{N},1\right]$, where $N = |X|$. It is easy to see that according to Theorem 5, μ_f determines a linear imprecision index if φ is infinitely differentiable on $[1,N)$ and $(-1)^{n-1} d^n\varphi(t)/dt^n \geq 0$, $n = 1,2,...$, for any $t \in [1,N)$.

Example 1. Let $\varphi(t) = \ln(t)$ then $\mu_f(A) = \ln(|A|)/\ln(|X|)$. In this case the corresponding linear imprecision index can be considered as the analog of generalized Hartley's measure. We see that $(-1)^{n-1} d^n \ln(t)/dt^n = (n-1)!t^{-n} \geq 0$ for $t \geq 1$, i.e. μ_f determines a linear imprecision index on M_{low}.

Example 2. Given two source of information about the object of our interest. These sources are described by lower probabilities g_1 and g_2. Assume that the pointed sources are consistent, i.e. $\max\{g_1, g_2\} \in M_{low}$. We are going to use one of the sources in further analysis. This situation may be caused that we work, for example, with necessity functions and the choice of more exact information $\max\{g_1, g_2\}$ pushes out from possibility theory. Assume that we make a choice from $\{g_1, g_2\}$ using the metric on M_{mon} defined by

$$d(g_1, g_2) = \sum_{B \in 2^X} m(B)|g_1(B) - g_2(B)|,$$

where $g_1, g_2 \in M_{mon}$, m is a weight function with $m(\varnothing) = m(X) = 0$, $m(B) > 0$ for all $B \in 2^X \setminus \{\varnothing, X\}$. We choose g_1 if $d(g_1, \max\{g_1, g_2\}) < d(g_2, \max\{g_1, g_2\})$, g_2 if $d(g_1, \max\{g_1, g_2\}) > d(g_2, \max\{g_1, g_2\})$, and if $d(g_1, \max\{g_1, g_2\}) = d(g_2, \max\{g_1, g_2\})$ then the additional analysis is needed for making a decision. Now we show how this metric is related to the notion of imprecision index. Simplifying the expression

$$d(g_1, \max\{g_1, g_2\}) - d(g_2, \max\{g_1, g_2\}) =$$
$$\sum_{B: g_2(B) > g_1(B)} m(B)(g_2(B) - g_1(B)) -$$
$$\sum_{B: g_1(B) > g_2(B)} m(B)(g_1(B) - g_2(B)) =$$
$$\sum_{A \in 2^X} m(B)(g_2(B) - g_1(B)) =$$
$$\left(1 - \sum_{B \in 2^X} m(B)g_1(B)\right) - \left(1 - \sum_{B \in 2^X} m(B)g_2(B)\right).$$

We get that the expressions $\left(1 - \sum_{B \in 2^X} m(B)g_i(B)\right)$, $i = 1,2$, can be considered as values of the linear imprecision index f if m obeys the condition 2) of Theorem 2. In this case we choose g_1 if $f(g_1) < f(g_2)$, i.e. the first source gives us more exact information than the second.

4 The algebraic structure of the set of linear imprecision indices

Let f_1, f_2 be linear functionals on M then their linear combination $f = af_1 + bf_2$, $a, b \in \mathbb{R}$ is also a linear functional. If we take into consideration set functions $\mu_{f_1}, \mu_{f_2}, \mu_f$, we see that $\mu_f = a\mu_{f_1} + b\mu_{f_2}$, i.e. the set of all linear functionals on M is a linear space and this space is isomorphic to the linear space M of all set functions on 2^X. It is easy to show that if f_1, f_2 are linear imprecision indices then their convex sum $f = af_1 + bf_2$, where $a, b \geq 0$, $a + b = 1$, is also linear

imprecision index, i.e. the set of all linear imprecision indices is a convex set. We denote by M_I the set of all set functions μ_f, which correspond to linear imprecision indices on M_{low}. One can say that we understand the algebraic structure of a convex set if we find its extreme points. The following theorem gives the necessary and sufficient condition on an arbitrary $\mu \in M_I$ to be an extreme point.

Theorem 6. *Let* $\mu \in M_I$, $\mu = \sum_{A\in\mathfrak{B}} m(A)\overline{\eta}_{\langle A\rangle} - b\overline{\eta}_{\langle X\rangle}$, *where* $\mathfrak{B} \subseteq 2^X \setminus \{\varnothing, X\}$, $m(A) > 0$ *for all* $A \in \mathfrak{B}$, $b > 0$, *then* μ *is an extreme point of* M_I *iff functions* $\{1_A\}_{A\in\mathfrak{B}}$ *are linearly independent.*

Proof. Notice first that any $\mu \in M_I$ has the representation $\mu = \sum_{A\in\mathfrak{B}} m(A)\overline{\eta}_{\langle A\rangle} - b\overline{\eta}_{\langle X\rangle}$ by Corollary 1, $b > 0$, and $\mathfrak{B}$ is not empty. Secondly, $\mu(\{x\}) = 0$ for all $x \in X$, i.e.

$$\sum_{A\in\mathfrak{B}} m(A)1_A = b1_X .$$

We will show that μ is not an extreme point of M_I iff functions $\{1_A\}_{A\in\mathfrak{B}}$ are linearly dependent. This implies evidently the theorem statement. Assume that functions $\{1_A\}_{A\in\mathfrak{B}}$ are linearly dependent. Then there exist two different solutions of $\sum_{A\in\mathfrak{B}} \alpha_A 1_A = 1_X$ w.r.t. α_A, $A \in \mathfrak{B}$. We choose one of them as $\alpha_A^{(1)} = m(A)/b$, $A \in \mathfrak{B}$. Since $\alpha_A^{(1)} > 0$ for all $A \in \mathfrak{B}$, we can choose another solution $\alpha_A^{(2)}$ with $\alpha_A^{(2)} \geq 0$, $A \in \mathfrak{B}$. Let $b_2 = 1\Big/\left(\left(\sum_{A\in\mathfrak{B}} \alpha_A^{(2)}\right) - 1\right)$, then it is easy to see that $b_2 > 0$ and the set function μ_2, defined by

$$\mu_2 = \sum_{A\in\mathfrak{B}} b_2\alpha_A^{(2)}\overline{\eta}_{\langle A\rangle} - b_2\overline{\eta}_{\langle X\rangle},$$

is in M_I. Defining

$$c = \sup\left\{r \in \mathbb{R} \mid rb_2\alpha_A^{(2)} \leq m(A), A \in \mathfrak{B}, rb_2 \leq b\right\},$$

we confirm that $c \in (0,1)$, $\mu \geq c\mu_2$. Then

$$\mu_1 = \frac{1}{1-c}(\mu - c\mu_2) = \sum_{A\in\mathfrak{B}} m_1(A)\overline{\eta}_{\langle A\rangle} - b_1\overline{\eta}_{\langle X\rangle}.$$

where $m_1(A) = \frac{1}{1-c}\left(m(A) - cb_2\alpha_A^{(2)}\right)$, $b_1 = \frac{1}{1-c}(b - cb_2)$, is in M_I. We see that $\mu = (1-c)\mu_1 + c\mu_2$, i.e. we have proved that μ is not an extreme point of M_I.

Vice versa; assume that μ is not an extreme point of M_I. Then there exist set functions $\mu_1, \mu_2 \in M_I$ such that $\mu = a\mu_1 + b\mu_2$, where $a, b > 0$ and $a + b = 1$. Since $\mu_1, \mu_2 \in M_I(X)$ we have $\sum_{A\in\mathfrak{B}} m_i(A)1_A = b_i 1_X$, where $b_i > 0$, $i = 1,2$. Therefore, the equation $\sum_{A\in\mathfrak{B}} \alpha_A 1_A = 1_X$ has more than one solution w.r.t. $\alpha_A \in \mathbb{R}$, $A \in \mathfrak{B}$, hence, functions $\{1_A\}_{A\in\mathfrak{B}}$ are linearly dependent if μ is not an extreme point of M_I .■

Theorem 6 implies that the set M_I has the finite number of extreme points. According to the Theorem by Krein-Milman [9], any $\mu \in M_I$ can be represented as a convex sum of extreme points. However, it is a very hard problem to describe such extreme points explicitly. Further we consider one convex subset of M_I, for which this problem can be solved.

Definition 3. Let f be a linear imprecision index on M_{low}, then we call it *complementarily symmetrical* if $m^{\mu_f}(A) = m^{\mu_f}(\overline{A})$ for all $A \in 2^X \setminus \{\varnothing, X\}$.

Important examples of complementarily symmetrical linear imprecision indices are primitive imprecision indices. We see that

$$\nu_B(g) = g(X) - g(B) - g(\overline{B}) + g(\varnothing),$$

$$\begin{aligned}\mu_{\nu_B}(A) &= \eta_{\langle A\rangle}(X) - \eta_{\langle A\rangle}(B) - \eta_{\langle A\rangle}(\overline{B}) + \eta_{\langle A\rangle}(\varnothing)\\ &= \eta^{\langle\varnothing\rangle}(A) - \eta^{\langle\overline{B}\rangle}(A) - \eta^{\langle B\rangle}(A) + \eta^{\langle X\rangle}(A).\end{aligned}$$

Therefore, $m^{\mu_{\nu_B}}(A) = 1$ if $A \in \{\varnothing, X\}$, $m^{\mu_{\nu_B}}(A) = -1$ if $A \in \{B, \overline{B}\}$, and $m^{\mu_{\nu_B}}(A) = 0$ otherwise. We can also express μ_{ν_B} through plausibility functions. In this case

$$\begin{aligned}\mu_{\nu_B}(A) &= 1 - \eta_{\langle\overline{B}\rangle}(\overline{A}) - \eta_{\langle B\rangle}(\overline{A}) + \eta_{\langle X\rangle}(\overline{A})\\ &= \overline{\eta}_{\langle B\rangle}(A) + \overline{\eta}_{\langle\overline{B}\rangle}(A) - \overline{\eta}_{\langle X\rangle}(A).\end{aligned}$$

By Theorem 6 it is easy to show that primitive indices ν_B, $B \in 2^X \setminus \{\varnothing, X\}$, are extreme points of M_I. Actually, it follows from the fact that functions $\{1_B, 1_{\overline{B}}\}$ are linearly independent.

The role of primitive indices for describing the set of all complementarily symmetrical linear indices shows the following theorem.

Theorem 7. *The set of all complementarily symmetrical linear indices is convex. Any complementarily symmetri-*

cal linear index can be uniquely represented by a convex sum of primitive indices.

Proof. The convexity of all complementarily symmetrical linear indices it is obvious. Now we will prove that any complementarily symmetrical linear index can be represented by a convex sum of primitive indices. Let f be a complementarily symmetrical linear index and $g \in M_{low}$ then

$$f(g) = \sum_{B \in 2^X} m^{\mu_f}(B) g(B),$$

where $m^{\mu_f}(B) = m^{\mu_f}(\overline{B})$ for all $B \in 2^X \setminus \{\varnothing, X\}$. Let $\mathfrak{D} = \{B \in 2^X \setminus \{X\} \mid x \in B\}$, $\overline{\mathfrak{D}} = \{B \in 2^X \mid \overline{B} \in \mathfrak{D}\}$ for some $x \in X$ then $\mathfrak{D} \cup \overline{\mathfrak{D}} = 2^X \setminus \{\varnothing, X\}$, $\mathfrak{D} \cap \overline{\mathfrak{D}} = \varnothing$.

$$\begin{aligned} f(g) &= m^{\mu_f}(X) g(X) + m^{\mu_f}(\varnothing) g(\varnothing) \\ &\quad + \sum_{B \in \mathfrak{D}} m^{\mu_f}(B) \big(g(B) + g(\overline{B}) \big) \\ &= -\sum_{B \in \mathfrak{D}} m^{\mu_f}(B) \big(g(X) - g(B) - g(\overline{B}) + g(\varnothing) \big) \\ &\quad + \sum_{B \in \mathfrak{D}} m^{\mu_f}(B) \big(g(X) + g(\varnothing) \big) \\ &\quad + m^{\mu_f}(X) g(X) + m^{\mu_f}(\varnothing) g(\varnothing). \end{aligned}$$

We see that $\sum_{B \in \mathfrak{D}} m^{\mu_f}(B) = \sum_{B: x \in B} m^{\mu_f}(B) - m^{\mu_f}(X) = -m^{\mu_f}(X) = -1$. The equality $\sum_{B \in 2^X} m^{\mu_f}(B) = 0$ implies that $m^{\mu_f}(\varnothing) = -\sum_{B \in \mathfrak{D}} \big(m^{\mu_f}(B) + m^{\mu_f}(\overline{B}) \big) - m^{\mu_f}(X) = 1$. Hence,

$$f(g) = \sum_{B \in \mathfrak{D}} (-1) m^{\mu_f}(B) \nu_B,$$

where $(-1) m^{\mu_f}(B) \geq 0$ for all $B \in \mathfrak{D}$, and $\sum_{B \in \mathfrak{D}} (-1) m^{\mu_f}(B) = 1$, i.e. f can be represented by a convex sum of primitive indices.

We prove that the found representation is unique if we show that system $\{\nu_B\}_{B \in \mathfrak{D}}$ of all primitive indices is linearly independent in the linear space of all linear functionals on M, or we show the same property for set functions $\{\mu_{\nu_B}\}_{B \in \mathfrak{D}}$. It is easy to see that set functions $\mu_{\nu_B} = \overline{\eta}_{\langle B \rangle} + \overline{\eta}_{\langle \overline{B} \rangle} - \overline{\eta}_{\langle X \rangle}$, $B \in \mathfrak{D}$, are linearly independent, this follows immediately from the fact that set functions $\{\overline{\eta}_{\langle B \rangle}\}_{B \in 2^X \setminus \{\varnothing\}}$ are also linearly independent in M.■

Example 3. Let $\xi : X \to \mathbb{R}$, $\max_{x \in X} \xi(x) - \min_{x \in X} \xi(x) = 1$. Then we can define the linear imprecision index by Choquet integral [4] $f(g) = \int_X \xi d\overline{g} - \int_X \xi dg$, where $g \in M_{low}$. Then $\mu_f(B) = \max_{x \in B} \xi(x) - \min_{x \in B} \xi(x)$ for $B \neq \varnothing$. It is easy to show that such defined an index f is complementarily symmetrical. It is worth to mention that in the theory of imprecise probabilities $\int_X \xi d\overline{g}$ can be viewed as an upper estimate of the expectation $E[\xi]$, and $\int_X \xi dg$ as a lower estimate of the expectation $E[\xi]$.

Example 4. Let g be a coherent lower probability, and $g(A) = \min\{P_1(A), P_2(A)\}$, where $P_1, P_2 \in M_{\Pr}$, $A \in 2^X$. Let f be a complementarily symmetrical imprecision index. Then it is easy to show that

$$f(g) = \sum_{A \in 2^X} m(A) \left| P_1(A) - P_2(A) \right|,$$

where m is a non-negative set function on 2^X with $m(\varnothing) = 0$, $m(X) = 0$ and $\sum_{A \in 2^X} m(A) = 1$. Therefore, in this case we express the value of the imprecision index through the metric

$$d(P_1, P_2) = \sum_{A \in 2^X} m(A) \left| P_1(A) - P_2(A) \right|, \; P_1, P_2 \in M_{\Pr},$$

on $M_{\Pr}$ if m has the property $m(A) + m(\overline{A}) > 0$ for all $A \in 2^X \setminus \{\varnothing, X\}$.

5 The extension of imprecision indices to the set of all non-additive measures

In this section we will try to extend the notion of imprecision index. We consider first one simple generalization of imprecision indices onto the set M_{up}.

Definition 4. A functional $f : M_{up} \to [0,1]$ is called *imprecision index* if the following conditions are fulfilled: 1) $g \in M_{\Pr}$ implies $f(g) = 0$; 2) $f(g_1) \leq f(g_2)$ for all $g_1, g_2 \in M_{up}$ such that $g_1 \leq g_2$; 3) $f\left(\overline{\eta}_{\langle X \rangle}\right) = 1$.

We call an imprecision index f on M_{up} linear if it has linear properties on M_{up}. We can define this linear functional on the set of all set functions, and we take by definition that $f\left(\eta_{\langle \varnothing \rangle}\right) = 0$.

The following proposition shows the connection between imprecision indices on M_{low} and M_{up}.

Proposition 3. *Let $f_1 : M_{low} \to [0,1]$ then f_1 is an imprecision index on M_{low} iff the functional $f_2 : M_{up} \to [0,1]$ defined by $f_2(g) = f_1(\overline{g})$, $g \in M_{up}$, is an imprecision*

index on M_{up}. *In addition,* f_1 *is a linear index on* M_{low} *iff* f_2 *is a linear imprecision index on* M_{up}.

Corollary 2. *Let* f *be a linear functional on* M *then* f *is an imprecision index on* M_{up} *iff the set function* μ^f, *defined by* $\mu^f(B) = f\left(\overline{\eta}_{\langle B\rangle}\right)$, *is in* M_1.

We see that using Proposition 3 and Corollary 2, we can formulate all results, proved for imprecision indices on M_{low}, through imprecision indices, defined on M_{up}. For example, Theorem 2 can be reformulated as follows.

Theorem 2*. *Any linear imprecision index* f *on* M_{up} *can be uniquely represented by*

$$f(g) = \sum_{B\in 2^X} m(B)g(B) - a,$$

where the set function m *obeys the following conditions:*
1) $m(\varnothing) = 0$, $m(X) = 0$, $m(B) \geq 0$ *for all* $B \in 2^X$;
2) $\sum_{B\in 2^X} m(B)1_B = 1_X$, $a = \sum_{B\in 2^X} m(B) - 1$.

Comparing Theorems 2 and 2*, we see that conditions 1), 2) are very close. If $a = 1$ then we can define an imprecision index on M_{low} and M_{up} by one linear functional. Namely, if the linear functional f defines the imprecision on M_{low}, then $-f$ defines the imprecision index on M_{up}, or $|f|$ defines an imprecision index on M_{low} and M_{up} simultaneously. In some cases, the sign of f may be useful, since it enables to check what the argument of f is: it is a lower or upper probability. If the argument g is not in $M_{low} \cup M_{up}$ we can say that g gives us rather lower estimations of probabilities than upper probabilities if $f(g) > 0$, and vice versa. In some cases, we should guarantee that $|f(g)| = |f(\overline{g})|$, in other words, the amount of imprecision is the same, if we describe uncertainty by lower or by upper probabilities. This situation is analyzed in the following proposition.

Proposition 4. *Let* f *be a linear functional on* M *and we use notations from Theorems 2, 2*. Then* $|f|$ *defines a linear imprecision index on* M_{low} *and* M_{up} *with* $|f(g)| = |f(\overline{g})|$ *for all* $g \in M_{low}$ *iff* f *is a complementarily symmetrical index on* M_{low}.

Proof. We see that $a = 1$ is the necessary condition, and this condition is fulfilled for complementarily symmetrical indices. Consider the sum

$$f(g) + f(\overline{g}) = 2 - \sum_{B\in 2^X} m(B)\overline{g}(B) - \sum_{B\in 2^X} m(B)g(B).$$

which has to be equal to zero for every $g \in M_{low}$.

$$f(g) + f(\overline{g}) = \sum_{B\in 2^X\setminus\{\varnothing,X\}} m(B)g(\overline{B}) - \sum_{B\in 2^X\setminus\{\varnothing,X\}} m(B)g(B) = \sum_{B\in 2^X\setminus\{\varnothing,X\}} \left(m(\overline{B}) - m(B)\right)g(B).$$

Since $m(\overline{B}) - m(B) = m^{\mu_f}(B) - m^{\mu_f}(\overline{B})$, for any complementarily symmetrical index $f(g) + f(\overline{g}) = 0$. We prove the proposition if we show that the condition $m(B) - m(\overline{B}) = 0$ is also necessary one. Let $g = \eta_{\langle D\rangle}$, $|D| = |X| - 1$ then $f(g) + f(\overline{g}) = m(\overline{D}) - m(D)$, i.e. $m(\overline{D}) - m(D) = 0$ for any $D \in 2^X$ with $|D| = |X| - 1$. Assume by induction the statement $m(\overline{D}) - m(D) = 0$ is true for any $D \in 2^X$ with $|D| = |X| - i$, $i = 1,...,k-1$, $k < |X| - 1$. We show that $m(\overline{D}) - m(D) = 0$ for any $D \in 2^X$, where $|D| = |X| - k$. Actually, choosing $g = \eta_{\langle D\rangle}$ with $|D| = |X| - k$, we get $f(g) + f(\overline{g}) = \sum_{D\subseteq B\subset X} \left(m(\overline{B}) - m(B)\right)g(B) = m(\overline{D}) - m(D)$, i.e. $m(\overline{D}) - m(D) = 0$ for any $D \in 2^X$ with $|D| = |X| - k$. ■

If we are going to generalize measuring imprecision for general case, i.e. imprecision indices are defined on M_{mon}, we should consider two types of uncertainty, caused by imprecision and inconsistency, and propose their interpretation. One possible interpretation consists in the following. Suppose that the set function $g \in M_{mon}$ should give us low estimates of probabilities, however, $g \notin M_{low}$. Then some of its values are greater than it is possible, and this implies that information contains some amount of inconsistency. Suppose that for measuring imprecision we use an index f on M_{low}. It seems to be logical to evaluate the amount of imprecision in g by the value

$$f_{\mathrm{Imp}}(g) = \inf_{q\in M_{low}|q\leq g} f(q).$$

We see that the functional f_{Imp} can be considered as an extension of f onto M_{mon}. Let $g \in M_{up}$ then $f_{\mathrm{Imp}}(g) = 0$, and we conclude that the amount of imprecision is equal to zero, i.e. we have exact information in our disposal, however, there is uncertainty caused by inconsistency. The amount of this uncertainty can be also evaluated. In this case we choose the same axiomatic for inconsistency index as for imprecision index for upper probabilities. Then we can measure inconsistency by

$f(\overline{g})$. If $g \in M_{mon}$ and $g \notin M_{up}$, we can introduce an inconsistency index by

$$f_{\mathrm{Inc}}(g) = \inf_{q \in M_{up} | q \geq g} f(\overline{q}).$$

We see that $f_{\mathrm{Inc}}(g) = 0$ if $g \in M_{low}$. It is clear that f_{Imp} is antimonotone on M_{mon}, i.e. $g_1 \leq g_2$ implies $f_{\mathrm{Imp}}(g_1) \geq f_{\mathrm{Imp}}(g_2)$ for $g_1, g_2 \in M_{mon}$. f_{Inc} is monotone on M_{mon}, i.e. $g_1 \geq g_2$ implies $f_{\mathrm{Imp}}(g_1) \geq f_{\mathrm{Imp}}(g_2)$ for $g_1, g_2 \in M_{mon}$. Further we will use the following notations: $g = \min\{g_1, g_2\}$ if $g(A) = \min\{g_1(A), g_2(A)\}$, for all $A \in 2^X$, $g, g_1, g_2 \in M_{mon}$. Next lemmas shows, how the problem of calculating f_{Imp} f_{Inc} can be simplified.

Lemma 1. $f_{\mathrm{Imp}}(g) = \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, g\})$.

Proof. Since $\min\{\alpha, g\} \in M_{low}$ for any $\alpha \in M_{\mathrm{Pr}}$, we conclude that $f_{\mathrm{Imp}}(g) \leq \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, g\})$. Let $q \in M_{low}$, $q \leq g$ then there is an $\alpha \in M_{\mathrm{Pr}}$ with $q \leq \alpha$. We see $q \leq \min\{\alpha, g\}$, i.e. $f_{\mathrm{Imp}}(g) \geq \inf_{\alpha \in M_{\mathrm{Pr}} | \alpha \leq g} f(\min\{\alpha, g\})$. So, there is one possibility $f_{\mathrm{Imp}}(g) = \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, g\})$. ■

The next result is proved analogously as Lemma 1.

Lemma 2. $f_{\mathrm{Inc}}(g) = \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, \overline{g}\})$.

Lemma 3. Let $g = 0.5q + 0.5\overline{q}$, $q \in M_{mon}$, then $f_{\mathrm{Imp}}(g) = f_{\mathrm{Imp}}(\overline{g})$.

Proof. It is true because $g = \overline{g}$ in this case. ■

If we take another interpretation that $g \in M_{mon}$ gives us upper estimations of probabilities then we can follow the proposed scheme for defining imprecision and inconsistency indices, assuming that $\overline{g}$ gives us lower estimates of probabilities, i.e. if f is an imprecision index on M_{low}, then in this case $f_{\mathrm{Imp}}(\overline{g})$ gives us the amount of imprecision, and $f_{\mathrm{Inc}}(\overline{g})$ gives us the amount of inconsistency. In some situations we do not know what information we have in our disposal, we know only that g gives us estimates of probabilities, and we have to decide – it is lower estimates of probabilities or upper estimates of probabilities. One way, based on an imprecision index f, defined on M_{low}, consists in the following. We can assume that in the analyzed information the amount of imprecision should be greater or equal than the amount of inconsistency. Then, calculating the value

$$f_{\mathrm{S}}(g) = \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, g\}) - \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, \overline{g}\}),$$

we suppose that g is rather lower probability than upper probability if $f_{\mathrm{S}}(g) \geq 0$, and rather upper probability then lower probability if $f_{\mathrm{S}}(g) < 0$.

Lemma 4. *Let f be a complementarily symmetrical linear imprecision index on M_{low} then $f_S(g) = f(g)$.*

Proof. Let all conditions of the lemma hold and $\mathfrak{D} = \{B \in 2^X \setminus \{X\} \mid x \in B\}$ for some $x \in X$ then by Theorem 7 $f(g) = \sum_{A \in \mathfrak{D}} m(A)(\overline{g}(A) - g(A))$, where $g \in M_{mon}$, $m(A) \geq 0$ for all $A \in \mathfrak{D}$, and $\sum_{A \in \mathfrak{D}} m(A) = 1$. Let $\alpha \in M_{\mathrm{Pr}}$, $g \in M_{mon}$ then

$$f(\min\{\alpha, g\}) - f(\min\{\alpha, \overline{g}\}) = \sum_{A \in \mathfrak{D}} m(A) q(A),$$

where $q(A) = \max\{\alpha(A), \overline{g}(A)\} - \min\{\alpha(A), g(A)\} - \max\{\alpha(A), g(A)\} + \min\{\alpha(A), \overline{g}(A)\} = \overline{g}(A) - g(A)$, i.e. $f(\min\{\alpha, g\}) = f(g) + f(\min\{\alpha, \overline{g}\})$, or

$$\inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, g\}) = f(g) + \inf_{\alpha \in M_{\mathrm{Pr}}} f(\min\{\alpha, \overline{g}\}),$$

and we get the result required. ■

Example 5. Let we have two source of information about the object of our interest in the form of possibility measures defined on the power set of the finite set X. These possibility measures are given by possibility distribution functions $\pi_i : X \to [0,1]$, $i = 1,2$, and values of the corresponding possibility and necessity measures Π_i, N_i, $i = 1,2$, are computed by formulas: $\Pi_i(A) = \max_{x \in X} \pi_i(x)$, $A \in 2^X \setminus \{\varnothing\}$, and $\Pi_i(\varnothing) = 0$; $N_i(A) = 1 - \Pi_i(\overline{A})$, $A \in 2^X$. By our assumption, the values of N_i give us lower estimates of probabilities, the values of Π_i give us lower estimates of probabilities. For our example we assume that $X = \{x_1, x_2, x_3\}$, and functions $\pi_i : X \to [0,1]$, $i = 1,2$, are given by Table 1. Combining information of these two sources, we get the measure $g = \max\{N_1, N_2\}$, which should be a lower probability by our assumption, but it is not really in M_{low} because $g(A) > g(\overline{A})$ for $A = \{x_1\}$ and $A = \{x_2, x_3\}$ (see Table 2, where values of Π_i, $i = 1,2$, g, and corresponding dual measures are shown).

	x_1	x_2	x_3
π_1	1	0.5	0.5
π_2	0.4	1	0.6

Table 1: Values of possibility distribution functions.

x_1	x_2	x_3	Π_1	Π_2	N_1	N_2	g	$\overline{g}$
0	0	0	0	0	0	0	0	0
1	0	0	1	0.4	0.5	0	0.5	0.4
0	1	0	0.5	1	0	0.4	0.4	0.5
1	1	0	1	1	0.5	0.4	0.5	1
0	0	1	0.5	0.6	0	0	0	0.5
1	0	1	1	0.6	0.5	0	0.5	0.6
0	1	1	0.5	1	0	0.6	0.6	0.5
1	1	1	1	1	1	1	1	1

Table 2: Values of monotone measures.

Now for measuring imprecision and inconsistency we will use the following imprecision indices on $M_{low}(X)$:

$$v_1(g) = \left(2^{|X|} - 2\right)^{-1} \sum_{B \in 2^X} \left|\overline{g}(B) - g(B)\right|,$$

$$v_\infty(g) = \max\left\{\left|\overline{g}(B) - g(B)\right| \mid B \in 2^X\right\},$$

$$GH(g) = \frac{1}{\ln(X)} \sum_{B \in 2^X \setminus \{\varnothing\}} m_g(B) \ln|B|.$$

Notice that v_1, GH are linear imprecision indices, and v_∞ is non-linear one. The results of measuring uncertainty by these indices are shown in Table 3.

	Imprecision			Inconsistency		
	v_1	v_∞	GH	v_1	v_∞	GH
N_1	0.5	0.5	0.5	0	0	0
N_2	0.5(3)	0.6	0.526	0	0	0
g	0.2	0.5	0.2	0.03(3)	0.1	0.0288

Table 3: Evaluation of uncertainty by imprecision indices.

6 Summary and Conclusions

Although, measuring uncertainty plays a central role in various uncertainty theories, there is no possibility to find one true uncertainty measure. This can be explained by the fact that there are many various types of uncertainty, they have different interpretations; it is very difficult to understand their mutual interaction. One way for overcoming this problem is to find families of suitable uncertainty measures, satisfying some justified properties. The choice of the best uncertainty measure considerably depends on the problem solved. In this paper we have proposed how imprecision can be measured if uncertain information is described by monotone measures, in particular lower or upper probabilities. We have treated the case, where uncertainty consists of some randomness, imprecision, and inconsistency. The introduced axiomatics enables us to give detailed description of linear imprecision indices, and investigate some of them with symmetrical properties.

Acknowledgements

The authors express their sincere thanks to the anonymous reviewers, whose excellent work allows to increase the paper's quality. Dr. Bronevich on his behalf expresses his gratitude to the Fulbright Program, Binghamton University, and Prof. George Klir for research opportunity provided.

References

[1] J. Abellan, G.J. Klir. Additivity of uncertainty measures on credal sets. *International Journal of General Systems,* 34: 691–713, 2005.

[2] A.G. Bronevich. On the closure of families of fuzzy measures under eventwise aggregations. *Fuzzy sets and systems*, 153: 45 – 70, 2005.

[3] A. Chateauneuf. Decomposable capacities, distorted probabilities and concave capacities. *Mathematical Social Sciences*, 31: 19 – 37, 1996.

[4] D. Denneberg *Non-additive measure and integral.* Dordrecht, Kluwer, 1997.

[5] D. Harmanec, G.J. Klir. Measuring total uncertainty in Dempster-Shafer theory: A novel approach. *International Journal of General Systems*, 22: 405 – 419, 1994.

[6] M. Higashi, G.J. Klir. Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems*, 9: 43 – 58, 1983.

[7] G.J. Klir. *Uncertainty and information: foundations of generalized information theory.* John Wiley & Sons, Inc., 2006.

[8] A.E. Lepskiy, A.G. Bronevich. An axiomatic approach to the definition of imprecision index of fuzzy measures. *In Proc. of the second International scientific seminar "Integrative models and soft computing in artificial intelligence"*, Science Press of mathematical literature, Kolomna, 2003, pp. 127-130. (in Russian)

[9] R.R. Phelps. *Lectures on Choquet's theorem.* Springer-Verlag, Heidelberg, 2000.

[10] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman & Hall, London, 1991.

Credal Nets with Probabilities Estimated with an Extreme Imprecise Dirichlet Model

A. Cano, M. Gómez-Olmedo, S. Moral
Dpto. Ciencias de la Computación
Universidad de Granada
18071 - Granada (Spain)
(acu,mgomez,smc)@decsai.ugr.es

Abstract

The propagation of probabilities in credal networks when probabilities are estimated with a global imprecise Dirichlet model is an important open problem. Only Zaffalon [21] has proposed an algorithm for the Naive classifier. The main difficulty is that, in general, computing upper and lower probability intervals implies the resolution of an optimization of a fraction of two polynomials. In the case of the Naive credal classifier, Zaffalon has shown that the function is a convex function of only one parameter, but there is not a similar result for general credal sets. In this paper, we propose the use of an imprecise global model, but we restrict the distributions to only the most extreme ones. The result is a model giving rise that in the case of estimating a conditional probability under independence relationships, it can produce smaller intervals than the global general model. Its main advantage is that the optimization problem is simpler, and available procedures can be directly applied, as the ones proposed in [7].

Keywords. Locally specified credal networks, global imprecise Dirichlet model, propagation algorithms, probability trees.

1 Introduction

Credal networks [12] are an extension of *Bayesian networks* where instead of having a joint precise global probability distribution we have a closed and convex set of possible distributions (a *credal set* [15]). This credal set produces a conditional credal set for each variable given its parents. There are two basic possibilities:

- The credal net is *separately specified* [12], i.e. the set of joint probability distributions is obtained by specifying a credal set of conditional probability distributions for each variable and each configuration of its parents, and then the joint credal set is the convex hull of the probability distributions obtained by multiplying the conditional probability distributions resulting by selecting one element from each conditional credal set (the joint credal set is the *strong extension* of the local conditional credal sets [12]).
- The credal net is *globally specified*, when only the joint credal set is given.

Most of the effort to design algorithms for computation in credal networks has been devoted to the case of separately specified credal nets. In general, this computation is equivalent to the resolution of a combinatorial optimization problem. One of the most promising approaches is based on the branch-and-bound technique [17, 7]. Also, there are several approximate algorithms, as the ones based on the simulated annealing technique [6] or the ones based on making the variables binaries in order to apply the efficient 2U algorithm [13, 3].

There is less work for globally specified credal networks. Preliminary models were proposed by Cozman [11, 12], but he followed a robust statistics methodology, considering credal sets that were neighborhoods of standard Bayesian networks. Recently, Antonucci and Zaffalon [2] have proposed a general method based on the use of auxiliary variables as in [5] to transform a globally specified credal network into a separately specified one. This allows the application of existing algorithms for separately specified networks to cases that initially were non-separately given.

However, the Antonucci and Zaffalon [2] transformation can not directly solve some important imprecise networks that can arise in practice. This is the case of credal nets in which conditional probabilities are estimated from a database of observations with an imprecise global Dirichlet model (IDM) [20]. The main problem is that in this situation we need auxiliary variables with infinite values (as the parameters can

have values in a continuum). If the IDM is locally applied to each conditional probability distribution (we consider a different IDM for each variable and each configuration of its parents), then there is no problem, as only the extreme parameters are relevant, and we can apply the transformation by Cano, Cano, and Moral [5]. This local application was initially proposed in Zaffalon [22]. But its main difficulty was that it has a tendency to produce too wide intervals that are too uninformative. For this reason Zaffalon [21] proposed[1] a global application of the IDM. This application has the problem that to compute lower and upper conditional probabilities, it is necessary the resolution of an optimization of a fraction of two polynomials in several parameters. In the case of the Naive credal classifier, Zaffalon [21] has shown that the function is a convex function of one parameter, and he proposes a numerical method for its optimization, but there is not a similar result for general networks.

In this paper, we propose the use of an imprecise global model, but we restrict the class IDM to the set of its extreme distributions. The result is a model giving rise to the same upper and lower probabilities, when estimating the uncertainty of a future simple event, but in the case of estimating a conditional probability under independence relationships, it can provide smaller intervals. Its main advantage is that the optimization problem is simpler, being possible to express the problem as a locally specified credal network for which standard algorithms for separately specified networks can be applied. In order to make the representation more efficient, we will represent conditional probability tables as probability trees as the ones used in [6, 7].

The paper is organized as follows: in Section 2 the basic concepts of credal sets and credal networks are given; in Section 3 we consider the IDM applied to estimating the probabilities in a credal network and introduce the extreme IDM; in Section 4 we show the transformation of a credal network with probabilities estimated with an IDM model (the general or the extreme one) into a locally specified credal network; in Section 5 the results of some preliminary experiments are shown; and finally Section 6 is devoted to the conclusions.

2 Credal Networks

Let $\mathbf{X}$ be a set of variables. Let us assume that each variable $X \in \mathbf{X}$ takes values on a finite set Ω_X (the frame of X). We shall use x to denote a generic value of X, $x \in \Omega_X$. If $\mathbf{Y} \subseteq \mathbf{X}$, then this variable will take values on he Cartesian product $\prod_{X \in \mathbf{Y}} \Omega_X$, denoted by $\Omega_{\mathbf{Y}}$. The elements of $\Omega_{\mathbf{Y}}$ are called *configurations* of $\mathbf{Y}$ and will be written as $\mathbf{y}$.

A *credal set* about $\mathbf{Y}$ is a closed and convex set of probability distributions on $\Omega_{\mathbf{Y}}$, denoted as $K_{\mathbf{Y}}$. If the number of extreme points is finite, then this convex set will be given by enumerating its extreme points: $K_{\mathbf{Y}} = \mathrm{CH}(\{P_1, \ldots, P_l\})$, where CH stands for the convex hull.

A *credal network* about variables $\mathbf{X}$ is a directed acyclic graph G, with a node for each $X \in \mathbf{X}$ and a credal set $K_{\mathbf{X}}$ such that every extreme distribution $P \in \mathrm{Ext}(K_{\mathbf{X}})$, factorizes according to the graph:

$$P(\mathbf{x}) = \prod_x P(x|\pi_X(\mathbf{x})) \tag{1}$$

where $\mathbf{\Pi}_X$ is the set of parents of X in G and $\pi_X(\mathbf{x})$ the configuration of these parents corresponding to $\mathbf{x}$.

A credal network is said to be *separately specified* [16] if the global credal set $K_{\mathbf{X}}$ can be obtained by giving a credal set, $K(X|\pi_X)$, for each variable X and each configuration of its parents π_X and then obtaining all the possible joint probabilities by expression (1).

A *locally specified credal set* [2] about $\mathbf{X}$ is composed of the following elements:

- The set of variables $\mathbf{X}$.
- An additional set of auxiliary variables which Antonucci and Zaffalon [2] call decision variables $\mathbf{D}$. Each variable $D \in \mathbf{D}$ takes values in a set Ω_D.
- A directed acyclic graph G with a node for each variable in $\mathbf{X} \cup \mathbf{D}$.
- A precise conditional probability distribution for each variable $X \in \mathbf{X}$ conditioned to its parents $\mathbf{\Pi}_X$ in G.
- A set $R_D(\pi_D) \subseteq \Omega_D$ for each decision variable D and each configuration of its parent variables π_D in G.

A locally specified credal net can define a credal net about $\mathbf{X} \cup \mathbf{D}$ and another about $\mathbf{X}$, by marginalization. These can be obtained by the following procedure:

- Consider for each $D \in \mathbf{D}$ the family of *decision functions* $f_D : \Omega_{\mathbf{\Pi}_D} \to \Omega_D$, such that $f_D(\pi_D) \in R_D(\pi_D), \forall \pi_D \in \Omega_{\pi_D}$.

[1] In the acknowledgments of the paper it is said that the model was suggested to him by Peter Walley

- Consider the set of *strategies*, where an strategy is given by a decision function for each decision variable.

- Each strategy defines a precise probability distribution: the one obtained by factorization according to G and given by the precise probability conditional distributions of each variable $X \in \mathbf{X}$ and the degenerate conditional probability distributions given by the decision functions of the decision variables of the given strategy: $P(d|\pi_D) = 1$, if $d = f_D(\pi_D)$ and 0, otherwise.

- The credal set about $\mathbf{X} \cup \mathbf{D}$ is the convex hull of the probabilities defined from the set of strategies. As the extreme points of this credal set factorize according to G, they define a credal network.

- The credal set about $\mathbf{X}$ is the one obtained by marginalization of the credal set about $\mathbf{X} \cup \mathbf{D}$ (equivalent to marginalizing each one of the probabilities). This set factorizes on the graph G' on $\mathbf{X}$, obtained by deleting nodes in $\mathbf{D}$ and connecting with arcs the parents of each decision node with all its children (if a decision node D has as parent another decision node D', then we also have to make a connection from the parents of D' to the children of D, and this also recursively applies to the parents of D').

The advantage of having a credal set about $\mathbf{X}$ locally expressed is that we can solve the computation of upper and lower conditional probabilities, or the dominance relationship [19], by means of an optimization problem in the set of strategies. If the sets Ω_D are finite, then both approximate [13, 3, 6] and exact [17, 7] algorithms able of solving medium size problems are currently available[2].

3 Credal Networks from the Imprecise Dirichlet Model

In this section we consider that we have a set of variables $\mathbf{X}$, a graph G and a database with N cases in which all the variables are observed (there are no missing data). For each configuration $\mathbf{y}$ of a subset of variables $\mathbf{Y} \subseteq \mathbf{X}$ we can measure the absolute frequency of it in the database $N_{\mathbf{y}}$. We want to estimate a credal set for graph G from the observations in the database.

[2]Most of these algorithms have been initially developed for separately specified nets, but with some small modifications they can be applied to locally specified ones. For example, for the model of this paper we did not need any modification of the algorithm in [7].

The *Imprecise Dirichlet Model* (IDM) [20] was introduced for estimating probability values from a set of observations and has been extensively used by its good theoretical properties and performance in experiments.

Assume the case of one variable X, if we want to estimate the probabilities $P(x)$ with the precise Dirichlet model, we have to assume a vector of positive parameters $(\alpha_x)_{x\in\Omega_X}$. The value $S = \sum_{x\in\Omega_X} \alpha_x$ is called the *equivalent sample size*. If we denote the probability $P(x)$ by θ_x, then the Dirichlet density is proportional to $\prod_x \theta_x^{\alpha_x - 1}$. In these conditions the estimation of the probability $P(x)$ for a future event is equal to $\frac{(N_x+\alpha_x)}{(N+S)}$ (the expected value of the posterior probability given the data).

The Imprecise Dirichlet Model (IDM) considers a set of prior distributions, those obtained by fixing a global sample size S and considering all the vectors of positive parameters $(\alpha_x)_{x\in\Omega_X}$ such that $S = \sum_{x\in\Omega_X} \alpha_x$. This gives rise to an interval estimation (corresponding to all the possible vectors compatible with a given S) of $P(x)$, which is given by:

$$\left[\frac{N_x}{N+S}, \frac{N_x+S}{N+S}\right] \tag{2}$$

Usually a parameter S in the interval $[1, 2]$ is considered, and recently some authors as Bernard [4] advocates for the use of $S = 2$.

When applying the IDM to obtain the credal set of a credal network, this can be done in two ways: local or global. In the local application we obtain a separable credal network. What we do is to apply an IDM to each variable X and each configuration of its parents π_X, considering only the part of the database compatible with configuration π_X, i.e. the cases that for variables $\mathbf{\Pi}_X$ have the same values than in configuration π_X. Then we obtain a local set for each variable and each configuration of its parents: the probabilities satisfying the intervals in equation (2) where the frequencies are measured in the restricted database (same values than the configuration of parent variables). The global credal set is obtained by strong extension (the convex hull of the set of all the probabilities equal to the multiplication of a conditional probability distribution for each variable given its parents, where this conditional probabilities are selected from the local conditional credal sets).

This was the method initially employed, but soon it was noticed that it can produce too wide posterior intervals [21], and a small imprecision in all the conditional probabilities can give rise to high degrees of imprecision in conditional probabilities that are a

function of all these conditional probabilities.

The other possible application is the global one [21]. According to it, the credal set is obtained by considering a global application of the IDM to all the variables $\mathbf{X}$. We consider the credal set given by the probabilities obtained from the IDM and that factorize according to G. When a global precise Dirichlet model is applied to $\mathbf{X}$ with parameters $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{X}}}$, then the estimated probabilities for any conditional probability of a variable X conditional to a configuration $\mathbf{Y}=\mathbf{y}$, coincides with the ones obtained with a Dirichlet model with a vector of parameters$(\alpha_{x,\mathbf{y}})_{x\in\Omega_X}$ which can be obtained from the original vector by adding in the non participating variables. If $\mathbf{Z}$ are the variables in $\mathbf{X}-\mathbf{Y}-\{X\}$, then we have that $\alpha_{x,\mathbf{y}}=\sum_{\mathbf{z}}\alpha_{x,\mathbf{y},\mathbf{z}}$.

When considering a global application of the IDM, the set of all conditional probability distributions for each single variable X given a configuration of its parents π_X is the same than in the local application. But in the global application restrictions in the parameters used in the different conditional probabilities. Imagine that we have two binary variables, X and Y, and that X is a parent of Y. If for the marginal of X we use a Dirichlet distribution with parameters $(\alpha_{x_1},\alpha_{x_2})$, then for the conditional probability of Y given $X=x_1$, we have to use a Dirichlet distribution with parameters $(\alpha_{y_1},\alpha_{y_2})$ with $\alpha_{y_1}+\alpha_{y_2}=\alpha_{x_1}$ and for the conditional probability of Y given $X=x_2$, the parameters has to verify $\alpha_{y_1}+\alpha_{y_2}=\alpha_{x_2}$. So the parameters use in one variable impose restrictions in the parameters used in the rest of variables. As a consequence, the joint credal set is not the one obtained by selecting an arbitrary conditional probability for each variable given its parents and multiplying them. We have to take into account the existing restrictions between the parameters which impose restrictions into the conditional probabilities for the different variables.

In the following section, we will show that it is possible to locally express the associated credal net, but there is an important problem: the decision variables are continuous and so we have to solve an optimization problem with continuous variables, which is not simple in general, and for which we do not know any paper reporting an implementation of a general algorithm to compute upper or lower conditional probabilities. Only Zaffalon [21] has reported an algorithm for the case of a Naive graph to compute the dominance relationship.

What we propose here is a modification of the IDM model that we will call the *extreme IDM*. In the extreme IDM, instead of considering all the prior Dirichlet with $S=\sum_{x\in\Omega_X}\alpha_x$ for a given S, only the extreme ones are considered: one for each $x_0\in\Omega_X$ given by parameters $(\alpha_x)_{x\in\Omega_X}$, where $\alpha_x=S$, if $x=x_0$ and 0.0, otherwise. This density will be called the *extreme density* concentrated in value x_0 with sample size S. These prior densities on the parameters are *improper* densities, i.e. their integral is not equal to 1.0, but infinite. Their use has been justified by the estimation they produce of the posterior probabilities after a sample. Some of them are the limit of proper density functions and have a simpler interpretation. Above density can be considered as the limit when ϵ approaches to 0 of the densities with parameters $(\alpha^{\epsilon}_x)_{x\in\Omega_X}$, where $\alpha^{\epsilon}_x=S$, if $x=x_0$ and ϵ, otherwise. The estimation of the future probabilities will be the limit of the estimation with the proper densities when epsilon tends to 0. When we consider the parameters $\alpha_x=0.0,\forall x\in\Omega_X$, the estimation we obtain for future probabilities coincide with the maximum likelihood estimation (relative frequencies), i.e. $P(x)$ is estimated by N_x/N.

The main fact about the new model is that instead of considering all the infinite densities determined by a simple size S, we only consider the extreme ones, in which all the sample size is concentrated in only one element[3]. This gives rise to one density for each one of the possible value of X.

When considering the extreme IDM for the estimation of future probabilities of a single variable X, what we obtain as estimation for $P(x)$ is the same interval than in formula (2). This is immediate, as the upper and lower limits of the intervals are obtained in the extreme densities. The densities in which the parameters are not concentrated in only one point, produce inner values of the intervals (2).

However, in a credal net we can have differences as we take into account the independence relationships represented by the graph. In general, we obtain intervals which are included into the intervals associated to the use of the global original IDM.

The global application of the extreme IDM with parameter S to a graph G and set of variables $\mathbf{X}$ is given by the credal set which is equal to the convex hull of all the probability distributions that factorizes according to the graph with conditional distributions obtained in the following way:

1. Consider a value $\mathbf{x}_0\in\Omega_{\mathbf{X}}$.
2. For each variable X and each conditional configuration of its parents π_X, estimate the probability distribution of X given this configuration in the following way:

[3]We consider that the use of improper densities is not essential for the extreme model.

(a) If the configuration π_X coincides with $\mathbf{x}_0$ in the set of parents of X then $P(x|\pi_X)$ is equal to $\frac{N_{x,\pi_X}+S}{N_{\pi_X}+S}$ if the value of X in configuration $\mathbf{x}_0$ is equal to x, and equal to $\frac{N_{x,\pi_X}}{N_{\pi_X}+S}$, otherwise; where N_{π_X} is the frequency of configuration π_X in the sample, and N_{x,π_X} the frequency of cases in which we have configuration π_X and $X = x$ in the sample.

(b) If the configuration π_X does not coincide with $\mathbf{x}_0$ in the set of parents of X then $P(x|\pi_X)$ is equal to $\frac{N_{x,\pi_X}}{N_{\pi_X}}$.

What we do is to consider all the extreme densities, one for each value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ given by parameters $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{X}}}$, where $\alpha_{\mathbf{x}_0} = S$ and 0.0, otherwise. With this vector of parameters, all the conditional probabilities are estimated. For a variable, X, and a configuration of its parents, π_X, if $\mathbf{x}_0$ coincides with this configuration in the set of parents of X, then the conditional probability about X is estimated with the extreme density concentrated in the value of X in configuration $\mathbf{x}_0$ with parameter S. If $\mathbf{x}_0$ does not coincide with this configuration in the set of parents of X, then we have to estimate the conditional probability with a vector of values which are all equal to 0.0, i.e. we apply maximum likelihood estimation. If applying the maximum likelihood estimation $N_{\pi_X} = 0$, then the estimation of the probability is not defined. We will consider the uniform distribution in this case[4].

Example 1 *We are going to show the differences between the global IDM model and the extreme IDM model in a very sample case.*

Assume three binary variables X, Y, Z and a single credal network in which X is a parent of Y and Z (as a Naive Bayes in which X is the root node). Consider a sample of size equal to 2 with observations:

X	Y	Z
x_1	y_1	z_1
x_2	y_1	z_1

Assume that we apply the extreme global IDM with global sample size $S = 2$ to estimate the conditional probabilities and we want to compute the upper probability of $X = x_1$ given that $Y = y_1, Z = z_1$. This probability, $P(x_1|y_1, z_1)$ is obtained by maximizing the result of Bayes rule, that taking into account the existing conditional independence relationships can be expressed as:

$$\frac{P(y_1|x_1).P(z_1|x_1).P(x_1)}{P(y_1|x_1).P(z_1|x_1).P(x_1) + P(y_1|x_2).P(z_1|x_2).P(x_2)}$$

The upper value with extreme prior densities is obtained when these probabilities are estimated with parameters $\alpha_{x_1,y_1,z_1} = 2$ and 0.0 otherwise, and the value of the upper probability is 0.75 (the value is obtained by estimating the probabilities with relative frequencies from a sample obtained from the original one by adding two cases in which $X = x_1, Y = y_1, Z = z_1$). This upper limit can be also obtained with another extreme parameter: $\alpha_{x_2,y_2,z_2} = 2$ and 0.0 otherwise.

If we consider the global IDM model, then more sets of parameters are allowed, and not only those concentrated in only one configuration of values. In particular, we can have $\alpha_{x_1,y_1,z_1} = 1, \alpha_{x_2,y_2,z_2} = 1$ and 0.0 otherwise. If we compute the conditional probability using this set of parameters (using relative frequencies to a sample in which two new cases are added: one in which $X = x_1, Y = y_1, Z = z_1$ and other in which $X = x_2, Y = y_2, Z = z_2$) we obtain a value of 0.8, which is the upper limit of the interval. So, in this case, when applying the global model, the upper limit is greater than when using the restricted model.

To give an idea of the differences between the two models, let us generalize above situation: imagine that we have a Naive Bayes model with X as root node and a number n of children variables ($n = 2$ in previous case). Assume that we also have a sample of size 2 similar to the above one (one in which $X = x_1$ and another in which $X = x_2$ and in both of them the first case of the remaining variables is observed), and that we want to compute the upper probability of $X = x_1$ conditioned to the first case of each variable. With variable, there is no difference between the models. With $n = 3$ the difference is very small, and with $n \geq 4$ both models produce again the same result.

4 Local Specification

In this section we will show that credal networks estimated with the IDM can be locally specified. First we will start with the complete model in which it will be necessary to use decision variables with infinite values.

Given a credal network with graph G learned with the IDM with global sample size S, we will consider the following credal network:

- For each variable X with parents $\mathbf{\Pi}_X$ in the graph, consider a decision variable D_X, which will be a parent of X. This variable

[4] Any conditional probability distribution will give rise to the same joint distribution, as these values are going to be multiplied by 0.0.

will have as set of values the set of vectors $(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}$, where $\alpha_{x,\pi_X} > 0$ and $\sum_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}} \alpha_{x,\pi_X} = S$.

- For each configuration π_X and vector $(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}$, the conditional probability of X is given by:

$$P(x|\pi_X,(\alpha_{x,\pi_X})_{x\in\Omega_X,\pi_X\in\Omega_{\mathbf{\Pi}_X}}) = \frac{N_{x,\pi_X}+\alpha_{x,\pi_X}}{N_{\pi_X}+S_{\pi_X}}$$

  where $S_{\pi_X} = \sum_{x\in\Omega_X} \alpha_{x,\pi_X}$.

- Consider an order of the variables which is compatible with the graph G. For each variable, X, in this order consider the set $\mathbf{T}_X = \mathbf{\Pi}_X \cup \{X\}$. Compute the intersections $\mathbf{R}_{X,Y} = \mathbf{T}_X \cap \mathbf{T}_Y$ with all the variables Y preceding X in the graph. Make as parents of D_X all the variables D_Y, for which $\mathbf{R}_{X,Y}$ is a non empty maximal set (there is not another $\mathbf{R}_{X,Y'}$ including it).

- f_{D_X} is defined as a function that associates to each configuration of its parents the set of possible values for D_X. This will be done, by determining the set of possible values for each one of its parents and then taking the intersection for all the parents. For a parent variable D_Y and a vector belonging to its domain $(\beta_{\mathbf{y}})_{\mathbf{y}\in\Omega_{\mathbf{T}_Y}}$, the set of possible values for D_X will be equal to the set of vectors $(\alpha_{\mathbf{x}})_{\mathbf{x}\in\Omega_{\mathbf{T}_X}}$ such that $\sum_{\mathbf{u}} \beta_{\mathbf{y}} = \sum_{\mathbf{v}} \alpha_{\mathbf{x}}$, where $\mathbf{U} = \mathbf{T}_Y - \mathbf{R}_{X,Y}$, $\mathbf{V} = \mathbf{T}_X - \mathbf{R}_{X,Y}$, i.e. the results of adding the vectors in the non common variables coincide.

With this procedure we only estimate conditional probabilities, considering that the joint probabilities can be obtained by multiplication. So all the probabilities factorize according to G.

In this local specification, for each variable X, the domain for the decision variable D_X is the set of possible parameters for the prior Dirichlet distributions if the joint probability has global parameter S. The conditional probability is determined for each parameter vector, by doing the corresponding estimation from the database and the given prior distribution. Finally, the role of functions f_{D_X} is to keep consistency among parameters taking into account the existing restrictions in the global application of the IDM. For that, we relate the vectors of parameters D_X and D_Y if the corresponding sets of variables $\mathbf{T}_X$ and $\mathbf{T}_X$ have non-empty intersection. Consistency is achieved if the marginalization of the vectors of parameters on the intersection of both sets of variables is the same, where the marginalization is computed by adding in the variables not in the intersection (in the same way than when computing a marginal probability). This is based on the properties of the Dirichlet densities (see [14, 4]).

The main problem of this description as a local network is that variables D_X take values in a continuous infinite set of parameters. This makes infeasible the application of existing algorithms for computing upper and lower conditional probabilities, which are designed for categorical variables. In the following, we will show that the use of the extreme IDM gives rise to a credal network that can be locally specified in a simple way by introducing categorical decision variables.

In the extreme IDM we have a prior density for each value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$, so in the posterior credal set after observing the database we will have a joint probability for each one of these values. We have to introduce decision auxiliary variables able of representing these values. This will be done by considering a decision variable R_X for each variable X with the same set of values than X: Ω_X. The set of values of variables $R_X, X \in \mathbf{X}$, will represent the configuration $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ in which the parameter S is concentrated.

Decision variables, R_X, do not have parents.

If we have variable X with parents $\mathbf{\Pi}_X$ in graph G, we add links from each variable R_Y where $Y = X$ or $Y \in \mathbf{\Pi}_X$ to X (we extend the parents of X by adding its decision variable and the decision variables of its parents). Let us call $\mathbf{R}_{\mathbf{\Pi}_X}$ the set of variables R_Y where $Y \in \mathbf{\Pi}_X$, and as usual (in lowercase), $\mathbf{r}_{\mathbf{\Pi}_X}$ will represent a configuration of this set of variables. The conditional probability of a variable X given $\mathbf{\Pi}_X = \pi_X, R_X = r_X$ and $\mathbf{R}_{\mathbf{\Pi}_X} = \mathbf{r}_{\mathbf{\Pi}_X}$ is computed as follows:

- If for one variable Y in Π_X, the value of Y in configuration π_X is not equal to the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, then

$$P(x|\pi_X,\mathbf{r}_{\mathbf{\Pi}_X},r_X) = \frac{N_{x,\pi_X}}{N_{\pi_X}} \qquad (3)$$

  where the conditional distribution is the uniform if $N_{\pi_X} = 0$.

- If for any variable Y in Π_X, the value of Y in configuration π_X is the same than the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, and the value of X is the same than the value of R_X ($x = r_X$), then

$$P(x|\pi_X,\mathbf{r}_{\mathbf{\Pi}_X},r_X) = \frac{N_{x,\pi_X}+S}{N_{\pi_X}+S} \qquad (4)$$

- If for any variable Y in Π_X, the value of Y in configuration π_X is the same than the value of R_Y in configuration $\mathbf{r}_{\mathbf{\Pi}_X}$, and the value of X is not equal to the value of R_X ($x \neq r_X$), then

$$P(x|\pi_X, \mathbf{r}_{\mathbf{\Pi}_X}, r_X) = \frac{N_{x,\pi_X}}{N_{\pi_X} + S} \qquad (5)$$

It is immediate that this specification determines the same credal set over G as the one defined in Section 3, taking into account that the values of variables $R_X, X \in \mathbf{X}$, represent the value $\mathbf{x}_0 \in \Omega_{\mathbf{X}}$ in which the extreme Dirichlet distribution is concentrated.

One important problem of this representation is that the number of variables in each conditional probability is duplicated, and as the size of conditional tables is exponential in the number of variables, then we can have tables of quadratic size with respect to the size of precise conditional probability tables in G. However, the size of the conditional probabilities can be smaller if we use an appropriate representation. In this paper we consider the use of the *probability tree* representation [9, 18, 7].

A *probability tree* $\mathcal{T}$ is a directed labelled tree, where each internal node represents a variable and each leaf represents a non-negative real number. Each internal node has one outgoing arc for each state of the variable associated with that node. The *size* of a tree $\mathcal{T}$, denoted by $size(\mathcal{T})$, is defined as its number of leaves.

A probability tree $\mathcal{T}$ on variables $\mathbf{Y}$ represents a potential (a joint or conditional probability distribution) in these variables $h : \Omega_{\mathbf{Y}} \to I\!R_0^+$ if for each $\mathbf{y} \in \Omega_{\mathbf{Y}}$ the value $h(\mathbf{y})$ is the number stored in the leaf node that is reached by starting from the root node and selecting the child corresponding to the value of Y in $\mathbf{y}$ for each internal node labelled Y.

A probability tree is usually a more compact representation of a potential than a table. This is illustrated in Figure 1, which displays a potential h and its representation using a probability tree. The tree contains the same information as the table, but using only five values instead of eight. Furthermore, trees enable even more compact representations to be obtained in exchange for loss of accuracy. This is achieved by pruning certain leaves and replacing them by the average value, as shown in the second tree in Figure 1.

All the necessary operations to compute with probability potentials in credal networks can be directly carried out in the probability tree representation, without transforming it into a table [9, 18, 7]. In the following we give the probability tree representation of the conditional probability distribution of a variable X in a local specification of an extreme IDM credal network. It is built with the following procedure $BuildTree(\mathcal{T}, \mathbf{Z}, \mathbf{y})$, where $\mathcal{T}$ is the tree we are building, $\mathbf{Z}$ is the set of variables from $\mathbf{\Pi}_X$ we have to consider and $\mathbf{y}$ the configuration of the variables already introduced in the tree and that corresponds to the path from the root to the present tree $\mathcal{T}$. Initially the procedure is called with $\mathcal{T}$ empty, $\mathbf{y}$ empty, and $\mathbf{Z} = \mathbf{\Pi}_X$. It performs the following steps:

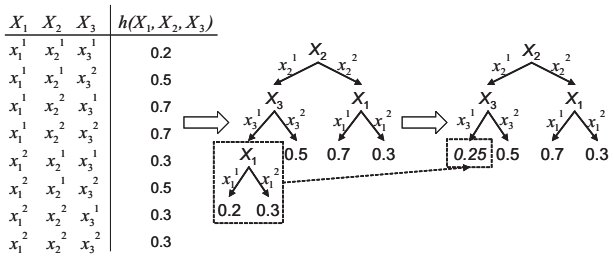


Figure 1: A probability potential h, its representation as a probability tree and its approximation after pruning various branches

- Take a variable $Z \in \mathbf{Z}$, branch $\mathcal{T}$ by Z, then branch also all its children by variable R_Z. Remove Z from $\mathbf{Z}$.
- For each one of the leaves $\mathcal{T}'$ of the resulting tree, consider the configuration $\mathbf{y}'$ equal to $\mathbf{y}$ plus the value of $Z = z$ corresponding to this leaf.
  - If for this leaf, the values of Z and R_Z are the same, then call recursively to $BuildTree(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$, continuing with the construction of the tree.
  - If for this leaf, the values of Z and R_Z are different, then call recursively to $BuildTree2(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$.
- If $\mathbf{Z} = \emptyset$, then the tree is finished by branching by X and all its children by R_X. For all the resulting leaves we store the conditional probability of $X = x$ given the configuration $\mathbf{y}$ (that now is a complete configuration of its parents). This probability is computed with expressions (4) or (5) depending on whether the leaf is obtained for the same value of R_X and X or for different values of these variables (in the corresponding expressions $\pi_{\mathbf{X}}$ is the current configuration $\mathbf{y}$).

$BuildTree2(\mathcal{T}, \mathbf{Z}, \mathbf{y})$ is a simpler procedure that obtains the conditional probability when the configuration of the parents is different of the configuration of the decision variables (by maximum likelihood):

- If $\mathbf{Z} = \emptyset$, then the tree is finished by branching by X and in its leaves we store the conditional probability of $X = x$ given the configuration $\mathbf{y}$. This probability is computed with expressions (3). As above, in the corresponding expression, $\pi_{\mathbf{X}}$ is the current configuration $\mathbf{y}$.
- If $\mathbf{Z} \neq \emptyset$, then take $Z \in \mathbf{Z}$, branch the tree by Z. Remove Z from $\mathbf{Z}$.
- For each one of the leaves $\mathcal{T}'$ of the resulting tree, consider the configuration $\mathbf{y}'$ equal to $\mathbf{y}$ plus the value of $Z = z$ corresponding to this leaf. Then, call recursively to $BuildTree2(\mathcal{T}', \mathbf{Z}, \mathbf{y}')$.

As example, assume two binary variables X and Y for which we have the following table of frequencies:

	Y=0	Y=1
X=0	1	3
X=1	2	1

The resulting tree for the conditional probability of Y given X and $S = 2$, is given in Figure 2.

It can be shown that if n is the size of a table of X given $\mathbf{\Pi}_X$, then the number of leaves of this tree representation will be $n.\left(|\Omega_X| + \sum_{Y \in \mathbf{\Pi}_X}(|\Omega_Y| - 1)\right)$. In this example, we have represented a table of size $n = 4$ with a tree of 12 leaves. This is obtained from the following fact: the number of cases in which the value of the decision variables coincides with the conditioning configuration is n, and each one of then is branched by R_X of cardinal $|\Omega_X|$. Now, each conditioning variable Y defines $(|\Omega_Y|-1)$ branches in which the complete probability table of size n is estimated by maximum likelihood (no coincidence of the conditioning variables and the value of parents variables).

5 Experiments

The local estimation algorithm with the extreme IDM has been implemented in Elvira environment [10] producing the local specification at the same time. With this we have been able of applying the existing algorithms for credal networks as the ones described in [7] which have also been implemented in Elvira. We have done a very simple and preliminary experiment. We have selected a Naive Bayes graph with a class variable and 10 attributes (all binary variables). We have simulated samples with different sizes (from 10 to 1000). We have selected a Naive Bayes, as with no independencies the results are the same than with the complete IDM. So, we do the experiments with a graph in which many independence relationships among the variables are represented. In these conditions, we have estimated the locally specified credal network and computed the conditional probability for the class when all the attributes have been observed. We have considered 3 different situations: the observations are random, for each attribute we observe the most frequent value, and finally the case in which for each attribute we observe the least frequent value. We report the length of the computed posterior intervals. The intervals are computed with a simple exact deletion algorithm with probability trees (see details in [8]). The sample generation is repeated 50 times for each sample size and set of observations and in Table 1 we report the average and standard deviations of the lengths interval probabilities (Evi1 corresponds to random observations, Evi2 to observing the most frequent cases, and Evi3 to observing the least frequent cases).

We observe that the intervals decrease in size when the sample size is increased. Also when we observe the most frequent values the intervals are smaller than when the least frequent values are observed. Random observations give rise to intermediate intervals. In this stage, we can not say much more, except that the intervals are very wide with the smaller sample size (10) but that the imprecision is small with sample sizes of 1000. To our opinion, this imprecision is *reasonable*.

6 Conclusions

In this paper, we have proposed a new model to estimate probabilities for a credal network. This model is a restriction of the general IDM, where only the extreme densities are considered. Its main advantage of the new one is that the resulting credal network allows a simple local specification with categorical decision variables and then it is suitable for the application of existing algorithms for the computation of posterior intervals or dominance relationships.

We have shown the results of the imprecision in the intervals in some very preliminary experiments. But, really it would be necessary to carry out more tests to see the behaviour in real classification problems and to study the differences with the complete IDM. We believe that the differences between the two models are less important than the selection of parameter S and, at present, there is no general agreement about which is the most suitable value of S. We do not expect meaningful differences between them. We have also to take into account that it is possible that the fact that the new model is more restrictive could be compensated with a greater S (using $S = 2$ in all the situations).

Another point we would like to raise is that, though the IDM is a widely accepted model with very good

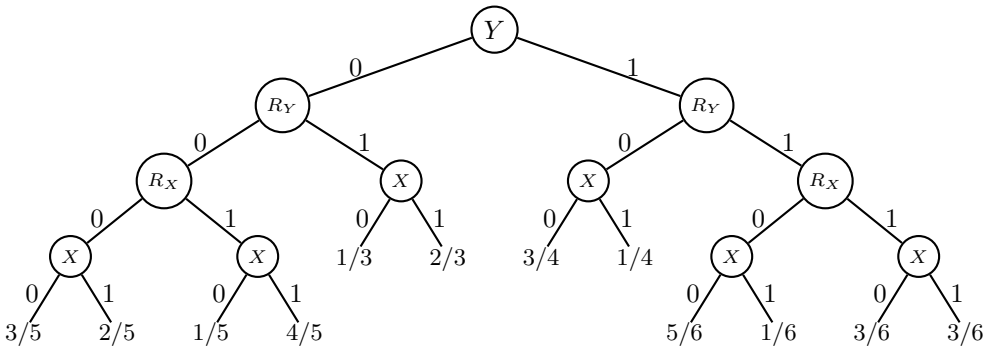


Figure 2: Tree representation of X given Y, R_Y, R_X

	Evi1		Evi2		Evi3	
Iter	aver.	dev.	aver.	dev.	ave.	dev.
10	0.948286	0.051853	0.608757	0.251884	0.999999	4.2475E-5
20	0.814695	0.172520	0.382576	0.247139	0.983606	0.115112
50	0.573315	0.144312	0.062716	0.067052	0.968920	0.121630
100	0.326327	0.126361	0.010081	0.007831	0.869229	0.238078
200	0.170638	0.053589	0.002283	0.001472	0.656434	0.209858
500	0.063706	0.017014	6.9051E-4	3.4956E-4	0.366218	0.123964
1000	0.032087	0.006731	3.0723E-4	1.1755E-4	0.181275	0.051277

Table 1: Average lengths standard deviations for the posterior conditional intervals ($S = 2$)

theoretical properties, it is not the only possible model for being used as prior information. In the problem we have studied in this paper, we see that the general model has computational problems. We also experimented difficulties with the global IDM when studying independence in [1] and we considered a different more restrictive IDM as it was impossible to make decisions about independence with the original IDM using a generalization of Bayesian scores (there was no dominance even with very large samples). So it is important to investigate alternative models for prior information, comparing their behaviour in solving different problems.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Technology under project Algra (TIN2004-06204-C03-02).

References

[1] J. Abellán and S. Moral. A new score for independence based on the imprecise Dirichlet model. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications (ISIPTA '05)*, pages 1–10. SIPTA, 2005.

[2] A. Antonucci and M. Zaffalon. Locally specified credal networks. In M. Studen and Vomlel, editors, *Proceedings of the third European Workshop on Probabilistic Graphical Models*, pages 25–34. Action M Agency, 2006.

[3] A. Antonucci, M. Zaffalon, J. S. Ide, and F. G. Cozman. Binarization algorithms for approximate updating in credal nets. In L. Penserini, P. Peppas, and A. Perini, editors, *Proceedings of the third European Starting AI Researcher Symposium*, pages 120–131. IOS Press, 2006.

[4] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39:123–150, 2005.

[5] A. Cano, J.E. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing. In *Proceedings of the Fith International Conference IPMU'94*, pages 4–8, Paris, 1994.

[6] A. Cano, J.M. Fernandez-Luna, and S. Moral. Computing probability intervals with simulated annealing and probability trees. *Journal of Applied Non-Classical Logics*, 12:151–171, 2002.

[7] A. Cano, M. Gómez, S. Moral, and J. Abellán. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44:261–280, 2007.

[8] A. Cano and S. Moral. Using probability trees to compute marginals with imprecise probabilities.

International Journal of Approximate Reasoning, 29:1–46, 2002.

[9] A. Cano, S. Moral, and A. Salmerón. Penniless propagation in join trees. *International Journal of Intelligent Systems*, 15:1027–1059, 2000.

[10] Elvira Consortium. Elvira: An environment for probabilistic graphical models. In J.A. Gámez and A. Salmerón, editors, *Proceedings of the 1st European Workshop on Probabilistic Graphical Models*, pages 222–230, 2002.

[11] F.G. Cozman. Robustness analysis of bayesian networks with global neighborhoods. Technical Report CMU-RI-TR96-42, Carnegie Mellon University, 1996.

[12] F.G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[13] E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106:77–107, 1998.

[14] D. Geiger and D. Heckerman. A characterization of the dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25.

[15] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[16] J.C.F. Rocha and F.G. Cozman. Inference with separately specified sets of probabilities in credal networks. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 430–437. Morgan & Kaufmann, 2002.

[17] J.C.F. Rocha and F.G. Cozman. Inference in credal networks with branch-and-bound algorithms. In *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications (ISIPTA03)*, pages 482–495, 2003.

[18] A. Salmerón, A. Cano, and S. Moral. Importance sampling in bayesian networks using probability trees. *Computational Statistics and Data Analysis*, 34:387–413, 2000.

[19] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[20] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[21] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T.L. Fine, and T. Seidenfeld, editors, *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, 2001.

[22] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105:5–21, 2002.

Comparative Probability Orders and the Flip Relation

Marston Conder
The University of Auckland,
Auckland NZ
m.conder@auckland.ac.nz

Dominic Searles
The University of Auckland
Auckland NZ
dnsearles@gmail.com

Arkadii Slinko
The University of Auckland,
Auckland NZ
a.slinko@auckland.ac.nz

Abstract

In this paper we study the flip relation on the set of comparative probability orders on n atoms introduced by Maclagan (1999). With this relation the set of all comparative probability orders becomes a graph $\mathcal{G}_n$. Firstly, we prove that any comparative probability order with an underlying probability measure is uniquely determined by the set of its neighbours in $\mathcal{G}_n$. This theorem generalises the theorem of Fishburn, Pekeč and Reeds (2002). We show that the existence of the underlying probability measure is essential for the validity of this result. Secondly, we obtain the numerical characteristics of the flip relation in $\mathcal{G}_6$. Thirdly, we prove that a comparative probability order on n atoms can have in $\mathcal{G}_n$ up to ϕ_{n+1} neighbours, where ϕ_n is the nth Fibonacci number. We conjecture that this number is maximal possible. This partly answers a question posed by Maclagan.

Keywords. comparative probability, flippable pair, probability elicitation, subset comparisons, simple game, weighted majority game, desirability relation

1 Introduction

Considering comparative probability orders from the combinatorial viewpoint, Maclagan [13] introduced the concept of a flippable pair of subsets. We show that the concept of flippable pair is important for several other reasons and adds richness to the whole theory of comparative probability orders. In particular, we show that comparisons of subsets in flippable pairs correspond to irreducible vectors in the discrete cone of a comparative probability order. Fishburn at al [9] showed that in any minimal set of comparisons that define a representable comparative probability order all pairs of subsets in those comparisons are critical. We strengthen this theorem by showing that they must be not only critical but also flippable.

We show that there is an important distinction in algebraic properties of discrete cones for representable and non-representable comparative probability orders. In the former case the cone has a basis of irreducible vectors and in the latter irreducible vectors may not generate the cone.

Maclagan formulated a number of very interesting questions (see [13, p. 295]) which we partly answer here. In particular, she asked how many flippable pairs a comparative probability order may have. In this paper we show that a representable comparative probability order may have up to ϕ_{n+1} flippable pairs, which is the $(n+1)$th Fibonacci number. We conjecture that this lower bound on maximal number of flippable pairs is sharp. The latter results was obtained by Dominic Searles in his summer scholarship project (2006) under supervision of the other two authors.

Section 2 contains preliminary results and formulates Maclagan's problem. Section 3 discusses the concept of a flippable pair and proves the aforementioned generalisation of Fishburn-Pekeč-Reeds theorem. Section 4 numerically characterises the flip relation on six atoms. In Section 5 we discuss Searles' conjecture in relation to Maclagan's problem and prove the aforementioned lower bound. Section 6 introduces a class of simple games related to comparative probability orders and Section 7 concludes with stating sveral open problems.

2 Preliminaries

2.1 Comparative Probability Orders and Probability Measures

Given a (weak) order[1] $\preceq$ on a set A, the symbols $\prec$ and $\sim$ will, as usual, denote the corresponding (strict) linear order and indifference, respectively.

Definition 1. *Let X be a finite set. A linear order $\preceq$ on 2^X is called a* comparative probability order *on*

[1] reflexive, complete and transitive binary relation

X if $\emptyset \prec A$ for every non-empty subset A of X, and $\preceq$ satisfies de Finetti's axiom, namely

$$A \preceq B \iff A \cup C \preceq B \cup C, \tag{1}$$

for all $A, B, C \in 2^X$ such that $(A \cup B) \cap C = \emptyset$.

As in [7, 8] at this stage of investigation we preclude indifferences between sets. For convenience, we will further suppose that $X = [n] = \{1, 2, \ldots, n\}$ and denote the set of all comparative probability orders on $2^{[n]}$ by $\mathcal{P}_n$.

If we have a probability measure $\mathbf{p} = (p_1, \ldots, p_n)$ on X, where p_i is the probability of i, then we know the probability of every event A by the rule $p(A) = \sum_{i \in A} p_i$. We may now define an order $\preceq_{\mathbf{p}}$ on 2^X by

$$A \preceq_{\mathbf{p}} B \quad \text{if and only if} \quad p(A) \leq p(B).$$

If probabilities of all events are different, then $\preceq_{\mathbf{p}}$ is a comparative probability order on X. Any such order is called *(additively) representable*. The set of representable orders is denoted by $\mathcal{L}_n$. It is known [10] that $\mathcal{L}_n$ is strictly contained in $\mathcal{P}_n$ for all $n \geq 5$.

Since a representable comparative probability order does not have a unique probability measure representing it but a class of them, any comparative probability order can be viewed as a credal set [12] of a very special type. We will return to this interpretation slightly later.

As in [7, 8], it is often convenient to assume that $1 \prec 2 \prec \ldots \prec n$, This reduces the number of possible orders under consideration by a factor of $n!$. The set of all comparative probability orders on $[n]$ that satisfy this condition, will be denoted by $\mathcal{P}_n^*$ and the set of all representable comparative probability orders on $[n]$ will be denoted by $\mathcal{L}_n^*$.

We can also define a representable comparative probability order by any vector of positive utilities $\mathbf{u} = (u_1, \ldots, u_n)$ by

$$A \preceq_{\mathbf{u}} B \quad \text{if and only if} \quad \sum_{i \in A} u_i \leq \sum_{i \in B} u_i.$$

We do not get anything new since this will be the order $\preceq_{\mathbf{p}}$ for the measure $\mathbf{p} = \frac{1}{S}\mathbf{u}$, where $S = \sum_{i=1}^n u_i$. However, sometimes it is convenient to have the coordinates of $\mathbf{u}$ integers. We will call $u(A) = \sum_{i \in A} u_i$ the *utility* of A.

2.2 Discrete Cones

To every linear order $\preceq \in \mathcal{P}_n^*$, there corresponds a *discrete cone* $C(\preceq)$ in T^n, where $T = \{-1, 0, 1\}$ (as defined in [11, 7]).

Definition 2. *A subset $\mathcal{C} \subseteq T^n$ is said to be a discrete cone if the following properties hold:*

D1. *$\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \subseteq \mathcal{C}$, where $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$,*

D2. *for every $\mathbf{x} \in T^n$, exactly one vector of the set $\{-\mathbf{x}, \mathbf{x}\}$ belongs to $\mathcal{C}$,*

D3. *$\mathbf{x} + \mathbf{y} \in C$ whenever $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\mathbf{x} + \mathbf{y} \in T^n$.*

We note that in [7] Fishburn requires $\mathbf{0} \notin \mathcal{C}$ because his orders are anti-reflexive. In our case, condition D2 implies $\mathbf{0} \in \mathcal{C}$.

For each subset $A \subseteq X$ we define the indicator vector χ_A of this subset by setting $\chi_A(i) = 1$, if $i \in A$, and $\chi_A(i) = 0$, if $i \notin A$. Given a comparative probability order $\preceq$ on X, we define the indicator vector $\chi(A, B) = \chi_B - \chi_A \in T^n$ for every possible comparison $A \preceq B$. The set of all indicator vectors $\chi(A, B)$, for $A, B \in 2^X$ such that $A \preceq B$, is denoted by $C(\preceq)$. The two axioms of comparative probability guarantee that $C(\preceq)$ is a discrete cone (see [7, Lemma 2.1]).

Definition 3. *A comparative probability order $\preceq$ satisfies the mth cancellation condition C_m if and only if there is no set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of non-zero vectors in $C(\preceq)$ for which there exist positive integers $a_1, \ldots, a_m$ such that*

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m = 0. \tag{2}$$

It is known [10, 7, 4] that a comparative probability order $\preceq$ is representable if and only if all cancellation conditions for $C(\preceq)$ are satisfied.

There is an interpretation of discrete cones in terms of gambles. Any vector of T^n represents a gamble. The gamble

$$\mathbf{x} = (x_1, \ldots, x_n) \in T^n$$

pays $x_i \in T$ if the state i materialises. On appearance of $\mathbf{0} \neq \mathbf{x} \in T^n$ a participating agent must be ready to accept either $\mathbf{x}$ or $-\mathbf{x}$. The basic rationality assumption requires that the set of acceptable gambles form a discrete cone.

One may measure rationality of an agent looking at how consistent she was in accepting and rejecting various gambles. We need the following concept.

Definition 4. *Let $\mathcal{C}$ be a discrete cone corresponding to a personal comparative probability of an agent. A multiset*

$$P = \{\mathbf{x}_1^{a_1}, \mathbf{x}_2^{a_2}, \ldots, \mathbf{x}_m^{a_m}\},$$

where $\mathbf{x}_i \in \mathcal{C}$ and $a_i \in \mathbb{N}$, is called a portfolio of acceptable gambles.

Gambles are like risky securities. You may own a different number of shares of the same company. Similarly, a portfolio can contain several identical gambles. If the personal comparative probability of an agent is representable by a measure, then all portfolios of acceptable gambles are (in the long run) profitable.

Definition 5. *The portfolio P is said to be* neutral *if (2) is satisfied.*

The criterion of representability given in [10] can be reformulated in terms of portfolios as follows

Theorem 1 ([10]). *Suppose $\preceq$ be the agent's comparative probability order on 2^Ω and $\mathcal{C}$ be the corresponding discrete cone. Then $\preceq$ is representable iff $\mathcal{C}$ has no neutral portfolios of acceptable gambles.*

One can measure the degree of rationality of the agent by the minimal size of the portfolio of gambles which she cannot handle correctly.

2.3 Generation of Cones and Preference Elicitation

Let us define a restricted sum for vectors in a discrete cone $\mathcal{C}$. Let $\mathbf{u}, \mathbf{v} \in \mathcal{C}$. Then

$$\mathbf{u} \oplus \mathbf{v} = \begin{cases} \mathbf{u}+\mathbf{v} & \text{if } \mathbf{u}+\mathbf{v} \in T^n, \\ \text{undefined} & \text{if } \mathbf{u}+\mathbf{v} \notin T^n. \end{cases}$$

This makes a discrete cone an algebraic object, first studied by Kumar [11].

Definition 6. *We say that the cone $\mathcal{C}$ is weakly generated by vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ if every non-zero vector $\mathbf{c} \in \mathcal{C}$ can be expressed as a restricted sum of $\mathbf{v}_1, \ldots, \mathbf{v}_k$, in which each generating vector can be used as many times as needed. We denote this by $\mathcal{C} = <\mathbf{v}_1, \ldots, \mathbf{v}_k>_w$.*

For the cone of a representable comparative probability order there is a much stronger tool to produce new vectors of the cone from a set of given ones. The following condition is a reformulation of Axiom 3 in [9] in terms of discrete cones associated with $\preceq$. See also [3].

Lemma 1. *Let $\prec \in \mathcal{L}_n^*$ be a representable comparative probability order and $C(\prec)$ the corresponding discrete cone. Suppose $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subseteq C(\prec)$ and suppose that for some positive rational numbers $a_1, \ldots, a_m$ and $\mathbf{x} \in T^n$,*

$$\mathbf{x} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m. \tag{3}$$

Then $\mathbf{x} \in C(\prec)$.

Definition 7. *Let $\preceq$ be a representable comparative probability order. We say that the cone $\mathcal{C} = C(\preceq)$ is* strongly generated *by vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ if every non-zero vector $\mathbf{c} \in \mathcal{C}$ can be obtained from $\mathbf{v}_1, \ldots, \mathbf{v}_k$ by taking linear combinations with positive rational coefficients. We denote this by $\mathcal{C} = <\mathbf{v}_1, \ldots, \mathbf{v}_k>$.*

These two latter concepts are important in the light of probability elicitation problem that Fishburn et al [9] considered. When we elicit a comparative probability without knowing that an underlying probability measure exists, then queries

$$A_1?B_1, \ldots, A_k?B_k, \tag{4}$$

resulting in comparisons $A_1 \prec B_1, \ldots, A_k \prec B_k$, determine the order $\preceq$ if and only if the vectors $\mathbf{v}_1 = \chi(A_1, B_1), \ldots, \mathbf{v}_k = \chi(A_k, B_k)$ weakly generate $C(\preceq)$. If it is already known that a representable order is being elicited, then (4) defines $\preceq$ if and only if the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ strongly generate $C(\preceq)$. Conder et al [2] give an example where the set of strong generators of the cone does not generate the cone weakly.

2.4 Geometric Representation of Representable Orders

Let $A, B \subseteq [n]$ be disjoint subsets, of which at least one is non-empty. Let $H(A, B)$ be a hyperplane consisting of all points $\mathbf{x} \in \mathbb{R}^n$ satisfying the equation

$$\sum_{a \in A} x_a - \sum_{b \in B} x_b = 0.$$

We denote the corresponding hyperplane arrangement by $\mathcal{A}_n$. Also let J be the hyperplane

$$x_1 + x_2 + \ldots + x_n = 1,$$

and let $\mathcal{H}_n = \mathcal{A}_n^J$ be the induced hyperplane arrangement.

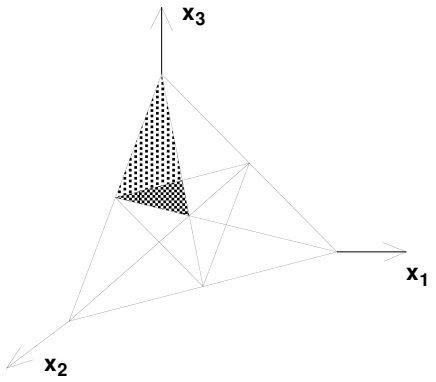


Figure 1

Fine and Gill [5] showed that the regions of $\mathcal{H}_n$ in the positive orthant $\mathbb{R}_+^n$ of $\mathbb{R}^n$ correspond to representable orders from $\mathcal{P}_n$.

Example 1. *The 12 regions of $\mathcal{H}_3$ on Figure 1 represent all 12 comparative probability orders on $\{1,2,3\}$. The two shaded triangular regions correspond to the two orders for which $1 \prec 2 \prec 3$, namely*

$$1 \prec 2 \prec 12 \prec 3 \prec 13 \prec 23 \prec 123, \quad (5)$$

$$1 \prec 2 \prec 3 \prec 12 \prec 13 \prec 23 \prec 123, \quad (6)$$

with the lighter one corresponding to the first order (the lexicographic order).

Now we can see what is special in the credal sets that correspond to comparative probability orders. They are not only convex, as credal sets must be, but they are in fact polytopes.

Problem 1 (Maclagan [13]). *How many facets do regions of $\mathcal{H}_n$ have?*

The minimal number of facets of a region in $\mathcal{H}_n$ is n [4, 2]. The maximal number of facets is not known. Searles' conjecture which we discuss in Section 5 states that the maximal number of facets is ϕ_{n+1}, the $(n+1)$th Fibonacci number.

3 Critical and flippable pairs

Definition 8. *Let A and B be disjoint subsets of $[n]$. The pair (A,B) is said to be* critical[2] *for $\preceq$ if $A \prec B$ and there is no $C \subseteq [n]$ for which $A \prec C \prec B$.*

Definition 9. *Let A and B be disjoint subsets of $[n]$. The pair (A,B) is said to be* flippable *for $\preceq$ if for every $D \subseteq [n]$, disjoint from $A \cup B$, the pair $(A \cup D, B \cup D)$ is critical.*

Since in the latter definition we allow the possibility that $D = \emptyset$, every flippable pair is critical.

We note that the set of flippable pairs is not empty, since the central pair of any comparative probability order is flippable [10]. Indeed, this consists of a certain set A and its complement $A^c = X \setminus A$, and there is no D which has empty intersection with both of these sets. It is not known whether this can be the *only* flippable pair of the order.

Suppose now that a pair (A,B) is flippable for a comparative probability order $\preceq$, and $A \neq \emptyset$. Then reversing each comparison $A \cup D \prec B \cup D$ (to $B \cup D \prec A \cup D$), we will obtain a new comparative probability order $\preceq'$, since the de Finetti axiom will still be satisfied. We say that $\preceq'$ is obtained from $\preceq$ by *flipping over* $A \prec B$. The orders $\preceq$ and $\preceq'$ are called *flip-related*. This flip relation turns $\mathcal{P}_n$ into a graph which we will denote $\mathcal{G}_n$.

[2] We follow Fishburn [9] in this definition, while Maclagan [13] calls such pairs *primitive*.

A pair (A,B) with $A = \emptyset$ can be flippable with no possibility of flipping over. Below we mark with an asterisk the three flippable pairs of the comparative probability order (5):

$$\emptyset \prec_* 1 \prec_* 2 \prec 12 \prec_* 3 \prec 13 \prec 23 \prec 123.$$

The first comparison $\emptyset \prec_* 1$ cannot be flipped over while the other two can be. For example, if we flip this order over $(12,3)$ we will obtain the order (6) which geometrically means passing from the lightly shaded triangle to the darkly shaded one. Or else we can say that flipping over takes us from one credal set to the adjacent one.

Definition 10. *An element $\mathbf{w}$ of the cone $\mathcal{C}$ is said to be* reducible *if there exist two other vectors $\mathbf{u}, \mathbf{v} \in \mathcal{C}$ such that $\mathbf{w} = \mathbf{u} \oplus \mathbf{v}$, and* irreducible *otherwise. The set of all irreducible elements of $\mathcal{C}$ will be denoted as $Irr(\mathcal{C})$.*

Theorem 2. *A pair (A,B) of disjoint subsets is flippable for $\preceq$ if and only if the corresponding indicator vector $\chi(A,B)$ is irreducible in $\mathcal{C}(\preceq)$.*

Proof. Suppose (A,B) is flippable but $\mathbf{w} = \chi(A,B)$ is reducible. Then $\mathbf{w} = \mathbf{u} \oplus \mathbf{v}$, where $\mathbf{u} = \chi(C,D)$ and $\mathbf{v} = \chi(E,F)$ for some C, D, E, F such that $C \prec D$ and $E \prec F$. We may assume without loss of generality that $C \cap D = E \cap F = \emptyset$. Since $\mathbf{u} + \mathbf{v} \in \mathcal{C}(\preceq) \subset T^n$ and $C \cap D = E \cap F = \emptyset$, we have $C \cap E = D \cap F = \emptyset$. Also since $\chi(A,B) = \chi(C,D) + \chi(E,F)$, it is easy to see that

$$A = (C \setminus F) \cup (E \setminus D) \quad \text{and} \quad B = (D \setminus E) \cup (F \setminus C).$$

Let $X = C \cap F$. Then $X \cap (A \cup B) = \emptyset$, and since $(C \cup D) \cap (E \setminus D) = (E \cup F) \cap (D \setminus E) = \emptyset$ we have

$$A \cup X = C \cup (E \setminus D) \prec D \cup (E \setminus D) =$$

$$(D \setminus E) \cup E \prec (D \setminus E) \cup F = B \cup X.$$

In particular, $A \cup X$ and $B \cup X$ are not neighbours in $\preceq$, so (A,B) is not flippable — contradiction.

Suppose now that $A \prec B$ but (A,B) is not flippable. Then there exist subsets C and D such that $(A \cup B) \cap C = \emptyset$ and

$$A \cup C \prec D \prec B \cup C.$$

We may assume that C is minimal for which such D exists. In this case we must have $C \cap D = \emptyset$, for otherwise the common elements in C and D can be removed and a contradiction with minimality of C follows. Now if $\mathbf{u} = \chi(A \cup C, D)$ and $\mathbf{v} = \chi(D, B \cup C)$, then

$$\mathbf{u} \oplus \mathbf{v} = \chi(A \cup C, B \cup C) = \chi(A,B) = \mathbf{w},$$

and so $\mathbf{w}$ is reducible. □

Fishburn et al [9, Theorem 3.7] proved that any smallest set of comparisons that determines a representable comparative probability order in $\mathcal{L}_n$ must consist of critical pairs. Here we prove a stronger result.

Theorem 3. *Let $\preceq$ be a representable comparative probability order. Then the set of irreducible elements of $\mathcal{C} = C(\preceq)$ is the smallest set that weakly generates $\mathcal{C}$.*

Proof. It is clear that the set of *all* irreducible elements $\mathrm{Irr}(\mathcal{C})$ of $\mathcal{C} = C(\preceq)$ is contained in any set of weak generators. Let $\mathbf{x} \in \mathcal{C}$. We will prove that either $\mathbf{x}$ belongs to $\mathrm{Irr}(\mathcal{C})$ or $\mathbf{x}$ can be represented as a restricted sum of elements of $\mathrm{Irr}(\mathcal{C})$. Suppose $\mathbf{x} \notin \mathrm{Irr}(\mathcal{C})$. Then $\mathbf{x} = \mathbf{x}_1 \oplus \mathbf{x}_2$ for some $\mathbf{x}_i \in \mathcal{C}$. If both of them belong to $\mathrm{Irr}(\mathcal{C})$, we are done. If at least one of them does not, then we continue representing both as restricted sums of vectors of $\mathcal{C}$. In this way, we obtain a binary tree of elements of $\mathcal{C}$. We claim that not a single branch of this tree can be longer than the cardinality of $\mathcal{C}$. If one of the branches were longer, then there would be two equal elements in it. Hence it would be possible to start a tree with some element and find the same element deep inside the tree. Without loss of generality, we can assume that $\mathbf{x}$ itself can be found in a tree generated by $\mathbf{x}$. If we stop when $\mathbf{x}$ has appeared for the second time, then we will have

$$\mathbf{x} = G(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m),$$

where G is some term in the algebra $\langle\, \mathcal{C}, \oplus \,\rangle$. Then if we express restricted addition through the ordinary one, the term $\mathbf{x}$ will cancel on both sides, and we will obtain an expression

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_m\mathbf{x}_m = \mathbf{0}$$

with all coefficients a_i positive integers. This will violate the mth cancellation condition. □

This theorem strengthens the aforementioned result of Fishburn, Pekec and Reeds in two directions. Firstly we prove a stronger property for pairs, secondly we prove this for a larger set of pairs.

We see that a minimal set of queries (4) that define a representable comparative probability order in $\mathcal{P}_n$ is unique. In contrast, a minimal set of queries (4) that define a representable comparative probability order in $\mathcal{L}_n$ is not unique. This can be seen, for example, from Example 2 of [2].

Theorem 3 does not hold for non-representable orderings as the following example shows.

Example 2. *In the following non-representable comparative probability order we mark all flippable pairs with an asterisk:*

$$\emptyset \prec 1 \prec 2 \prec 3 \prec 12 \prec 13 \prec_* 4 \prec 14 \prec_* 23 \prec 5$$

$$\prec_* 123 \prec 24 \prec 34 \prec_* 15 \prec 124 \prec 25 \prec_* 134 \ldots \ .$$

There are five such pairs. Let $\mathbf{f}_1 = \chi(13, 4)$, $\mathbf{f}_2 = \chi(14, 23)$, $\mathbf{f}_3 = \chi(5, 123)$, $\mathbf{f}_4 = \chi(34, 15)$, $\mathbf{f}_5 = \chi(25, 134)$, and also let $\mathbf{x} = \chi(23, 5)$. Then it is easy to check that

$$\mathbf{x} = \mathbf{f}_1 \oplus ((\mathbf{f}_5 \oplus (\mathbf{f}_2 \oplus \mathbf{x})) \oplus \mathbf{f}_4). \tag{7}$$

But on the other hand, $\mathbf{x}$ cannot be represented as a restricted sum of $\mathbf{f}_1, \ldots, \mathbf{f}_5$ since it is not in the subspace spanned by $\mathbf{f}_1, \ldots, \mathbf{f}_5$. The reason for (7) is of course the equation $\mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_4 + \mathbf{f}_5 = \mathbf{0}$, which is a violation of the fourth cancellation condition C_4.

There is a marked difference in algebraic properties of representable cones (Theorem 3) and the cone of the non-representable comparative probability order in the previous example. We wonder if this can be made a criterion of representability.

Problem 2. *Is it true that a discrete cone is representable if and only if it is generated by its irreducible vectors?*

4 Characteristics of the flip relation and Maclagan's problem

It is clear that it is sufficient to solve Maclagan's problem (Problem 1) for comparative probability orders in $\mathcal{L}_n^*$. For $n = 5$ and $n = 6$ we can find a solution computationally, using the following fact:

Proposition 1 ([2]). *Let $\preceq$ be a representable comparative probability order in $\mathcal{L}_n$, and let P be the corresponding convex polytope, which is a region of the hyperplane arrangement $\mathcal{H}_n$. Then the number of facets of P equals the number of representable comparative probability orders that are flip-related to $\preceq$ (plus one if the pair $\emptyset \prec 1$ is flippable).*

As we know, the flip relation turns $\mathcal{P}_n$ into a graph. Let $\preceq$ and $\preceq'$ be two comparative probability orders which are connected by an edge in this graph (and so are flip-related). We say that $\preceq$ and $\preceq'$ are in *friendly* relation if they are either both representable or both non-representable.

In the following tables, by the number of flips of the order $\preceq$ we mean the number of flippable pairs of $\preceq$. Let $A \prec B$ be a flippable pair of $\preceq$ such that $A \neq \emptyset$. We say that the flip of the pair $A \prec B$ is *friendly* if

the given order $\preceq$ and the order $\preceq'$ resulting from this flip are in friendly relation.

Let $\preceq \in \mathcal{P}_n^*$ be a representable comparative probability order. There are two situations when a flip of $\preceq$ fails to be friendly: either the corresponding flippable pair is $\emptyset \prec 1$, or the order $\preceq'$ resulting from this flip is of a type different to $\preceq$.

The characteristics of the flip relation for $n = 5$ are given in the following table

Representable orders in $\mathcal{P}_5^*$		
# flips	# friendly flips	# of orders
5	5	169
	4	11 (11)
6	6	159
	5	82 (3)
7	7	65
	6	15
	5	6
8	8	9

Non-representable orders in $\mathcal{P}_5^*$		
# flips	# friendly flips	# of orders
5	3	6
	2	2
	1	16
6	2	6

Note that the numbers in parentheses are the numbers of orders for which the pair $\emptyset \prec 1$ is flippable. The total number of comparative probability orders in $\mathcal{P}_5$ in each category can be obtained by multiplying by $5! = 120$. The corresponding indicators of the flip relation for $n = 6$ are given in [2].

The number of facets of the regions of $\mathcal{H}_5$ corresponding to orders of $\mathcal{L}_5^*$ are given here:

# facets	5	6	7	8	all
# regions	265	177	65	9	512

The number of facets of the regions of $\mathcal{H}_6$ corresponding to $\mathcal{L}_6^*$ is given in [2]. Here will only notice that the smallest number of facets is 6 with 38,025 regions and the maximal is 13 with 20 regions.

It is worth paying attention to the fact that for $n = 5$ and $n = 6$, all comparative probability orders with the largest possible number of flips (namely 8 for $n = 5$, and 13 for $n = 6$) are representable, and all of their flips are friendly. This does not always happen, when an order has the smallest possible number of flips. Nevertheless, this is true for any representable order with smallest possible number of flips: all its flips are friendly [4, 2].

Maclagan [13] gave an example of a non-representable comparative probability order in $\mathcal{P}_6$ whose set of flippable pairs was a subset of the set of all flippable pairs of a representable comparative probability order. She concluded that for $n \geq 6$, an order might not be determined by the set of its flippable pairs.

Strictly speaking, we have to talk about the sequence of flippable pairs of an order $\preceq$, since these pairs may occur in $\preceq$ in different order. Strengthening the result of Maclagan, we have found eight sequences of comparisons with the property that each is the sequence of flippable pairs for two different non-representable comparative probability orders in $\mathcal{P}_6$ [2]. We list one such sequence below: $14 \prec 5$, $15 \prec 24$, $125 \prec 34$, $45 \prec 16$, $26 \prec 145$, $1245 \prec 36$. These eight sequences were found with the help of the MAGMA [1] system, which we used to determine and analyse several examples of orderings on sets of small order.

5 Searles' Conjecture

Let us summarise what we know about the cardinality of $|\mathrm{Irr}(\mathcal{C})|$ in the following

Theorem 4 ([4, 2]). *Let $\preceq$ be a comparative probability order on 2^X with $|X| = n$, and $\mathcal{C}$ be the corresponding discrete cone. Then*

- *if $\preceq$ is representable, then the set of all irreducible elements $\mathrm{Irr}(\mathcal{C})$ generates $\mathcal{C}$ and $|\mathrm{Irr}(\mathcal{C})| \geq n$, while*
- *if $\preceq$ is non-representable, then the set of all irreducible elements $\mathrm{Irr}(\mathcal{C})$ may not generate $\mathcal{C}$ and it may be that $|\mathrm{Irr}(\mathcal{C})| < n$.*

As we mentioned, MAGMA computations show that

- in $\mathcal{G}_5$: $5 \leq |\mathrm{Irr}(\mathcal{C})| \leq 8$, and
- in $\mathcal{G}_6$: $5 \leq |\mathrm{Irr}(\mathcal{C})| \leq 13$,

and all intermediate values are attainable.

Searles noticed that $8 = \phi_6$ and $13 = \phi_7$, where ϕ_n is the nth Fibonacci number, that is the nth member of the sequence defined by $\phi_1 = \phi_2 = 1$ and $\phi_{n+2} = \phi_{n+1} + \phi_n$. Its initial values are: $1, 1, 2, 3, 5, 8, 13, 21, \ldots$ He conjectured that

Conjecture 1. *The maximal number of facets of regions of $\mathcal{H}_n$ is equal to the maximal cardinality of $Irr(\mathcal{C}(\preceq))$ for $\preceq \in \mathcal{L}_n^*$, and equal to the Fibonacci number ϕ_{n+1}.*

The first part of this conjecture will be proved if we show that for some representative comparative probability order $\preceq$, for which $|\mathrm{Irr}(\mathcal{C}(\preceq))|$ is maximal, all

flips of $\preceq$ are friendly. The existence of such order was checked for all $n \leq 12$.

Searles made the following advance towards proving the second part of this conjecture.

Theorem 5. *In $\mathcal{P}_n$ there exists a comparative probability order with a discrete cone $\mathcal{C}$ for which $|\mathrm{Irr}(C)| = \phi_{n+1}$, where ϕ_n is the nth Fibonacci number.*

The proof will be split into several observations. Let us introduce the following notation first. Let $\mathbf{u} = (u_1, \ldots, u_n)$ be a vector such that $0 < u_1 < \ldots < u_n$ and $q > 0$ be a number such that $u_j < q < u_{j+1}$ for some j (we assume that $u_{n+1} = \infty$). In this case we set $(\mathbf{u}, q)$ to be the vector of $\mathbb{R}^{n+1}$ such that

$$(\mathbf{u}, q) = (u_1, \ldots, u_j, q, u_{j+1}, \ldots, u_n).$$

We also denote $\ell_n = (1, 2, 4, \ldots, 2^{n-1})$ and $2\ell_n = (2, 4, 8, \ldots, 2^n)$. An easy observation is this:

Proposition 2. *$\preceq_{\ell_n}$ is the lexicographic order, and the utilities of any two consecutive terms in it differ by 1. These utilities cover the whole range between 0 and $2^n - 1$.*

Proof. We leave the verification to the reader. □

Proposition 3. *Let q be an odd number such that $q < 2^{n+1}$ and $\mathbf{m} = (2\ell_n, q)$. Then the difference between the utilities of any two consecutive terms of $\preceq_{\mathbf{m}}$ is not greater than 2.*

Proof. Suppose $2^j < q < 2^{j+1}$, that is, q is the utility of j in $\preceq_{\mathbf{m}}$. Suppose now that (A, B) is a critical pair for $\preceq_{\mathbf{m}}$. If $j \notin B$, then the statement follows from Proposition 2. If $j \in B$ and $j \in A$, the statement follows from the same proposition. Assume now that $j \in B$ but $j \notin A$. Then $B = \{j\} \cup B'$, where B' does not contain j. If $B' \neq \emptyset$, then by Proposition 2 there exists A', not containing j, such that $0 \leq u(B') - u(A') \leq 2$. Then A must be $\{j\} \cup A'$ and the proposition is true. Finally, if $B = \{j\}$, then since $u(j) \leq 2^{n+1} - 1$ by Proposition 2 there will be an A, not containing j, such that $u(B) - u(A) = 1$. □

Let us denote by $\mathcal{S}_{n+1}$ the class of orderings on $X = \{1, 2, \ldots, n+1\}$ of the type $\preceq_{\mathbf{m}}$, where $\mathbf{m} = (2\ell_n, q)$ for some odd $q < 2^n$. And let j denote the number such that $2^j < q < 2^{j+1}$. Obviously, $j < n + 1$.

Proposition 4. *From the position at which the subset $\{j\}$ appears in the order $\preceq_{\mathbf{m}}$ until the position at which all subsets contain j, subsets not containing j alternate with those containing j, with the difference in utilities for any two consecutive terms being 1.*

Proof. All subsets not containing j have even utility and all those containing j have odd utilities. If we consider these two sequences separately, by Proposition 2 the difference of utilities of neighboring terms in each sequence will be equal to 2. Hence they have to alternate in $\preceq_{\mathbf{m}}$. □

Lemma 2. *Let $\preceq_{\mathbf{m}}$ be an order from the class $\mathcal{S}_{n+1}$ and let (A, B) be a critical pair for $\preceq_{\mathbf{m}}$. Then the following conditions are equivalent:*

(a) *(A, B) is flippable;*

(b) *either A or B contains j;*

(c) *$u(B) - u(A) = 1$.*

Proof. (a) $\Longrightarrow$ (b): Suppose (A, B) is flippable. As (A, B) is critical, it is impossible for A and B each to contain j. We only have to prove that it is impossible for both of them not to contain j. If $j \notin A$ and $j \notin B$, then $u(A) + 2 = u(B) < u(j)$. Then $u(A) < u(B) < u(n+1) = 2^n$, hence neither A nor B contains $n + 1$. But then for $A' = A \cup \{n+1\}$ and $B' = B \cup \{n+1\}$ we have $u(j) < u(A') < u(B')$. Both A' and B' do not contain j, hence they are in the alternating part of the ordering, and since $u(B') - u(A') = 2$, they cannot be consecutive terms. As (A, B) is flippable, this is impossible, which proves that either A or B contain j.

(b) $\Longrightarrow$ (c): This follows from Proposition 4.

(c) $\Longrightarrow$ (a): This is true not only for orders from our class, but also for all orders defined by integer utility vectors. Indeed, if $u(B) - u(A) = 1$, then for any $C \cap (A \cup B) = \emptyset$ we have $u(B \cup C) - u(A \cup C) = 1$, and $B \cup C$ and $A \cup C$ are consecutive. □

Up to now, the utility of q and j did not matter. Now we will try to maximise the number of flippable pairs in $\preceq_{\mathbf{m}}$, so we will need to choose q carefully. It should come as no surprise that the optimal choice of q will depend on n, so we will talk about q_n now. For the rest of the proof we will set

$$q_n = \frac{(-1)^{n+1} + 2^n}{3}. \tag{8}$$

An equivalent way of defining q_n would be by the recurrence relation

$$q_n = q_{n-1} + 2q_{n-2} \tag{9}$$

with the initial values $q_3 = 3$, $q_4 = 5$. We also note:

Proposition 5. $q_n \equiv 2 + (-1)^{n+1} \pmod 4$.

Proof. Easy induction using (9). □

Let us now consider a flippable pair (A, B) for $\preceq_{\mathbf{m}}$. Since $j = n-2$, we have either $A = A' \cup \{n-2\}$ or $B = B' \cup \{n-2\}$. In the first case, (A', B) is a pair of nonintersecting subsets of the lexicographic order on $[n+1] \setminus \{n-2\}$ with $u(B) - u(A') = q+1$. In the second, the pair will be (B', A) with $u(A) - u(B') = q-1$.

Let g_n be the number of pairs $A \prec B$ with $u(B) - u(A') = q+1$ in the lexicographic order $\preceq_{2\ell_n}$, and let h_n be the number of pairs $A \prec B$ with $u(B) - u(A') = q-1$ in the same order. We have proved the following:

Lemma 3. *The number of flippable pairs in $\preceq_{\mathbf{m}}$ is $g_n + h_n$.*

This reduces our calculations to a rather understandable lexicographic order $\preceq_{2\ell_n}$.

For convenience we will denote $q_n^+ = q_n + 1$ and $q_n^- = q_n - 1$. We note that Proposition 5 implies

Proposition 6. $q_n^- \equiv 1 + (-1)^{n+1} \pmod 4$, *and* $q_n^+ \equiv 3 + (-1)^{n+1} \pmod 4$.

A direct calculation also shows that the following equations hold:

Proposition 7.

$$q_{n+1}^- = 2q_n^- \quad \textit{for all odd } n \geq 3, \tag{10}$$
$$q_{n+1}^- = 2q_n^- + 2 \quad \textit{for all even } n \geq 4, \tag{11}$$
$$q_{n+1}^+ = 2q_n^+ - 2 \quad \textit{for all odd } n \geq 3, \tag{12}$$
$$q_{n+1}^+ = 2q_n^+ \quad \textit{for all even } n \geq 4. \tag{13}$$

Lemma 4. *For any odd $n \geq 3$ the following recurrence relations hold:*

$$g_{n+1} = g_n + h_n, \tag{14}$$
$$h_{n+1} = h_n, \tag{15}$$

and for any even $n \geq 4$

$$g_{n+1} = g_n, \tag{16}$$
$$h_{n+1} = g_n + h_n. \tag{17}$$

Proof. Firstly we assume that n is odd. Then $n+1$ is even. We know from (10) that $q_{n+1}^- = 2q_n^-$. Given any nonintersecting pair $A < B$ in $\preceq_{2\ell_n}$, we may shift it to the right, replacing each element i with the element $i+1$, to obtain a nonintersecting pair $\bar{A} < \bar{B}$ of $\preceq_{2\ell_{n+1}}$. This procedure of shifting doubles the difference in utilities, so $u(\bar{B}) - u(\bar{A}) = 2q_n^- = q_{n+1}^-$. This proves $h_{n+1} \geq h_n$. Moreover, by Proposition 6, $q_{n+1}^- \equiv 0 \pmod 4$ hence no nonintersecting pair $C < D$ of $\preceq_{2\ell_{n+1}}$ with difference q_{n+1}^- can involve 1, either in C or in D. Therefore $C = \bar{A}$ and $D = \bar{B}$ for some nonintersecting pair $A < B$, and so $h_{n+1} = h_n$.

We can also use h_n nonintersecting pairs of $\preceq_{2\ell_{n+1}}$ as described above to construct the same number of nonintersecting pairs of $\preceq_{2\ell_{n+1}}$ with utility difference $q_{n+1}^+ = q_{n+1}^- + 2$. If $A < B$ is such a pair, we notice that 1 belongs neither to A nor to B. Adding 1 to B will create a pair $A < B \cup \{1\}$ with the utility difference q_{n+1}^+. We can also use (12) and a shifting technique to create another g_n nonintersecting pairs with utility difference q_{n+1}^+. Indeed, if $A < B$ is a nonintersecting pair in $\preceq_{2\ell_n}$ with utility difference q_n^+, then the pair $\{1\} \cup \bar{A} < \bar{B}$ will be nonintersecting in $\preceq_{2\ell_{n+1}}$ with utility difference $2q_n^+ - 2 = q_{n+1}^+$. Thus $g_{n+1} \geq g_n + h_n$.

We have now two ways of obtaining nonintersecting pairs from $\preceq_{2\ell_{n+1}}$ with utility difference q_{n+1}^+. The first method gives us pairs $C < D$ with $1 \in D$, while the second method gives us pairs $C < D$ with $1 \in C$. Now, let $C < D$ be a nonintersecting pair in $\preceq_{2\ell_{n+1}}$ with utility difference q_{n+1}^+. As $n+1$ is even, Proposition 6 gives $q_{n+1}^+ \equiv 2 \pmod 4$. This implies that either $1 \in C$ or $1 \in D$. Now as above, we can show that $C < D$ can be obtained by the second or the first method, respectively. Thus $g_{n+1} = g_n + h_n$.

For even n, the statement can be proved similarly, using the other two equations in Proposition 7. □

Proof of Theorem 5. Let us consider the case $n = 3$. We have $q_3 = 3$, so $q_3^- = 2$ and $q_3^+ = 4$. We have three nonintersecting pairs in $\preceq_{2\ell_3}$ with utility difference two, namely $\emptyset < 1$, $1 < 2$, and $12 < 3$, and two nonintersecting pairs with utility difference four, namely, $\emptyset < 2$ and $2 < 3$. Thus $g_3 = 2$ and $h_3 = 3$. Alternatively, we may say that $(g_3, h_3) = (\phi_3, \phi_4)$. It is also easy to check that $(g_4, h_4) = (5, 3) = (\phi_5, \phi_4)$. A simple induction argument now shows that $(g_n, h_n) = (\phi_n, \phi_{n+1})$ for odd n and $(g_n, h_n) = (\phi_{n+1}, \phi_n)$ for even n. By Lemma 3 we find that the number of flippable pairs of $\preceq_{\mathbf{m}}$ is

$$g_n + h_n = \phi_{n+1} + \phi_n = \phi_{n+2}.$$

It remains to notice that $\preceq_{\mathbf{m}}$ is in $\mathcal{G}_{n+1}$. □

6 Simple games related to comparative probability orders

Let us consider a finite set X consisting of n elements (which are called *players*). For convenience, X can be taken to be the set $[n] = \{1, 2, \ldots, n\}$.

Definition 11 ([18, 16]). *A simple game is a pair $G = (X, W)$, where W is a subset of the power set 2^X satisfying the* monotonicity condition*: if $A \in W$ and $A \subset B \subseteq X$, then $B \in W$.*

Elements of the set W are called *winning coalitions*. We also define the complement $L = 2^X \setminus W$, and call the elements of this set *losing coalitions*. A winning coalition is said to be *minimal* if each of its proper subsets is losing. By the monotonicity condition, every simple game is fully determined by its set of minimal winning coalitions. Also for $A \subseteq X$, we will denote its complement $X \setminus A$ by A^c.

Definition 12. *A simple game is called* proper *if $A \in W$ implies that $A^c \in L$, and* strong *if $A \in L$ implies that $A^c \in W$. A simple game which is proper and strong is also called a* constant-sum game.

In a constant-sum game, there are exactly 2^{n-1} winning coalitions and exactly 2^{n-1} losing coalitions.

Definition 13. *A simple game G is called a* weighted majority game *if there exists a weight function $w\colon X \to \mathbb{R}^+$ (where $\mathbb{R}^+$ is the set of all non-negative reals) and a real number q, called the* quota*, such that $A \in W$ if and only if $\sum_{i \in A} w_i \geq q$.*

Associated with every simple game $G = (X, W)$ is a *desirability relation* $\preceq_G$ on X. This was defined by Lapidot and actively studied by Peleg (see [16]).

Definition 14. *Given a simple game G we say that a coalition $A \in 2^X$ is* less desirable *than a coalition $B \in 2^X$ if it has the property that whenever the coalition $A \cup C$ is winning for some coalition $C \in 2^X$ such that $C \cap (A \cup B) = \emptyset$, the coalition $B \cup C$ is winning as well. We denote this by $A \preceq_G B$, or by $A \preceq B$ when the game is clear from the context. Let us also write $A \sim_G B$ whenever $A \preceq_G B$ and $B \preceq_G A$.*

For an arbitrary simple game G, the relation $\preceq_G$ satisfies the following weak version of the de Finetti condition: for any subsets $A, B, C \in 2^X$ such that $C \cap (A \cup B) = \emptyset$,

$$A \preceq_G B \Longrightarrow A \cup C \preceq_G B \cup C. \tag{18}$$

(Note that the arrow is only one-sided.) In other respects, this might not be a well-behaved relation. It might not be complete, and its strict companion $\prec_G$ could be cyclic (see [16]). For the class of games we will define, however, this relation is as nice as it can be. It is also quite natural in the light of (18).

Any (strict) comparative probability order $\leq$ on $X = [n]$ defines a constant-sum simple game $G(\leq)$. Indeed, all subsets of X are ordered according to $\leq$, say

$$\emptyset < A_1 < \ldots < A_{2^{n-1}-1} < A_{2^{n-1}} < \ldots < A_{2^n-1} < X.$$

Let us take $W = \{A_{2^{n-1}}, \ldots, X\}$, to obtain a constant-sum game $G(\leq)$. The pair $(A_{2^{n-1}-1}, A_{2^{n-1}})$ is the central pair of $\leq$, and as shown in [10], we have $A^c_{2^{n-1}-1} = A_{2^{n-1}}$. Also this pair is always flippable.

Proposition 8. *If $\leq$ is defined as above, then $\leq \subseteq \preceq_{G(\leq)}$. In particular, the desirability relation of such a game is complete, and the strict desirability relation is acyclic.*

Proof. Let A and B be two subsets of X, and suppose without of loss of generality that $A \leq B$. Now suppose also that $A \cup T \in W$ for some $T \cap (A \cup B) = \emptyset$. Then by de Finetti's axiom, $A \cup T \leq B \cup T$, which implies that $B \cup T \in W$ by definition of $G(\leq)$. Thus $A \preceq_{G(\leq)} B$. □

If $\leq$ is a representable comparative probability order, then $G(\leq)$ is a weighted majority game.

Peleg asked if any constant-sum simple game with complete desirability relation and acyclic strict desirability relation is a weighted majority game. This question was answered negatively in [17] (see also [16, Section 4.10]), but the cardinality of X in that counter-example is large (and not even specified). If our previous question is answered, it could provide us with a natural way of constructing such examples for smaller n. As we will see below, however, any non-representable comparative probability order that can be used for this purpose must have some very special properties in $\mathcal{P}_n$ relative to the flip relation. The following lemma explains why.

Lemma 5. *If the comparative probability order $\leq'$ is obtained from a comparative probability order $\leq$ by a flip over a flippable pair which is not central, then $G(\leq') = G(\leq)$.*

Proof. Suppose we flip over the flippable pair (A, B). Then for any $C \subset X$ such that $C \cap (A \cup B) = \emptyset$, the sets $A \cup C$ and $B \cup C$ are neighbours and cannot split the central pair. Hence in $G(\leq)$, either both $A \cup C$ and $B \cup C$ are winning, or both are losing, and thus $A \sim_{G(\leq)} B$. The same will happen in $G(\leq')$, and so $G(\leq') = G(\leq)$. □

Corollary 1. *Let $\leq$ be any comparative probability order in $\mathcal{P}_n$. If $\leq$ is connected to a representable comparative probability order by a sequence of flips, none of which changes the central pair of $\leq$, then $G(\leq)$ is a weighted majority game.*

Theorem 6. *If $\leq \in \mathcal{P}_5$ or $\leq \in \mathcal{P}_6$, then $G(\leq)$ is a weighted majority game.*

Proof. It is known (see [18]) that every constant sum game with five players is a weighted majority game. In the case of $\leq \in \mathcal{P}_5$, we can deduce this directly from our results. First, it can be seen from Table 1 that every comparative probability order $\leq$ in $\mathcal{P}_5$ has at least two representable neighbours. At least one of

these must be flip-related to $\leq$ via a non-central pair, and hence the above lemma applies. This deals with the case $n = 5$. For $n = 6$, we have used MAGMA [1] to verify that for every $\leq$ in $\mathcal{P}_6$, the probability measure $\mathbf{p}$ of some representable order $\preceq \in \mathcal{L}_6$ gives a weight function w that makes $G(\leq)$ a weighted majority game. □

7 More Open Problems

Problem 3. *Is it true that $G(\leq)$ is always a weighted majority game?*

Problem 4. *Is Searles' conjecture true?*

Problem 5. *What is the minimum value of $|\mathrm{Irr}(C)|$ in $\mathcal{G}_n$?*

Problem 6. *Is $\mathcal{G}_n$ connected?*

It was checked in [13], and independently by us, that $\mathcal{G}_6$ is connected. As all representative orders form a connected subgraph in $\mathcal{G}_n$, it would be natural to try to prove that any order in $\mathcal{G}_n$ is connected to a representable order. This is not obvious. In $\mathcal{G}_6$, for example, there are vertices (orders) without representable neighbours. A stronger version of this problem which is required for extending Theorem 6 to all n is as follows.

Problem 7. *Is any non-representable order in $\mathcal{G}_n$ connected to a representable order by a sequence of non-central flips?*

References

[1] W. Bosma, J. Cannon and C. Playoust. The MAGMA Algebra System I: The User Language, *J. Symbolic Comput.* 24: 235–265, 1997.

[2] R. Christian, M.Conder and A. Slinko. Flippable Pairs and Subset Comparisons in Comparative Probability Orderings and Related Simple Games, The Centre for Interuniversity Research in Qualitative Economics (CIREQ), Cahier 15-2006. University of Montreal, 2006.

[3] R. Christian and A. Slinko. Answers to Two Questions of Fishburn on Subset Comparisons in Comparative Probability Orderings, Proceedings of The 4th Int. Symposium on Imprecise Probabilities and Their Applications (ISIPTA 05), Pittsburgh, Pensylvania, 117–124, 2005.

[4] M. Conder and A. Slinko. A counterexample to Fishburn's conjecture on finite linear qualitative probability. *Journal of Mathematical Psychology* 48: 425–431, 2004.

[5] T. Fine and J. Gill. The enumeration of comparative probability relations. *Annals of Probability* 4: 667–673, 1976.

[6] B. de Finetti. Sul significato soggetivo della probabilità, *Fundamenta Mathematicae* 17: 298–329, 1931.

[7] P.C. Fishburn. Finite Linear Qualitative Probability, *Journal of Mathematical Psychology* 40: 64–77, 1996.

[8] P.C. Fishburn. Failure of Cancellation Conditions for Additive Linear Orders, *Journal of Combinatorial Designs* 5: 353–365, 1997.

[9] P.C. Fishburn, A. Pekeč and J.A. Reeds. Subset Comparisons for Additive Linear Orders, *Mathematics of Operations Research* 27: 227–243, 2002.

[10] C.H. Kraft, J.W. Pratt and A. Seidenberg. Intuitive Probability on Finite Sets, *Annals of Mathematical Statistics* 30: 408–419, 1959.

[11] A. Kumar. *Lower Probability on Infinite Spaces and Instability of Stationary Sequences*. A PhD Thesis. Cornell University, 1982.

[12] I. Levi. Hard Choices: Decision Making under Unresolved Conflict. Cambridge University Press, Cambridge, 1986.

[13] D. Maclagan. Boolean Term Orders and the Root System B_n. *Order* 15: 279–295, 1999.

[14] P. Orlik and H. Terao. *Arrangements of Hyperplanes*. Springer-Verlag, Berlin, 1992.

[15] D. Scott. Measurement structures and inequalities, *Journal of Mathematical Psychology* 1: 233–247, 1964.

[16] A.D. Taylor and W.S. Zwicker. Simple games. Princeton University Press. Princeton. NJ, 1999.

[17] A.D. Taylor and W.S. Zwicker. Simple games and Magic Squares, *Journal of Combinatorial Theory*, ser. A **71**, 67–88, 1995.

[18] J. von Neumann and O. Morgenstern. Theory of games and economic behavior. Princeton University Press. Princeton. NJ, 1944.

Multinomial nonparametric predictive inference with sub-categories

F.P.A. Coolen
Durham University
Frank.Coolen@durham.ac.uk

T. Augustin
Ludwig-Maximilians University, Munich
thomas@stat.uni-muenchen.de

Abstract

Nonparametric predictive inference (NPI) is a powerful tool for predictive inference under nearly complete prior ignorance. After summarizing our NPI approach for multinomial data, as presented in [8, 9], both for situations with and without known total number of possible categories, we illustrate how this approach can be generalized to deal with sub-categories, enabling consistent inferences at different levels of detail for the specification of observations. This approach deals with main categories and sub-categories in a logical manner, directly based on the powerful probability wheel representation for multinomial data that is central to our method and that ensures strong internal consistency properties. Detailed theory for such inferences, enabling for example more layers of sub-categories as might occur in tree-like data base structures, has yet to be developed, but is conceptually straightforward and in line with the illustrations for more basic inferences presented in this paper.

Keywords. CA model, imprecise Dirichlet model, nonparametric predictive inference, probability wheel representation.

1 Introduction

Statistical data in various application areas are often multinomial, i.e. the observations fall into one of several unordered categories. Recently, the current authors have developed a nonparametric predictive inferential approach for such data [8, 9]. This approach provides lower and upper probabilities for a future observation, on the basis of observed multinomial data, and it adds only few modelling assumptions to the data. The method has been presented both for situations in which one has no information about the number of possible categories [8], and for situations with at most K possible categories [9], where the additional knowledge in the latter case leads to less imprecision for some events of interest. In this paper, we will refer to the general NPI approach for multinomial data, by Coolen and Augustin [8, 9], as the 'CA model'[1]. In the earlier papers, the advantages of the CA model are discussed and illustrated in detail, and the resulting lower and upper probabilities are also compared to those based on Walley's Imprecise Dirichlet Model (IDM) [23], which has attracted considerable attention in a variety of application areas [4].

The CA model fits in the framework of 'Nonparametric Predictive Inference' (NPI) [2, 7], which is generally based on Hill's assumption $A_{(n)}$ [18]. However, for multinomial data, a variation of this assumption is required, which was introduced by Coolen and Augustin [8] and called 'circular-$A_{(n)}$', and which is very close in nature to Hill's $A_{(n)}$ as both are post-data versions of exchangeability [14]. Coolen [7] illustrated the natural use of circular-$A_{(n)}$ for circular data.

A key assumption for the CA model as presented before [8, 9], as well as for most models for multinomial data including Walley's IDM, is that the different categories are in no way related. Not only should the categories not be ordered, but there should also not be other possible links between some of the categories. For example, such methods are not fully suited for situations where two or more categories may be considered as sub-categories of a larger category, for example[2] one may be interested in situations where one distinguishes between main colours such as green, red and blue, but in addition distinguishes between light-blue and dark-blue within the latter category. An interesting property, called the Representation Invariance Principle (RIP), of Walley's IDM [23] is that this distinction has no effect on probabilities for events which do not directly involve 'blue', this property does not hold in general for the CA model [8, 9]. In this pa-

[1] CA: *Circulus Alearius* and/or *Circular-$A_{(n)}$*.

[2] The use of colours as different categories in illustrative examples might be considered inappropriate, as one could consider an existing natural ordering of colours, but it has become somewhat of a tradition in this field following Walley [23].

per, we call categories such as light-blue and dark-blue 'sub-categories' of the category blue, and we present the basic way in which the CA model can deal explicitly with such sub-categories.

In Section 2 of this paper, we present an overview of the CA model as presented before, both for a known and unknown number of categories [8, 9]. Section 3 illustrates how the CA model should be generalized in order to deal with sub-categories, which is mostly explained via an example as the main theory is under development, and Section 4 provides some concluding remarks.

2 The CA model

2.1 The basic setting

Hill [18] introduced the assumption $A_{(n)}$ as a basis for predictive inference in case of real-valued observations. Suppose we have n observations ordered as $z_1 < z_2 < \ldots < z_n$, which partition the real-line into $n+1$ intervals (z_{j-1}, z_j) for $j = 1, \ldots, n+1$, where we use notation $z_0 = -\infty$ and $z_{n+1} = \infty$. Hill's assumption $A_{(n)}$ is that a future observation, represented by a random quantity Z_{n+1}, falls into any such interval with equal probability, so we have $P(Z_{n+1} \in (z_{j-1}, z_j)) = \frac{1}{n+1}$ for $j = 1, \ldots, n+1$. This assumption implies that the rank of Z_{n+1} amongst the n observed data has equal probability to be any value in $\{1, \ldots, n+1\}$. This clearly is a post-data assumption, related to exchangeability [14], which provides direct posterior predictive probabilities [13]. Hill [18, 19] argued that $A_{(n)}$ is a reasonable basis for inference in the absence of any further process information beyond the data set, when actually predicting a future random quantity. Augustin and Coolen [2] prove generally that Nonparametric Predictive Inference (NPI) based on $A_{(n)}$ has strong consistency properties in the theory of interval probability [22, 24, 25]. Interestingly, as NPI is based on $A_{(n)}$, such inference is fully in line with 'perfectly calibrated' inference along the lines of Lawless and Fredette [20], who however restricted attention to precise probability.

In the CA model, multinomial data are represented as observations on a probability wheel, and hence as circular data. A straightforward variation of $A_{(n)}$ that is suitable for inference based on such data, and again linked to exchangeability of $n+1$ observations, is the assumption *circular-$A_{(n)}$*, denoted by $\circledA_{(n)}$ [7, 8]: Let ordered circular data $x_1 < x_2 < \ldots < x_n$ create n intervals on a circle, denoted by $I_j = (x_j, x_{j+1})$ for $j = 1, \ldots, n-1$, and $I_n = (x_n, x_1)$. The assumption $\circledA_{(n)}$ is that a future observation X_{n+1} falls into each of these n intervals with equal (classical) probability, so

$$P(X_{n+1} \in I_j) = \frac{1}{n}, \quad \text{for } j = 1, \ldots, n. \qquad (1)$$

Clearly, $\circledA_{(n)}$ is again a post-data assumption, related to the appropriate exchangeability assumption for such circular data, in exactly the same way as $A_{(n)}$ was related to exchangeability of $n+1$ values on the real-line. NPI based on $\circledA_{(n)}$ has the same consistency properties as shown in [2] for such inference based on $A_{(n)}$.

In the CA model [8, 9], $\circledA_{(n)}$ is combined with the assumed underlying representation of multinomial data as outcomes of spinning a probability wheel. Without additional assumptions about the probability mass $1/n$ per interval I_j, the predictive inferences based on the CA model are again in the form of interval probabilities [2, 22, 24, 25], where a lower probability for an event A is represented by $\underline{P}(A)$, and the corresponding upper probability by $\overline{P}(A)$. Effectively, the lower probability is the maximum lower bound for the classical probability for A that is consistent with the probabilities as assigned by $\circledA_{(n)}$ and in accordance with the probability wheel model, according to De Finetti's fundamental theorem of probability [14], and the upper probability is the minimum upper bound consistent in this way.

The predictive lower and upper probabilities presented in [8, 9], and reviewed in this section, are based on an underlying assumed model, ensuring that they not only make sense for one specific set of data, which they do being F-probability [24, 25] and due to the fact that they bound the observed relative frequencies, but they are also consistent if more observations are added to the data. We now give a brief summary of the key aspects of this model and its properties.

The CA model underlying the nonparametric predictive lower and upper probabilities presented below, is based on a probability wheel representation, with each observation category represented by a single segment of the probability wheel. The idea of such a probability wheel is as follows (see [15] for use of the same concept as a reference experiment underlying subjective probability). An arrow, fixed at the center of a circle, spins around, such that the arrow is equally likely to stop at any segment of the same size, where a segment is an area between two lines from the center of the circle to its circumference. In our model for multinomial data, we assume explicitly that each possible observation category is represented by only a single segment on the circle. Even more, we assume that there is no natural (or assumed) ordering of the observation categories, and therefore also no such ordering of the segments on the circle. Clearly, if we had perfect knowledge of the sizes of all seg-

ments on the probability wheel, we would have full knowledge of the probability distribution for future observations from this multinomial setting. The CA model can deal both with situations where the number of possible categories is unknown [8] and where it is known that there are K possible categories [9], and it only assumes a finite number of exchangeable multinomial observations, $\mathcal{A}_{(n)}$, and the probability wheel representation. As this probability wheel is only an abstract model, we have no information about the configuration of different segments on it. This is important for our nonparametric predictive inferences based on $\mathcal{A}_{(n)}$ once we consider unions of two or more categories, and adds to imprecision of our inferences, in the sense that our lower and upper probabilities are optimal bounds over all configurations of the possible segments on the probability wheel. In Section 3 we change this perspective a little, by allowing categories to be subdivided into sub-categories, in such a way that both inferences at the category and at the sub-category level can be considered. We will show how the CA model can deal with sub-categories by explicitly representing sub-categories within the corresponding category in the probability wheel representation. Each sub-category is again assumed to be represented by a single segment on the probability wheel.

When we combine the concept of a probability wheel, with each observation category represented by a single segment, with the assumption $\mathcal{A}_{(n)}$, on the basis of n observations, then we can represent this situation as if the n observations are represented by n lines, which partition the circle into n equally sized slices, representing that the next observation is equally likely to fall into each one of these slices. The assumption that each observation category is represented by only one segment on the probability wheel, implies that the lines representing observations in the same category are 'next to each other'. For example, if precisely two observations fall into one category, then our current inferences with regard to the next observation falling into this category, are based on the current representation with two lines next to each other which both represent this category, and the other lines, in case of more than 2 observations, representing different categories. Under the assumption $\mathcal{A}_{(n)}$, the probability $\frac{1}{n}$ for the line on the probability wheel corresponding to the next observation to be in between the two lines representing these observations in the same category, is the lower probability that the next observation belongs to that same category as well. For the upper probability, we consider all possible configurations of segments on the probability wheel, which are consistent with the observations and their corresponding lines on the wheel. The upper probability is then the maximum amount of probability, under $\mathcal{A}_{(n)}$ and these data and configurations, that can be assigned to the segments corresponding to the event of interest.

The assumption that each observation category is represented by a single segment on the probability wheel is crucial to the imprecision in the lower and upper probabilities, and is essential as without this assumption the CA model would lead to vacuous lower and upper probabilities for all non-trivial events.

2.2 Inference for an unknown number of categories

Our inferences in this paper are restricted to a single future observation, which is assumed to be exchangeable with the n observations so far. We will refer to such a future observation as the 'next observation', and will denote it by Y_{n+1}. We will assume that each observation can be assigned to a category with certainty, but we do not require these categories to be defined prior to the observations. We assume that available data consist of n_j observations in category c_j, for $j = 1, \ldots, k$, with $\sum_{j=1}^{k} n_j = n$. If the categories are defined upon observation, we have that $n_j \geq 1$, and hence that $1 \leq k \leq n$. We could include further specifically defined categories to our data description, to which no observations belong, but doing so will not influence any of our inferences (as is easily confirmed), so we will not consider this possibility further. For the general setting with unknown total number of possible categories, we must include notation for new, as yet unseen, categories. We distinguish between *D*efined *N*ew categories, of which we need to take the possibility of having several different such categories into account, denoted by DN_i for $i = 1, \ldots, l$ for $l \geq 1$, and the possibility that the next observation belongs to any not yet observed category (including categories DN_i), which we describe as an *U*nobserved *N*ew outcome and denote as $Y_{n+1} = UN$. By allowing $l \geq 0$ and $0 \leq r \leq k$ in this notation, we can define two types of events that comprise the most generally formulated events that need to be considered for Y_{n+1} in our multinomial setting. These two general events are

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i \tag{2}$$

and

$$Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i \tag{3}$$

Excluding one or more defined new categories in the event of interest, as in (2), can only affect our inferences for events including UN.

The general CA model results for nonparametric predictive inference for the next observation, Y_{n+1}, based on multinomial data, with complete absence of knowledge on the number of possible categories apart from the information provided by $n > 0$ observations, and based on $\mathcal{A}_{(n)}$ and the probability wheel model representation, were presented in [8]. For the first of the general events, the lower probability[3] is

$$\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i) = \begin{cases} \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r\right), \text{ for } k \geq 2r \\ \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r + \max(2r-k-l, 0)\right), \\ \qquad\qquad \text{for } r \leq k \leq 2r \end{cases} \quad (4)$$

and the corresponding upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup UN \setminus \bigcup_{i=1}^{l} DN_i) = \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + k - r\right) \quad (5)$$

For the second of these general events, the lower probability is

$$\underline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i) = \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} - r\right) \quad (6)$$

and the corresponding upper probability is

$$\overline{P}(Y_{n+1} \in \bigcup_{s=1}^{r} c_{j_s} \cup \bigcup_{i=1}^{l} DN_i) = \begin{cases} \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + k - r\right), \text{ for } r \leq k \leq 2r, \\ \frac{1}{n}\left(\sum_{s=1}^{r} n_{j_s} + r + \min(k-2r, l)\right), \\ \qquad\qquad \text{for } k \geq 2r \end{cases} \quad (7)$$

2.3 Inference with a known number of possible categories

If we assume, for the same multinomial setting, that there is a known number of possible categories, denoted by K, then this extra assumption has an effect on the lower and upper probabilities in the CA model [9]. We restrict attention to $K \geq 3$, as for the binomial situation with $K = 2$ NPI can be based on an assumed data representation on a line, as presented by Coolen [6], which leads to slightly less imprecision than a representation on a circle as in this paper. We can now denote the $K \geq 3$ possible categories by $C_1, \ldots, C_K$, even if their precise definition might only be possible following observations. Without loss of generality, we assume that the first k of these, $C_1, \ldots, C_k$ for $1 \leq k \leq K$, have already been observed and the last $K-k$, $C_{k+1}, \ldots, C_K$ have not yet been observed. Let n_j be the number of observations in C_j, so $n_j \geq 1$ for $j \in \{1, \ldots, k\}$ and $n_j = 0$ for $j \in \{k+1, \ldots, K\}$, and $n = \sum_{j=1}^{k} n_j$. The two general events of interest introduced before, when K was not known, are now reduced to a single general event,

$$Y_{n+1} \in \bigcup_{j \in J} C_j \quad (8)$$

with $J \subseteq \{1, \ldots, K\}$, but except where mentioned explicitly we exclude the trivial events $J = \emptyset$ and $J = \{1, \ldots, K\}$ from our considerations. Let $OJ = J \cap \{1, \ldots, k\}$ denote the index-set for the categories in the event of interest that have already been observed, and $UJ = J \cap \{k+1, \ldots, K\}$ the corresponding index-set for the categories in the event of interest that have not yet been observed. Let r be the number of elements of OJ and l the number of elements of UJ, so $0 \leq r \leq k$ and $0 \leq l \leq K-k$. This implies that $k-r$ observed categories and $K-k-l$ unobserved categories are not included in the event of interest.

The lower and upper probabilities for event (8), according to the CA model with K known, are [9]

$$\underline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n}\left(\sum_{j \in OJ} n_j - r + \max(2r + l - K, 0)\right) \quad (9)$$

and

$$\overline{P}(Y_{n+1} \in \bigcup_{j \in J} C_j) = \frac{1}{n}\left(\sum_{j \in OJ} n_j - r + \min(2r + l, k)\right) \quad (10)$$

For the two trivial events, the NPI-based lower and upper probabilities are obvious. If $J = \{1, \ldots, K\}$, the upper probability of event (8) is equal to 1, in line with (10), and also the lower probability (9) is trivially defined as 1, which is fully in line with the probability wheel representation which underlies the

[3] All probabilities in this paper are predictive given the first n observations, we do not explicitly mention the dependence on the first n observations in the notation.

CA model. Similarly, if $J = \emptyset$, the lower probability of event (8) is equal to 0, in line with (9), and the upper probability (10) is defined as 0. In our further discussion, we will not explicitly mention these trivial events anymore. At the end of this section, we briefly illustrate (9) and (10) via an example, which will be generalized to include sub-categories in Section 3.

2.4 Fundamental properties of the inferences

To derive all the above lower and upper probabilities, we consider all possible configurations σ on the probability wheel, apply $\mathcal{A}_{(n)}$ to each of these to obtain lower and upper predictive probabilities $\underline{P}_\sigma(\cdot)$ and $\overline{P}_\sigma(\cdot)$, and then take the lower and upper envelope with respect to the set Σ of all configurations [8, 9]. In case of known K, there are fewer configurations on the probability wheel possible for some events of interest, but never more, than when no maximum number of possible categories is known or assumed, hence lower and upper probabilities can be less imprecise if K is known then for the corresponding event in the more general case, but they can never be more imprecise. Actually, such lower and upper probabilities are nested in the logical manner, as the optimization procedures to derive the lower and upper probabilities also take all configurations into account corresponding to known K in the case of an unknown number of categories. All these lower and upper probabilities satisfy a number of important and attractive properties [8, 9]: **(a)** they satisfy the conjugacy property in interval probability theory, and beyond that they are, by applying arguments along the line of [1] (see also [12]), actually F-probability in the sense of Weichselberger [24, 25] and they are coherent in the sense of Walley [22]; **(b)** Corresponding lower and upper probabilities always contain the empirical probability for the event of interest; **(c)** in the limiting situation with $(n \to \infty)$, corresponding lower and upper probabilities become identical. The properties named in (a) imply that the CA model provides sound interval-probabilistic statistical inferences, with strong internal consistency properties. Properties (b) and (c) ensure that these inferences are sensible from classical statistical ('frequency') perspective. Convenient expressions to calculate the lower and upper expectations as simply weighted sums instead of solutions to linear optimization problems are presented in [9].

Properties of the CA model have been discussed in detail before, both for the situations with an unknown total number of possible categories [8] and with at most K possible categories [9], in those papers the resulting inferences were also compared with corresponding inferences based on Walley's Imprecise Dirichlet Model (IDM) [23]. We advocate in particular the fact that the inferences from the CA model do not generally satisfy Walley's 'Representation Invariance Principle' [23], as there are particular situations where for example the number of different categories observed so far would logically have an impact on predictive inference for some events of interest, including events involving categories that have not yet been observed. The CA model provides an attractive alternative to the IDM, and is particularly different on details which were remarked upon by many discussants of Walley's paper [23]. Of course, in situations with substantial data available and a limited number of categories, inferences based on the CA model and the IDM are very similar, in the limit these all agree with empirical probabilities converging to the underlying probabilities (derived from the sizes of the segments). An obvious advantage of the IDM is the fact that it is directly based on a parametric model, with a class of priors used in a similar manner as common in robust Bayesian methods [3]. This implies that inferences can be both in terms of the model parameters and of the future observations, the latter via the class of posterior predictive distributions corresponding to the class of priors chosen [4]. However, as many inferences can be formulated predictively in an attractive and natural manner [7, 16], this apparent advantage of the IDM over the CA model does not hinder applicability of the latter too much.

2.5 An illustrative example

Example 1 briefly illustrates multinomial NPI with a known number of categories, hence formulae (9) and (10) are used.

Example 1.
Suppose that there are $K = 6$ possible categories, namely Blue, Red, Yellow, Green, White, Other, henceforth also indicated by their first letter. Suppose that $n = 9$ observations are available, with the following numbers per category: $B-3$, $R-1$, $Y-2$, $G-3$, $W-0$, $O-0$. We illustrate NPI for the 10th observation, Y_{10}, under the usual assumptions for NPI for multinomial data, as discussed in this section and in more detail in [8, 9]. Some lower and upper probabilities for the events concerning Y_{10} are given in Table 1, it is easy to check that these results illustrate (9) and (10).

$Y_{10} \in \{\cdot\}$	$[\underline{P}, \overline{P}]$
B	$[2/9, 4/9]$
B, R	$[2/9, 6/9]$
B, R, Y	$[3/9, 7/9]$
B, R, Y, G	$[7/9, 1]$
B, R, Y, G, W	$[8/9, 1]$

Table 1. Some lower and upper probabilities (Ex. 1)

For the illustration of our inferences by this example it is helpful to look at the increasing sequence of events described in Table 1. As a consequence of Theorem 2 in [9], where two-monotonicity of $\underline{P}(\cdot)$ was proven, there exists a "least favorable configuration" producing all the lower probabilities of the elements of the sequence as well as a "most favorable configuration" related to all the upper probabilities. For the lower probability note that the probability assigned to a colour that has been observed $n_j - 1$ times is at least $(n_j - 1)/n$. This already gives the whole contribution of the colour to the lower probability as long as there are enough colours not in the event of interest to separate the segments, in order to avoid having to attribute further probability mass $1/n$ to the segment connecting two neighbouring colours in the event of interest. Consequently, we obtain the lower probabilities by the following configuration, where B and R are separated by O and R and Y by W, while Y and G can not be separated anymore, and so additional masses contribute to the lower probability of the event $\{B, R, Y, G\}$, i.e. its lower probability exceeds $\sum_{j \in OJ} n_j - r$.

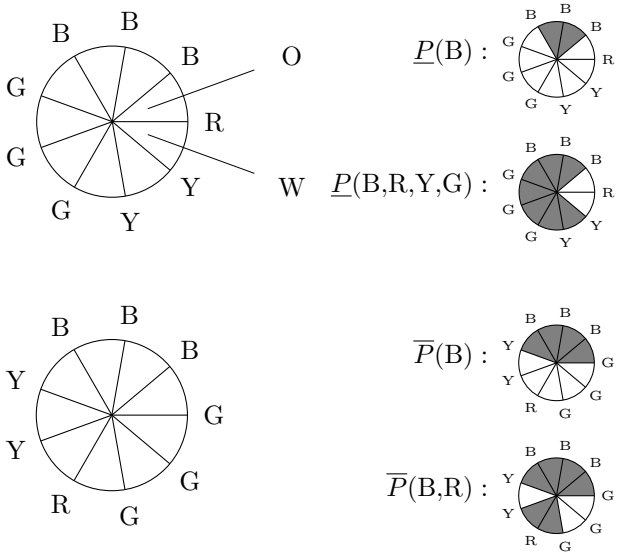


Figure 1. Configurations leading to the lower and upper probabilities in Example 1

Similar arguments apply to the derivation of the upper probability. The main difference is that we now want to assign as much probability mass as possible to the colours in the event of interest, and so we assume that not yet observed colours that are also not in the event of interest do not occur on the probability wheel at all. Again we separate the colours in the event of interest as far as possible, but now with the aim to add probability mass $1/n$ as much as possible. This leads to the configuration at the bottom of Figure 1.

3 Sub-categories

3.1 The modelling of sub-categories

In this section we present the basic principle for dealing with sub-categories in the CA model, and we illustrate this via a basic example. The highest level of categories, in line with the categories as presented in Section 2, will occasionally be referred to as 'main categories', where it is relevant to distinguish these from sub-categories. It is assumed that a main category might be divided into several sub-categories, in such a way that sub-categories are not overlapping and that each sub-category is only related to a single main category. We will assume that each observation belongs to a single main category, and where applicable also to a single sub-category. Such a setting with sub-categories appears, for example, in hierarchical classifications (e.g.[17]). As for the basic CA model (Section 2), both variations with known and unknown total number of possible sub-categories per main category can be dealt with, we restrict our discussion mostly to situations where these numbers are known. We briefly discuss some generalizations in Section 4.

The general principle for dealing with sub-categories is as follows. Each main category is assumed to be represented, in the CA model [8, 9], by a single segment on a probability wheel, with no information about the configuration of all segments representing observed and other relevant categories (those that play a role in predictions). Lower and upper predictive probabilities for the next observation, based on the CA model, are computed by combining this assumed representation with the appropriate $A_{(n)}$ assumption, and via minimization and maximization, respectively, over all configurations that are possible for the given data and categories considered. Suppose now that a particular category (e.g. 'Blue') is divided into sub-categories (e.g. 'Light Blue', 'Dark Blue', 'Other Blue'), then this is included in the probability wheel representation underlying the CA model by assuming that the single segment representing the main category is divided into sub-segments, where it is again assumed that each sub-category is represented by a single sub-segment. We have no knowledge, and wish to make no assumptions, about any particular ordering of such sub-segments, hence for events involving one or more sub-segments, the predictive lower and upper probabilities are again derived via the usual $A_{(n)}$ assumption, which remains unaffected by the appearance of sub-categories, and minimization and maximization over all possible configurations, now also considering all possible configurations of sub-categories within each corresponding main category. Of course, if one has at least two main

categories for which observations are available, then the segments representing the sub-categories of one main category do not form the full circle of the probability wheel, so the combinatorical arguments and computations involved with the sub-categories differ slightly from those for the main categories, yet the principle is straightforward. Note that, if one were to add a new single 'higher-level' category, with the main categories all considered to be sub-categories of this higher-level category, than this makes no difference to the CA model inferences as via the optimisation over all possible configurations that single 'higher-level' category would have no effect whatsoever. It is again possible, as for the events in Section 2 which only considered one level of categories, to derive general expressions for lower and upper probabilities for general events, these have yet to be derived and hence they will be presented at a later stage. It is easily seen that the key properties for NPI for multinomial data discussed in Section 2, including the F-probability property and coherence, can again be rigorously proven in the same way as was done in [9], when sub-categories are included in the model.

We illustrate the general and natural manner in which the CA model can deal with sub-categories in Example 2. For ease of presentation, we restrict attention in this example to a situation with known number K of possible categories [9], as we focus on inferences involving sub-categories. For the case with an unknown number of possible categories, the manner in which the CA model enables sub-categories to be taken into account is identical. We also mostly consider only the case of a known number of sub-categories, this can be generalized to an unknown number of sub-categories in a manner that logically combines the presented way for dealing with sub-categories and the general method for dealing with an unknown total number of categories [8]. As a final restriction to keep presentation at a basic level, we only consider sub-categories of a single main category, of course sub-categories of other main categories are dealt with in the same manner, and one can, for example, generally also consider predictive inference for events involving sub-categories of different main categories. Detailed general results for all such situations will be presented elsewhere.

3.2 Example continued

Example 2.
As in Example 1, suppose that there are six possible categories, Blue, Red, Yellow, Green, White, Other, also indicated by their first letter. In addition, let us assume that observations in Blue are further specified in the sub-categories Light Blue (LB), Dark Blue (DB), or Other Blue (OB). Suppose that 9 observations are available, with the following numbers per (sub-)category: $LB-1$, $DB-2$, $OB-0$, $R-1$, $Y-2$, $G-3$, $W-0$, $O-0$. Of course, these data still imply that there are 3 observations in the main category B, so the lower and upper probabilities for the event $Y_{10} = B$ are as before,

$$[\underline{P}, \overline{P}](Y_{10} = B) = [2/9, 4/9]$$

If we consider events such that Y_{10} belongs to a single sub-category, the resulting lower and upper probabilities are no different from what they would have been if these sub-categories had been main categories, as each is still represented by a single segment on the probability wheel. However, for events involving the union of two sub-categories, the possible configurations of all three sub-categories LB, DB, OB within the main category B must be taken into account. For example, the upper probability for the event $Y_{10} \in \{LB, OB\}$ corresponds to the configurations where DB separates LB, OB within the main category B, while it is irrelevant where B is in the overall configuration with regard to the other main categories, as this is only relevant when events involving unions of main categories are considered, or, as we will discuss later, unions of one or more main categories and sub-categories of other main categories. This separation of LB, OB ensures that of the probability masses that have to be in the main category B, namely two probabilities of 1/9 each, only the probability 1/9 between the two lines representing DB observations has to be assigned to DB, and as LB and OB are on the two extreme sides within the category B, they can now be assigned maximum probabilities of 2/9 and 1/9, respectively. Hence, the upper probability for the event $Y_{10} \in \{LB, OB\}$ is 3/9. The lower probability for this event is 0, as it is easily seen that it is possible (for several configurations) that no actual segment of the probability wheel as created by the data and reflecting the probability masses as assigned by $\mathcal{A}_{(n)}$ in the CA model, [8, 9] must belong to either LB or OB. With similar derivations the lower and upper probabilities presented in Table 2 are derived (see also the example in Figure 2).

The last event in Table 2 is, of course, identical to $Y_{10} = B$. If we had introduced multiple 'Other Blue' sub-categories (OB_i), with no observations for each as yet, then the upper probability that Y_{10} was in any of such sub-category would be equal to 2/9 in case of two such OB_i, and 3/9 in case of three or more of such OB_i, the latter case in agreement with the possible use of UN (see Section 2 and [8]) for such sub-categories if we had not made any assumptions on the number of sub-categories of Blue.

$Y_{10} \in \{\cdot\}$	$[\underline{P}, \overline{P}]$
LB	$[0, 2/9]$
DB	$[1/9, 3/9]$
OB	$[0, 1/9]$
LB, DB	$[1/9, 4/9]$
LB, OB	$[0, 3/9]$
DB, OB	$[1/9, 4/9]$
LB, DB, OB	$[2/9, 4/9]$

Table 2. Some lower and upper probabilities (Ex. 2)

$$\underline{P}(\mathrm{LB} \cup \mathrm{DB}) = \frac{1}{9},$$

$$\overline{P}(\mathrm{LB} \cup \mathrm{DB}) = \frac{4}{9}.$$

Figure 2. On the lower and upper probability of $LB \cup DB$

Let us also briefly consider unions of these sub-categories with other main categories. It should be emphasized that considering the information at sub-category level within the main category Blue, or just at the main category level, has no effect whatsoever on events which do not involve B or any of its sub-categories, due to the fact that all considerations of more detailed configurations to deal with the sub-categories only took account of the segment representing Blue, and did not affect the configurations at main categories level. Lower and upper probabilities for events such as $Y_{10} \in \{DB, Y\}$ are derived as usual, as a single sub-category is involved they are identical to corresponding lower and upper probabilities that would correspond to the situation with DB considered as a main category, so this event has lower probability 2/9 and upper probability 6/9. If more sub-categories of the same main category are included in the event of interest, then the same considerations as discussed above must be taken into account, so all configurations of the sub-categories within the main category must be included in the analysis. For such events including other main categories, however, we must combine this with the configurations at the main categories level, which again becomes mainly important in the case of a known total number of main categories and events involving more than half of these [9]. For example, the lower probability for the event $Y_{10} \in \{LB, OB, R, G, W, O\}$ is equal to 3/9, as only the main category Y and sub-category DB are not included, and as long as Y and DB are not next to each other in the configuration, they can both get a maximum of 3 segments assigned out of the 9 in which the observations have divided the probability wheel, where each such segment represents predictive probability 1/9. The upper probability for this event is 7/9, as all but two segments can be assigned to all (sub-)categories in this event of interest. Of course, this also illustrates the conjugacy property with regard to the complementary event $Y_{10} \in \{DB, Y\}$ considered above. For this event involving 2 of the 3 specified sub-categories of B, and 4 of the 5 main categories other than B, clearly there are no possible configurations with all these 6 (sub-)categories included in the event separated from each other by other categories with each at least one observation in them. If this last situation were the case, the upper probability would have been identical to the sum of the upper probabilities of the events $Y = X$ with $X \in \{LB, OB, R, G, W, O\}$ [9].

To emphasize the difference between sub-categories and main categories, let us compare the event $Y_{10} \in \{LB, DB\}$, which has lower and upper probabilities 1/9 and 4/9, with the event $Y_{10} \in \{R, Y\}$. For both sets of (sub-)categories in these events, we have one (sub-)category with a single observation, and one with two observations. However, the lower and upper probabilities for the event $Y_{10} \in \{R, Y\}$ are equal to 1/9 and 5/9, so this upper probability is larger than that for $Y_{10} \in \{LB, DB\}$. This results from the fact that the main categories R and Y can be fully separated, in the configurations for the probability wheel representation, by categories with positive numbers of observations in them, whereas the sub-categories LB and DB can only be separated by OB, in which there are no observations. If one of the observations in G had actually been in OB, then both these events considered here would have had the same upper probability 5/9.

It will be clear from this example that the CA model, as before, does not satisfy Walley's 'Representation Invariance Principle' (RIP) [23], a fact which we have commented on in detail before [8, 9], and which we do perceive as an advantage of our model. One could argue, however, that the fact that, in the CA model, it does not matter whether one uses information at main category level, or at sub-category level, as long as this category is not involved in the event of interest, is very close in nature to the underlying idea of Walley's RIP.

4 Concluding remarks

In this paper we have reviewed the CA model as presented so far [8, 9], and we have outlined the general manner in which the CA model can deal with data at

sub-category level, to get consistent inferences at both main and sub-category levels. Detailed expressions for lower and upper probabilities, for general events in a variety of situations with regard to assumed knowledge of numbers of (sub-)categories will be presented elsewhere, but all follow the basic concept outlined in Section 3 and illustrated in Example 2. This generalization of the CA model is of great practical use, as interest is often explicitly at sub-category levels, with potentially even more layers of sub-categories playing a role. As long as such different layers are representable by tree structures, the same approach as outlined here can be used, guaranteeing strong internal consistency of inferences at varying levels due to the use of the probability wheel representation. It remains important here that no actual ordering of (sub-)categories is known. If one wishes to use a multinomial approach with categories ordered, as for example Coolen [5] did for lifetime data on the basis of Walley's IDM, then the CA model with the probability wheel representation might not be suitable. In particular if one models time categories, with a natural one-dimensional ordering, the general framework of NPI offers more suitable modelling opportunities, as Coolen and Yan [10] presented for grouped lifetime data, using another variation of Hill's $A_{(n)}$ for dealing with right-censored data [11].

Throughout this paper, and in [8, 9], we assume to have perfect information on each observation, that is we know with certainty which unique (sub-)category it belongs to. If only partial information is available, in the sense that it is only known for a particular observation to belong to a subset of (sub-)categories [21, 27], then the CA model is easily adapted to deal with such information in a consistent manner, taking all possibilities of the values of that particular observation into account and again optimizing over all possible corresponding configurations of the observations on the probability wheel. However, all such generalizations make it harder to derive general expressions for the lower and upper probabilities for events of interest, as the combinatorial problems in deriving analytic solutions of the optimization processes involved become ever more complex.

In the CA model, as in NPI in general [2, 7], updating in the light of new observations is straightforward, as simply new lower and upper probabilities are calculated on the basis of the entire data set. Conditioning, however, is more complex [2], where conditioning is understood as taking additional information into account on the particular random quantity of interest, in contrast to information in the form of further observed exchangeable random quantities in updating. Generalization of the classical, precise probabilistic, concept of conditioning is acknowledged to be a complex issue in theory of lower and upper probability [24, 26], and this is not any different in conditioning within the CA model. For example, suppose that for the situation in Example 2, one learns that Y_{10} is actually Blue, but that one then is interested in which of the three specified sub-categories it belongs to. Following the basics of the NPI approach, and of the CA model, a correct way of arguing is that of the nine observations so far, only three can still be assumed to satisfy the post-data exchangeability assumption that is key for any inference based on $A_{(n)}$ and its variations such as $\mathcal{A}_{(n)}$, namely the three already observed Blue outcomes, of which one was LB and two were DB, with OB as only other sub-category assumed. Hence, instead of considering Y_{10} with a post-data exchangeability assumption with 9 available observations, one should now redefine the random quantity of interest as, say, $\tilde{Y}_4$, with 3 observations available, and (if deemed appropriate) one can use $\mathcal{A}_{(n)}$ with the three sub-categories now functioning as main categories, in which case the lower and upper probabilities for events involving $\tilde{Y}_4$ are easily derived using (9) and (10). Generally, the lower and upper probabilities for $\tilde{Y}_4$ derived in this manner are not proportional to those for the corresponding events involving Y_{10} and based on all 9 observations, before taking the information $Y_{10} = B$ into account. Although this is not a surprise due to the complex general nature of conditional lower and upper probabilities, detailed study of properties of such conditioning within the CA model is an important topic for future research.

Acknowledgements

We are grateful to referees for helpful suggestions on presentation. We were extremely impressed by the efforts of one referee who was kind enough to provide Latex code based on the tikz-package for the creation of the probability wheel plots used in this paper.

References

[1] T. Augustin. Generalized basic probability assignments. *International Journal of General Systems*, 34: 451-463, 2005.

[2] T. Augustin and F.P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124: 251-272, 2004.

[3] J.O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25: 303-328, 1990.

[4] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39: 123-150, 2005.

[5] F.P.A. Coolen. An imprecise Dirichlet model for Bayesian analysis of failure data including right-censored observations. *Reliability Engineering and System Safety*, 56: 61-68, 1997.

[6] F.P.A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36: 349-357, 1998.

[7] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15: 21-47, 2006.

[8] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. In: *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, F.G. Cozman, R. Nau and T. Seidenfeld (Eds), pp. 125-134, 2005.

[9] F.P.A. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. In revision for *International Journal of Approximate Reasoning*. SFB-Disc. Paper 489: `http://www.stat.uni-muenchen.de/sfb386/`

[10] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference for grouped lifetime data. *Reliability Engineering and System Safety*, 80: 243-252, 2003.

[11] F.P.A. Coolen and K.J. Yan. Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126: 25-54, 2004.

[12] G. de Cooman, E. Miranda, I. Couso, Lower previsions induced by multi-valued mappings, Journal of Statistical Planning and Inference 133: 173-197, 2004.

[13] A.P. Dempster. On direct probabilities. *Journal of the Royal Statistical Society, Series B*, 25: 100-110, 1963.

[14] B. De Finetti. *Theory of Probability*. Wiley, Chichester, 1974.

[15] S. French and D. Rios Insua. *Statistical Decision Theory*. Arnold, 2000.

[16] S. Geisser. *Predictive Inference: an Introduction*. Chapman and Hall, New York, 1993.

[17] A.D. Gordon. *Classification* (2nd Edition). Chapman and Hall, Boca Raton, 1999.

[18] B.M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63: 677-691, 1968.

[19] B.M. Hill. De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: *Bayesian Statistics 3*, Bernardo et al. (eds.), Oxford University Press, pp. 211-241, 1988.

[20] J.F. Lawless and M. Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92: 529-542, 2005.

[21] L.V. Utkin, T. Augustin. Decision making under incomplete data using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44: 322-338, 2007.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[23] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society B*, 58: 3-57, 1996.

[24] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24: 149-170, 2000.

[25] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

[26] K. Weichselberger, T. Augustin. On the competition and symbiosis of two concepts of conditional interval probability. In: *Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, J.M. Bernard, T. Seidenfeld and M. Zaffalon (Eds), pp. 608-629, 2003.

[27] M. Zaffalon. Conservative rules for predictive inference with incomplete data. In: *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, F.G. Cozman, R. Nau and T. Seidenfeld (Eds), pp. 406-415, 2005.

Jury size and composition - a predictive approach

F.P.A. Coolen, B. Houlding, S.G. Parkinson
Durham University, UK

Abstract

We consider two basic aspects of juries that must decide on guilt verdicts, namely the size of juries and their composition in situations where society consists of sub-populations. We refer to the actual jury that needs to provide a verdict as the 'first jury', and as their judgement should reflect that of society, we consider an imaginary 'second jury' to represent society. The focus is mostly on a lower probability of a guilty verdict by the second jury, conditional on a guilty verdict by the first jury, under suitable exchangeability assumptions between this second jury and the first jury. Using a lower probability of a guilty verdict naturally provides a 'benefit of doubt to the defendant' robustness of the inference. By use of a predictive approach, no assumptions on the guilt of a defendant are required, which distinguishes this approach from those presented before. The statistical inferences used in this paper are relatively straightforward, as only cases are considered where the lower probabilities according to Coolen's Nonparametric Predictive Inference for Bernoulli random quantities [5] and Walley's Imprecise Beta Model [24, 25] coincide.

Keywords. Imprecise Beta Model, lower probability, Nonparametric Predictive Inference, representation of sub-populations.

1 Introduction

In law, the use of juries is often regarded as a natural manner for reaching a verdict, mostly used when a defendant is charged with a serious crime. In such situations, there is typically uncertainty about the guilt of the defendant, and most civilized societies only wish to convict the defendant if there is considered to be very strong evidence that the defendant committed the crime: in case there is remaining doubt, the defendant should normally be given the benefit of the doubt, and should not be convicted. Due to the presence of uncertainty, it is natural that probabilistic and statistical methods have been used to analyze several theoretical aspects of juries (e.g. [11]), and of uncertainty in law more generally (e.g. [12]). During a trial, an enormous amount of information is typically presented to a jury. Such information may consist of many facts brought alight, with different emphasis on their relevance and circumstances under which these facts did or might have occurred (or not), and the manner in which this all is presented can be very confusing to members of the jury. Clearly, this makes it difficult to translate all such information into suitable data for a statistical approach based on a full model, and the one-off nature of specific court cases appears to prevent a classical frequentist statistical approach to support jurors in reaching a verdict. From a Bayesian perspective, it would be extremely difficult to provide a detailed model a priori, as one would have to foresee all possible information that might appear in a court case, in the right order (as e.g. the defence will often adapt its strategy to counter arguments presented by the prosecutor), and based on detailed expert judgements (as, effectively, only one realization of the whole process is actually observed, so any prior information is likely to remain influential). Of course, some aspects of 'uncertainty in law' have been discussed frequently, e.g. the so-called 'prosecutor's fallacy', which is a mistake due to confusion of conditional probabilities [1]. If one would wish to use Bayesian statistical reasoning to decide on a defendant's guilt, one would also require prior probabilities on his guilt. It would not only be very difficult to assess such prior probabilities meaningfully, but any explicit quantification of a juror's prior beliefs that the defendant is guilty would be considered to be highly inappropriate. Jurors are typically not trained in law, statistics or probability, so such an approach would be deemed to fail even if suggested. It is, therefore, very difficult to even consider a suitable general way in which statistics could assist jurors with their possibly very difficult task, namely that of deducing whether or not the defendant is guilty on the basis of

all evidence presented.

In this paper, we are certainly not attempting the impossible. However, we emphasize the complexity of the use of statistical methods to support jurors on deciding their verdict, as any such use of statistics is explicitly absent in the approach presented in this paper. We do not propose a method for quantifying a 'level of certainty about guilt', and we do not require any prior thoughts about the defendant's guilt. We focus our attention on juries, and we study size of juries from a novel perspective, from which we also consider composition of juries if a population consists of recognized sub-populations. The main novelty in our approach is that nowhere any assumptions are made about the defendant's guilt, and also no attempt is made to model the complex stream of information jurors have to consider during a process. By considering a predictive criterion, which is introduced and explained in Section 2, we can still comment meaningfully on appropriateness of jury sizes from a theoretical perspective. It is important to emphasize here that we do not take practicalities of the processes used by juries to reach an overall verdict into account [14], we assume throughout this paper that each juror takes the evidence presented into account and reaches a decision without conferring with other jurors. Actually, our approach even allows the latter to take place, but as outcomes of such deliberations might depend on particular personality characteristics of individual jurors, it would make the appropriateness of the key exchangeability assumption underlying our approach (Section 2) less clear.

Section 2 provides a short discussion of a typical statistical method for inference on jury verdicts and jury size, as presented in the literature. Then it presents the main criterion and assumptions underlying our novel approach, as well as the results of our approach on jury size. In Section 3 we show how this approach can be used to decide on optimal representations of 'independent' sub-populations in a jury, our approach as presented here also has some attractive features when compared to e.g. statistical methods for stratified sampling, which we will discuss briefly in Section 4 together with some further comments. Throughout this paper, uncertainty is quantified via lower and upper probabilities, where it is particularly attractive to use lower probabilities as, for the events considered, these effectively 'give the benefit of doubt' to the defendant. As we only consider a relatively straightforward statistical model with lower and upper probabilities, we use these without many further comments. For the events considered, lower and upper probabilities from Coolen's Nonparametric Predictive Inference for Bernoulli random quantities [5] coincide with those from Walley's Imprecise Beta Model [24], which is the special case of Walley's Imprecise Dirichlet Model for the situation with only two categories [25][1].

2 Jury size

Friedman [11] discusses different jury sizes and criteria for convictions, focussing on 12 jurors, with either a 12-out-of-12 or 10-out-of-12 criterion (the latter leading to a guilty verdict if supported by at least 10 of the 12 jurors), and on 6 jurors (6-out-of-6). He emphasizes that his analysis is not based on whether or not a person is actually guilty, and he also does not make any assumptions about guilt. Instead, he focusses on the degree to which the person appears to be legally guilty or the inverse, the degree to which he can defend himself. Friedman suggests that this appearance of guilt may be considered as equivalent to the probability that an individual juror would consider the defendant guilty, and assumes that the defendant affects each of the jurors equally and independently. This allows the use of the Binomial distribution, for given number of jurors and given degree of apparent guilt, to calculate the probability of conviction. Friedman then considers the probability of conviction as a function of this degree of apparent guilt, and discusses some characteristics of several jury systems from this perspective. Clearly, the unanimous 12-out-of-12 system has a relatively low probability of conviction for values of the degree of apparent guilt which are not close to 1. Friedman's discussion is in well-known statistical terms of errors of Type I, i.e. conviction of innocent individuals, and errors of Type II, i.e. failure to convict guilty individuals. This discussion is somewhat informal due to the change from assumed (non-) guilt to degree of apparent guilt. Friedman mentions that this statistical model is based on the assumption that all jurors are unbiased and equivalent in their perception. He briefly discusses the possibility of an atypical juror, which may be a strong argument in favour of jury systems that do not require unanimity. Essential in this approach is the introduction of a parameter, ϕ say, which, although not directly observable, is assumed to have a meaningful and unambiguous interpretation, in Friedman's work it is the degree of apparent guilt and $\phi \in [0, 1]$, with $\phi = 0$ meaning that the defendant is certainly not guilty, in the sense that his innocense is absolutely certain to every juror, and $\phi = 1$ meaning that every juror is absolutely certain of the defendant's guilt.

[1] For Walley's model, the value of a further parameter s in the notation of [25] must be chosen: throughout this paper we set $s = 1$ without further mentioning, as this is the value for which the lower and upper probabilities for the events considered coincide with those from Coolen's NPI approach.

Bayesian methods in statistics provide a framework for dealing with uncertainty about parameters in a consistent manner, namely by expressing subjective beliefs about such parameters, for an assumed statistical model, via prior probability distributions, which are then combined with observed data to give the posterior probability distribution of the parameters. In many situations this seems highly sensible, although it does explicitly require information about the parameters to be taken into account. Clearly, with the parameter used by Friedman, representing the defendants degree of apparent guilt, it may be a far from trivial task to model subjective beliefs about this parameter via a probability distribution. Nevertheless, it might be considered attractive to attempt a Bayesian approach to problems on adequate jury size and composition, with a parameter representing either the defendant's guilt, or Friedman's 'appearance of guilt'. However, in addition to the need for a prior distribution on such a parameter, any such an approach would require further assumed probabilities, namely for the variety of events which can be summarized as 'juror gives correct judgement'. Not only is it extremely difficult to have meaningful information on such events, let alone to quantify the uncertainty about them, these events are also (normally) unobservable and any assigned probability values will be influential on the overall inferential results.

In this paper, we present a different approach to considerations of jury size, and jury composition (Section 3). Let us consider the main reason for the very existence of a jury: it is assumed to represent the population in the sense that its final verdict should, ideally, be in line with that of 'the population', if 'the population' were confronted with the same information from the whole process. Of course, it is difficult to formulate any such a 'verdict of an entire population', we propose the following solution. Throughout this paper, we will refer to the actual jury as JA, and we consider a second, imaginary jury JI, also selected from the general population in a similar manner as JA. We now study aspects of JA by making some suitable exchangeability assumptions, and considering predictive inferences on JI's verdict based on information from JA's verdict. In particular, we will consider the lower probability of a guilty verdict by JI, given a guilty verdict by JA. We discuss this idea in more detail at the end of this section, we first develop the idea further and consider its implications for jury size considerations.

A first possible approach would be to assume exchangeability at the level of the juries, which may be most natural if JA and JI consist of the same number of jurors and the same conviction rule (required number of jurors' guilty votes to provide an overall jury guilty verdict) applies for both. In this setting, the precise conviction rule is of no actual relevance. We consider the JA verdict as one observation of a Bernoulli random quantity, and the JI verdict as a second Bernoulli random quantity which we wish to predict, and which we assume to be exchangeable with the JA verdict. Let us denote a guilty verdict of JA (JI) by $JA{-}G$ ($JI{-}G$). Both Coolen's NPI approach for Bernoulli random quantities [5], and Walley's IBM [24, 25] give $\underline{P}(JI-G|JA-G) = 1/2$, which does not provide much useful insight in this setting, and is certainly not very strong evidence that 'the population' would consider the guilty verdict appropriate. Of course, by conjugacy the corresponding upper probability of a not-guilty verdict by JI is $1/2$, so one could argue that this would support a guilty verdict as a fair representation of the population's judgement in such a case, but as it is generally accepted (in societies that like to consider themselves 'civilized') that a defendant is only convicted in case of strong evidence, and hence that the defendant should get the benefit of the doubt, this result based on assumed exchangeability at the jury level does not appear to be strong enough as a basis for decisions. For completeness, let us also mention the corresponding upper probability $\overline{P}(JI-G|JA-G) = 1$, which seems logical in such cases where there is no evidence in the available data (here the single observation $JA-G$) that there has to be any level of doubt about the defendants guilt.

A logical alternative approach to this problem is by focussing on the votes of individual jurors, and to assume exchangeability between jurors in JA and jurors in JI. From here on, we assume such exchangeability at the level of individual jurors. Focussing on individual jurors' votes, it becomes important to consider the conviction rule applied. From a mathematical perspective, it might be of interest to study all conviction rules that can be defined, in relation to real-world law scenarios it makes sense to restrict attention to k-out-of-K rules (with $k > K/2$), where the jury verdict is 'guilty' if at least k of the K jurors vote 'guilty'. Actually, we will focus on the unanimity conviction rule ($k = K$) for guilty verdicts of JA. It will be relevant, however, to consider more general k-out-of-K rules for JI, as we use JI to reflect the population at large, and as such it might for example be of interest to know the lower probability that JI would reach a guilty verdict under a specific conviction rule, given that the jurors in JA voted 'guilty' unanimously. For even wider flexibility, we will consider scenarios under which JA and JI are not required to consist of the same number of jurors, with n jurors in JA and m jurors in JI. It should be emphasized here that $n = 12$ is the present situation in many jury systems,

although studies of effectiveness of juries consisting of 6 or 8 jurors have been reported [9, 20, 23]. We will discuss below what unanimous guilty verdicts of juries JA of some sizes other than 12 imply for juries JI.

Coolen [5] derived and justified the following general results for nonparametric predictive inference (NPI) for $m+n$ exchangeable Bernoulli random quantities. Suppose that we have a sequence of $n+m$ exchangeable Bernoulli trials, each with 'success' and 'failure' as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n, then a sufficient representation of the data for our inferences is $Y_1^n = s$, due to the assumed exchangeability of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$, and, for ease of notation, let us define $\binom{s+r_0}{s} = 0$. Then the NPI-based upper probability for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

$$\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = \binom{n+m}{n}^{-1} \times \sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s}$$

The corresponding lower probability is derived via the conjugacy property

$$\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c | Y_1^n = s)$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

For the setting in the current paper, we are only considering data consisting of a unanimous guilty verdict of JA, so $s = n$, in which case we will denote $Y_1^n = n$ by (n, n), and the event that there are at least y successes in the following m observations, so $Y_{n+1}^{n+m} \geq y$ for which we will use the notation $Y_m \geq y$, leading to NPI lower probability, for $y = 0, 1, \ldots, m$

$$\underline{P}(Y_m \geq y | (n, n)) = 1 - \frac{(n+y-1)!m!}{(y-1)!(n+m)!} \qquad (1)$$

The corresponding upper probability is equal to 1, which is fully in line with intuition yet of little relevance for the rest of this paper. For Walley's imprecise Beta model [24], which is the special case of his Imprecise Dirichlet Model with only 2 categories [25], the lower probability for this event is identical to (1). This also coincides with the 'cautious' or 'conservative' Bayesian inference advocated by Hartigan [17] for such cases. It should be emphasized that Coolen's NPI and Walley's imprecise Beta model do not generally give identical upper and lower probabilities, with Coolen's NPI leading to slightly more imprecision for many events, due to the fact that it only assumes exchangeability of the $m+n$ random quantities involved whereas Walley's approach stays close to the robust Bayes framework [3], using a Binomial model which requires assumed embedding in an infinite sequence of such exchangeable random quantities [10].

For our jury problem, with the n jurors in JA all voting guilty and m jurors in JI, (1) provides the lower probability (according to NPI and Walley's model) that JI would also reach a guilty verdict under a y-out-of-m rule. If $m = n$ and $y = m$, so JI also requires a unanimous guilty vote to reach a guilty verdict, with JI and JA having the same number of jurors, then this lower probability is 1/2. This is naturally in agreement with the situation, briefly described above, where only exchangeability at jury level is assumed, which holds of course here due to the assumed exchangeability of jurors, and the same number of jurors and conviction rules for JA and JI in this situation[2]. More generally, if we assume that unanimity is also required for a guilty verdict of JI (so $y = m$), but we do not restrict the value of m, then the lower probability (1) is equal to $n/(n+m)$, which is increasing in n and decreasing in m, also in line with intuition.

Let us consider some numerical values of (1) for situations of relevance to our discussion on jury size. These values are explicitly given in the text below, and to aid the discussion they are also presented in Table 1.

$n = 12,\ m = 12$:					
$y =$	11	10	9	8	7
$\underline{P} =$	.761	.891	.953	.981	.993
$n = 6,\ m = 6$:					
$y =$	5	4			
$\underline{P} =$	.773	.909			
$n = 6,\ m = 12$:					
$y =$	8	7			
$\underline{P} =$	.908	.950			
$n = 24,\ m = 24$:					
$y =$	23	20	13		
$\underline{P} =$	.755	.975	.99996		
$n = 24,\ m = 12$:					
$y =$	7				
$\underline{P} =$	.9995				
$n = 12,\ m = 100$:					
$y =$	100	99	95	51	
$\underline{P} =$	.107	.204	.502	.9995	

Table 1: Some values of $\underline{P} = \underline{P}(Y_m \geq y | (n, n)))$

[2]For simplicity, we assume here that the unanimity rule actually applied for JA: if this is not the case, then one might not learn the exact number of jurors in JA that voted 'guilty' - both NPI and Walley's method can, of course, also deal with information that would appear in such cases, but we do not discuss this explicitly in this paper.

First of all, for $m = n = 12$, these lower probabilities for the values $y = 11, 10, 9, 8, 7$ are 0.761, 0.891, 0.953, 0.981, 0.993, respectively. This means that, if a 12-person jury JA reaches a unanimous guilty verdict, then the lower probability that the majority of members of a second 12-person jury, with all 24 individual jurors involved assumed to be exchangeable (with regard to their individual votes - a further discussion of this assumption is provided near the end of this section), would have reached the same guilty verdict, is very high indeed (0.993). One can interpret this as a reflection of the strength of evidence of the information in the JA guilty verdict. However, one could argue that, ideally, a substantial majority of the population should (be expected to) agree with the guilty verdict, so perhaps the values 0.891 (for $y = 10$) or 0.953 ($y = 9$) are more natural to focus on. As mentioned above, several studies have focussed, both from theoretical and practical perspectives, on juries of smaller sizes, in particular 6-person juries have been considered [9, 20, 23]. For the case where both JA and JI are 6-person juries, so with $m = n = 6$, the lower probability (1) is equal to $0.773, 0.909$ for $y = 5, 4$, respectively. So, the unanimous guilty vote of the 6-person jury JA now only implies a lower probability of 0.909 for the event that a majority of the 6-person jury JI would agree with this verdict, which is a substantial reduction from the 0.993 for the corresponding lower probability if both JA and JI consisted of 12 persons. Another method that could be used to compare actual jury sizes 12 and 6, is by considering the lower probability (1) for $n = 6$, but with $m = 12$ and $y = 7$, which is equal to 0.950. However, due to the discrete nature of these events comparisons are slightly complicated, as $m = 6$ and $y = 4$ more naturally relates to the case with $m = 12$ and $y = 8$, for the latter (still with $n = 6$) the lower probability (1) is equal to 0.908, which is very close to the 0.909 for the former case. Studies of the performance of juries of size 6 are mostly initiated by practical aspects of 12-person juries, unfortunately also including considerations of costs. From such a perspective, the increased risk of getting a JA guilty verdict under the unanimity rule for a 6-person jury, which would not be in line with the verdict of the majority of the larger population, and when compared to a 12-person jury, might need to be balanced with such cost considerations, although this would involve consideration of utilities at a level that many might find 'unethical' to do explicitly as it would require balancing between utilities of an individual (the defendant) and of society at large.

It is also interesting to see what could be gained, in terms of the lower probability (1), by doubling the JA size to $n = 24$. For $m = 24$, (1) is equal to 0.755 for $y = 23$, 0.975 for $y = 20$ and 0.99996 for $y = 13$, while for $m = 12$ the lower probability of a majority of JI jurors agreeing with the guilty verdict (so $y = 7$) is equal to 0.9995 (which one may wish to compare to the corresponding values, as mentioned above, of 0.993 and 0.950 for $n = 12$ and $n = 6$, respectively).

When considering the role of JI in representing the society at large, one can also argue that a substantially larger value of m would be appropriate. In Section 3, when considering jury composition in case of a population consisting of subgroups, we will find the use of size $m = 100$ for JI convenient. For the current scenario, $m = 100$ leads, for $n = 12$, to the following values for the lower probability (1): for $y = 100, 99, 95, 51$ we get $0.107, 0.204, 0.502, 0.9995$, respectively. Notice that this lower probability for $y = 51$, reflecting that a majority of JI will vote guilty given the unanimous guilty vote of the 12 jurors in JA, is greater than the corresponding lower probabilities for a majority of guilty votes in JI in the situations discussed above, with smaller values of m. Of course, for increasing m, NPI [5] requires an exchangeability assumption over an increasing number of random quantities. This raises the question of what happens if $m \to \infty$. For a meaningful answer, let us consider the limit of the right-hand side of (1) with $y = \theta m$ for $0 < \theta \leq 1$:

$$\lim_{m \to \infty} \left(1 - \frac{(n + \theta m - 1)! m!}{(\theta m - 1)! (n + m)!} \right) = 1 - \theta^n \qquad (2)$$

In this limiting situation (see [5] for a similar argument), the exchangeability assumption in NPI becomes 'infinite exchangeability', for which case De Finetti's Representation Theorem [10] shows that one could represent the random quantities involved as conditionally independent given a parameter, where the parameter is also a random quantity. One might see parallels between such a parameter and the above θ, but they are different, as our θ only has a meaning in the predictive inference considered, that is to specify events of interest, and is not considered to be an unknown property of the infinite sequence of future observations considered in this inference. Our inferences do not require the use of a prior distribution for θ, which would necessarily have required additional assumptions which we try to avoid. This limit $1 - \theta^n$ of $\underline{P}(Y_m \geq \theta m | (n, n))$ is decreasing in θ, which makes immediately clear that θ should not be interpreted as a limit for the proportion of guilty votes for the m jurors considered in JI. For illustration, this limiting lower probability (2) is given in Table 2, for some values of n and θ. Although these limiting values provide some insight, we find the actual inferences quite confusing as populations from which juries are selected will never be of infinite size, so restricting attention

to JI of smaller sizes, as discussed above, seems more in line with intuition.

n	$\theta = 0.50$	0.75	0.90	0.95	0.99
6	0.9844	0.8220	0.4686	0.2649	0.0585
12	0.9998	0.9683	0.7176	0.4596	0.1136
24	1.0000	0.9990	0.9292	0.7080	0.2143

Table 2: Some limiting lower probabilities (2)

Before we consider corresponding inferences on appropriate representative subgroups of different subpopulations (Section 3), we discuss the underlying exchangeability assumption between jurors in a bit more detail, also from the perspective of NPI [5] and Hill's assumption $A_{(n)}$ which is implicit in NPI.

It seems sensible to assume exchangeability of the individual jurors in JA and JI, as we did above (apart from the first considerations, when we only assumed exchangeability of the two juries JA and JI). For NPI [5], this exchangeability is actually assumed with regard to an assumed underlying representation of the Bernoulli random quantities which is very similar to the representation used by Thomas Bayes [2]. It is assumed that, corresponding to the Bernoulli random quantities, there are real-valued random quantities which are not observable, but which are so that if they exceed an unknown threshold they are 'successes', else they are 'failures'. Coolen's NPI for Bernoulli data [5] uses this representation together with Hill's assumption $A_{(n)}$[3], which effectively for this real-valued setting is a 'post-data exchangeability' assumption, meaning that the exchangeability assumption on $n + m$ random quantities still holds, for as far as prediction of m random quantities is concerned, once the values of the first n are known. This representation might be quite appropriate in a jury setting, as one could consider an underlying process where each individual juror reaches a conclusion on the strength of their believe in the guilt of the convicted person, and compares this strength to an individual 'guilt threshold value' to reach the individual vote. For the exchangeability assumption used in our approach, one could assume that the differences between each individual's strength of believe in guilt and corresponding individual guilt threshold value would be the unobservable real-valued random quantity in the assumed representation underlying NPI. Hence, we do not need to assume that all jurors would actually have the same guilt threshold value, nor that the strengths of their beliefs of guilt must be comparable. The fact that such concepts are not measurable in a meaningful manner supports the appropriateness of $A_{(n)}$ in this setting [6, 18, 19], as one never gets information that could be used to counter the underlying exchangeability assumption.

The question whether or not the exchangeability assumption is really appropriate here is quite subtle. It is again important to emphasize that we only assume exchangeability of the m and n jurors in JA and JI, which is reasonable if we have no specific information on these individuals and if we would assume that jurors in JI would be selected from the large population by the same process as used to select the jurors in JA. This, however, might not imply that these jurors are exchangeable with all members of the population, as the selection process is likely to favour or exclude some in the population. However, we believe that this issue is inherent to any selection procedure for juries, and therefore to any legal system that uses juries, and we consider it an advantage that our method does not actually need to assume such exchangeability between all members of society (the above discussion involving the limit for $m \to \infty$ was included more for its theoretical value than for its real-world relevance).

At the beginning of this section, we reviewed the approach by Friedman [11], which in a classical statistical manner focusses on errors of Type I and Type II for jury verdicts, and which makes clear the inherent difficulty when representing the defendant's guilt, or a corresponding 'degree of apparent guilt', in the statistical reasoning. The method presented in this section does not make use of any of these concepts, and only looks at jury verdicts under assumed exchangeability of jurors, so it explicitly does not add any assumption or inference on whether or not the jury is correct. It is important to emphasize this, as many might consider this a disadvantage. However, in most individual situations it will by the nature of court cases not be known whether or not the defendant is guilty, and avoiding any attempt to quantify beliefs about actual (or apparent) guilt seems to simplify the discussion in a straightforward and fair manner. Of course, methods such as Friedman [11] presented have their merits, but we believe that our method provides useful additional insights and possible arguments on appropriate jury sizes. We have only considered our approach under assumed unanimous guilty verdicts by JA. The approach is easily extended to also consider more general k-out-of-K conviction rules for JA, but as we have no ambition to propose, or even consider, an optimal rule, we do not address such different rules for JA further in this paper.

[3] We use the notation $A_{(n)}$ here generically, for inference on m future observations the actual assumption made is, in notation of Hill [18, 19] $A_{(n+m-1)}$, which also implies $A_{(l)}$ for all $l < n + m - 1$ [5].

3 Jury composition

In this section, we briefly consider the interesting question of how to select representative juries from populations that consist of known separate sub-populations, where we assume independence of these sub-populations with regard to the individual verdicts of jurors from different sub-populations. We assume that the number and (relative) sizes of the sub-populations are known, and also that for each member of the population the sub-population to which they belong is known. We use the same general approach as in Section 2, with the actual jury JA and the imaginary jury JI, where the use of JI provides a convenient way for taking the relative sizes of the sub-populations into account. We assume that the individual verdicts of jurors belonging to the same sub-population are exchangeable, as before, and per sub-population we use the same lower (and upper) probabilities as in Section 2. In most of this section, we consider only two sub-populations. For more sub-populations, the general conclusion remains valid.

Let the two sub-populations be denoted by A and B, with $p_A \in (0,1)$ the proportion of the whole population that belongs to A. Let jury JA consist of n_A jurors from A and n_B from B, with $n_A + n_B = n$, and jury JI of m_A jurors from A and m_B from B, with $m_A + m_B = m$. An intuitive way to choose the numbers of jurors from each sub-population in JA, assuming that n has already been chosen, is by taking n_A as close as possible to $p_A n$, so to achieve proportional representation of the sub-populations in JA. However, if again we consider the jurors as representatives of the population, and hence of the sub-populations, this choice might not be optimal from a similar perspective as used in Section 2, namely when considering the lower probability that a second jury JI would also provide a guilty verdict if JA does so. A natural manner in which to reflect the relative sizes of the sub-populations is by choosing (approximately) the same proportions for the numbers of representatives in JI, as throughout our approach the role of the imaginary jury JI is to reflect the larger population. We saw in Section 2 that the actual choice of the size m of JI affects the predictive inferences of interest, but as we just want to introduce our approach for this setting, we will use $m = 100$ for illustrations in this section. So JI will be assumed to consist of $100p_A$ (rounded to nearest integer to give m_A) jurors from A, and $m_B = 100 - m_A$ jurors from B. For this JI, which clearly reflects the sub-populations, we now wish to choose n_A and n_B, under the assumption that $n_A + n_B = n$ and n is predetermined, such that a verdict of guilty by JA leads to maximum lower probability of a guilty verdict by JI. In this paper, we only consider unanimity conviction rules for both JA and JI in this situation, the approach is easily generalized to more general conviction rules for JA, JI or both. Due to the assumed independence of individual jurors' verdicts between jurors from JA and from JI, the lower probability of the event that all $m_A + m_B$ jurors in JI vote guilty, given all $n_A + n_B$ jurors in JA voted guilty (and under the same exchangeability assumptions per sub-population as used throughout this paper), is equal to

$$\frac{n_A}{n_A + m_A} \times \frac{n_B}{n_B + m_B}$$

By a basic exercise one can derive a general expression for the optimal choices of n_A and n_B which achieve the maximum value for this lower probability, but these do not provide much general insight, apart from the fact that the optimal fraction n_A/n is equal to $1/2$ if $p_A = 1/2$ (this is of course logical by symmetry), but will be closer to $1/2$ than p_A is in all other cases. In other words, the smaller of the two sub-populations will relatively be over-represented in JA, of course with this all under the constraint due to the discrete nature of n_A and the fact that n is likely to be small. For example, the optimal number n_A in a $n = 20$ person jury JA, for $m = 100$ (under the unanimity conviction rule for both juries), is equal to 8 for $p_A = 0.1$, 9 for $p_A = 0.2$ and for $p_A = 0.3$, and 10 for $p_A = 0.4$ and for $p_A = 0.5$. The optimal values of n_A for p_A greater than $1/2$ follow by symmetry. It might be considered to be remarkable that, for $p_A = 0.1$ and the imaginary jury JI consisting of 100 jurors (so 10 from A and 90 from B), $n_A = 8$ and $n_B = 12$ would give the optimal 20-person jury according to this predictive criterion. The lower probability optimised here is actually pretty robust if one varies n_A a little from this optimum, but it is substantially larger than if one would only select 2 jurors from A and 18 from B ('proportional representation'), namely 0.0523 versus 0.0278 for the latter case. Of course, these lower probabilities are pretty small as m is quite large, but if one relaxes the conviction rule for JI, similar results are achieved. Overall, this over-representation of smaller sub-groups is not really surprising, as the additional information from an extra juror added to a small number of jurors for a particular subgroup, in terms of the predictive power of the total information, is stronger than the corresponding information lost by reducing a larger number of jurors for the different subgroup accordingly.

We do not wish to provide a more detailed study of this approach to decisions on jury composition, as the main goal here is the introduction of this criterion using the predictive lower probability of a guilty verdict by JI, given a guilty verdict by JA, and to emphasize

the attractive role of JI in representing the population. Naturally, there are many related topics that can be studied, and for some of these we did some preliminary analyses and calculations. For example, in the situation of two sub-populations, the influence of particular choices of m and n can be considered (the over-representation of the smaller sub-population to achieve optimality holds generally), and more general conviction rules can also be studied. We calculated several cases, only relaxing the conviction rule of JI, and the over-representation of the smaller sub-population was always present, be it to a lesser extent than for the unanimity rule for JI. For example, corresponding to the case discussed above with $m = 100$ and $n = 20$, if we use the 97-out-of-100 rule for conviction by JI, then for $p_A = 0.1$ the optimal n_A is equal to 6 (instead of 8 for unanimity as discussed above). This effect seems logical, as the loss of detailed information about the sub-population A is less likely to have a substantial influence on JI's overall verdict in the latter situation. We also performed some calculations for three sub-populations, in which case also the smallest (largest) sub-population is over-represented (under-represented) in the optimal jury composition. For example, again with $n = 20$ and $m = 100$, if sub-populations A, B and C consist of 10, 10 and 80 percent of the population, then the optimal representations are 6, 6 and 8, respectively, under the unanimity conviction rule for both juries JA and JI.

4 Concluding remarks

There is a considerable literature on the use of statistical methods in relation to aspects of law, including attention to specific problems involving juries which particularly received much attention in the seventies [9, 11, 13, 14, 15, 16, 23]. In addition to these mostly theoretical studies on jury size and conviction rules, there are also many studies of actual jury behaviour, see for example Ellsworth [8] who reports on a detailed observational study with attention to a variety of practical aspects, consideration of which goes far beyond the theoretical goals of the current paper. However, the use of lower and upper probabilities [24, 26, 27] in law scenarios is, unfortunately, still pretty rare, whereas it provides an attractive method to deal with the 'benefit of doubt to the defendant' issue which in law seems to be quite generally accepted, and more appealing than perhaps in many other areas where uncertainty is quantified to enable inference and decision making. In the discussion to Walley's paper which introduced the Imprecise Dirichlet Model [25], one discussant remarked that the first ever recorded use of lower and upper probability was actually in a law problem, by Ostrogradsky. The current authors have not been able to verify this claim, yet it is of interest to mention that Ostogradsky [21, 22] did consider two types of judge ('juror' in our terminology), namely 'condemning judges' and 'acquitting judges', and assumed different probability distributions for these, considering the propensity to render a guilty verdict when the person on trial is actually innocent. He then proceeded to calculate the probability of erroneous majority judgement, and using the 'principle of insufficient reason' for the prior probability of guilt, he showed that this probability of erroneous majority judgement only depends on the difference between the numbers of condemning and acquitting judges involved. Although this does not involve, neither explicitly nor in its nature, lower and upper probabilities, the idea to study the influence on different-natured jurors would be of interest to also study from our perspective, although it could not be embedded naturally in an NPI approach as such juror characteristics would typically not be observable.

The major contributions of this paper are the novel use of an imaginary 'second' jury JI to represent the larger population in a predictive statistical framework, with the corresponding opportunity to study appropriateness of real jury (JA) sizes and conviction rules, and the fact that the inferences do not make any assumptions on actual (or apparent) guilt of the defendant and also do not even attempt to conclude on such guilt. This work can be extended in many ways, most clearly of course by studying other conviction rules for JA, JI or both. In Section 3, the predictive approach was suggested for decisions on appropriate representations of sub-populations. This problem can also be considered from the classical perspective of 'stratified sampling' [4], where one often uses criteria considering the overall variance of a random outcome. The predictive approach presented here is an attractive alternative to classical stratified sampling, and could be studied in detail for more general sampling scenarios.

This approach could also allow an alternative to traditional Type I and Type II errors, with the former formulated as the event that JI would not convict the defendant when JA does reach a guilty verdict, and the latter as the event that JI would convict the defendant when JA does not. One would be particularly interested in the upper probabilities for these events. In this paper we have focussed on the lower probability of a guilty verdict by JI, given a guilty verdict by JA, which would correspond via the conjugacy property to the upper probability of a Type I error, if the latter was defined as suggested. We have not considered the Type II error, but we acknowledge that detailed study of its upper probability could provide

useful insights into this predictive approach to issues related to juries. The goal of this paper was not to present such a detailed study, but to propose a new approach to a classical theoretical problem. The paper was also not aimed at specific real-world jury scenarios, where far more complicated issues often play a role. Nevertheless, we believe that the results from this theoretical exercise can provide new insights into practical issues related to the use of juries.

In a study of jury size and composition, one might expect a general conclusion on 'best choices'. We do not pretend to be well placed to give such advice, as our only ambition has been to introduce a novel manner for study of jury size and composition that has the advantages described above. Practical limitations make it unlikely that jury sizes in law would increase, and of course from the perspective of the defendant it seems best (under the jurors' exchangeability assumptions) to have the maximum possible number of jurors and the strictest conviction rule. However, although we addressed this problem from the perspective of juries in law, a similar approach can be used for other decision problems involving representative groups. If there is not such a clear direction in which 'benefit of doubt' should be applied, one may wish to take both lower and upper probabilities into account, but even then the predictive approach proposed in this paper appears to provide sufficient promise to warrant further study.

Acknowledgements

We are grateful to Colin Aitken, Minh Ha-Duong and Eugene Seneta for providing useful suggestions on relevant literature and copies of relevant papers, and to referees for helpful suggestions on presentation. Steven Parkinson's contribution to this research was supported by an Undergraduate Research Bursary from The Nuffield Foundation.

References

[1] C.G.G. Aitken. Interpretation of evidence, and sample size determination. In: Gastwirth, J.L. (Ed.). *Statistical Science in the Courtroom*. Springer, New York, pp. 1-24, 2000.

[2] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370-418; 54: 296-325, 1763.

[3] J.O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25: 303-328, 1990.

[4] W.G. Cochran. *Sampling Techniques* (3rd Ed.). Wiley, New York, 1977.

[5] F.P.A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36: 349-357, 1998.

[6] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15: 21-47, 2006.

[7] F.P.A. Coolen and P. Coolen-Schrijner. Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, 137: 23-33, 2007.

[8] P.C. Ellsworth. Are twelve heads better than one? *Law and Comtemporary Problems*, 52: 205-224, 1989.

[9] V. Fabian. On the effect of jury size. *Journal of the American Statistical Association*, 72: 535-536, 1977.

[10] B. De Finetti. *Theory of Probability*. Wiley, Chichester, 1974.

[11] H. Friedman. Trial by jury: criteria for convictions, jury size and type I and type II errors. *The American Statistician*, 26: 21-23, 1972.

[12] J.L. Gastwirth (Ed.). *Statistical Science in the Courtroom*. Springer, New York, 2000.

[13] A.E. Gelfand and H. Solomon. A study of Poisson's models for jury verdicts in criminal and civil trials. *Journal of the American Statistical Association*, 68: 271-278, 1973.

[14] A.E. Gelfand and H. Solomon. Modeling jury verdicts in the American legal system. *Journal of the American Statistical Association*, 69: 32-37, 1974.

[15] A.E. Gelfand and H. Solomon. Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association*, 70: 305-310, 1975.

[16] A.E. Gelfand and H. Solomon. Comments on 'On the effect of jury size'. *Journal of the American Statistical Association*, 72: 536-537, 1977.

[17] J.A. Hartigan. *Bayes Theory*. Springer, New York, 1983.

[18] B.M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63: 677-691, 1968.

[19] B.M. Hill. De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: *Bayesian Statistics 3*, J.M. Bernardo, et al. (eds.), Oxford University Press, pp. 211-241, 1988.

[20] M. Hunter. Improving the jury system: reducing jury size. *Public Law Research Institute*, 1996. (w3.uchastings.edu/ plri/spr96tex/jurysiz.html)

[21] M.V. Ostrogradsky. Extrait d'un mémoire sur la probabilité des erreurs des tribunaux. Bulletin Scientifique, No. 3. Sciences Mathématiques et Physiques. L'Académie Impériale des Sciences de Saint-Pétersbourg, 1, pp. xix-xxv, 1834 (published in 1838).

[22] E. Seneta. M.V. Ostrogradsky as probabilist. In: *Mikhail Ostrogradsky - Honoring his Bicentenary*, A. Samoilenko and H. Syta (eds.), Institute of Mathematics, National Academy of Sciences of Ukraine, pp. 69-81, 2001.

[23] D.A. Vollrath and J.H. Davis. Jury size and decision rule. In: *The Jury: Its Role in American Society*, R.J. Simon (ed.), 1980.

[24] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[25] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society B*, 58: 3-57, 1996.

[26] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24: 149-170, 2000.

[27] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

On coherent immediate prediction: Connecting two theories of imprecise probability

Gert de Cooman and **Filip Hermans**
SYSTeMS Research Group
Ghent University
Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium
{gert.decooman,filip.hermans}@UGent.be

Abstract

We give an overview of two approaches to probability theory where lower and upper probabilities, rather than probabilities, are used: Walley's behavioural theory of imprecise probabilities, and Shafer and Vovk's game-theoretic account of probability. We show that the two theories are more closely related than would be suspected at first sight, and we establish a correspondence between them that (i) has an interesting interpretation, and (ii) allows us to freely import results from one theory into the other. Our approach leads to an account of immediate prediction in the framework of Walley's theory, and we prove an interesting and quite general version of the weak law of large numbers.

Keywords. Game-theoretic probability, imprecise probabilities, coherence, conglomerability, event tree, lower prevision, immediate prediction, Prequential Principle, law of large numbers, Hoeffding's inequality.

1 Introduction

In recent years, we have witnessed the growth of a number of theories of uncertainty, where imprecise (lower and upper) probabilities and previsions, rather than precise (or point-valued) probabilities and previsions, have a central part. Here we consider two of them, Glenn Shafer and Vladimir Vovk's game-theoretic account of probability [18], which is introduced in Section 2, and Peter Walley's behavioural theory [20], outlined in Section 3. These seem to have a rather different interpretation, and they certainly have been influenced by different schools of thought: Walley follows the tradition of Frank Ramsey [10], Bruno de Finetti [4] and Peter Williams [24] in trying to establish a rational model for a subject's beliefs in terms of her behaviour. Shafer and Vovk follow an approach that is strongly coloured by ideas about gambling systems and martingales. They use Cournot's Principle to interpret lower and upper probabilities (see [17]; and [18, Chapter 2] for a nice historical overview), whereas on Walley's approach, lower and upper probabilities are defined in terms of a subject's betting rates.

What we set out to do here, and in particular in Sections 4 and 5, is to show that in many practical situations, the two approaches are strongly connected.[1] This implies that quite a few results, valid in one theory, can automatically be converted and reinterpreted in terms of the other. Moreover, we shall see that we can develop an account of coherent immediate prediction in the context of Walley's behavioural theory, and prove, in Section 6, a weak law of large numbers with an intuitively appealing interpretation. We use this weak law in Section 7 to suggest a way of scoring a predictive model that satisfies A. Philip Dawid's *Prequential Principle* [1, 2].

2 Shafer and Vovk's game-theoretic approach to probability

In their game-theoretic approach to probability [18], Shafer and Vovk consider a game with two players, World and Skeptic, who play according to a certain *protocol*. They obtain the most interesting results for what they call *coherent probability protocols*. This section is devoted to explaining what this means.

G1. The first player, World, can make a number of moves, where the possible next moves may depend on the previous moves he has made, but do not in any way depend on the previous moves made by Skeptic.

This means that we can represent his game-play by an event tree (see also [14, 16] for more information about event trees). We restrict ourselves here to the discussion of *bounded protocols*, where World makes only a finite and bounded number of moves from the beginning to the end of the game, whatever happens. But we do not exclude the possibility that at some point in the tree, World has the choice between an infinite number of next moves.

[1] Our line of reasoning here should be compared to the one in [17], where Shafer *et al.* use the game-theoretic framework developed in [18] to construct a theory of predictive upper and lower previsions whose interpretation is based on Cournot's Principle. See also the comments near the end of Section 5.

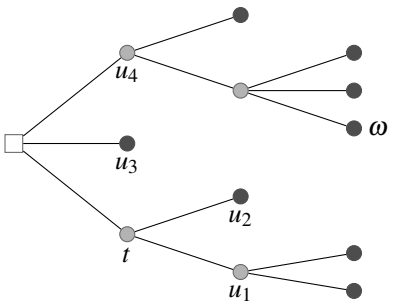


Figure 1: A simple event tree for World, displaying the initial situation $\Box$, other non-terminal situations (such as t) as grey circles, and paths, or terminal situations, (such as ω) as black circles. Also depicted is a cut of $\Box$, consisting of the situations u_1, u_2, u_3 and u_4.

Let us establish some terminology related to World's event tree. A *path* in the tree represents a possible sequence of moves for World from the beginning to the end of the game. We denote the set of all possible paths ω by Ω, the *sample space* of the game. A *situation* t is some connected segment of a path that is *initial*, i.e., starts at the root of the tree. It identifies the moves World has made up to a certain point, and it can be identified with a node in the tree. We denote the set of all situations by $\Omega^\diamondsuit$. It includes the set Ω of *terminal* situations, which can be identified with paths. All other situations are called *non-terminal*; among them is the *initial* situation $\Box$, which represents the empty initial segment. See Figure 1 for a simple graphical example explaining these notions.

If for two situations s and t, s is a(n initial) segment of t, then we say that s *precedes* t or that t *follows* s, and write $s \sqsubseteq t$. If ω is a path and $t \sqsubseteq \omega$ then we say that the path ω *goes through* situation t. We write $s \sqsubset t$, and say that s *strictly precedes* t, if $s \sqsubseteq t$ and $s \neq t$. Denote by $\uparrow t := \{\omega \in \Omega \colon t \sqsubseteq \omega\}$ the set of all paths that go through t. If we call any subset of Ω an *event*, then $\uparrow t$ is the event that corresponds to World getting to a situation t. It is clear that not all events will be of the type $\uparrow t$.[2]

Any (partial) function on $\Omega^\diamondsuit$ is called a *process*, and any process whose domain includes all situations that follow a situation t is called a *t-process*. A special t-process is the *distance* $d(t,\cdot)$ which for any situation $s \sqsupseteq t$ returns the number of steps $d(t,s)$ along the tree from t to s. In the bounded protocols we are considering here, there is a natural number D such that $d(t,s) \leq D$ for all $s \sqsupseteq t$.

Similarly, any (partial) function on Ω is called a *variable*, and any variable on Ω whose domain includes all paths that go through a situation t is called a *t-variable*. If we restrict a t-process $\mathscr{F}$ to the set $\uparrow t$ of all terminal situations that follow t, we obtain a t-variable, which we denote by $\mathscr{F}_\Omega$.

Call a *cut* U of a situation t any set of situations that (i) follow t, and (ii) such that for all paths ω through t [$t \sqsubseteq \omega$], there is a unique $u \in U$ that ω goes through [$u \sqsubseteq \omega$]; see also Figure 1. A set U of situations is a cut of t if and only if the corresponding set $\{\uparrow u \colon u \in U\}$ is a partition of $\uparrow t$. A cut can be interpreted as a (complete) stopping time.

If a situation $s \sqsupseteq t$ precedes (follows) some element of a cut U of t, then we say that s *precedes* (*follows*) U, and we write $s \sqsubseteq U$ ($s \sqsupseteq U$). Similarly for 'strictly precedes (follows)'. For two cuts U and V of t, we say that U *precedes* V if each element of U is followed by some element of V.

A *child* of a non-terminal situation t is a situation that immediately follows it. The set $C(t)$ of children of t constitutes a cut of t, called its *children cut*. Also, the set Ω of terminal situations is a cut of $\Box$, called the *terminal cut*. $\uparrow t$ is the corresponding terminal cut of a situation t.

If U is a cut of t, then we call a t-variable g *U-measurable* if for all u in U, g assumes the same value $g(u) := g(\omega)$ for all ω that go through u. In that case we can also consider g as a variable on U, which we can denote as g_U.

If $\mathscr{F}$ is a t-process, then with any cut U of t we can associate a t-variable $\mathscr{F}_U$, which assumes the same value $\mathscr{F}_U(\omega) := \mathscr{F}(u)$ in all ω that follow $u \in U$. This t-variable is clearly U-measurable, and can be considered as a variable on U. This notation is consistent with the notation $\mathscr{F}_\Omega$ introduced earlier. Similarly, we can associate with $\mathscr{F}$ a new, *U-stopped*, t-process $U(\mathscr{F})$, as follows:

$$U(\mathscr{F})(s) := \begin{cases} \mathscr{F}(s) & \text{if } t \sqsubseteq s \sqsubseteq U \\ \mathscr{F}(u) & \text{if } u \in U \text{ and } u \sqsubseteq s. \end{cases}$$

The t-variable $U(\mathscr{F})_\Omega$ is U-measurable, and is actually equal to $\mathscr{F}_U$.

We call a *move* $\mathbf{w}$ for World in a non-terminal situation t any arc that connects t to one of its children $s \in C(t)$, meaning that $s = t\mathbf{w}$ is the concatenation of the segment t and the arc $\mathbf{w}$. World's *move space* in t is the set $\mathbf{W}_t$ of those moves $\mathbf{w}$ that World can make in t: $\mathbf{W}_t = \{\mathbf{w} \colon t\mathbf{w} \in C(t)\}$. We have already mentioned that $\mathbf{W}_t$ may be infinite. But it should contain at least two elements (otherwise there is no choice for World to make).

We now turn to the other player, Skeptic. His possible moves may well depend on the previous moves that World has made, in the following sense. In each non-terminal situation t, he has some set $\mathbf{S}_t$ of moves $\mathbf{s}$ available to him, called Skeptic's *move space* in t.

G2. In each non-terminal situation t, there is a (positive or negative) gain for Skeptic associated with each of the possible moves $\mathbf{s}$ in $\mathbf{S}_t$ that Skeptic can make. This gain depends only on the situation t and the next move $\mathbf{w}$ that World will make.

[2] Shafer [15] calls events of this type *exact*. Further on, in Section 4, exact events will be the only events that can be legitimately conditioned on, because only they may occur as part of World's game-play.

This means that for each non-terminal situation t there is a *gain function* $\lambda_t : \mathbf{S}_t \times \mathbf{W}_t \to \mathbb{R}$, such that $\lambda_t(\mathbf{s},\mathbf{w})$ represents the change in Skeptic's capital in situation t when he makes move $\mathbf{s}$ and World makes move $\mathbf{w}$.

Let us introduce some further notions and terminology related to Skeptic's game-play. A *strategy* $\mathscr{P}$ for Skeptic is a partial process defined on the set $\Omega^{\Diamond} \setminus \Omega$ of non-terminal situations, such that $\mathscr{P}(t) \in \mathbf{S}_t$ is the move that Skeptic will make in each non-terminal situation t. With each such strategy $\mathscr{P}$ there corresponds a *capital process* $\mathscr{K}^{\mathscr{P}}$, whose value in each situation t gives us Skeptic's capital accumulated so far, when he starts out with zero capital and plays according to the strategy $\mathscr{P}$. It is given by the recursion relation

$$\mathscr{K}^{\mathscr{P}}(t\mathbf{w}) = \mathscr{K}^{\mathscr{P}}(t) + \lambda_t(\mathscr{P}(t),\mathbf{w}), \quad \mathbf{w} \in \mathbf{W}_t,$$

with initial condition $\mathscr{K}^{\mathscr{P}}(\Box) = 0$. Of course, when Skeptic starts out (in $\Box$) with capital α and uses strategy $\mathscr{P}$, his corresponding accumulated capital is given by the process $\alpha + \mathscr{K}^{\mathscr{P}}$. In the terminal situations, his accumulated capital is then given by the real variable $\alpha + \mathscr{K}^{\mathscr{P}}_{\Omega}$.

If we start in a non-terminal situation t, rather than in $\Box$, then we can consider t-strategies $\mathscr{P}$ that tell Skeptic how to move starting from t, and the corresponding capital process $\mathscr{K}^{\mathscr{P}}$ is then also a t-process, that tells us how much capital Skeptic has accumulated since starting with zero capital in situation t and using t-strategy $\mathscr{P}$.

Assumptions G1 and G2 determine so-called *gambling protocols*. They are sufficient for us to be able to define lower and upper prices for real variables. Consider a non-terminal situation t and a real t-variable f. Then the *upper price* $\overline{\mathbb{E}}_t(f)$ *for* f *in* t is defined as the infimum capital α that Skeptic has to start out with in t in order that there would be some t-strategy $\mathscr{P}$ such that his accumulated capital $\alpha + \mathscr{K}^{\mathscr{P}}$ allows him, at the end of the game, to hedge f, whatever moves World makes after t:

$$\overline{\mathbb{E}}_t(f) := \inf\left\{\alpha \colon \alpha + \mathscr{K}^{\mathscr{P}}_{\Omega} \geq f \text{ for some } t\text{-strategy } \mathscr{P}\right\}, \tag{1}$$

where $\alpha + \mathscr{K}^{\mathscr{P}}_{\Omega} \geq f$ is taken to mean that $\alpha + \mathscr{K}^{\mathscr{P}}(\omega) \geq f(\omega)$ for all terminal situations ω that go through t. Similarly, for the *lower price* $\underline{\mathbb{E}}_t(f)$ *for* f *in* t:

$$\underline{\mathbb{E}}_t(f) := \sup\left\{\alpha \colon \alpha - \mathscr{K}^{\mathscr{P}}_{\Omega} \leq f \text{ for some } t\text{-strategy } \mathscr{P}\right\}, \tag{2}$$

so $\underline{\mathbb{E}}_t(f) = -\overline{\mathbb{E}}_t(-f)$. If we start from the initial situation $t = \Box$, we simply get the *upper and lower prices* for a real variable f, which we also denote by $\overline{\mathbb{E}}(f)$ and $\underline{\mathbb{E}}(f)$.

A gambling protocol is called a *probability protocol* when besides G1 and G2, two more requirements are satisfied.

P1. For each non-terminal situation t, Skeptic's move space $\mathbf{S}_t$ is a convex cone in some linear space: $a_1\mathbf{s}_1 + a_2\mathbf{s}_2 \in \mathbf{S}_t$ for all non-negative real numbers a_1 and a_2 and all $\mathbf{s}_1$ and $\mathbf{s}_2$ in $\mathbf{S}_t$.

P2. For each non-terminal situation t, Skeptic's gain function λ_t has the following linearity property: $\lambda_t(a_1\mathbf{s}_1 + a_2\mathbf{s}_2, \mathbf{w}) = a_1\lambda_t(\mathbf{s}_1,\mathbf{w}) + a_2\lambda_t(\mathbf{s}_2,\mathbf{w})$ for all non-negative real numbers a_1 and a_2, all $\mathbf{s}_1$ and $\mathbf{s}_2$ in $\mathbf{S}_t$ and all $\mathbf{w}$ in $\mathbf{W}_t$.

Finally, a probability protocol is called *coherent*[3] when moreover

C. For each non-terminal situation t, and for each $\mathbf{s}$ in $\mathbf{S}_t$ there is some $\mathbf{w}$ in $\mathbf{W}_t$ such that $\lambda_t(\mathbf{s},\mathbf{w}) \leq 0$.

It is clear what this last requirement means: in each non-terminal situation, World has a strategy for playing from t onwards such that Skeptic cannot (strictly) increase his capital from t onwards, whatever t-strategy he might use.

For such coherent probability protocols, Shafer and Vovk prove a number of interesting properties for the corresponding lower (and upper) prices. We list a number of them here. For any real t-variable f, we can associate with a cut U of t another special U-measurable real t-variable $\underline{\mathbb{E}}_U$ by $\underline{\mathbb{E}}_U(f)(\omega) = \underline{\mathbb{E}}_u(f)$, for all paths ω through t, where u is the unique situation in U that ω goes through. For any two real t-variables f_1 and f_2, $f_1 \leq f_2$ is taken to mean that $f_1(\omega) \leq f_2(\omega)$ for all paths ω that go through t.

Proposition 1 (Properties of lower and upper prices in a coherent probability protocol [18])**.** *Consider a coherent probability protocol, let t be a non-terminal situation, f, f_1 and f_2 real t-variables, and U a cut of t. Then*

1. $\inf_{\omega \in \uparrow t} f(\omega) \leq \underline{\mathbb{E}}_t(f) \leq \overline{\mathbb{E}}_t(f) \leq \sup_{\omega \in \uparrow t} f(\omega)$ *[convexity];*
2. $\underline{\mathbb{E}}_t(f_1 + f_2) \geq \underline{\mathbb{E}}_t(f_1) + \underline{\mathbb{E}}_t(f_2)$ *[super-additivity];*
3. $\underline{\mathbb{E}}_t(\lambda f) = \lambda \underline{\mathbb{E}}_t(f)$ *for all real* $\lambda \geq 0$ *[non-negative homogeneity];*
4. $\underline{\mathbb{E}}_t(f + \alpha) = \underline{\mathbb{E}}_t(f) + \alpha$ *for all real* α *[constant additivity];*
5. $\underline{\mathbb{E}}_t(\alpha) = \alpha$ *for all real* α *[normalisation];*
6. $f_1 \leq f_2$ *implies that* $\underline{\mathbb{E}}_t(f_1) \leq \underline{\mathbb{E}}_t(f_2)$ *[monotonicity];*
7. $\underline{\mathbb{E}}_t(f) = \underline{\mathbb{E}}_t(\underline{\mathbb{E}}_U(f))$ *[law of iterated expectation].*

What is more, Shafer and Vovk use specific instances of such coherent probability protocols to prove various limit theorems (such as the law of large numbers, the central limit theorem, the law of the iterated logarithm), from which they can derive, as special cases, the well-known measure-theoretic versions. We shall come back to this in Section 6.

[3] For a discussion of the use of 'coherent' here, we refer to [17, Appendix C].

3 Walley's behavioural approach to probability

In his book on the behavioural theory of imprecise probabilities [20], Walley considers many different types of related uncertainty models. We shall restrict ourselves here to the most general and most powerful one, which also turns out to be the easiest to explain, namely coherent sets of really desirable gambles; see also [21].

Consider a non-empty set Ω of possible alternatives ω, only one of which actually obtains (or will obtain); we assume that it is possible, at least in principle, to determine which alternative does so. Also consider a subject who is uncertain about which possible alternative actually obtains (or will obtain). A *gamble*[4] on Ω is a real-valued map on Ω. It is interpreted as an uncertain reward, expressed in units of some predetermined linear utility scale: if ω actually obtains, then the reward is $f(\omega)$, which may be positive or negative. If a subject *accepts* a gamble f, this means that she is willing to engage in the transaction where, (i) first it is determined which ω obtains, and then (ii) she receives the reward $f(\omega)$. We can try and model the subject's beliefs about Ω by considering which gambles she accepts.

Suppose our subject specifies some set $\mathscr{R}$ of gambles she accepts, called a *set of really desirable gambles*. Such a set is called *coherent* if it satisfies the following *rationality requirements*:

D1. if $f < 0$ then $f \notin \mathscr{R}$ [avoiding partial loss];

D2. if $f \geq 0$ then $f \in \mathscr{R}$ [accepting partial gain];

D3. if f_1 and f_2 belong to $\mathscr{R}$ then their (point-wise) sum $f_1 + f_2$ also belongs to $\mathscr{R}$ [combination];

D4. if f belongs to $\mathscr{R}$ then its (point-wise) scalar product λf also belongs to $\mathscr{R}$ for all non-negative real numbers λ [scaling].

Here '$f < 0$' means '$f \leq 0$ and not $f = 0$'. Walley has also argued that sets of really desirable gambles should satisfy an additional axiom, where I_B denotes the *indicator* of the event B [a gamble that assumes the value one on B and zero elsewhere]:

D5. $\mathscr{R}$ is $\mathscr{B}$-conglomerable for any partition $\mathscr{B}$ of Ω: if $I_B f \in \mathscr{R}$ for all $B \in \mathscr{B}$, then also $f \in \mathscr{R}$ [full conglomerability].

Full conglomerability is a very strong requirement, and it is not without controversy. If a model $\mathscr{R}$ is $\mathscr{B}$-conglomerable, this means that certain inconsistency problems when conditioning on elements B of $\mathscr{B}$ are avoided; see [20, Section 6.8] for more details and examples. Conglomerability of belief models was not required by forerunners of Walley, such as Williams [24],[5] or de Finetti [4]. While we agree with Walley that conglomerability is a desirable property for sets of really desirable gambles, we do not believe that *full* conglomerability is always necessary: it seems that we only need to require conglomerability with respect to those partitions that we actually intend to condition our model on.[6] This is the path we shall follow in Section 4.

Given a coherent set of really desirable gambles, we can define *conditional lower and upper previsions* as follows: for any gamble f and any non-empty subset B of Ω, with indicator I_B,

$$\overline{P}(f|B) := \inf\{\alpha \colon I_B(\alpha - f) \in \mathscr{R}\} \tag{3}$$

$$\underline{P}(f|B) := \sup\{\alpha \colon I_B(f - \alpha) \in \mathscr{R}\}, \tag{4}$$

so $\underline{P}(f|B) = -\overline{P}(-f|B)$, and $\underline{P}(f|B)$ is the supremum price α for which the subject will buy the gamble f, i.e., accept the gamble $f - \alpha$, contingent on the occurrence of B. For any event A, we define the conditional lower probability $\underline{P}(A|B) := \underline{P}(I_A|B)$, i.e., the subject's supremum rate for betting on the event A, contingent on the occurrence of B, and similarly for $\overline{P}(A|B) := \overline{P}(I_A|B)$.

We want to stress here that by its definition [Eq. (4)], $\underline{P}(f|B)$ is a conditional lower prevision on what Walley [20, Section 6.1] has called the *contingent interpretation*: it is a supremum acceptable price for buying the gamble f *contingent* on the occurrence of B, meaning that the subject accepts the contingent gambles $I_B(f - \underline{P}(f|B) + \varepsilon)$, $\varepsilon > 0$, which are called off unless B occurs. This should be contrasted with the *updating interpretation* for the conditional lower prevision $\underline{P}(f|B)$, which is a subject's *present* (before the occurrence of B) supremum acceptable price for buying f after receiving the information that B has occurred (and nothing else!). Walley's *Updating Principle* [20, Section 6.1.6], which we shall accept, and use further on in Section 4, (essentially) states that conditional lower previsions should be the same on both interpretations. There is also a third way of looking at a conditional lower prevision $\underline{P}(f|B)$, which we shall call the *dynamic interpretation*, and where $\underline{P}(f|B)$ stands for the subject's supremum acceptable buying price for f *after she gets to know that* B has occurred. For precise conditional previsions, this seems to be the interpretation considered in [6, 11, 12, 17]. It is

[4] Walley [20] assumes gambles to be bounded. We make no such assumption here. It seems the concept of a really desirable gamble (at least formally) allows for such a generalisation, because the coherence axioms for real desirability, as opposed to those for Walley's related notions of almost- and strict desirability, nowhere hinge on such a boundedness assumption, at least not from a technical mathematical point of view.

[5] Axioms (D1)–(D4), but not (D5), were actually suggested by Williams. But it seems that we need at least some weaker form of (D5), namely the cut conglomerability (D5') considered further on, to derive our main results: Theorems 3 and 6.

[6] The view expressed here seems related to Shafer's, as sketched near the end of [13, Appendix 1].

far from obvious that there should be a relation between the first two and the third interpretations.[7] We shall briefly come back to this distinction in the following sections.

For a partition $\mathscr{B}$ of Ω, we let $\underline{P}(f|\mathscr{B}) := \sum_{B\in\mathscr{B}} I_B \underline{P}(f|B)$ be the gamble on Ω that in any element ω of B assumes the value $\underline{P}(f|B)$, where B is any element of $\mathscr{B}$.

The following properties of conditional lower and upper previsions associated with a coherent set of really desirable gambles were (essentially) proven by Walley.

Proposition 2 (Properties of conditional lower and upper previsions [20])**.** *Consider a coherent set of really desirable gambles $\mathscr{R}$, let B be any non-empty subset of Ω, and let f, f_1 and f_2 be gambles on Ω. Then*[8]

1. $\inf_{\omega\in B} f(\omega) \le \underline{P}(f|B) \le \overline{P}(f|B) \le \sup_{\omega\in B} f(\omega)$ *[convexity];*
2. $\underline{P}(f_1+f_2|B) \ge \underline{P}(f_1|B)+\underline{P}(f_2|B)$ *[super-additivity];*
3. $\underline{P}(\lambda f|B) = \lambda \underline{P}(f|B)$ *for all real $\lambda \ge 0$ [non-negative homogeneity];*
4. $\underline{P}(f+\alpha|B) = \underline{P}(f|B)+\alpha$ *for all real α [constant additivity];*
5. $\underline{P}(\alpha|B) = \alpha$ *for all real α [normalisation];*
6. $f_1 \le f_2$ *implies that* $\underline{P}(f_1|B) \le \underline{P}(f_2|B)$ *[monotonicity];*
7. *if $\mathscr{B}$ is a partition of Ω that refines the partition $\{B, B^c\}$ and $\mathscr{R}$ is $\mathscr{B}$-conglomerable, then* $\underline{P}(f|B) \ge \underline{P}(\underline{P}(f|\mathscr{B})|B)$ *[conglomerative property].*

The analogy between Propositions 1 and 2 is striking, even if there is an equality in Proposition 1.7 and only an inequality in Proposition 2.7.[9] We now set out to identify the exact correspondence between the two models.[10]

4 Connecting the two approaches

In order to lay bare the connections between the game-theoretic and the behavioural approach, we enter Shafer and Vovk's world, and consider another player, called Subject, who, *in situation* $\Box$, has certain *piece-wise* beliefs about what moves World will make.

More specifically, for each non-terminal situation $t \in \Omega^\Diamond \setminus \Omega$, she has beliefs (in situation $\Box$) about which move $\mathbf{w}$ World will choose from the set $\mathbf{W}_t$ of moves available to him in t. We suppose she represents those beliefs in the form of a *coherent*[11] set $\mathscr{R}_t$ of really desirable gambles on $\mathbf{W}_t$. These beliefs are conditional on the updating interpretation, in the sense that they represent Subject's beliefs in situation $\Box$ about what World will do *immediately after he gets to situation t*. We call any specification of such coherent $\mathscr{R}_t$, $t \in \Omega^\Diamond \setminus \Omega$, an *immediate prediction model* for Subject. It should be stressed here that $\mathscr{R}_t$ should *not* be interpreted dynamically, i.e., as a set of gambles on $\mathbf{W}_t$ that Subject accepts in situation t.

We can now ask ourselves what the behavioural implications of these conditional assessments $\mathscr{R}_t$ in the immediate prediction model are. For instance, what do they tell us about whether or not Subject should accept certain gambles[12] on Ω, the set of possible paths for World? In other words, how can these beliefs (in $\Box$) about which next move World will make in each non-terminal situation t be combined coherently into beliefs (in $\Box$) about World's complete sequence of moves?

In order to investigate this, we use Walley's very general and powerful method of *natural extension*, which is just *conservative coherent reasoning*. We shall construct, using the local pieces of information $\mathscr{R}_t$, a set of really desirable gambles on Ω for Subject in situation $\Box$ that is (i) coherent, and (ii) as small as possible, meaning that no more gambles should be accepted than is actually required by coherence.

First, we collect the pieces. Consider any non-terminal situation $t \in \Omega^\Diamond \setminus \Omega$ and any gamble h_t in $\mathscr{R}_t$. Then with h_t we can associate a t-gamble,[13] also denoted by h_t, and defined by

$$h_t(\omega) := h_t(\omega(t)),$$

for all $\omega \sqsupseteq t$, where we denote by $\omega(t)$ the unique element of $\mathbf{W}_t$ such that $t\omega(t) \sqsubseteq \omega$. The t-gamble h_t is U-measurable for any cut U of t that is non-trivial, i.e., such that $U \neq \{t\}$. This implies that we can interpret h_t as a map on U. In fact, we shall write $h_t(s) := h_t(\omega(t))$, for any $t \sqsubset s$, where ω is any terminal situation that follows s.

$I_{\uparrow t} h_t$ represents the gamble on Ω that is called off unless World ends up in situation t, and which, when it is not called off, depends only on World's move immediately after t, and gives the same value $h_t(\mathbf{w})$ to all paths ω that go through

[7] We may be wrong, but it seems to us that in [17], the authors confuse the updating interpretation with the dynamic interpretation when they claim that "[their new understanding of lower and upper previsions] justifies Peter Walley's updating principle".

[8] Here, as in Proposition 1, we implicitly assume that whatever we write down is well-defined, meaning that for instance no sums of $-\infty$ and $+\infty$ appear, and that the function $\underline{P}(f|\mathscr{B})$ is real-valued, and nowhere infinite. Shafer and Vovk do not seem to mention the need for this.

[9] Concatenation inequalities for lower prices do appear in the more general context described in [17].

[10] We shall find a specific situation where applying Walley's theory leads to equalities rather than the more general inequalities of Proposition 2.7. This seems to happen generally for what is called *marginal extension* in a situation of immediate prediction, meaning that we start out with, and extend, an initial model where we condition on increasingly finer partitions, and where the initial conditional model for any partition deals with gambles that are measurable with respect to the finer partitions; see [20, Theorem 6.7.2] and [9].

[11] Since we do not immediately envisage conditioning this local model on subsets of $\mathbf{W}_t$, we impose no extra conglomerability requirements here, only the coherence conditions D1–D4.

[12] In Shafer and Vovk's language, gambles are real variables.

[13] Just as for variables, we can define a t-gamble as a partial gamble whose domain includes $\uparrow t$.

$t\mathbf{w}$. The fact that Subject, in situation $\Box$, accepts h_t on $\mathbf{W}_t$ conditional on World's getting to t, translates immediately to the fact that Subject accepts the contingent gamble $I_{\uparrow t}h_t$ on Ω, by Walley's Updating Principle. We thus end up with a set of gambles on Ω

$$\mathscr{R} := \bigcup_{t\in\Omega^\Diamond\setminus\Omega} \{I_{\uparrow t}h_t\colon h_t \in \mathscr{R}_t\}$$

that Subject accepts in situation $\Box$. The only thing left to do now, is to find the smallest coherent set $\mathscr{E}_{\mathscr{R}}$ of really desirable gambles that includes $\mathscr{R}$ (if indeed there is any such coherent set). Here we take coherence to refer to conditions D1–D4, together with D5', a variation on D5 which refers to conglomerability with respect to those partitions that we actually intend to condition on, as suggested in Section 3.

These partitions are what we call *cut partitions*. Consider any cut U of the initial situation $\Box$. Then the set of events $\mathscr{B}_U := \{\uparrow u\colon u \in U\}$ is a partition of Ω, called the *U-partition*. D5' requires that our set of really desirable gambles should be *cut conglomerable*, i.e., conglomerable with respect to every cut partition $\mathscr{B}_U$.[14]

Why do we only require conglomerability for cut partitions? Simply because we are interested in *predictive inference*: we eventually will want to find out about the gambles on Ω that Subject accepts in situation $\Box$, conditional (contingent) on World getting to a situation t. This is related to finding lower previsions for Subject conditional on the corresponding events $\uparrow t$. A collection $\{\uparrow t\colon t \in T\}$ of such events constitutes a partition of the sample space Ω if and only if T is a cut of $\Box$.

Because we require cut conglomerability, it follows in particular that $\mathscr{E}_{\mathscr{R}}$ will contain the sums of gambles $g := \sum_{u\in U} I_{\uparrow u}h_u$ for all *non-terminal* cuts U of $\Box$ and all choices of $h_u \in \mathscr{R}_u$, $u \in U$. This is because $I_{\uparrow u}g = I_{\uparrow u}h_u \in \mathscr{R}$ for all $u \in U$. Because moreover $\mathscr{E}_{\mathscr{R}}$ should be a convex cone [by D3 and D4], any sum of such sums $\sum_{u\in U} I_{\uparrow u}h_u$ over a finite number of non-terminal cuts U should also belong to $\mathscr{E}_{\mathscr{R}}$. But, since in the case of bounded protocols we are discussing here, World can only make a bounded and finite number of moves, $\Omega^\Diamond \setminus \Omega$ is a finite union of such non-terminal cuts, and therefore the sums $\sum_{u\in\Omega^\Diamond\setminus\Omega} I_{\uparrow u}h_u$ should belong to $\mathscr{E}_{\mathscr{R}}$ for all choices $h_u \in \mathscr{R}_u$, $u \in \Omega^\Diamond \setminus \Omega$.

Call therefore, for any non-terminal situation t, a *t-selection* any partial process $\mathscr{S}$ defined on the non-terminal situations $s \sqsupseteq t$ such that $\mathscr{S}(s) \in \mathscr{R}_s$. With such a t-selection, we can associate a t-process, called a *gamble process* $\mathscr{G}^{\mathscr{S}}$, with value

$$\mathscr{G}^{\mathscr{S}}(s) = \sum_{t\sqsubseteq u\sqsubset s} \mathscr{S}(u)(s)$$

in all situations s that follow t, where it should be recalled that $\mathscr{S}(u)(s) = \mathscr{S}(u)(\omega(u))$ for all $\omega \sqsupseteq s$ (see above). Alternatively, $\mathscr{G}^{\mathscr{S}}$ is given by the recursion relation $\mathscr{G}^{\mathscr{S}}(s\mathbf{w}) = \mathscr{G}^{\mathscr{S}}(s) + \mathscr{S}(s)(\mathbf{w})$ for all non-terminal $s \sqsupseteq t$ and all $\mathbf{w} \in \mathbf{W}_s$, with initial value $\mathscr{G}^{\mathscr{S}}(t) = 0$. In particular, this leads to the t-gamble $\mathscr{G}^{\mathscr{S}}_\Omega$ defined on all terminal situations ω that follow t, by letting

$$\mathscr{G}^{\mathscr{S}}_\Omega = \sum_{t\sqsubseteq u, u\in\Omega^\Diamond\setminus\Omega} I_{\uparrow u}\mathscr{S}(u).$$

We have just argued that the gambles $\mathscr{G}^{\mathscr{S}}_\Omega$ should belong to $\mathscr{E}_{\mathscr{R}}$ for all non-terminal situations t and all t-selections $\mathscr{S}$. As before for strategy and capital processes, we call a $\Box$-selection $\mathscr{S}$ simply a *selection*, and a $\Box$-gamble process simply a *gamble process*. It is now but a technical step to prove Theorem 3 below. It is a significant generalisation, in terms of sets of really desirable gambles rather than coherent lower previsions,[15] of the Marginal Extension Theorem first proven by Walley [20, Theorem 6.7.2] and subsequently extended by De Cooman and Miranda [9].

Theorem 3 (Marginal Extension Theorem)**.** *There is a smallest set of gambles that satisfies D1–D4 and D5' and includes $\mathscr{R}$. This natural extension of $\mathscr{R}$ is given by*

$$\mathscr{E}_{\mathscr{R}} := \left\{g\colon g \geq \mathscr{G}^{\mathscr{S}}_\Omega \text{ for some selection } \mathscr{S}\right\}.$$

Moreover, for any non-terminal situation t and any t-gamble g, it holds that $I_{\uparrow t}g \in \mathscr{E}_{\mathscr{R}}$ if and only if there is some t-selection $\mathscr{S}_t$ such that $g \geq \mathscr{G}^{\mathscr{S}_t}_\Omega$, where as before, $g \geq \mathscr{G}^{\mathscr{S}_t}_\Omega$ is taken to mean that $g(\omega) \geq \mathscr{G}^{\mathscr{S}_t}_\Omega(\omega)$ for all terminal situations ω that follow t.

We now use the coherent set of really desirable gambles $\mathscr{E}_{\mathscr{R}}$ to define special lower (and upper) previsions $\underline{P}(\cdot|t) := \underline{P}(\cdot|\uparrow t)$ for Subject in situation $\Box$, conditional on an event $\uparrow t$, i.e., on World getting to situation t, indicated in Section 3.[16] We shall call such conditional lower previsions *predictive* lower previsions. We then get, using Theorem 3, that for any non-terminal situation t,

$$\underline{P}(f|t) := \sup\left\{\alpha\colon I_{\uparrow t}(f-\alpha) \in \mathscr{E}_{\mathscr{R}}\right\} \tag{5}$$

$$= \sup\left\{\alpha\colon f-\alpha \geq \mathscr{G}^{\mathscr{S}}_\Omega \text{ for some } t\text{-selection } \mathscr{S}\right\}. \tag{6}$$

Eq. (5) is also valid in terminal situations t, whereas Eq. (6) clearly isn't.

Besides the properties in Proposition 2, which hold in general for conditional lower and upper previsions, the predictive lower and upper previsions we consider here also satisfy a number of additional properties, listed in Propositions 4 and 5.

[14] When all of World's move spaces $\mathbf{W}_t$ are finite, cut conglomerability (D5') is a consequence of D3, and therefore needs no extra attention.

[15] The difference in language may obscure that this is indeed a generalisation. But see Theorem 7 for expressions in terms of predictive lower previsions that should make the connection much clearer.

[16] We stress again that these are conditional lower and upper previsions on the contingent/updating interpretation.

Proposition 4 (Additional properties of predictive lower and upper previsions)**.** *Let t be any situation, and let f, f_1 and f_2 be gambles on Ω.*

1. *if t is a terminal situation ω, then $\underline{P}(f|\omega) = \overline{P}(f|\omega) = f(\omega)$;*
2. *$\underline{P}(f|t) = \underline{P}(fI_{\uparrow t}|t)$ and $\overline{P}(f|t) = \overline{P}(fI_{\uparrow t}|t)$;*
3. *$f_1 \leq f_2$ (on $\uparrow t$) implies that $\underline{P}(f_1|t) \leq \underline{P}(f_2|t)$ [monotonicity].*

Before we go on, there is an important point that must be stressed and clarified. It is an immediate consequence of Proposition 4.2 that when f and g are any two gambles that coincide on $\uparrow t$, then $\underline{P}(f|t) = \underline{P}(g|t)$. This means that $\underline{P}(f|t)$ is completely determined by the values that f assumes on $\uparrow t$, and it allows us to define $\underline{P}(\cdot|t)$ on gambles that are only necessarily defined on $\uparrow t$, i.e., on t-gambles. We shall do so freely in what follows.

For any cut U of a situation t, we may define the t-gamble $\underline{P}(f|U)$ as the gamble that assumes the value $\underline{P}(f|u)$ in any $\omega \sqsupseteq t$, where u is the unique element of U that ω goes through. This t-gamble is U-measurable by construction, and it can be considered as a gamble on U.

Proposition 5 (Separate coherence)**.** *Let t be any situation, let U be any cut of t, and let f and g be t-gambles, where g is U-measurable.*

1. *$\underline{P}(\uparrow t|t) = 1$;*
2. *$\underline{P}(g|U) = g_U$;*
3. *$\underline{P}(f+g|U) = g_U + \underline{P}(f|U)$;*
4. *if g is moreover non-negative, then $\underline{P}(gf|U) = g_U \underline{P}(f|U)$.*

There appears to be a close correspondence between the expressions [such as (2)] for lower prices $\underline{\mathbb{E}}_t(f)$ associated with coherent probability protocols and those [such as (6)] for the predictive lower previsions $\underline{P}(f|t)$ based on an immediate prediction model. Say that a given coherent probability protocol and given immediate prediction model *match* whenever they lead to identical corresponding lower prices $\underline{\mathbb{E}}_t$ and predictive lower previsions $\underline{P}(\cdot|t)$ for all *non-terminal* $t \in \Omega^\Diamond \setminus \Omega$.

Theorem 6 (Matching Theorem)**.** *For every coherent probability protocol there is an immediate prediction model such that the two match, and conversely, for every immediate prediction model there is a coherent probability protocol such that the two match.*

It is interesting to indicate here how matching is actually achieved. If we have a coherent probability protocol with move spaces $\mathbf{S}_t$ and gain functions λ_t for Skeptic, define the immediate prediction model for Subject to be (essentially) $\mathscr{R}_t := \{-\lambda(\mathbf{s},\cdot)\colon \mathbf{s} \in \mathbf{S}_t\}$. If, conversely, we have an immediate prediction model for Subject consisting of the sets $\mathscr{R}_t$, define the move spaces for Skeptic by $\mathbf{S}_t := \mathscr{R}_t$, and his gain functions by $\lambda_t(h,\cdot) := -h$ for all h in $\mathscr{R}_t$.

Theorem 7 (Concatenation Formula)**.** *Consider any two cuts U and V of a situation t such that U precedes V. Then for all t-gambles f on Ω,*[17]

1. *$\underline{P}(f|t) = \underline{P}(\underline{P}(f|U)|t)$;*
2. *$\underline{P}(f|U) = \underline{P}(\underline{P}(f|V)|U)$.*

This theorem, in combination with the following two propositions (8 and 9), tells us that all predictive lower (and upper) previsions can be calculated using backwards recursion, by starting with the trivial predictive previsions $\overline{P}(f|\Omega) = \underline{P}(f|\Omega) = f$ for the terminal cut Ω, and using only the local models $\mathscr{R}_t$. To see this, observe in addition that in the above theorem, the t-gamble $\underline{P}(f|V)$ is V-measurable, and therefore actually a gamble on V.

To make clear what the following Proposition 8 implies, consider any t-selection $\mathscr{S}$, and define the *U-called off t-selection* $\mathscr{S}^U$ as the selection that mimics $\mathscr{S}$ until we get to U, where we begin to select the zero gambles: for any non-terminal situation $s \sqsupseteq t$, let $\mathscr{S}^U(s) := \mathscr{S}(s)$ if s strictly precedes (some element of) U, and let $\mathscr{S}^U(s) := 0 \in \mathscr{R}_s$ otherwise. Then

$$U(\mathscr{G}^{\mathscr{S}}) = \mathscr{G}^{\mathscr{S}^U} \quad \text{and therefore} \quad \mathscr{G}_U^{\mathscr{S}} = \mathscr{G}_\Omega^{\mathscr{S}^U}, \qquad (7)$$

so we see that stopped gamble processes are gamble processes themselves, that correspond to selections being 'called-off' after a cut. This also means that we can actually restrict ourselves to selections $\mathscr{S}$ that are U-called off in Proposition 8.

Proposition 8. *Let t be a non-terminal situation, and let U be a cut of t. Then for any U-measurable t-gamble f, $I_{\uparrow t} f \in \mathscr{E}_{\mathscr{R}}$ if and only is there is some t-selection $\mathscr{S}$ such that $I_{\uparrow t} f \geq \mathscr{G}_\Omega^{\mathscr{S}^U}$, or equivalently, $f_U \geq \mathscr{G}_U^{\mathscr{S}}$. Consequently,*

$$\begin{aligned}\underline{P}(f|t) &= \sup\left\{\alpha\colon f - \alpha \geq \mathscr{G}_\Omega^{\mathscr{S}^U} \text{ for some } t\text{-selection } \mathscr{S}\right\} \\ &= \sup\left\{\alpha\colon f_U - \alpha \geq \mathscr{G}_U^{\mathscr{S}} \text{ for some } t\text{-selection } \mathscr{S}\right\}.\end{aligned}$$

If a t-gamble h is measurable with respect to the children cut $C(t)$ of a non-terminal situation t, then we can interpret it as gamble on $\mathbf{W}_t$. For such gambles, the following immediate corollary of Proposition 8 tells us that the predictive lower previsions $\underline{P}(h|t)$ are completely determined by the local modal $\mathscr{R}_t$.

[17] Here too, it is implicitly assumed that all expressions are well-defined, e.g., that in the second statement, $\underline{P}(f|v)$ is a real number for all $v \in V$, making sure that $\underline{P}(f|V)$ is indeed a gamble.

Proposition 9. *Let t be a non-terminal situation, and consider a $C(t)$-measurable gamble h. Then*

$$\underline{P}(h|t) = \underline{P}_t(h) := \sup\{\alpha\colon h-\alpha \in \mathscr{R}_t\}.$$

5 Interpretation

The Matching Theorem has a very interesting interpretation. In Shafer and Vovk's approach, World is sometimes decomposed into two players, Reality and Forecaster. It is Reality whose moves are characterised by the above-mentioned event tree, and Forecaster who determines what Skeptic's move space $\mathbf{S}_t$ and gain function λ_t are, in each non-terminal situation t. We now make Shafer and Vovk's model a bit more involved, by adding something to it.

Suppose that Forecaster has certain beliefs, *in situation* $\square$, about what move Reality will make next in each non-terminal situation t,, and suppose she models those beliefs by specifying a coherent set $\mathscr{R}_t$ of really desirable gambles on $\mathbf{W}_t$. In other words, *we identify Forecaster with Subject*.[18]

When Forecaster specifies such a set, she is making certain behavioural commitments. In fact, she is committing herself to accepting, in situation $\square$, any gamble in $\mathscr{R}_t$, contingent on World getting to situation t, and to accepting any combination of such gambles according to the combination axioms D3, D4 and D5'. This implies that we can derive predictive lower previsions $\underline{P}(\cdot|t)$, with the following interpretation: in situation $\square$, $\underline{P}(f|t)$ is the supremum price Forecaster can be made to buy the t-gamble f for, conditional on World's getting to t, and on the basis of the commitments she has made in the initial situation $\square$.

What Skeptic can now do, is take Forecaster up on her commitments. This means that in situation $\square$, he can use a selection $\mathscr{S}$, which for each non-terminal situation t, selects a gamble (or equivalently, any non-negative linear combination of gambles) $\mathscr{S}(t) = h_t$ in $\mathscr{R}_t$ and offer the corresponding gamble $\mathscr{G}_\Omega^{\mathscr{S}}$ on Ω to Forecaster, who is bound to accept it. If Reality's next move in situation t is $\mathbf{w} \in \mathbf{W}_t$, this changes Skeptic's capital by (the positive or negative amount) $-h_t(\mathbf{w})$. In other words, his move space $\mathbf{s}_t$ can then be identified with the convex set of gambles $\mathscr{R}_t$ and his gain function λ_t is then given by $\lambda_t(h_t,\cdot) = -h_t$. But then the *selection* $\mathscr{S}$ can be identified with a *strategy* $\mathscr{P}$ for Skeptic, and $\mathscr{K}_\Omega^{\mathscr{P}} = -\mathscr{G}_\Omega^{\mathscr{S}}$ (this is the essence of the proof of Theorem 6), which tells us that we are led to a coherent probability protocol, and that the corresponding lower prices $\underline{\mathbb{E}}_t$ for Skeptic coincide with Forecaster's predictive lower previsions $\underline{P}(\cdot|t)$.

[18] The germ for this idea, in the case that Forecaster's beliefs can be expressed using precise probability models on the $\mathscr{L}(\mathbf{W}_t)$, is already present in Shafer's work, see for instance [18, Chapter 8] and [13, Appendix 1]. We extend this idea here to Walley's imprecise probability models.

In a very nice paper [17], Shafer, Gillett and Scherl discuss ways of introducing and interpreting lower previsions in a game-theoretic framework, not in terms of prices that a subject is willing to pay for a gamble, but in terms of whether a subject believes he can make a lot of money (utility) at those prices. They consider such conditional lower previsions both on a contingent and on a dynamic interpretation, and argue that there is equality between them in certain cases. Here, we have decided to stick to the more usual interpretation of lower and upper previsions, and concentrated on the contingent/updating interpretation. We see that also on our approach, the game-theoretic framework is useful.

This is of particular relevance to the laws of large numbers that Shafer and Vovk derive in their game-theoretic framework, because such laws can now be given a behavioural interpretation in terms of Forecaster's (or any Subject's) (predictive) lower and upper previsions. To give an example, we now turn to deriving a very general weak law of large numbers.

6 A more general weak law of large numbers

Consider a non-terminal situation t and a cut U of t. Define the t-variable n_U such that $n_U(\omega)$ is the distance $d(t,u)$, measured in moves along the tree, from t to the unique situation u in U that ω goes through. n_U is clearly U-measurable, and $n_U(u)$ is simply the distance $d(t,u)$ from t to u. We assume that $n_U(u) > 0$, or in other words that $U \neq \{t\}$. Of course, in the bounded protocols we are considering here, n_U is bounded, and we denote its minimum by N_U.

Now consider for each s between t and U a *bounded* gamble h_s and a real number m_s such that $h_s - m_s \in \mathscr{R}_s$, meaning that Forecaster in situation $\square$ accepts to buy h_s for m_s, contingent on Reality getting to situation s. Let $B > 0$ be any common upper bound for $\sup h_s - \inf h_s$, for all $t \sqsubseteq s \sqsubset U$. Then it follows from the coherence of $\mathscr{R}_s$ [D1] that $m_s \leq \sup h_s$. To make things interesting, we shall also assume that $\inf h_s \leq m_s$, because otherwise $h_s - m_s \geq 0$ and accepting this gamble represents no real commitment on Forecaster's part. As a result, we see that $|h_s - m_s| \leq B$.

We are interested in the following t-gamble G_U, given by

$$G_U = \frac{1}{n_U} \sum_{t \sqsubseteq s \sqsubset U} I_{\uparrow s}[h_s - m_s],$$

which provides a measure for how much, on average, the gambles h_s yield an outcome above Forecaster's accepted buying price m_s, along segments of the tree starting in t and ending right before U. In other words, G_U measures the average gain for Forecaster along segments from t to U, associated with commitments she has made and is taken up on, because Reality has to move along these segments.

This gamble G_U is U-measurable too. We may therefore interpret G_U as a gamble on U. Also, for any h_s and any $u \in U$, we know that because $s \sqsubset u$, h_s has the same value $h_s(u) := h_s(\omega(s))$ in all ω that go through u. This allows us to write

$$G_U(u) = \frac{1}{n_U(u)} \sum_{t \sqsubseteq s \sqsubset u} [h_s(u) - m_s].$$

We would like to study Forecaster's beliefs (in the initial situation $\Box$ and contingent on Reality getting to t) in the occurrence of the event

$$\{G_U \geq -\varepsilon\} := \{\omega \in \uparrow t \colon G_U(\omega) \geq -\varepsilon\},$$

where $\varepsilon > 0$. In other words, we want to know $\underline{P}(\{G_U \geq -\varepsilon\}|t)$, which is Forecaster's supremum rate for betting on the event that his average gain from t to U will be at least $-\varepsilon$, contingent on Reality's getting to t.

Theorem 10 (Weak Law of Large Numbers)**.** *For all $\varepsilon > 0$,*

$$\underline{P}(\{G_U \geq -\varepsilon\}|t) \geq 1 - \exp\left(-\frac{N_U \varepsilon^2}{4B^2}\right).$$

We see that as N_U increases this lower bound increases to one, so the theorem can be very loosely formulated as follows: *As the horizon recedes, Forecaster, if she is coherent, should believe increasingly more strongly that her average gain along any path from the present to the horizon will not be negative*. Of course, this is a very general version of the weak law of large numbers. It significantly extends the result mentioned in Section 5. Perhaps surprisingly, it can be seen as generalisation of Hoeffding's inequality for martingale differences [7] (see also [22, Chapter 4] and [19, Appendix A.7]) to coherent lower previsions on event trees.

7 Scoring a predictive model

Suppose Reality follows a path up to some situation u_o in U, which leads to an average gain $G_U(u_o)$ for Forecaster. Suppose this average gain is negative: $G_U(u_o) < 0$.

Then we see that $\uparrow u_o \subseteq \{G_U < -\varepsilon\}$ for all $0 < \varepsilon < -G_U(u_o)$, and therefore all these events $\{G_U < -\varepsilon\}$ have actually occurred (because $\uparrow u_o$ has). On the other hand, Forecaster's upper probability (in $\Box$) for their occurrence satisfies $\overline{P}(\{G_U < -\varepsilon\}) \leq \exp(-\frac{N_U \varepsilon^2}{4B^2})$, by Theorem 10. Coherence then tells us that Forecaster's upper probability (in $\Box$) for the event $\uparrow u_o$, which has actually occurred, is then at most $S_{N_U}(\gamma_U(u_o))$, where

$$S_N(x) = \exp\left(-\frac{N}{4}x^2\right) \quad \text{and} \quad \gamma_U(u) := \frac{G_U(u_o)}{B}.$$

By assumption, $\gamma_U(u_o)$ is a number in $[-1, 0)$. Coherence requires that Forecaster, because of her local predictive commitments, can be forced (by Skeptic, if he chooses his strategy well) to bet against the occurrence of the event $\uparrow u_o$ at a rate that is at least $1 - S_{N_U}(\gamma_U(u_o))$. So we see that Forecaster is losing utility because of her local predictive commitments. Just how much depends on how close $\gamma_U(u_o)$ lies to -1 , and on how large N_U is; see Figure 2.

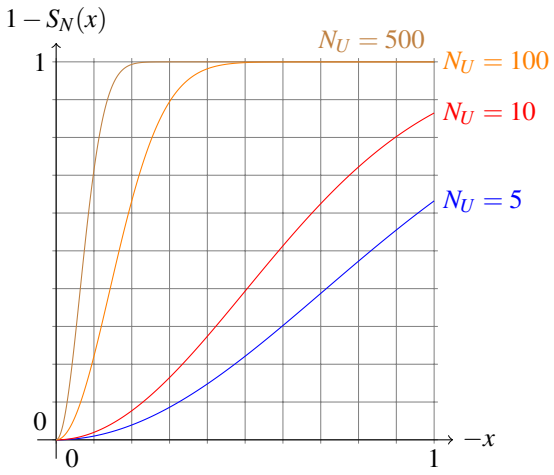


Figure 2: What Forecaster can be made to pay, $1 - S_N(x)$, as a function of $x = \gamma_U(u)$, for different values of $N = N_U$.

The upper bound $S_{N_U}(\gamma_U(u_o))$ we have constructed for the upper probability of $\uparrow u_o$ has a very interesting property, which we now try to make more explicit. Indeed, if we were to calculate Forecaster's upper probability $\overline{P}(\uparrow u_o)$ for $\uparrow u_o$ directly using Eq. (6), this value would generally depend on Forecaster's predictive assessments $\mathscr{R}_s$ for situations s that do not precede u_o, and that Reality therefore never got to. We shall see that such is not the case for the upper bound $S_{N_U}(\gamma_U(u_o))$ constructed using Theorem 10.

Consider any situation s before U but not on the path through u_o, meaning that Reality never got to this situation s. Therefore the corresponding gamble $h_s - m_s$ in the expression for G_U is not used in calculating the value of $G_U(u_o)$, so we can change it to anything else, and still obtain the same value of $G_U(u_o)$.

Indeed, consider any other predictive model, where the only thing we ask is that the $\mathscr{R}'_s$ coincide with the $\mathscr{R}_s$ for all s that precede u_o. For other s, the $\mathscr{R}'_s$ can be chosen arbitrarily, but still coherently. Now construct a new average gain gamble G'_U for this alternative predictive model, where the only restriction is that we let $h'_s = h_s$ and $m'_s = m_s$ if s precedes u_o. Then we know from the reasoning above that $G'_U(u_o) = G_U(u_o)$, so the new upper probability that the event $\uparrow u_o$ will be observed is at most

$$S_{N_U}\left(\frac{G'_U(u_o)}{B}\right) = S_{N_U}\left(\frac{G_U(u_o)}{B}\right) = S_{N_U}(\gamma_U(u_o)).$$

In other words, the upper bound $S_N(\gamma_U(u))$ we found for Forecaster's upper probability of Reality getting to a situation u_o *depends only on Forecaster's local predictive*

assessments $\mathscr{R}_s$ for situations s that Reality has actually got to, and not on her assessments for other situations. This means that this method for scoring a predictive model satisfies Dawid's *Prequential Principle* [1, 2].

8 Additional Remarks

We have proven the correspondence between the two approaches only for event trees with a bounded horizon. For games with infinite horizon, the correspondence becomes less immediate, because Shafer and Vovk implicitly make use of coherence axioms that are stronger than D1–D4 and D5', leading to lower prices that dominate the corresponding predictive lower previsions. Exact matching would be restored of course, provided we could argue that these additional requirements are rational for any subject to comply with. This could be an interesting topic for further research.

Acknowledgements

We would like to thank Enrique Miranda, Marco Zaffalon, Glenn Shafer, Vladimir Vovk and Didier Dubois for discussing and questioning the views expressed here, even though some of these discussions took place more than a few years ago. Sébastien Destercke and Erik Quaeghebeur have read and commented on earlier drafts of this paper.

References

[1] A. Ph. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:278–292, 1984.

[2] A. Ph. Dawid and V. G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5:125–162, 1999.

[3] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[4] B. de Finetti. *Theory of Probability*. John Wiley & Sons, Chichester, 1974–1975. English translation of [3], two volumes.

[5] P. Gärdenfors and N.-E. Sahlin. *Decision, Probability, and Utility*. Cambridge University Press, Cambridge, 1988.

[6] M. Goldstein. The prevision of a prevision. *Journal of the American Statistical Society*, 87:817–819, 1983.

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the Americal Statistical Association*, 58:13–30, 1963.

[8] H. E. Kyburg Jr. and H. E. Smokler, editors. *Studies in Subjective Probability*. Wiley, New York, 1964. Second edition (with new material) 1980.

[9] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 2006. In press.

[10] F. P. Ramsey. Truth and probability (1926). In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter VII, pages 156–198. Kegan, Paul, Trench, Trubner & Co., London, 1931. Reprinted in [8] and [5].

[11] G. Shafer. Bayes's two arguments for the Rule of Conditioning. *The Annals of Statistics*, 10:1075–1089, 1982.

[12] G. Shafer. A subjective interpretation of conditional probability. *Journal of Philosophical Logic*, 12:453–466, 1983.

[13] G. Shafer. Conditional probability. *International Statistical Review*, 53:261–277, 1985.

[14] G. Shafer. *The Art of Causal Conjecture*. The MIT Press, Cambridge, MA, 1996.

[15] G. Shafer. The significance of Jacob Bernoulli's *Ars Conjectandi* for the philosophy of probability today. *Journal of Econometrics*, 75:15–32, 1996.

[16] G. Shafer, P. R. Gillett, and R. Scherl. The logic of events. *Annals of Mathematics and Artificial Intelligence*, 28:315–389, 2000.

[17] G. Shafer, P. R. Gillett, and R. B. Scherl. A new understanding of subjective probability and its generalization to lower and upper prevision. *International Journal of Approximate Reasoning*, 33:1–49, 2003.

[18] G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[19] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a Random World*. Springer, New York, 2005.

[20] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[21] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[22] L. Wasserman. *All of Statistics*. Springer, New York, 2004.

[23] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975.

[24] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007. Revised journal version of [23].

Immediate prediction under exchangeability and representation insensitivity

Gert de Cooman
Ghent University
SYSTeMS Research Group
gert.decooman@ugent.be

Enrique Miranda
Rey Juan Carlos University
Dept. of Statistics and O.R.
enrique.miranda@urjc.es

Erik Quaeghebeur
Ghent University
SYSTeMS Research Group
erik.quaeghebeur@ugent.be

Abstract

We consider immediate predictive inference, where a subject, using a number of observations of a finite number of exchangeable random variables, is asked to coherently model his beliefs about the next observation, in terms of a predictive lower prevision. We study when such predictive lower previsions are representation insensitive, meaning that they are essentially independent of the choice of the (finite) set of possible values for the random variables. Such representation insensitive predictive models have very interesting properties, and among such models, the ones produced by the Imprecise Dirichlet-Multinomial Model are quite special in a number of ways.

Keywords. Predictive inference, immediate prediction, lower prevision, coherence, exchangeability, representation invariance, representation insensitivity, Imprecise Dirichlet-Multinomial Model, Johnson's sufficientness postulate.

1 Introduction

Consider a subject who is making $N > 0$ successive observations of a certain phenomenon. We represent these observations by N random variables $X_1, \ldots, X_N$. By *random variable*, we mean a variable about whose value the subject may entertain certain beliefs. We assume that at each successive instant k, the actual value of the random variables X_k can be determined in principle. To fix ideas, our subject might be drawing balls without replacement from an urn, in which case X_k could designate the colour of the k-th ball taken from the urn.

In the type of predictive inference we consider here, our subject in some way uses zero or more observations $X_1, \ldots, X_n$ made previously, i.e., those up to a certain instant $n \in \{0, 1, \ldots, N-1\}$, to predict, or make inferences about, the values of the future, or as yet unmade, observations $X_{n+1}, \ldots, X_N$. Here, we only consider the problem of *immediate prediction*: he is only trying to predict, or make inferences about, the value of the next observation X_{n+1}.

We are particularly interested in the problem of making such predictive inferences under prior ignorance: initially, *before making any observation, our subject knows very little or nothing about what produces these observations*. In the urn example, this is the situation where he doesn't know the composition of the urn, e.g., how many balls there are, or what their colours are. What we do assume, however, is that our subject makes an assessment of *exchangeability* to the effect that the order in which a sequence of observations has been made does not matter for his predictions.

What a subject usually does, in such a situation, is to determine, beforehand, a (finite and non-empty) set $\mathscr{X}$ of possible values, also called *categories*, for the random variables X_k. It is then sometimes held, especially by advocates of a logical interpretation for probability, that our subject's beliefs should be represented by some given family of predictive probability mass functions. Such a predictive family is made up of real-valued maps $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$, which give, for each $n = 0, \ldots, N-1$ and each $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathscr{X}^n$, the (so-called *predictive*) probability mass function for the $(n+1)$-th observation, given the values $(X_1, \ldots, X_n) = (x_1, \ldots, x_n) = \boldsymbol{x}$ of the n previous observations. Any such family should in particular reflect the above-mentioned exchangeability assessment. Cases in point are the Laplace–Bayes Rule of Succession in the case of two categories [10], or Carnap's more general λ-calculus [2].

The inferences in Carnap's λ-calculus, to give but one example, can strongly depend on the number of elements in the set $\mathscr{X}$. This may well be considered undesirable. If for instance, we consider drawing balls from an urn, predictive inferences about whether the next ball will be '*red or green*' ideally should not depend on whether we assume beforehand that the possible categories are '*red*', '*green*', '*blue*' and '*any other colour*', or whether we take them to be '*red or green*', '*blue*', '*yellow*' and '*any other colour*'. This desirable property was called *representation invariance* by Peter Walley [14], who argued that it cannot be satisfied by a *precise* probability model, i.e., by a system consisting of a family of predictive probability mass functions $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for every $\mathscr{X}$, but that it is satisfied by the so-

called Imprecise Dirichlet-Multinomial Model (or IDMM for short [15]). The IDMM can be seen as a special system of predictive *lower previsions* and it is a (predictive) cousin of the parametric Imprecise Dirichlet Model (or IDM [14]). Lower previsions are behavioural belief models that generalise the more classical Bayesian ones, such as probability mass functions, or previsions. We assume that the reader is familiar with at least the basic aspects of the theory of coherent lower previsions [13].

Here, we intend to study general systems of such predictive lower previsions. In Section 2, we give a general definition of such predictive systems and study a number of properties they can satisfy, such as coherence and exchangeability. In Section 3, we study the property of representation insensitivity for predictive systems, which is a stronger version of Walley's representation invariance, tailored to making inferences under prior ignorance. We show in Section 4 that there are representation insensitive and exchangeable predictive systems, by giving two examples. These two can be used to generate the mixing predictive systems, studied in Section 5. Among these, the ones corresponding to an IDMM take a special place, as they are the only ones to satisfy all the above-mentioned properties and an extra *specificity* property, related to behaviour under conditioning. In the Conclusions (Section 6), we list a number of interesting, but as of yet unresolved, questions.

2 Predictive families and systems

2.1 Families of predictive lower previsions

First assume that, before the subject starts making the observations X_k, he fixes a non-empty and finite set $\mathscr{X}$ of possible values for all the random variables X_k. Now suppose that he has observed the sequence of values $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n$ of the first n random variables, or in other words, he knows that $X_k = x_k$ for $k = 1, \ldots, n$. We want to represent his beliefs about the value of the next observation X_{n+1}, and the model we propose for this is a lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on the set $\mathscr{L}(\mathscr{X})$ of all gambles on $\mathscr{X}$. Let us first make clear what this means (see Walley's book [13] for more information).

A *gamble* f on $\mathscr{X}$ is a real-valued map on $\mathscr{X}$. It represents an uncertain reward, expressed in terms of some predetermined linear utility scale. So a gamble f yields a (possibly negative) reward of $f(x)$ utiles if the value of the next variable X_{n+1} turns out to be the category x in $\mathscr{X}$. The set of all gambles on $\mathscr{X}$ is denoted by $\mathscr{L}(\mathscr{X})$. The *lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$ of any gamble f on $\mathscr{X}$ is the subject's supremum acceptable price for buying this gamble, or in other words, the highest s such that he accepts the uncertain reward $f(X_{n+1}) - p$ for all $p < s$, conditional on his having observed the values $\boldsymbol{x} = (x_1, \ldots, x_n)$ for the first n variables $(X_1, \ldots, X_n)$. His corresponding *predictive upper prevision*, or infimum selling price for f, is then given by the conjugacy relationship: $\overline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = -\underline{P}^{n+1}_{\mathscr{X}}(-f|\boldsymbol{x})$.

A specific class of gambles is related to *events*, i.e., subsets A of $\mathscr{X}$. This is the class of indicators I_A that map any element of A to one and all other elements of $\mathscr{X}$ to zero. We identify events A with their indicators I_A. A lower prevision that is defined on (indicators of) events only is called a *lower probability*, and we often write $\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{x})$ instead of $\underline{P}^{n+1}_{\mathscr{X}}(I_A|\boldsymbol{x})$.

The *predictive lower prevision* $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$, which models beliefs about the value of the random variable X_{n+1} given the observations $(X_1, \ldots, X_n) = \boldsymbol{x}$, is the real-valued functional on $\mathscr{L}(\mathscr{X})$ that assigns to any gamble f its predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x})$. We assume that the subject has such a predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^n$ and all $n \in \{0, \ldots, N-1\}$, where $N > 0$ is some fixed positive integer, representing the total number of observations we are interested in. For $n = 0$, there is some slight abuse of notation here, because we then actually have an unconditional predictive lower prevision $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$ for the first observation X_1, and no observations have yet been made.

Definition 1 (Family of predictive lower previsions)**.** *Consider a finite and non-empty set of categories $\mathscr{X}$. An $\mathscr{X}$-*family of predictive lower previsions*, or* predictive $\mathscr{X}$-family *for short, for up to $N > 0$ observations is a set of predictive lower previsions*

$$\boldsymbol{\sigma}^N_{\mathscr{X}} := \left\{\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) \colon \boldsymbol{x} \in \mathscr{X}^n \text{ and } n = 0, \ldots, N-1\right\}.$$

It is useful to consider the special case, quite common in the literature, of a family of predictive lower previsions of which all members $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ are actually *linear previsions* $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$. This means that for each $n = 0, \ldots, N-1$ and $\boldsymbol{x}$ in $\mathscr{X}^n$ there is some predictive *(probability) mass function* $p^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ on $\mathscr{X}$ such that $\sum_{z \in \mathscr{X}} p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) = 1$, for all z in $\mathscr{X}$, $p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}) \geq 0$, and for all gambles f on $\mathscr{X}$

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}) = \sum_{z \in \mathscr{X}} f(z) p^{n+1}_{\mathscr{X}}(z|\boldsymbol{x}).$$

Such linear previsions are the Bayesian belief models usually encountered in the literature (see for instance de Finetti's book [7]). We can use Bayes's rule to combine these predictive mass functions into unique *joint mass functions* $p^n_{\mathscr{X}}$ on $\mathscr{X}^n := \times_{i=1}^n \mathscr{X}$, given by

$$p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(x_1, \ldots, x_n) = \prod_{k=0}^{n-1} p^{k+1}_{\mathscr{X}}(x_{k+1}|x_1, \ldots, x_k),$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathscr{X}^n$ and all $n = 1, \ldots, N$. This also results in unique corresponding linear previsions (expectation operators) $P^n_{\mathscr{X}}$ defined for all f in $\mathscr{L}(\mathscr{X}^n)$ by

$$P^n_{\mathscr{X}}(f) = \sum_{\boldsymbol{x} \in \mathscr{X}^n} f(\boldsymbol{x}) p^n_{\mathscr{X}}(\boldsymbol{x}). \tag{1}$$

For $n = N$, we call $P_{\mathscr{X}}^N$ the *joint linear prevision* associated with the given predictive family of linear previsions. It models beliefs about the values that the random variables $(X_1, \ldots, X_N)$ assume *jointly* in $\mathscr{X}^N$.

2.2 Systems of predictive lower previsions

When a subject is using a family of predictive lower previsions $\sigma_{\mathscr{X}}^N$, this means he has assumed beforehand that the random variables $X_1, \ldots, X_N$ all take values in the set $\mathscr{X}$. It cannot, therefore, be excluded at this point that his inferences, as represented by the predictive lower previsions $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$, strongly depend on the choice of the set of possible values $\mathscr{X}$. Any initial choice of $\mathscr{X}$ may lead to an essentially very different predictive family $\sigma_{\mathscr{X}}^N$. In order to be able to deal with this possible dependence mathematically, we now define predictive systems as follows.

Definition 2 (System of predictive lower previsions)**.** *Fix $N > 0$. If we consider for any finite and non-empty set of categories $\mathscr{X}$ a corresponding $\mathscr{X}$-family $\sigma_{\mathscr{X}}^N$ of predictive lower previsions $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$, we get a new collection*

$$\sigma^N := \left\{\sigma_{\mathscr{X}}^N : \mathscr{X} \text{ is a finite and non-empty set}\right\},$$

called a system of predictive lower previsions*, or* predictive system *for short, for up to N observations. We denote the set of all predictive systems for a given (fixed) N by Σ^N.*

It is such predictive systems that we are interested in, and whose properties we intend to study. Consider the set Σ^N of all predictive systems for up to N observations. For two such predictive systems σ^N and λ^N we say that σ^N is *less committal*, or *more conservative*, than λ^N, and we denote this by $\sigma^N \preceq \lambda^N$, if each predictive lower prevision $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ in σ^N is *point-wise dominated* by the corresponding predictive lower prevision $\underline{Q}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ in λ^N:

$$\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x}) \leq \underline{Q}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x})$$

for all gambles f on $\mathscr{X}$. The reason for this terminology should be clear: a subject using a predictive system λ^N will then be buying gambles f on $\mathscr{X}$ at supremum prices $\underline{Q}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x})$ that are at least as high as the supremum prices $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x})$ of a subject using predictive system σ^N.

The binary relation $\preceq$ on Σ^N is a partial order. A non-empty subset $\left\{\sigma_\gamma^N : \gamma \in \Gamma\right\}$ of Σ^N (where Γ is some index set) may have an infimum with respect to this partial order, and whenever it exists, this infimum corresponds to taking *lower envelopes*: if we fix $\mathscr{X}$, n and $\boldsymbol{x}$, then the corresponding predictive lower prevision in the infimum predictive system is the lower envelope $\inf_{\gamma\in\Gamma} \underline{P}_{\mathscr{X},\gamma}^{n+1}(\cdot|\boldsymbol{x})$ of the corresponding predictive lower previsions $\underline{P}_{\mathscr{X},\gamma}^{n+1}(\cdot|\boldsymbol{x})$ in the predictive systems σ_γ^N, $\gamma \in \Gamma$.

2.3 Coherence requirements

We impose some consistency, or rationality, requirements on the members $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ of a system σ^N of predictive lower previsions.

Definition 3 (Coherence)**.** *A system of predictive lower previsions is called* coherent *if it is the infimum (or lower envelope) of a collection of systems of predictive linear previsions.*

This condition is equivalent to requiring, for each choice of $\mathscr{X}$, that the conditional lower previsions $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ for $n = 0, \ldots, N-1$ and $\boldsymbol{x} \in \mathscr{X}^n$ should satisfy Walley's (joint) coherence condition.[1] This condition is in the present context also equivalent [12] to requiring that the predictive lower previsions $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ by themselves should be *(separately) coherent*, meaning that for each finite and non-empty set $\mathscr{X}$, $n = 0, \ldots, N-1$ and $\boldsymbol{x}$ in $\mathscr{X}^n$, $\underline{P}_{\mathscr{X}}^{n+1}(\cdot|\boldsymbol{x})$ should satisfy

(C1) $\underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x}) \geq \inf f$;

(C2) $\underline{P}_{\mathscr{X}}^{n+1}(f+g|\boldsymbol{x}) \geq \underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x}) + \underline{P}_{\mathscr{X}}^{n+1}(g|\boldsymbol{x})$;

(C3) $\underline{P}_{\mathscr{X}}^{n+1}(\lambda f|\boldsymbol{x}) = \lambda \underline{P}_{\mathscr{X}}^{n+1}(f|\boldsymbol{x})$;

for all gambles f and g on $\mathscr{X}$ and all real $\lambda \geq 0$.

2.4 Exchangeability and regular exchangeability

Next, we show how to formulate an assessment of *exchangeability* of the random variables $X_1, \ldots, X_N$ in terms of a system of predictive lower previsions. A subject would make such an assessment if he believed that the order in which these variables are observed is not important. Let us make this idea more precise.

We begin with the definition of exchangeability for a *precise* predictive system, i.e., a system of predictive linear previsions. For each choice of $\mathscr{X}$, the precise $\mathscr{X}$-family $\sigma_{\mathscr{X}}^N$ has a unique joint linear prevision $P_{\mathscr{X}}^N$ on $\mathscr{L}(\mathscr{X}^N)$, defined by Equation (1). *We then call the precise predictive system exchangeable if all the associated joint linear previsions $P_{\mathscr{X}}^N$ are.* Formally [5, 7], consider the set of all permutations of $\{1, \ldots, N\}$. With any such permutation π we can associate a permutation of $\mathscr{X}^N$, also denoted by π, that maps any $\boldsymbol{x} = (x_1, \ldots, x_N)$ in $\mathscr{X}^N$ to $\pi\boldsymbol{x} := (x_{\pi(1)}, \ldots, x_{\pi(N)})$. Similarly, with any gamble f on $\mathscr{X}^N$, we can consider the permuted gamble $\pi f := f \circ \pi$, or in other words $(\pi f)(\boldsymbol{x}) = f(\pi\boldsymbol{x})$. We then require that $P_{\mathscr{X}}^N(\pi f) = P_{\mathscr{X}}^N(f)$ for any such permutation π and any gamble f on $\mathscr{X}^N$. Equivalently, in terms of the joint mass

[1] See Chapters 6 and 7, and also Section K3 (Williams's Theorem) in Walley's book [13]. Since the random variables X_k are assumed to only take on a finite number of values, Walley's coherence condition coincides with the one first suggested by Williams [16].

function $p^N_{\mathscr{X}}$, we require that $p^N_{\mathscr{X}}(\pi\boldsymbol{x}) = p^N_{\mathscr{X}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in $\mathscr{X}^N$ and all permutations π.

We adopt the following definition of exchangeability for general predictive systems.

Definition 4 (Exchangeability)**.** *A system of predictive lower previsions is called* exchangeable *if it is the infimum (or lower envelope) of a collection of exchangeable systems of predictive linear previsions. We denote by $\langle\Sigma^N_{\mathrm{e}}, \preceq\rangle$ the set of all exchangeable predictive systems for up to N observations, with the same order relation $\preceq$ that we defined on $\langle\Sigma^N, \preceq\rangle$.*

The infimum (lower envelope) of any non-empty collection of exchangeable predictive systems is still exchangeable. This means that the partially ordered set $\langle\Sigma^N_{\mathrm{e}}, \preceq\rangle$ is a complete semi-lattice [3, Sections 3.19–3.20]. For reasons of mathematical convenience, we also introduce a stronger requirement.

Definition 5 (Regular exchangeability)**.** *A system of predictive lower previsions is called* regularly exchangeable *if it is the infimum (or lower envelope) of some collection σ^N_γ, $\gamma \in \Gamma$ of exchangeable systems of predictive linear previsions, such that for all finite and non-empty $\mathscr{X}$, all $\boldsymbol{x}$ in $\mathscr{X}^{N-1}$, and all γ in Γ,*

$$\begin{aligned} p^{N-1}_{\mathscr{X},\gamma}(\boldsymbol{x}) :=& P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}) \\ =& \prod_{k=0}^{N-2} p^{k+1}_{\mathscr{X},\gamma}(x_{k+1}|x_1,\ldots,x_k) > 0. \end{aligned}$$

Of course, all regularly exchangeable predictive systems are in particular also exchangeable and coherent. A *precise exchangeable predictive system* is regularly exchangeable if and only if $p^{N-1}_{\mathscr{X}}(x_1,\ldots,x_{N-1}) > 0$ for all $(x_1,\ldots,x_{N-1}) \in \mathscr{X}^{N-1}$ and all finite and non-empty sets $\mathscr{X}$. This shows that regular exchangeability is a stricter requirement than exchangeability.

The term *regular* here reminds of the notion of regular extension considered by Walley in [13]. In a regularly exchangeable predictive system every predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ is the lower envelope of the predictive linear previsions $P^{n+1}_{\mathscr{X},\gamma}(\cdot|\boldsymbol{x})$, which can be uniquely derived from the joint linear previsions $P^N_{\mathscr{X},\gamma}$ by applying Bayes's rule:

$$P^{n+1}_{\mathscr{X},\gamma}(f|\boldsymbol{x}) = \frac{P^N_{\mathscr{X},\gamma}(fI_{\{\boldsymbol{x}\}\times\mathscr{X}^{N-n}})}{P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})}$$

for every gamble $f \in \mathscr{L}(\mathscr{X})$ and sample $\boldsymbol{x} \in \mathscr{X}^n$, or equivalently,

$$p^{n+1}_{\mathscr{X},\gamma}(z|\boldsymbol{x}) = \frac{p^{n+1}_{\mathscr{X},\gamma}(\boldsymbol{x},z)}{p^n_{\mathscr{X},\gamma}(\boldsymbol{x})}$$

for all $z \in \mathscr{X}$ and $\boldsymbol{x} \in \mathscr{X}^n$, because the probability $p^n_{\mathscr{X},\gamma}(\boldsymbol{x}) := P^N_{\mathscr{X},\gamma}(\{\boldsymbol{x}\}\times\mathscr{X}^{N-n})$ of the conditioning event is non-zero.

In regularly exchangeable predictive systems, the number of times

$$T_z(\boldsymbol{x}) := |\{k \in \{1,\ldots,n\} \colon x_k = z\}|$$

that a given category z in $\mathscr{X}$ has been observed in some sample $\boldsymbol{x} \in \mathscr{X}^n$ of length $0 \leq n \leq N$, is of special importance. This leads us to consider the *counting map* $\boldsymbol{T}_{\mathscr{X}}$ that maps samples $\boldsymbol{x}$ of length n to the $\mathscr{X}$-tuple $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ whose components are $T_z(\boldsymbol{x})$, $z \in \mathscr{X}$. $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ tells us how many times each of the elements of $\mathscr{X}$ appears in the sample $\boldsymbol{x}$, and as $\boldsymbol{x}$ varies over $\mathscr{X}^n$, $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$ assumes all values in the set of *count vectors* $\mathscr{N}^n_{\mathscr{X}} := \left\{\boldsymbol{m} \in \mathbb{N}_0^{\mathscr{X}} : \sum_{z\in\mathscr{X}} m_z = n\right\}$. It is easy to see that any two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n have the same count vector $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$ if and only if there is some permutation π of $\{1,\ldots,n\}$ such that $\boldsymbol{y} = \pi\boldsymbol{x}$.

Proposition 1. *In any precise exchangeable predictive system σ^N, consider any finite and non-empty set $\mathscr{X}$, $0 \leq n \leq N-1$, and samples $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathscr{X}^n$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y})$. Then $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y})$ and moreover, if $p^n_{\mathscr{X}}(\boldsymbol{x}) = p^n_{\mathscr{X}}(\boldsymbol{y}) > 0$, then also $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$.*

In any regularly exchangeable predictive system, the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ only depend on the sample $\boldsymbol{x}$ through its count vector $\boldsymbol{m} = \boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x})$: for any other sample $\boldsymbol{y}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{y}) = \boldsymbol{m}$, it holds that $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{y})$ and we use the notation $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ for $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{x})$ in order to reflect this. In fact, from now on we only consider predictive systems—be they regularly exchangeable or not—for which the predictive lower previsions only depend on the observed samples through their count vectors, i.e., for which the count vectors are *sufficient statistics*.

One important reason for introducing regular exchangeability, is that it allows us to prove the following inequality, which has far-reaching consequences and which shall be used in Section 5.2. We denote by $\boldsymbol{e}_z$ the count vector in $\mathscr{N}^1_{\mathscr{X}}$ whose z-component is one and all of whose other components are zero; it corresponds to the case where we have a single observation which is of a category z in $\mathscr{X}$.

Proposition 2. *In any regularly exchangeable predictive system, it holds that*

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) \geq \underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m})$$

for all finite and non-empty sets $\mathscr{X}$, all $0 \leq n \leq N-2$, all $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$.

Here $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)$ denotes the gamble on $\mathscr{X}$ that assumes the value $\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_z)$ in $z \in \mathscr{X}$. It can be checked that the above inequality is an equality for precise regularly exchangeable predictive systems. The result follows then by taking lower envelopes.

3 Representation invariance and representation insensitivity

We are ready to consider Walley's notion of representation invariance; see his IDM paper [14] for more detailed discussion and motivation. While its definition seems to be fairly involved in case of general predictive inference, we shall see that it takes on a remarkably simple and intuitive form in the more special case of immediate prediction.

Representation invariance could also, and perhaps preferably so, be called *pooling invariance*. Consider a set of categories $\mathscr{X}$, and a partition $\mathscr{S}$ of $\mathscr{X}$. Each element S of such a partition corresponds to a single new category, that consists of all the elements $x \in S$ being pooled, i.e., considered as one. Denote by $S(x)$ the unique element of the partition $\mathscr{S}$ that a category $x \in \mathscr{X}$ belongs to. Now consider a gamble f on $\mathscr{X}$ that doesn't differentiate between pooled categories, or in other words, that is constant on the elements of $\mathscr{S}$. This f can be seen as a gamble $\tilde{f}$ on the set of categories $\mathscr{S}$, such that $\tilde{f}(S(x)) := f(x)$ for all $x \in \mathscr{X}$. Similarly, with a sample $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n$, there corresponds a sample $S(\boldsymbol{x}) := (S(x_1), \ldots, S(x_n)) \in \mathscr{S}^n$ of pooled categories. We consider $\mathscr{S}$ as a new set of categories, and representation invariance now requires that

$$\underline{P}^{n+1}_{\mathscr{S}}(\tilde{f}|S(\boldsymbol{x})) = \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{x}),$$

i.e., for gambles that do not differentiate between pooled categories, it should not matter whether we consider predictive inferences for the set of original categories $\mathscr{X}$, or for the set of pooled categories $\mathscr{S}$.

We are especially interested in predictive inference where a subject starts from a state of prior ignorance. In such a state, he has no reason to distinguish between the different elements of any set of categories $\mathscr{X}$ he has chosen. How can this be expressed in terms of predictive lower previsions? Consider a permutation ϖ of the elements of $\mathscr{X}$.[2] With any gamble f on $\mathscr{X}$, there corresponds a permuted gamble $\varpi f = f \circ \varpi$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a permuted sample $\varpi\boldsymbol{x} = (\varpi(x_1), \ldots, \varpi(x_n))$. If a subject has no reason to distinguish between categories z and their images ϖz, this means that

$$\underline{P}^{n+1}_{\mathscr{X}}(\varpi f|\boldsymbol{x}) = \underline{P}^{n+1}_{\mathscr{X}}(f|\varpi\boldsymbol{x}).$$

We call this property *category permutation invariance*.[3]

We call *representation insensitivity* the combination of both representation invariance and category permutation invariance. It means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation, i.e., category set. To make this more explicit, consider two non-empty and finite sets of categories $\mathscr{X}$ and $\mathscr{Y}$, and a so-called *relabeling map* $\rho\colon \mathscr{X} \to \mathscr{Y}$ that is *onto*, i.e., such that $\mathscr{Y} = \rho(\mathscr{X}) := \{\rho(x)\colon x \in \mathscr{X}\}$. Then with any gamble f on $\mathscr{Y}$ there corresponds a gamble $\rho f := f \circ \rho$ on $\mathscr{X}$. Similarly, with an observed sample $\boldsymbol{x}$ in $\mathscr{X}^n$, there corresponds a transformed sample $\rho\boldsymbol{x} = (\rho(x_1), \ldots, \rho(x_n))$ in $\mathscr{Y}^n$. *Representation insensitivity for immediate prediction then means that* $\underline{P}^{n+1}_{\mathscr{X}}(\rho f|\boldsymbol{x})$ *should be equal to* $\underline{P}^{n+1}_{\mathscr{Y}}(f|\rho\boldsymbol{x})$.

3.1 Definition and basic properties

For any gamble f on a finite and non-empty set of categories $\mathscr{X}$, its range $f(\mathscr{X}) := \{f(x)\colon x \in \mathscr{X}\}$ can again be considered as a finite and non-empty set of categories, and f itself can be considered as a relabeling map. With any $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ there corresponds a count vector $\boldsymbol{m}^f$ in $\mathscr{N}^n_{f(\mathscr{X})}$ defined by

$$m^f_r := \sum_{f(x)=r} m_x$$

for all $r \in f(\mathscr{X})$. Clearly, if $\boldsymbol{x}$ is a sample with count vector $\boldsymbol{m}$, then the relabeled sample $f\boldsymbol{x} = (f(x_1), \ldots, f(x_n))$ has count vector $\boldsymbol{m}^f$. Representation insensitivity is then equivalent to the following requirement, which we take as its definition, because of its simplicity and elegance.

Definition 6 (Representation insensitivity)**.** *A predictive system* σ^N *is* representation insensitive *if for all* $0 \leq n \leq N-1$*, for any finite and non-empty sets* $\mathscr{X}$ *and* $\mathscr{Y}$*, for any* $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ *and* $\boldsymbol{m}' \in \mathscr{N}^n_{\mathscr{Y}}$*, and for any gambles* f *on* $\mathscr{X}$ *and* g *on* $\mathscr{Y}$ *such that* $f(\mathscr{X}) = g(\mathscr{Y})$*, the following implication holds:*

$$\boldsymbol{m}^f = \boldsymbol{m}'^g \Rightarrow \underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{\mathscr{Y}}(g|\boldsymbol{m}').$$

Clearly, a predictive system σ^N is representation insensitive if and only if for all finite and non-empty sets $\mathscr{X}$, all $0 \leq n \leq N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all $f \in \mathscr{L}(\mathscr{X})$:

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{f(\mathscr{X})}(\mathrm{id}_{f(\mathscr{X})}|\boldsymbol{m}^f), \tag{2}$$

where $\mathrm{id}_{f(\mathscr{X})}$ denotes the identity map (gamble) on $f(\mathscr{X})$. The predictive lower prevision $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ then depends on $f(\mathscr{X})$ and $\boldsymbol{m}^f$ only, and not directly on $\mathscr{X}$, f and $\boldsymbol{m}$. More explicitly, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ *only depends on the values that* f *may assume, and on the number of times each value has been observed.*

We denote by $\Sigma^N_{\mathrm{e,ri}}$ the set of all exchangeable predictive systems that are representation insensitive. It is a subset of the class Σ^N_{e} of all exchangeable predictive systems, and

[2] This permutation ϖ of the elements of $\mathscr{X}$, or in other words of the *categories*, should be contrasted with the permutation π of the order of the observations, i.e., of the time set $\{1, \ldots, N\}$, considered in Section 2.4 in order to define exchangeability.

[3] It is related to the notion of (weak) permutation invariance that two of us studied in much detail in a paper [4] dealing with general issues of symmetry in uncertainty modelling.

it inherits the order $\preceq$. Clearly, taking (non-empty) infima preserves representation insensitivity, so $\langle \Sigma^N_{\text{e,ri}}, \preceq \rangle$ is a complete semi-lattice as well. We shall see in Theorem 5 that these two structures have the same bottom (the vacuous representation insensitive and exchangeable predictive system).

The remainder of this paper is devoted to the predictive systems in $\langle \Sigma^N_{\text{e,ri}}, \preceq \rangle$. So we are interested in finding, and studying the properties of, predictive systems that are both exchangeable (and therefore coherent) *and* representation insensitive. We believe performing such a study to be quite important, and we here report on our first attempts.

3.2 The lower probability function

With any predictive system σ^N, we can associate a map φ_{σ^N} that is defined on the subset $\{(n,m)\colon 0 \leq m \leq n \leq N-1\}$ of $\mathbb{N}_0^2$ by

$$\varphi_{\sigma^N}(n,m) := \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}} \,|n-m,m).$$

Why this map is important, becomes clear if we look at predictive systems that are representation insensitive. Consider any proper event $\emptyset \neq A \subsetneq \mathscr{X}$, then it follows by applying Equation (2) with $f = I_A$, that

$$\begin{aligned} \underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) &= \underline{P}^{n+1}_{\{0,1\}}(\mathrm{id}_{\{0,1\}} \,|n-m_A,m_A) \\ &= \varphi_{\sigma^N}(n,m_A) \end{aligned} \tag{3}$$

where $m_A := \sum_{z\in A} m_z$. So we see that in a representation insensitive predictive system, the lower probability of observing an event (that is neither considered to be impossible nor necessary) does not depend on the embedding set $\mathscr{X}$ nor on the event itself, but only on the total number of previous observations n, and on the number of times m that the event has been observed before, and is given by $\varphi_{\sigma^N}(n,m)$. Something similar holds of course for the upper probability of observing a non-trivial event. Indeed, by conjugacy,

$$\begin{aligned} \overline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) &= 1 - \underline{P}^{n+1}_{\mathscr{X}}(A^c|\boldsymbol{m}) = 1 - \varphi_{\sigma^N}(n,m_{A^c}) \\ &= 1 - \varphi_{\sigma^N}(n,n-m_A). \end{aligned} \tag{4}$$

This property of representation insensitive predictive systems is reminiscent of *Johnson's sufficientness postulate* [9] (we use Zabell's terminology [17]), which requires that the probability that the next observation will be a category x is a function $f_x(n,m_x)$ that depends only on the category x itself, on the number of times m_x that this category has been observed before, and on the total number of previous observations n. Representation insensitivity is stronger, because it entails that the function φ_{σ^N} that 'corresponds to' the f_x is the same for all categories x in all possible finite sets and non-empty $\mathscr{X}$.

We call φ_{σ^N} the *lower probability function* of the predictive system σ^N. We shall simply write φ instead of φ_{σ^N}, whenever it is clear from the context which predictive system we are talking about. Let us give a number of interesting properties for the lower probability function φ associated to a representation insensitive and coherent predictive system σ^N.

Proposition 3. *Let $N > 0$ and let σ^N be a representation insensitive and coherent predictive system with lower probability function φ. Then*

1. *φ is $[0,1]$-bounded:*
   $0 \leq \varphi(n,k) \leq 1$ for all $0 \leq k \leq n \leq N-1$.
2. *φ is super-additive in its second argument:*
   $\varphi(n,k+\ell) \geq \varphi(n,k) + \varphi(n,\ell)$ for all non-negative integers n, k and ℓ such that $k+\ell \leq n \leq N-1$.
3. *$\varphi(n,0) = 0$ for all $0 \leq n \leq N-1$.*
4. *$\varphi(n,k) \geq k\varphi(n,1)$ for $1 \leq k \leq n \leq N-1$, and $0 \leq n\varphi(n,1) \leq 1$ for $1 \leq n \leq N-1$.*
5. *φ is non-decreasing in its second argument:*
   $\varphi(n,k+1) \geq \varphi(n,k)$ for $0 \leq k < n \leq N-1$.

If σ^N is moreover regularly exchangeable, then

6. *$\varphi(n+1,k) + \varphi(n,k)[\varphi(n+1,k+1) - \varphi(n+1,k)] \leq \varphi(n,k)$ for $0 \leq k \leq n \leq N-2$.*
7. *φ is non-increasing in its first argument:*
   $\varphi(n+1,k) \leq \varphi(n,k)$ for $0 \leq k \leq n \leq N-2$.
8. *$\varphi(n,1) \geq \varphi(n+1,1)[1+\varphi(n,1)]$ for $1 \leq n \leq N-2$.*
9. *Suppose that $\varphi(n,1) > 0$ and define $s_n := \frac{1}{\varphi(n,1)} - n$ for $1 \leq n \leq N-1$.*[4] *Then $s_n \geq 0$, s_n is non-decreasing and $\varphi(n,1) = 1/(s_n+n)$.*

In particular, these results, together with Equations (3) and (4), allow us to draw interesting and intuitively appealing conclusions about predictive lower and upper probabilities, which are valid in any representation insensitive and coherent predictive system: (i) the lower probability of observing an event that hasn't been observed before is zero, and the upper probability of observing an event that has always been observed before is one [Proposition 3.3]; and (ii) if the number of observations remains fixed, then both the lower and the upper probability of observing an event again do not decrease if the number of times the event has already been observed increases [Proposition 3.5]. In predictive systems that are moreover regularly exchangeable, we also see that (iii) if the number of times an event has been observed remains the same as the number of observations increases, then the lower probability for observing the event again does not increase [Proposition 3.7].

[4] This s_n will later, in Section 5.2 turn out to be a constant (independent of the number of observations n) under special additional assumptions, and will play the rôle of the hyper-parameter s in the IDM.

When the predictive system consists solely of families of predictive linear previsions (apart from predictive lower previsions for dealing with zero previous observations, see Section 4), we can use the additivity of linear previsions, instead of the mere super-additivity of coherent lower previsions used previously, to get stronger versions of parts of Proposition 3[5]. Such predictive systems will be characterised in Theorem 6 further on.

Corollary 4. *Consider a representation insensitive and coherent predictive system σ^N, with a lower probability function φ, and such that all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ for $0 < n \leq N-1$ are linear previsions. Then for all $0 < n \leq N-1$ and all $k, \ell \geq 0$ such that $k+\ell \leq n$:*

1. $\varphi(n, k+\ell) = \varphi(n,k) + \varphi(n,\ell)$.
2. $\varphi(n,k) = k\varphi(n,1)$.

4 Are there representation insensitive exchangeable predictive systems?

We don't know yet if there are any predictive systems that are both representation insensitive and exchangeable. We remedy this situation here by establishing the existence of two 'extreme' types of representation insensitive and exchangeable predictive systems, one of which is also regularly exchangeable.

Consider, for any predictive system σ^N that is both representation insensitive and exchangeable, the predictive lower previsions for $n = 0$. These are actually unconditional lower previsions $\underline{P}^1_{\mathscr{X}}$ on $\mathscr{L}(\mathscr{X})$, modelling our beliefs about the first observation X_1, i.e., when no observations have yet been made. It follows right away from Proposition 3 and Equations (3) and (4) that for any proper subset A of $\mathscr{X}$, $\underline{P}^1_{\mathscr{X}}(A) = \varphi(0,0) = 0$. Since $\underline{P}^1_{\mathscr{X}}$ is assumed to be a (separately) coherent lower prevision, it follows that $\underline{P}^1_{\mathscr{X}}(f) = \min f$, for any gamble f on $\mathscr{X}$. So *all the $\underline{P}^1_{\mathscr{X}}$ in a representation insensitive and exchangeable predictive system must be so-called vacuous lower previsions.*[6] This means that there is no choice for the first predictions. It also means that it is impossible to achieve representation insensitivity in any precise predictive system (but see Theorem 6 for a predictive system that comes close).

This leads us to consider the so-called *vacuous* predictive system ν^N where all predictive previsions are vacuous: for all $0 \leq n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m}$ in $\mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$, $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \min f$.

Theorem 5. *The vacuous predictive system ν^N is regularly exchangeable and representation insensitive. It is the bottom (smallest element) of the complete semi-lattice $\langle \Sigma^N_{\text{e,ri}}, \preceq \rangle$. Its lower probability function is given by $\varphi(n,m) = 0$ for $0 \leq m \leq n \leq N-1$.*

In the vacuous predictive system the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are all vacuous, and therefore do not depend on the number of observations n, nor on the observed count vectors $\boldsymbol{m}$. A subject who is using the vacuous predictive system is not learning anything from the observations. Representation insensitivity and (regular) exchangeability do not guarantee that we become more committal as we have more information at our disposal. Indeed, with the vacuous predictive system, whatever our subject has observed before, he always remains fully uncommittal. If we want a predictive system where something is really being learned from the data, it seems we need to make some 'leap of faith', and add something to our assessments that is not a mere consequence of exchangeability and representation insensitivity.

So are there less trivial examples of exchangeable and representation insensitive predictive systems? We must make the vacuous choice for $n = 0$, but is there, for instance, a way to make the predictive lower previsions *precise*, or linear, for $n > 0$? The following theorem tells us there is only one such exchangeable and representation insensitive predictive system.

Theorem 6. *Consider a predictive system where for any $0 < n \leq N-1$ all the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are actually linear previsions $P^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$. If this predictive system is representation insensitive, then*

$$P^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \sum_{z \in \mathscr{X}} f(z)\frac{m_z}{n} \qquad (5)$$

for all $0 < n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. For its lower probability function φ, we then have $\varphi(n,k) = \frac{k}{n}$ for all $0 \leq k \leq n$ and $n > 0$. Moreover, the predictive previsions given by Equation (5)*, together with the vacuous lower previsions for $n = 0$, constitute a representation insensitive and exchangeable (but not regularly so) predictive system π^N.*

We call the predictive system π^N described in Theorem 6 the *Haldane* predictive system. The name refers to the fact that a Bayesian inference model with a multinomial likelihood function using Haldane's (improper) prior (see, e.g., Jeffreys [8, p. 123]) leads to these predictive previsions for $n > 0$.

It is a consequence of Walley's Marginal Extension Theorem [13, Section 6.7.3] that for any finite and non-empty $\mathscr{X}$, the only joint lower prevision on $\mathscr{L}(\mathscr{X}^N)$

[5] Note that the equalities in this corollary will also hold for some non-linear predictive systems, such as the mixing ones we shall consider in Section 5

[6] This result was proven, in another way, by Walley [13, Section 5.5.1], when he argued that his Embedding and Symmetry Principles under coherence only leave room for the vacuous lower prevision. When there are no prior observations ($n = 0$), the Embedding Principle is related to representation invariance, and the Symmetry Principle with what we have called category permutation invariance.

that is coherent with the Haldane predictive $\mathscr{X}$-family is given by $\underline{P}^N_{\mathscr{X}}(f) = \min_{z\in\mathscr{X}} f(z,\ldots,z)$. This implies that the Haldane predictive system is not regularly exchangeable: any dominating precise exchangeable predictive system satisfies $p^{N-1}_{\mathscr{X}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathscr{X}^{N-1}$ such that $\boldsymbol{T}_{\mathscr{X}}(\boldsymbol{x}) = \boldsymbol{m} \neq (N-1)\boldsymbol{e}_z$ for all $z \in \mathscr{X}$, and for any such $\boldsymbol{x}$, the requirements for regular exchangeability cannot be satisfied.

The Haldane predictive system only seems to be coherent with a joint lower prevision $\underline{P}^N_{\mathscr{X}}$ which expresses that our subject is certain that all variables X_k will assume *the same value*, but where he is completely ignorant about what that common value is. This is related to another observation: we deduce from Proposition 3.3 that in the Haldane predictive system, when $n > 0$ then not only the lower probability but also the upper probability of observing an event that hasn't been observed before is zero! This models that a subject is practically certain (because prepared to bet at all odds on the fact) that any event that hasn't been observed in the past will not be observed in the future either. The *sampling prevision* $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ for a gamble f in this predictive system is the expectation of f with respect to the observed (sampling) probability distribution on the set of categories. The Haldane predictive system is too strongly tied to the observations, and does not allow us to make 'reasonable' inferences in a general context.

5 Mixing predictive systems

So we have found two extreme representation insensitive and exchangeable predictive systems, both of which are not very useful: the first, because it doesn't allow us to learn from past observations, and the second, because its inferences are too strong and we seem to infer too much from the data. A natural question then is: can we find 'intermediate' representation insensitive and exchangeable predictive systems whose behaviour is stronger than the vacuous predictive system and weaker than the Haldane predictive system? The first idea that comes to mind, is to look at convex mixtures. Let us, therefore, consider a finite sequence ε, of N numbers $\varepsilon_n \in [0,1]$, $0 \leq n \leq N-1$, and study the *mixing predictive system* σ^N_ε whose predictive lower previsions are given by

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) := \varepsilon_n S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + (1-\varepsilon_n)\min f, \quad (6)$$

for all $0 \leq n \leq N-1$, all finite and non-empty sets of categories $\mathscr{X}$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$. As $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$ is only defined for $n > 0$, and since representation insensitivity and coherence require that $\underline{P}^1_{\mathscr{X}}$ should be vacuous, we always let $\varepsilon_0 = 0$ implicitly. We call any such sequence ε a *mixing sequence*, and we denote by φ_ε the lower probability function of the corresponding mixing predictive system σ^N_ε.

We are mainly interested in finding mixing predictive systems that are representation insensitive and (regularly) exchangeable. The following proposition tells us that the only real issue lies with exchangeability.

Proposition 7. *For any mixing sequence ε, the predictive system σ^N_ε is still representation insensitive. Moreover, let $0 \leq k \leq n \leq N-1$. Then $\varphi_\varepsilon(n,k) = \varepsilon_n \frac{k}{n}$, and if $\varepsilon_n > 0$ then $s_n = n\frac{1-\varepsilon_n}{\varepsilon_n}$ and $\varepsilon_n = \frac{n}{n+s_n}$. In particular $\varphi_\varepsilon(n,1) = \varepsilon_n/n$ is the lower probability of observing a non-trivial event that has been observed once before in n trials, $\varepsilon_n = n\varphi_\varepsilon(n,1)$ is the lower probability $\varphi_\varepsilon(n,n)$ of observing a non-trivial event that has always been observed before (n out of n times), and $s_n = \frac{1-\varphi_\varepsilon(n,n)}{\varphi_\varepsilon(n,1)}$ is the ratio of the upper probability of observing an event that has never been observed before to the lower probability of observing an event that has been observed once before, in n trials.*

We have already argued that in order to get away from making vacuous inferences, and in order to be able to learn from observations, we need to make some 'leap of faith' and go beyond merely requiring exchangeability and representation insensitivity. *One of the simplest ways to do so, is to specify the numbers $\varphi(n,1)$ for $n = 1,\ldots,N-1$, or in other words, to specify, beforehand, the lower probability of observing any non-trivial event that has been observed only once in n trials*. We can then ask for the most conservative representation insensitive predictive system that exhibits these lower probabilities. The following theorem tells us that mixing predictive systems play this part.

Theorem 8. *Consider $N > 0$ and a mixing sequence ε. Let σ^N be a representation insensitive coherent predictive system such that its associated lower probability function φ satisfies*

$$\varphi(n,1) \geq \varphi_\varepsilon(n,1) = \varepsilon_n/n$$

for all $0 < n \leq N-1$. Then $\sigma^N_\varepsilon \preceq \sigma^N$.

Mixing predictive systems have a special part in this theory, because they are quite simple, and in some sense most conservative. They are quite simple because all that is needed to specify them is the values $\varphi(n,1)$ of the lower probability function, or in other words, the lower probabilities that an event will occur that has been observed once in n observations. They are the most conservative coherent and representation insensitive predictive systems with the given values for $\varphi(n,1)$. In the following subsections we shall see that there are mixing predictive systems with a non-trivial mixing sequence ε that are also regularly exchangeable, and we derive a necessary condition on the mixing sequence ε for this to be the case.

5.1 The regular exchangeability of mixing predictive systems

Consider any mixing sequence ε and the corresponding mixing predictive system σ^N_ε. For the corresponding lower probability function φ_ε it holds by Proposition 7 that

$\varphi_\varepsilon(n,k) = \varepsilon_n \frac{k}{n}$; if we substitute this in the inequality of Proposition 3.8 we see that it is necessary for regular exchangeability that

$$\frac{\varepsilon_n}{n} \geq \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1,\ldots,N-1. \tag{7}$$

If one ε_n is zero, then all of the subsequent ε_{n+k} are zero as well: if inferences are vacuous after $n>0$ observations, they should also remain vacuous after subsequent ones. Or, to put it more boldly, in regularly exchangeable mixing predictive systems, if we are going to learn at all from observations, we have to start doing so from the first observation.

5.2 Predictive inferences for the IDMM

It is of particular interest to investigate for which types of mixing predictive systems, or in other words, for which mixing sequences ε, we generally have an equality rather than only an inequality in the condition of Proposition 2, i.e., for which

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \underline{P}^{n+1}_{\mathscr{X}}(\underline{P}^{n+2}_{\mathscr{X}}(f|\boldsymbol{m}+\boldsymbol{e}_\cdot)|\boldsymbol{m}), \tag{8}$$

for all finite and non-empty $\mathscr{X}$, all $0 \leq n \leq N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$ and all gambles f on $\mathscr{X}$, where the predictive lower previsions $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ are given by Equation (6). Using the definition of $S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m})$, and the coherence of $\underline{P}^{n+1}_{\mathscr{X}}(\cdot|\boldsymbol{m})$ we find, after some rearranging, that Equation (8) holds if and only if

$$\frac{\varepsilon_n}{n} = \frac{\varepsilon_{n+1}}{n+1}\left(1+\frac{\varepsilon_n}{n}\right), \quad n=1\ldots,N-1,$$

i.e., we have the equality in (7). Clearly, one ε_n is zero if and only if all of them are, which leads to the vacuous predictive system ν^N. We already know this vacuous system to be regularly exchangeable (and representation insensitive). If we assume on the other hand that $\varepsilon_n > 0$ for $n=1,\ldots,N$, and let $\zeta_n := n/\varepsilon_n = n+s_n \geq 1$, then the above equality can be rewritten as $\zeta_{n+1} = \zeta_n + 1$, which implies that there is some $s \geq 0$ such that $\zeta_n = n+s$, or equivalently, $s_n = s$ and consequently, $\varepsilon_n = \frac{n}{n+s}$, and

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) = \frac{n}{n+s} S^{n+1}_{\mathscr{X}}(f|\boldsymbol{m}) + \frac{s}{n+s}\min f \tag{9}$$

for $n=0,1,\ldots,N-1$. The predictive lower previsions in Equation (9) are precisely the ones that can be associated with the so-called Imprecise Dirichlet-Multinomial Model (or IDMM) with hyper-parameter s [15, Section 4.1]. We call mixing predictive systems of this type IDMM-*predictive systems*. The vacuous predictive system corresponds to letting $s \to \infty$.

Theorem 9. *The vacuous predictive system, and the* IDMM*-predictive systems for $s>0$ are regularly exchangeable and representation insensitive, and they are the only mixing predictive systems for which the equality* (8) *holds.*

Among the mixing predictive systems, the ones corresponding to the IDMM are also special in another way. which points to a quite peculiar, but intuitively appealing, property of predictive inferences produced by the IDMM. Indeed, assume that in addition to observing a count vector $\boldsymbol{m}$ of n observations, we know in some way that the $(n+1)$-th observation will belong to a proper subset A of $\mathscr{X}$—we might suppose for instance that the observation X_{n+1} has been made, but that it is imperfect, and only allows us to conclude that $X_{n+1} \in A$. Then we can ask what the updated beliefs are, i.e., what $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is. Since $\underline{P}^{n+1}_{\mathscr{X}}(A|\boldsymbol{m}) = \varepsilon_n m_A/n > 0$ if and only if $m_A > 0$ and $\varepsilon_n > 0$, let us assume that indeed $m_A > 0$ and $\varepsilon_n > 0$, in which case the requirements of coherence allow us to determine $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ uniquely, using the so-called Generalised Bayes Rule [13, Section 6.4]. This implies that $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is then the unique real μ such that

$$\underline{P}^{n+1}_{\mathscr{X}}(I_A(f-\mu)|\boldsymbol{m}) = 0.$$

We now have the following characterisation of IDMM-predictive systems.

Theorem 10 (Specificity)**.** *The* IDMM*-predictive systems with $s>0$ are the only mixing predictive systems with all $\varepsilon_n > 0$, $n=1,\ldots,N-1$ that satisfy*

$$\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A) = \underline{P}^{m_A+1}_{A}(f_A|\boldsymbol{m}_A) \tag{10}$$

for all $n=1,\ldots,N-1$, all $\boldsymbol{m} \in \mathscr{N}^n_{\mathscr{X}}$, all gambles f on $\mathscr{X}$ and all proper subsets A of $\mathscr{X}$ such that $m_A > 0$.

We have denoted by f_A the restriction of the gamble f to the set A, by $\boldsymbol{m}_A$ the A-tuple obtained from $\boldsymbol{m}$ by dropping the components that correspond to elements outside A. The sum of the components of $\boldsymbol{m}_A$ is m_A.

This so-called *specificity* property of inferences characterised by Equation (10) is quite peculiar. Suppose that you have observed n successive outcomes, leading to a count vector $\boldsymbol{m}$. If you know in addition that $X_{n+1} \in A$, then Equation (10) tells you that *the updated value $\underline{P}^{n+1}_{\mathscr{X}}(f|\boldsymbol{m},A)$ is the same as the one you would get by discarding all the* previous *observations producing values outside A, and in effect only retaining the m_A observations that were inside A!* Knowing that the $(n+1)$-th observation belongs to A allows you to ignore all the previous observations that happened to lie outside A. This is intuitively appealing, because it means that if you know that the outcome of the next observation belongs to A, only the related behaviour (the values of f on A and the previous observations of this set) matters for your prediction.

The name 'specificity' for this property was suggested to us by Jean-Marc Bernard. In one of his papers [1], he calls 'specific' any type of inference that has this particular property.

6 Conclusions

More work is needed in order to be able to draw a reasonably complete picture of the issue of representation insensitivity in predictive systems. Indeed, while doing research for this paper, we came across a multitude of questions that we haven't yet been able to answer, and we list only a few of them here.

(i) Are there (regularly) exchangeable and representation insensitive predictive systems that are not mixing predictive systems?

(ii) Related questions are: are there (regularly) exchangeable and representation insensitive predictive systems that, unlike the mixing systems, are not completely determined by the probabilities $\varphi(n,1)$ of observing an event that has been observed only once before in n observations; are there such predictive systems whose behaviour on gambles, unlike that of mixing systems, is not completely determined by the lower probability function φ; and are there such predictive systems whose lower probability function φ, unlike that of mixing systems, is not additive in the sense that $\varphi(n,k+\ell) = \varphi(n,k) + \varphi(n,\ell)$?

(iii) Are there (regularly) exchangeable and representation insensitive mixing predictive systems that are not of the IDMM-type? And if so,

(iv) are there (regularly) exchangeable, representation insensitive non-mixing predictive systems that satisfy Equation (10)?

(v) Can we arrive at stronger conclusions if we consider that the observations X_n make up an infinite exchangeable sequence?

(vi) Can more definite answers be given if we consider the general, rather than the immediate, prediction problem?

Acknowledgements

We thank Jean-Marc Bernard, Frank Coolen, Thomas Augustin, and the reviewers for useful discussions and comments.

References

[1] J.-M. Bernard. Bayesian analysis of tree-structured categorized data. *Revue Internationale de Systémique*, 11:11–29, 1997.

[2] R. Carnap. *The continuum of inductive methods*. The University of Chicago Press, 1952.

[3] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, 1990.

[4] G. de Cooman and E. Miranda. Symmetry of models versus models of symmetry. In W. L. Harper and G. R. Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.* King's College Publications, 2007.

[5] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937.

[6] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[7] B. de Finetti. *Theory of Probability*. John Wiley & Sons, Chichester, 1974–1975. English translation of [6], two volumes.

[8] H. Jeffreys. *Theory of Probability*. Oxford Classics series. Oxford University Press, 1998. Reprint of the third edition (1961), with corrections.

[9] W. E. Johnson. *Logic, Part III. The Logical Foundations of Science*. Cambridge University Press, 1924. Reprinted by Dover Publications in 1964.

[10] P.-S. Laplace. *Philosophical Essay on Probabilities*. Dover Publications, 1951. English translation of [11].

[11] P.-S. Laplace. *Essai philosophique sur les probabilités*. Christian Bourgois Éditeur, 1986. Reprinted from the fifth edition (1825).

[12] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 2007. Accepted for publication.

[13] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[14] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[15] P. Walley and J.-M. Bernard. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Laboratoire Cognition et Activitées Finalisés, Université de Paris 8, January 1999.

[16] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Reprinted in a revised form in the International Journal of Approximate Reasoning, 44(3), 366-383, 2007.

[17] S. L. Zabell. W. E. Johnson's "sufficientness" postulate. *The Annals of Statistics*, 10:1090–1099, 1982.

On the Explanatory Power of Indeterminate Probabilities

Horacio Arlo Costa
Carnegie Mellon University
hcosta@andrew.cmu.edu

Jeffrey Helzner
Columbia University
jh2239@columbia.edu

Abstract

Building on work that we reported in [1] we revisit the claims made by Fox and Tversky in [3] concerning their *comparative ignorance* hypothesis for decision making under uncertainty.

Keywords. normative, descriptive, indeterminate, rationality.

1 Introduction

The purpose of this paper is to report on recent developments in the research program that we introduced in [1] and continued in [2]. The motivating questions behind this program concern the extent to which certain normative theories of decision making that are based on indeterminate probabilities can be used to rationalize some observed deviations from the orthodox Bayesian account. This program may be contrasted with the so-called 'heuristics and biases' program which seeks descriptive theories of decision making that are capable of accommodating such deviations. Where our program is focused on the explanatory power of indeterminate probabilities as they are employed in certain normative theories, the heuristics and biases program introduces clearly non-normative explanatory devices such as psychological effects in order to accommodate certain deviations from orthodox Bayesian account.

Fox and Tversky's 'comparative ignorance' hypothesis of [3] is an important example of the sort of theoretical work that has been advanced within the heuristics and biases framework. Roughly, the basic idea behind comparative ignorance is that 'uncertainty aversion' is mainly driven by comparative contexts in which the decision maker is made aware of their lack of knowledge concerning a given uncertain event as a result of the salience of another uncertain event about which they are better informed. Fox and Tversky investigate the comparative ignorance hypothesis through a series of experiments that employ both 'clear' and 'vague' prospects in isolation and jointly. Following Fox and Tversky's terminology we will say that a comparative context obtains when the subject is presented with both a clear and a vague prospect.

The following illustrates a type of prospect that Fox and Tversky use in their work: An urn has been filled with 100 balls, where m of the balls are known to be solid black and $n \leq 100 - m$ are known to be solid white. What is the most that you would be willing to pay for a ticket that pays \$100 if a given random selection from the urn yields a black ball and pays \$0 dollars if the random selection yields a white ball? Prospects of this type for which $m+n = 100$ are said to be clear while those for which $m+n < 100$ are said to be vague.

Now suppose that we fix two prospects of the indicated type. Prospect A is clear and is determined by setting $m = 50 = n$. Prospect B is vague and is determined by setting $m = 0 = n$. According to the comparative ignorance hypothesis subjects who are presented with both A and B, thereby constituting a comparative context, will tend to exhibit a significant difference in their willingness to pay for these prospects while the difference between the maximum purchase price for those subjects who are offered A in isolation and those who are offered B in isolation will tend to be relatively insignificant.

In [1] we employed a rather novel methodology whereby subjects were asked to price their ticket on mixtures of the basic chance setups that were considered in Fox and Tversky's original experiment. For example, subjects were asked to state the most they would be willing to pay for their ticket given that the payoff would be determined by a two-stage chance setup where the first stage consists of a flip of a fair coin and the second stage, which depends on the outcome of the first, consists of a random draw from either the 50:50 urn or the urn that has a completely unknown ratio of black balls to white balls. With this

methodology we introduced gradations between Fox and Tversky's clear and vague bets; increasing the bias of the coin in the first stage in favor of the 50:50 urn results in a more clear bet while increasing the bias of the coin in the first stage in favor of the completely indeterminate urn results in a more vague bet. As we reported in [1] increasing amounts of vagueness had significance for the maximum buying prices even in the absence of a comparative context of the sort that was discussed by Fox and Tversky.

2 Experiment on Mixtures of Chance Setups

In our initial use of mixtures of chance setups we assumed that these setups were, at least in principle, reducible to one-stage setups. Given that an overarching goal of our project has been to investigate the explanatory power of indeterminate probabilities we have a clear interest in whether or not a play on a mixed chance setup can be exchanged for a play on a particular urn for which the subject is given some, although not necessarily complete, information concerning the ratio of black balls to white balls. We now report on study that takes some first steps towards an understanding of this reduction.

Our subjects, 56 undergraduates at Carnegie Mellon University, were presented with a questionnaire that began with the following description of the relevant chance setups:

Urn A contains exactly 100 balls. 50 of these balls are solid black and the remaining 50 are solid white.

Urn B contains exactly 100 balls. Each of these balls is either solid black or solid white, although the ratio of black balls to white balls is unknown.

Urn X contains exactly 100 balls. Each of these balls is either solid black or solid white. Further assumptions concerning this urn will be considered in the questions below.

The next item on the questionnaire was the following presentation of the alternatives that the subjects were asked to consider:

Alternative 1 We flip a fair coin. If the coin lands heads, then we draw a ball at random from Urn A. If the coin lands tales, then we draw a ball at random from Urn B. In either case, if the ball that is drawn is black, then you get \$100. However, if the ball that is drawn is white, then you get \$0.

Alternative 2 We draw a ball at random from Urn X. If the ball that is drawn is black, then you get \$100. If the ball that is drawn is white, then you get \$0.

Finally, we attempted to elicit a reduction with the following two questions:

Question 1: What is the smallest number m, between 0 and 100, such that if you were to learn that X contains at least m black balls then you would be willing to choose Alternative 2 when offered a choice between Alternative 1 and Alternative 2?

Question 2: Let m be your answer from Question 1. Assume that all you know about the distribution of black balls and white balls in Urn X is that there are at least m black balls in Urn X so that, according to your answer to Question 1, you would be willing to choose Alternative 2 when offered a choice between Alternative 1 and Alternative 2. Is there a number n, between 0 and $100-m$, such that if you were to learn that Urn X contains at least n white balls then you would no longer be willing to choose Alternative 2 when offered a choice between Alternative 1 and Alternative 2? If there is such a number, then write the least such number in the space below.

There are at least two rather obvious theoretical candidates that might be considered in connection with the questionnaire that has just been presented. According to the first of these, a reduction is given by the following operation:

$$U \oplus_\lambda V = \{\lambda p + (1-\lambda)q \mid p \in U \text{ and } q \in V\} \quad (1)$$

where U and V are each sets of probability distributions over a common state space and $\lambda \in [0,1]$ is a mixture weight. Interpret A as the set that contains exactly one distribution, namely the one that assigns a probability to drawing a black ball that is equal to that of drawing a white ball. Interpret B as the set of all distributions on $\{\text{Black}, \text{White}\}$. Taking λ as $\frac{1}{2}$, $A \oplus_{\frac{1}{2}} B$ evaluates to the set of all distributions p on the indicated set of states such that $p(\text{Black}) \geq \frac{1}{4}$ and $p(\text{White}) \geq \frac{1}{4}$. This suggests a reduction of the two-stage chance setup to a single-stage setup where the subject is told that there are at least 25 black balls and at least 25 white balls in the urn. For our second theoretical candidate imagine a subject who applies the principle of insufficient reason when considering B, the maximally indeterminate urn, and thus interprets the flip of the coin as leading to a play on a 50:50 urn in either case. This second account suggests a reduction of the two-stage setup to a single stage setup

	> 37.5	< 37.5
Black min	39	17
White max	47	9

Table 1:

> 37.5	< 37.5
34	2

Table 2:

where the subject is told that the urn has 50 black balls and 50 white balls.

It turns out that relatively few of the subjects were in complete agreement with either of the theoretical candidates. As noted, the first (second) theoretical candidate suggests a value of 25 (50) as the minimum number of known black balls required for Alternative 2 to be admissible and the first (second) theoretical candidate suggests a value of 25 (50) as the maximum number of known white balls such that Alternative 2 retains its status as an admissible option. [1] Now, if we suppose that each of these candidates is playing a role to some degree, then we can try to consider the extent to which one of the two seems to dominate by making a cut at 37.5, i.e. the midpoint ($\frac{25+50}{2}$) between the theoretical predictions regarding each of the two bounds.

Table 1 shows the number of subjects who reported a value above (below) the midpoint for each of the two bounds. Table 1 suggests that the second model, the one based on the principle of insufficient reason, is dominant, at least when the two questions are considered individually. Table 2 shows the breakdown when fit is considered with respect to both of the bounds. The first column of Table 2 shows the number of subjects who gave values above 37.5 for both of the bounds (i.e. the minimum for black and the maximum for white). Similarly, the second column shows the number of subjects who gave values below 37.5 for both of the bounds. Again the model that is based on the principle of insufficient reason appears dominant, 34 of 56 compared to 2 of 56.

While the analysis above suggests that the data favors the account based on the principle of indifference, it is important to note that relatively few subjects returned values that are in complete agreement with this account. As a possible explanation of this some might suggest that the mixed chance setup is itself a comparative context in that it makes salient a comparison between the maximally indeterminate urn and the 50:50 urn. To be sure, this is not quite a comparative context in the sense of Fox and Tversky: there is but one alternative being played against the mixed chance setup. Moreover, it is unclear how Fox and Tversky can in general interpret the upper and lower bounds that are reported. Of course we are open to interpreting these bounds as upper and lower probabilities for a potential credal state, but such an interpretation does not seem to be an option for Fox and Tversky. Nonetheless, let us allow a very generous interpretation of 'comparative context' so that the results of [1], which employed mixed chance setups, can be countered by objecting that we did in fact employ a comparative context. The following studies attempt to reinforce our point without appealing to mixed chance setups.

[1] This first set of values (i.e. the minimum number of known black balls) may be computed directly from Question 1, while the second set of values (i.e. the maximum number of known white balls) can be computed from Question 2 as $n - 1$ if a value is supplied and $100 - m$ if no value is supplied by the subject.

3 Two Experiments That Do Not Use Mixtures of Chance Setups

Fox and Tversky predict essentially Bayesian behavior in the absence of a comparative context. Moreover, in keeping with much of the heuristics and biases tradition, they seem to interpret all deviations from the Bayesian standard as instances of irrationality; deviations from this standard are predicted when the subject is under the spell of various psychological effects, which in the sort of cases that we have been considering are those resulting from the presence of a comparative context. While our primary focus in [1] was to argue against some of the core claims in [3], we will now consider data that has direct relevance to the manner in which Fox and Tversky seem to interpret deviations from the Bayesian standard.

Using a protocol that was derived from an example in [4] we attempted to ascertain the extent to which violations of the Bayesian standard could be accommodated by certain normative alternatives based on indeterminate probabilities. For the purposes of these studies we focused on the following decision rules:

E-admissibility: Assume that the decision maker's credal state can be represented by a set P of probability distributions. If X is a set of alternatives, then a is *E-admissible* in X iff $a \in X$ and there is some distribution in $p \in P$ for which $E_p[a] \geq E_p[b]$ for all $b \in X$, where $E_p[x]$ is the the expected utility of x against distribution p..

MMEU: Assume that the decision maker's credal state can be represented by a set P of probability distributions. For each alternative x, let x_- be

the greatest lower bound of $\{E_p[x] \mid p \in P\}$. If X is a set of alternatives, then $x \in X$ satisfies the *maximin criterion for expected utility* (MMEU) on X iff $x_- \geq y_-$ for all $y \in X$.

E-admissibility followed by MMEU: Assume that the decision kakers credal state can be represented by a set P of probability distributions. If X is a set of alternatives, then a is admissible in X according to this rule iff a satisfies MMEU on the set of alternatives that are E-admissible in X.

E-admissibility is discussed in [6] where it is taken as the first tier in Levi's two-tiered decision theory. Alternatively, taken as a free standing decision rule, E-admissibility corresponds to Levi's theory when the security, the second tier, is vacuous. MMEU has a long history in the statistics literature, e.g. in discussions of 'gamma minimax', and continues to receive attention in decision contexts [4, 5]. The third criterion is essentially an instance of Levi's decision theory where the second tier security rule is given by MMEU.

3.1 Study 1

Our subjects, 56 undergraduates at Carnegie Mellon University, were presented with a questionnaire that began with the following description of the underlying chance setup.

An urn has been filled with several balls, each of which is either solid black or solid white. While the exact ratio of black balls to white balls is unknown, the following statistical information is available:

Black The probability of selecting a black ball on a single random draw from the urn is at least %40 but not more than %60.

White The probability of selecting a white ball on a single random draw from the urn is at least %40 but not more than %60.

The next section of the questionnaire introduced the following choice problem, which, as noted above, is based on an example from [4]:

Consider the three alternatives in the table below. Note that the payoffs for these alternatives are in dollars. So, for example, Alternative pays $-10 if a black ball is drawn (i.e. you lose $10 if a black ball is drawn from the urn) and pays $12 dollars if a white ball is drawn.

	Black	**White**
A	-10	12
B	11	-9
C	0	0

A,B	B	C	Other
12	9	22	13

Table 3:

In the final section of the questionnaire the subjects were given the following prompt and then asked to indicate the alternatives that they would be willing to choose.

Suppose that you are offered the opportunity to specify the alternatives above that you are willing to choose, with the understanding that we will pick one of these alternatives *before the random selection from the urn* and you will receive the winnings, or pay the losses, that are generated by the alternative that we pick.

Before turning to the data from this initial study, let us apply the three decision rules from the previous section to the choice problem that is presented in the above questionnaire. Assume that utilities are determinate and linear in dollars. Assume that the agent's credal state is given by the description at the beginning of the protocol. That is, assume that the agent's credal state can be represented by the set $P = \{p : .4 \leq p(\text{Black}) \leq .6\}$. Under these assumptions it follows that A and B are the only E-admissible alternatives in $\{A, B, C\}$ while C is the only alternative that satisfies MMEU in $\{A, B, C\}$. The two-tiered rule, E-admissibility followed by MMEU, counts B as the lone admissible alternative since B is the only alternative that satisfies MMEU in $\{A, B\}$.

Table 3 shows that those subjects who regarded C as uniquely admissible constitute the largest group by a rather wide margin, with the total for C being approximately equal to the combined totals for A, B and B. It is worth noting that C is the only alternative that fails to be a bayes solution under the assumption that utilities in this range are determinate and essentially linear in dollars. Given that this is a noncomparative problem, C's dominant position seems to disconfirm Fox and Tversky's prediction of essentially Bayesian behavior in absence of a comparative context. This point can be strengthened if we note that Fox and Tversky seem to have in mind that the appropriate Bayesian model for predicting behavior in noncomparative contexts is one that appeals to the principle of insufficient reason. That is, Fox and Tversky seem to predict that subjects who are faced with the given noncomparative choice problem will choose as though they are maximizing expectations against the distribution that assigns the two states an equal probability. Assuming that utilities are determinate and linear in dollars, subjects who are in accordance with this prediction must be willing to choose A and B. Hence,

in addition to those who judged C to be uniquely admissible, those who judged B to be uniquely admissible fail to confirm Fox and Tversky's prediction for noncomparative choice under uncertainty.

Although their presence seems to disconfirm Fox and Tversky's predictions, at least under the assumption that utilities are determinate and linear in dollars, those who judged either B or C to be uniquely admissible in the triple are consistent with one of the normative alternatives discussed above under this very same linearity assumption. Of course we recognize that by appealing to other considerations, e.g. nonlinear utilities or perhaps various psychological effects, one might formulate decidedly non-normative decision models that are capable of reproducing the admissibility judgments that were reported by these subjects. It is for this reason that we decided to conduct a further investigation in order to determine the extent to which subjects reasoned in the manner suggested by the normative theory that reproduced their admissible choices. Details of this second study are the subject of the next section.

3.2 Study 2

Our subjects, 27 undergraduates at Carnegie Mellon University, were presented with a questionnaire that began exactly as the one that was employed in the previous study but continued with the following illustrations of each of the three decision rules that were discussed at the beginning of Section 3.

Albert's reasoning: Note that the only probability distributions that are consistent with the information that is given are those for which the probability assigned to drawing a black (white) chip is at least .4 and no more than .6. Among the distributions that satisfy these conditions, there are some for which A maximizes expected value. For example, if the probability of drawing a black ball is .4, and so the probability of drawing a white ball is .6, the following table gives the expected valued of each of the alternatives.

	p(**Black**) = .4	p(**White**) = .6	Expected Val.
A	-10	12	3.2
B	11	-9	-1.0
C	0	0	0

From the table it is clear that A maximizes expected value against the probability distribution that assigns a probability of .6 to drawing a white ball. Similarly, B maximizes expected value against the probability distribution that assigns a probability of .6 to drawing a black ball and a probability of .4 to drawing a white ball. On the other hand there is no distribution that is consistent with the information that is given and against which C maximizes expected value. With this reasoning I eliminated C from further consideration. I would be willing to choose A and B, but not C.

Bob's reasoning: Well, I eliminated C along the same lines as Albert suggested, but then I appealed to some additional considerations. Since the minimal expected value of A is -1.2, which occurs when the probability assigned to drawing a black ball is .6, and the minimal expected value of B is -1.0, which occurs when the probability assigned to drawing a white ball is .6, I decided to eliminate A from further consideration. I would be willing to choose B, but not A or C.

Carol's reasoning: My reasoning was essentially like Bob's, except for the part where he followed Albert. That is, I simply considered the minimal expectation of each of the three alternatives. Since the minimal expected value of A is -1.2 and the minimal expected value of B is -1.0, while the minimal expected value of C is 0, I eliminated A and B from further consideration. C has the largest minimal expectation. So, I would be willing to choose C, but not A or B.

Finally, the subjects in this study were asked to indicate their level of agreement with each of the statements below. We instructed the subjects to indicate their level of agreement on a scale from 1 to 10 (i.e. $1, 2, 3, ..., 10$) with 1 being 'not at all' and 10 being 'as much as possible':

- Albert's reasoning is compelling.
- Bob's reasoning is compelling.
- Carol's reasoning is compelling.
- Albert's reasoning resembles the reasoning that I used in formulating my own response to the question.
- Bob's reasoning resembles the reasoning that I used in formulating my own response to the question.
- Carol's reasoning resembles the reasoning that I used in formulating my own response to the question.

First, before turning the data obtained from the additional questions, we recall that the subjects in this second study also answered the questions that were given to those in the first study. Table 4 shows the breakdown of this group of subjects in terms of the same partition that was employed in Table 3. As was reported in Table 3 in connection with the first study, Table 4 shows that the group of subjects who judged C to be uniquely admissible in the triple is the largest of the four groups and, as before, is roughly the size of the groups for A, B and B combined.

A,B	B	C	Other
7	3	12	5

Table 4:

It is important to remember that the first part of the questionnaire that was employed in Study 2 is identical to the questionnaire that was used in Study 1. One of the referees who commented on an earlier version of this paper suggested that the additional questions that were used in Study 2 might have led the subjects. There are at least two reasons to believe that this is not the case. First, the additional questions, i.e those concerning the three character sketches, were posed at the end of the questionnaire. The subjects were instructed to respond to the questions in the order that they were presented and were told not to go back to revise their answers to earlier questions. Second, Table 4, which shows the data from the part of the questionnaire that duplicated what was used in Study 1, suggests a very similar breakdown to what was observed in Study 1.

Returning to the matter that prompted this second study, we note that 15 of the 27 subjects reported B or C. The issue that prompted this second study concerns the extent to which these subjects determine admissibility by appealing to the considerations that are suggested by one of three non-Bayesian, normative rules described above. In terms of the additional questions that were employed in this second study we can attempt to address this question by isolating those subjects who reported a high level of resemblance between their own reasoning and the appropriate non-Bayesian norm. Interpreting a high level of resemblance to be a value of 8 or above for the subject's response to the relevant question we observed that 11 of these 15 appealed to considerations that had a high level of resemblance to those suggested by the appropriate non-Bayesian norm. Finally, although the numbers are getting rather small at this point it is perhaps worth noting that although A, B is consistent with Fox and Tversky's predictions the majority of the subjects in the A, B group reported that their reasoning had a high resemblance to Albert's E-admissibility considerations.

4 Conditional Support

Subjects evaluate the resemblance of their own form of reasoning to the theories exemplified by Albert, Bob and Carol's reasoning at the end of the questionnaire. Previous to this, and after receiving the information about Albert, Bob and Carol, they assess the validity of these theories. One minimal desideratum here is that subjects who judge their own reasoning to have a strong resemblance to theory X (with X varying over the theories advanced by Albert, Bob and Carol) rank the theory X with a score superior to at least 5 in the scale from 0 to 10. Otherwise we would have a situation where subjects see themselves as judging according to a theory that they themselves consider to have dubious validity.

Not all subjects obey these minimal desiderata and we propose to filter them out in order to consider unconditional and conditional support for the three theories under consideration. In particular there is a subject who chooses C and sees himself as choosing according to Carols considerations but gives Carol's theory a score of 4.

If we consider unconditional support for C after this subject is eliminated form the pool of respondents the average unconditional support for C has a value of 8.54 (in comparison with a value of 8.1 before eliminating subjects who do not obey the aforementioned desideratum).

There is also a separate measure of interest which is given by the amount of support that a theory X received conditional on the fact that the subject chooses what X recommends and that the subject sees himself as choosing in accordance to X. We will consider that a subject sees herself as choosing in accordance with X if she ranks X as resembling her form of reasoning with a score of at least 8.

The average conditional support for C in these circumstances has a value of 9.25. Similarly the average conditional support for A, B has a value of 8.25. And the corresponding average conditional (and unconditional) support for B has a maximal value of 10. So, all the values of average conditional support are relatively high.

The average unconditional support for A, B is, nevertheless, lower than the average conditional support (7.71) indicating that there are some subjects who choose A, B but do not see themselves as choosing in accordance with Albert's form of reasoning. It would be interesting in future research to consider alternative forms of reasoning consistent with choosing A, B even when they might not be articulated in terms of indeterminate probabilities (applications of the principle of insufficient reasoning might be a salient option here).

In the case of option C the gap between unconditional and conditional support is less significant (8.54 after sensitivity analysis as opposed to 9.25) indicating that this form of non-Bayesian reasoning is very robust. Finally there is no gap between average unconditional

and conditional support in the case of B. This form of non-Bayesian reasoning occurs in a minority of cases as opposed to A, B and B, but it is supported in a very strong manner when it occurs.

5 Future work

The comparative ignorance hypothesis advanced by Fox and Tversky explains deviations from Bayesian behavior in cases where there is indeterminacy in terms of a psychological effect, namely *uncertainty aversion* driven by comparative contexts. But as we tried to make clear in this paper there are frequent cases of decision contexts where there is indeterminacy but no comparison is being made. The scenario in Study 1 is such a case. As we stressed above this is a case where there is no comparison between clear and vague bets. All bets are vague. In a situation of this sort it seems that there is no psychological effect, at least along the lines that were suggested by Fox and Tversky's account, that one might invoke to predict a deviation from Bayesian standards of rationality.

There are, nevertheless, several decision models that take indeterminacy seriously and might be used to explain the behavior verified in the experiments that we have reported. Of course there could be other models that take indeterminacy into account but in a way that is very different from what is suggested in the decision rules that we have considered. Alternatively, there might be an entirely different psychological effect that is driving the behavior of subjects. The existence of all these possibilities is what motivated our second experiment, where some theoretical options were presented to the subjects for their appraisal. Subjects had the option of saying that none of the presented options represented their reasoning adequately. Nevertheless we verified that 11 out of 15 subjects selected one of the theoretical options as closely resembling their reasoning. So, this seems to indicate that one of the theoretical options that takes indeterminacy seriously (MMEU) figures among the reasoning strategies of actual subjects.

Is it possible that psychological effects that have nothing to do with indeterminacy motivate the non-Bayesian behavior verified in the experiments? One referee pointed out that the fact that our example uses negative payoffs might be the cause of some of the behavior observed in the experiments. The idea is that agents might be motivated by loss aversion and that this explains the selection of option C in Study 1 (2). This nevertheless does not explain why subjects selected Carol's reasoning as resembling as much as possible their own reasoning. One needs to assume here that a majority of subjects were mistaken in assessing their own reasoning.

One experiment that can settle the issue (as suggested by the referee) is to run a version of Study 1 (2) where 15 dollars is added uniformly to all payoffs in the matrix used in both experiments. The referee predicts that in this case option C will lose its appeal and that options A and B (a Bayes solution) would be chosen. Notice that even if this behavior were observed this solution is also compatible with Albert's reasoning (E-admissibility). To settle this issue we propose to add a theoretical option along the lines of the principle of insufficient reason to the salient theoretical options offered to the subjects in a new version of experiment two. Fox and Tversky seemed to have predicted that in a situation of this sort agents will appeal to insufficient reasoning. Other methods of dealing with indeterminacy (like Albert's reasoning) remain possible as well. So, the problem of determining which one of these methods constitutes an empirically robust response to indeterminacy remains as open in this new experimental set up as it was in the scenario investigated in this paper.

6 Conclusions

In Section 2 we reported on an experiment that was conducted in order to investigate the manner in which subjects reduce mixtures of chance setups, of the sort that we employed in [1], to indeterminate probabilities. This was important to us because part of our overall research program is an exploration of the explanatory power of indeterminate probabilities, especially as this stands in contrast to the purely descriptive agenda that is articulated within the heuristics and biases paradigm. As we discussed in Section 2, the data that were generated by this experiment raised the possibility that mixtures of chance setups might constitute a comparative context of sorts. If so, then such a thing could be offered as an objection to the arguments that we advanced in [1], e.g. one could object that we had smuggled in a comparative context by using mixtures of chance setups. Anticipating this objection we conducted the two experiments that are reported in Section 3. These two experiments address the core of Fox and Tversky's claims without appealing to mixtures of chance setups. The results from the first of these studies suggests a significant amount of non-Bayesian behavior occurring in a noncomparative context. We were able to rationalize much of this non-Bayesian behavior in terms of three well-known normative rules that are based on indeterminate probabilities. The second study in Section 3 suggests that such rationalizations of the indicated non-Bayesian behavior are not merely of the 'as if' variety but rather approximate a sub-

stantial portion of the reasoning that is driving this behavior. Thus, despite the claims of Fox and Tversky, it appears that there is a significant amount of non-Bayesian behavior even in the absence of a comparative context.

References

[1] H. Arlo Costa and J. Helzner, *Comparative ignornace and the ellsberg phenomenon*, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications (F. Cozman, R. Nau, and T. Seidenfeld, eds.), 2005.

[2] ———, *Risk, indeterminacy and decision*, Proceedings of the ESSLI 2006 Workshop on Rationality and Knowledge (R. Parikh and S. Artemov, eds.), 2006.

[3] C.R. Fox and A. Tversky, *Ambiguity aversion and comparative ignorance*, The Quarterly Journal of Economics (1991).

[4] P. Gardenfors and N.E. Sahlin, *Unreliable probabilities, risk taking, and decision making*, Synthese (1982).

[5] I. Ghilboa and D. Schmeidler, *Maxmin expected utility with non-unique prior*, J. Math. Econ. (1989).

[6] I. Levi, *On indeterminate probabilities*, Journal of Philosophy (1974).

Independence concepts in evidence theory

Inés Couso
Dep. of Statistics and O.R.
University of Oviedo, Spain
couso@uniovi.es

Abstract

We study three conditions of independence within Evidence Theory framework. First condition refers to the selection of pairs of focal sets. The remaining two are related to the choice of a pair of elements, once a pair of focal sets has been selected. These three concepts allow us to formalize the ideas of lack of interaction between variables and between their (imprecise) observations. We illustrate the difference between both types of independence with simple examples about drawing balls from urns. We show that there are not implication relationships between both of them. We derive interesting conclusions about the relationships between the concepts of "independence in the selection" and "random set independence".

Keywords. Evidence Theory, Independence, Random Sets, Sets of Probabilities.

1 Introduction

The concept of stochastic independence is essential in probability theory. Factorization allows us to decompose complex problems into simpler components. When generalizing to imprecise probabilities, the concept of independence, which is unique in probability theory, can be extended in different ways. Different definitions of independence for imprecise probabilities are studied and compared in [1], [2] and [7].

Evidence theory ([5]) falls within the theory of imprecise probabilities. This way, definitions of independence for imprecise probabilities can be transferred to this context. In [3], for instance, sets of joint probability measures associated to joint mass assignments are constructed. Different ways of choosing the weights of the joint focal sets and the probability measures inside these sets are considered. Depending on these conditions, different sets of joint probability measures are obtained. The author shows that some of these cases lead to types of independence described in [2] such as strong independence, random set independence and unknown interaction. The author initially considers the class of all probability measures on a product space whose marginals are dominated by a pair of plausibility measures. Next he establishes three rules to construct probabilities within that class. Each rule is related to a particular aspect of independence and it determines a subclass in the initial set of probability measures. First rule refers to the choice of weights of the joint focal sets, and it is related to the concept of random set independence. Second and third rules are referred to the choice of the probability measures inside the focal sets. The author shows that the class of probability measures based on these three rules satisfies independence in the selection. We will go further on this study. First, we will recall these notions under a different framework. Then we will give an intuitive meaning for each rule, by means of simple examples about drawing balls from urns. Our main goal is showing that none of these rules is strictly necessary to get independence in the selection. In fact, we will construct product probabilities without using some of these rules. This will be possible because the same probability measure can be constructed by using different procedures. In fact, we can choose weights of the joint focal sets and/or the probability measures inside the focal sets and finally get the same probability measure.

We will also go into further details about the relationships between random set independence ([2]) and type 1 independence [1]. It is well known that the class of probability measures associated to random set independence includes the class of probability measures satisfying type 1 independence (see [2], for instance). We will check in the paper that this is a strict inclusion, except for trivial situations (precise probabilities).

Our analysis does not apply to all interpretations of Evidence Theory, but only when the pair of plausibility and belief functions is regarded as a family of

probability measures. Different interpretations of Evidence Theory as the Transferable Belief Model ([6]) lead to different approaches (see [8], for instance) to the concept of independence.

The paper is organized as follows. Section 2 provides the necessary technical background about upper probabilities, evidence theory and independence notions for imprecise probabilities. Section 3 is devoted to different representations of the class of probability measures dominated by a particular plausibility function. We end the paper with some general concluding remarks and open problems.

2 Preliminary concepts and notation

Let us introduce some notation and recall some definitions needed in the rest of the paper.

2.1 Sets of probability measures

Consider a finite universe Ω We will denote $\mathcal{P}_\Omega$ the class of all probability measures we can define on $\wp(\Omega)$. Let $\mathcal{P} \subseteq \mathcal{P}_\Omega$ an arbitrary subset. It induces upper and lower probability functions respectively defined by

$$P^*(A) = \sup_{Q\in\mathcal{P}} Q(A);\ P_*(A) = \inf_{Q\in\mathcal{P}} Q(A) \qquad (1)$$

The set of probability measures dominated by an upper probability P^* is denoted by $\mathcal{P}(P^*) = \{Q : Q(A) \leq P^*(A), \forall A \subseteq \Omega\}$. If the upper probability measure P^* is generated by the family $\mathcal{P}$, then $\mathcal{P}(P^*)$ is generally a proper superset of $\mathcal{P}$.

Mathematical evidence theory of Shafer extends classical probability theory. In this framework, a *basic mass assignment*, m, is a mass of probability defined over the power set of Ω. It assigns a positive mass to a family of subsets of Ω called the set $\mathcal{F}_m$ of focal subsets. Generally, $m(\emptyset) = 0$ and $\sum_{E\in\mathcal{F}_m} m(E) = 1$. This mass assignment induces set functions called plausibility and belief measures, respectively denoted by Pl and Bel, and defined by Shafer [5] as follows:

$$\mathrm{Pl}(A) = \sum_{E\cap A\neq\emptyset} m(E) \qquad \mathrm{Bel}(A) = \sum_{E\subseteq A} m(E).$$

2.2 Independence concepts for imprecise probabilities

Consider two variables or uncertain values which may be regarded as the outcomes of two experiments. Assume that the two outcomes are known to belong to the universes Ω_1 and Ω_2 which are finite. Assume that the set of possible joint outcomes is the cartesian product $\Omega_1 \times \Omega_2$. Let us respectively represent by $\mathcal{P}_1 \subseteq \mathcal{P}_{\Omega_1}$ and $\mathcal{P}_2 \subseteq \mathcal{P}_{\Omega_2}$ our knowledge about the true distribution of probability that models each marginal experiment. Let $\mathcal{P} \subseteq \mathcal{P}_{\Omega_1\times\Omega_2}$ represent our (imprecise) knowledge about the joint probability distribution associated to the joint experiment. Given a joint probability measure, P on $\Omega_1 \times \Omega_2$ we will respectively denote P_1 and P_2 its marginals on Ω_1 and Ω_2, i.e., $P_1(A) = P(A \times \Omega_2)$, and $P_2(B) = P(\Omega_1 \times B),\ \forall\, A \subseteq \Omega_1, B \subseteq \Omega_2$.

We say that there is *type 1 independence* [1] when every joint probability $P \in \mathcal{P}$ factorizes as $P = P_1 \otimes P_2$, i.e., $P(A \times B) = P(A \times \Omega_2)\, P(\Omega_1 \times B),\ \forall\, A \subseteq \Omega_1, B \subseteq \Omega_2$. In other words, when

$$\mathcal{P} \subseteq \{P_1 \otimes P_2 : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2\}.$$

This concept is closely related to the notion of *independence in the selection* studied in [2].

Suppose that $\mathcal{P}_1 = \mathcal{P}(\mathrm{Pl}_{m_1})$ and $\mathcal{P}_2 = \mathcal{P}(\mathrm{Pl}_{m_2})$. We say that there is *random set independence* if $\mathcal{P} = \mathcal{P}(\mathrm{Pl}_m)$, where $m = m_1 \odot m_2$, i.e.,

$$m(A \times B) = m_1(A)\, m_2(B),\ \forall\, A \subseteq \Omega_1,\ B \subseteq \Omega_2.$$

3 Probability measures dominated by a plausibility function

In this section we will deal with representations of the class of probability measures dominated by a particular plausibility function. Let Ω represent the (finite) universe of discourse and let $\mathcal{F}_m = \{A_1, \ldots, A_q\}$ be the class of focal sets associated to a basic mass assignment m. Let Pl_m denote the associated plausibility measure. Grabisch et al. ([4]) consider the family of tuples $Z(\mathcal{F}_m) = \{\vec{\alpha} = (\alpha_1, \ldots, \alpha_q) : \alpha_i : A_i \to [0, 1], \sum_{\omega\in A_i} \alpha_i(\omega) = m(A_i),\ i = 1, \ldots, q\}$. For each particular tuple $\vec{\alpha} \in Z(\mathcal{F}_m)$, they consider the associated probability measure $Q_{\vec{\alpha}} : \wp(\Omega) \to [0, 1]$ such that $Q_{\vec{\alpha}}(\{\omega\}) = \sum_{i\,:\,A_i\ni\omega} \alpha_i(\omega),\ \forall\, \omega \in \Omega$. Under this construction, they easily check that each $Q_{\vec{\alpha}}$ is dominated by Pl_m. Furthermore, for each $A \subseteq \Omega$, there exists $\vec{\alpha}^* \in Z(\mathcal{F}_m)$ such that $Q_{\vec{\alpha}^*}(A) = \mathrm{Pl}_m(A)$. Let the reader notice that these conditions do not suffice[1] to prove that the class $\mathcal{J}_m = \{Q_{\vec{\alpha}} : \vec{\alpha} \in Z(\mathcal{F}_m)\}$ coincides with $\mathcal{P}(\mathrm{Pl}_m)$. But, in fact, it does, as we will check at the end of this section.

Fetz independently considers in [3] the class of probability measures

$$\mathcal{K}_m := \left\{ \sum_{i=1}^{q} m(A_i) P^i : P^i \in \mathcal{K}^i \right\}, \text{ where}$$

[1] For instance, the class of extreme points of $\mathcal{P}(\mathrm{Pl}_m)$, $\mathrm{Ext}(\mathcal{P}(\mathrm{Pl}_m))$, satisfies the above conditions, but it does not coincide with the convex set $\mathcal{P}(\mathrm{Pl}_m)$.

$$\mathcal{K}^i = \{P^i \in \mathcal{P}_\Omega, : P^i(A_i) = 1, \ \forall\, i = 1, \ldots, q\}$$

In other words, each probability measure in $\mathcal{K}_m$ is a linear convex combination of q probability measures, $P^1, \ldots, P^q$. Each P^i is a probability measure on the focal A_i.

The family $\mathcal{K}_m$ coincides with $\mathcal{J}_m$. In fact, each tuple $\vec{\alpha} = (\alpha_{A_1}, \ldots, \alpha_{A_q})$ is associated to the tuple of probability measures $(P^1, \ldots, P^q)$ defined as

$$P^i(\{\omega\}) = \frac{\alpha_{A_i}(\omega)}{m(A_i)}, \ \forall\, \omega \in A_i, \ \forall\, i = 1, \ldots, q.$$

We can give an additional alternative description of the class $\mathcal{K}_m$. In fact a joint probability measure, $I\!P : \wp(\wp(\Omega) \times \Omega) \to [0, 1]$, can be associated to each $Q \in \mathcal{K}_m$. Its marginals on $\wp(\Omega)$ and Ω are respectively related to m and Q, as follows:

$$I\!P_1(\{A\}) = m(A) \text{ and } I\!P_2(A) = Q(A), \forall\, A \subseteq \Omega.$$

(In other words, Q coincides with the second marginal probability, $I\!P_2$, while m is the mass function associated to the first marginal probability, $I\!P_1$.) In fact, let us define

$$I\!P(\mathcal{C}) = \sum_{(i,\omega)\,:\,(A_i,\omega)\in\mathcal{C}} \alpha_i(\omega), \ \forall\, \mathcal{C} \subseteq \wp(\Omega) \times \Omega.$$

Remark 1. *For each particular pair (i, ω), the quantity $\alpha_i(\omega)$ represents the mass on the "point" (A_i, ω), i.e. $\alpha_i(\omega) = I\!P(\{(A_i, \omega)\})$.*

On the other hand, each probability P^i in Fetz's construction ([3]) coincides with the conditional probability measure:

$$P^i = I\!P(\,\cdot\,|\{A_i\} \times \Omega), \ \forall\, i = 1, \ldots, q.$$

Furthermore, the second marginal probability measure $Q(A) = I\!P_2(A)$ can be written as the linear convex combination:

$$Q = \sum_{i=1}^{q} m(A_i)\, P^i.$$

Remark 2. *We easily check that $I\!P$ is univocally determined by the pair $(m, (P^i)_{i=1}^q)$, since m represents the first marginal $I\!P_1$ and $(P^i)_{i=1}^q$ represents a family of conditional distributions, as we have checked in last remark. From now on, we will write $I\!P \equiv (m, (P^i)_{i=1}^q)$.*

Next we will show that the family $\mathcal{I}_m = \mathcal{K}_m$ coincides with the class of probability measures dominated by the plausibility measure, $\mathcal{P}(\mathrm{Pl}_m)$.

Theorem 1. *Let $\Omega = \{x_1, \ldots, x_n\}$ be a finite universe and let $m : \wp(\Omega) \to [0, 1]$ a basic mass assignment on it. Let $\mathrm{Pl}_m : \wp(\Omega) \to [0, 1]$ be a plausibility measure associated to m and let $Q : \wp(\Omega) \to [0, 1]$ be a probability measure dominated by Pl_m, $Q \in \mathcal{P}(\mathrm{Pl}_m)$. Then there exists a family of mappings $\{\alpha_A : A \to [0, 1]\}_{A\in\wp(\Omega)}$ such that*

$$m(A) = \sum_{\omega\in A} \alpha_A(\omega), \text{ and}$$

$$Q(\{\omega\}) = \sum_{A\ni\omega} \alpha_A(\omega), \ \forall\, \omega \in \Omega, A \subseteq \Omega.$$

Proof: Let us denote by $\mathcal{F}_m = \{A_1, \ldots, A_q\}$ the family of focal sets associated to m. Let us define the tuple $\vec{\alpha} = (\alpha_{A_1}, \ldots, \alpha_{A_q})$ as follows. For each $i = 1, \ldots, q$, let $\alpha_{A_i} : A_i \to [0, 1]$ be defined as:

$$\alpha_{A_i}(x_j) = \begin{cases} \min\{a_{ij}, b_{ij}\} & \text{if } x_j \in A_i \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where } a_{ij} = Q(\{x_j\}) - \sum_{k=1}^{i-1} \alpha_{A_k}(x_j)$$

$$\text{and } b_{ij} = m(A_i) - \sum_{l=1}^{j-1} \alpha_{A_i}(x_l).$$

On the other hand, let $\alpha_A(x_j) = 0$, $\forall\, j = 1, \ldots, n, A \notin \mathcal{F}_m$. We easily check that the required equalities hold.

Remark 3. *For an arbitrary $Q \in \mathcal{P}(\mathrm{Pl})$, there exists at least one tuple $\vec{\alpha}$ such that $Q = Q_{\vec{\alpha}}$. But this association is not necessarily unique. Let us consider, for instance, the universe $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and the mass assignment $m : \wp(\Omega) \to [0, 1]$ such that $\mathcal{F}_m = \{A_1, A_2\}$ where $A_1 = \{\omega_1, \omega_2\}$, $A_2 = \{\omega_1, \omega_2, \omega_3\}$, and $m(A_1) = 0.5 = m(A_2)$. Let us now consider the probability measure $Q : \wp(\Omega) \to [0, 1]$ such that $Q(\{\omega_1\}) = Q(\{\omega_2\}) = 5/12$ and $Q(\{\omega_3\}) = 1/6$. Let $\vec{\alpha} = (\alpha_{A_1}, \alpha_{A_2})$ and $\vec{\beta} = (\beta_{A_1}, \beta_{A_2})$ the tuples of mappings defined as follows:*

$$\alpha_{A_1}(\omega_1) = \alpha_{A_2}(\omega_2) = 0.25,$$

$$\alpha_{A_2}(\omega_1) = \alpha_{A_2}(\omega_2) = \alpha_{A_2}(\omega_3) = 1/6.$$

$$\beta_{A_1}(\omega_1) = 5/12, \beta_{A_2}(\omega_2) = 1/12,$$

$$\beta_{A_2}(\omega_1) = 0, \beta_{A_2}(\omega_2) = 1/3, \beta_{A_2}(\omega_3) = 1/6.$$

We easily check that

$$m(A) = \sum_{\omega\in A} \alpha_A(\omega) = \sum_{\omega\in A} \beta_A(\omega), \ \forall\, A \text{ and}$$

$$Q(\{\omega\}) = \sum_{A\ni\omega} \alpha_A(\omega) = \sum_{A\ni\omega} \beta_A(\omega), \ \forall\, \omega \in \Omega.$$

4 Independence concepts in evidence theory

The notion of independence in evidence theory is studied from different points of view in the literature. In [8], for instance, the ideas of decomposability and irrelevance are studied and compared within the Theory of Evidence. In this paper, we will distinguish between independence of variables and independence of their observations. First one is related to the concept of "type 1 independence" ([1]) and the second one is associated to "random set independence" [2].

In [3], Fetz establishes three different restrictions to the elements in $\mathcal{P}(\mathrm{Pl}_m)$. Each one of them is related to some aspect of independence. Fetz shows some relationships between these restrictions and some other notions of independence considered in [2]. In this section, we will continue this investigations. First of all, we will recall the notions given by Fetz, but we will use a different nomenclature. For each definition, we will give an intuitive interpretation and an example of of an urn model to which the definition is applied.

4.1 Three conditions of independence

Let $m_1 : \wp(\Omega_1) \to [0,1]$ and $m_2 : \wp(\Omega_2) \to [0,1]$ be two arbitrary basic mass assignments. Let us respectively denote by $\mathcal{F}_{m_1} = \{A_1, \ldots, A_q\}$ and $\mathcal{F}_{m_2} = \{B_1, \ldots, B_r\}$ their families of focal elements. Let us now consider a basic mass assignment on $\Omega_1 \times \Omega_2$, $m : \wp(\Omega_1 \times \Omega_2) \to [0,1]$ satisfying the following conditions:

- The family of focal elements associated to m coincides with (or it is included in) $\mathcal{F}_m = \{A_i \times B_j : i = 1, \ldots, q,\ j = 1, \ldots, r\}$.
- $m_1(A_i) = \sum_{j=1}^{r} m(A_i \times B_j),\ i = 1, \ldots, q.$
- $m_2(B_j) = \sum_{i=1}^{q} m(A_i \times B_j),\ j = 1, \ldots, r.$

Let $P \in \mathcal{P}(\mathrm{Pl}_m)$ and let $I\!P : \wp(\wp(\Omega_1 \times \Omega_2), \Omega_1 \times \Omega_2) \to [0,1]$ be a probability measure satisfying $I\!P_1(\{C\}) = m(C), \forall\, C \in \wp(\Omega_1 \times \Omega_2)$ and $I\!P_2 = P$.

For each pair $(i,j) \in \{1, \ldots, q\} \times \{1, \ldots, r\}$, let $P^{ij} : \wp(\Omega_1 \times \Omega_2) \to [0,1]$ be defined as follows:

$$P^{ij}(C) = I\!P(\wp(\Omega_1 \times \Omega_2) | \{A_i \times B_j\} \times C).$$

P^{ij} is a probability measure on $\Omega_1 \times \Omega_2$ and it satisfies the equality $P^{ij}(A_i \times B_j) = 1$. According to Remark 2, $I\!P$ is univocally determined by the pair $(m, (P^{ij})_{i=1\ j=1}^{q\ r})$ so we can identify them. Furthermore, the probability measure P can be written as

$$P = \sum_{i=1}^{q} \sum_{j=1}^{r} m(A_i \times B_j) P^{ij}.$$

Let us now show three different definitions of independence. They can be applied to probability measures of the form $I\!P \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$ and they are closely related to three restrictions established in [3] to the elements in the class $\mathcal{K}_m$. Each condition reflects a different aspect associated to the notion of independence, as we will check below.

Definition 1. *A probability measure* $I\!P \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$ *satisfies* first independence condition *if* $m = m_1 \odot m_2$, *i.e.*

$$m(A_i \times B_j) = m_1(A_i) \cdot m_2(B_j)$$
$$\forall\, i = 1, \ldots, q,\ j = 1, \ldots, r.$$

This notion is associated to the concept of random set independence recalled in Section 2. Let us illustrate this type of independence.

Example 1. *Suppose that we have two urns, each of them with 10 balls. First urn has five red, two white and three unpainted balls. Second urn has three red, three white and 4 unpainted balls. We select one ball from each urn in a stochastically independent way, and if either one the selected balls are not coloured, then they are painted white or red by a completely unknown procedure. There can be arbitrary correlation between the colours they are finally assigned.*

In this example, we are interested in the colours of the balls we draw from the urns. So, the universe of discourse is $\Omega_1 \times \Omega_2 = \{r, w\} \times \{r, w\}$. *The focal elements associated to both selections are* $\mathcal{F}_{m_1} = \{A_1, A_2, A_3\}$ and $\mathcal{F}_{m_2} = \{B_1, B_2, B_3\}$, where $A_1 = B_1 = \{r\}$, $A_2 = B_2 = \{w\}$ *and* $A_3 = B_3 = \{r, w\}$. *The marginal mass assignments for the colours of the selected balls are:*

$$\begin{array}{lll} m_1(A_1) = 0.5 & m_1(A_2) = 0.2 & m_1(A_3) = 0.3 \\ m_2(B_1) = 0.3 & m_1(B_2) = 0.3 & m_2(B_3) = 0.4 \end{array}$$

The mass assignment associated to the joint experiment satisfies the equalities:

$$m(A_i \times B_j) = m_1(A_i)\, m_2(B_j),\ \forall\, i, j.$$

The class of probability measures representing our (imprecise) information about the joint experiment is $\mathcal{P}(\mathrm{Pl}_m) = \mathcal{K}_m$. *Each one of them is associated to a probability measure* $I\!P$ *satisfying first condition of independence.*

Definition 2. *A probability measure* $I\!P \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$ *is said to satisfy* second independence condition *if* $P^{ij} = P_1^{ij} \otimes P_2^{ij},\ \forall\, i = 1, \ldots, q,\ j = 1, \ldots, r$, *i.e.,*

$$P^{ij}(A \times B) = P_1^{ij}(A) \cdot P_2^{ij}(B),$$
$$\forall\, A \subseteq \Omega_1, B \subseteq \Omega_2,\ \forall\, i = 1, \ldots, q, \forall\, j = 1, \ldots, r.$$

Example 2. *Consider the same urns as in example 1 and assume again that we select one ball from each urn in a stochastically independently way. Let us also assume that, when both selected balls are not painted, there is no correlation between the colours they are assigned. If we have no additional information, our knowledge about the joint experiment is described by the class of probability measures of the form* $P = \sum_{i=1}^{3}\sum_{j=1}^{3} m(A_i \times B_j)P^{ij}$, *where* m *is the mass assignment from Example 1, and* P^{ij} *is a probability measure on* $\Omega_1 \times \Omega_2$ *satisfying:*

- $P^{ij}(A \times B) = P_1^{ij}(A) \times P_2^{ij}(B), \quad \forall A \in \wp(\Omega_1),\ B \in \wp(\Omega_2)$,
- $P^{ij}(A_i \times B_j) = 1$, *for each* $i = 1, 2, 3$ *and each* $j = 1, 2, 3$.

Every probability measure $\mathbb{P} \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$ *associated to this information satisfies first and second independence conditions. As we pointed out above, both balls are selected in a stochastically independent way. Furthermore, when both selected balls have no colour, we use separate procedures to paint them. Nevertheless, there can remain some dependence relation. Let us, for instance assume the following procedure to assign each colour:*

- *If only one of the selected balls is coloured, we will draw a dice to choose the colour of the other one. If the number in the dice is "5", we will paint it with the same colour. Otherwise, we will choose the opposite.*
- *If both selected balls have no colour we will draw two coins, each one for each ball.*

The probability measure, $P : \wp(\Omega_1 \times \Omega_2) \to [0, 1]$, *associated to the joint experiment satisfies both conditions given in definitions 1 and 2. However, it cannot be expressed as a product. In fact, there exists an stochastic dependence between the colours of both balls. Let us notice, for instance, that*

- $P(\{(r, r)\}) = 0.15 + 0.2 \cdot \frac{1}{4} + 0.09 \cdot \frac{1}{6} + 0.12 \cdot \frac{1}{4}$
- $P(\{r\} \times \Omega_2) = 0.5 + 0.09 \cdot \frac{1}{6} + 0.09 \cdot \frac{5}{6} + 0.12 \cdot \frac{1}{2}$, *and*
- $P(\Omega_1 \times \{r\}) = 0.3 + 0.2 \cdot \frac{1}{6} + 0.06 \cdot \frac{5}{6} + 0.12 \cdot \frac{1}{2}$

Thus, $P(\{(r, r)\}) = 0.245$ *does not coincide with* $P(\{r\} \times \Omega_2) \cdot P(\Omega_1 \times \{r\}) = 0.65 \cdot 0.46$.

Definition 3. *A probability measure* $\mathbb{P} \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$ *satisfies* third independence condition *when*

$$P_1^{i1} = \ldots = P_1^{ir} = P_1^{i},\ \forall\, i = 1, \ldots, q \quad \text{and}$$

$$P_2^{1j} = \ldots = P_1^{qj} = P_2^{j},\ \forall\, j = 1, \ldots, r.$$

Example 3. *Suppose again we have the urns in example 1. Let us draw a ball from each urn. If some of the balls is uncoloured, we decide its colour without checking whether the other one is red, white or uncoloured. Nevertheless, there can be some dependence relationship between both colours. Let us, for instance, consider the following procedure to assign each colour:*

- *If only one of the balls is coloured, we will toss a dice. If the number in the dice is "5", we will paint it red. Otherwise, we will paint it white.*
- *If both balls are uncoloured, we will toss the same dice to decide their colour. If the number in the dice is 5, we will paint both of them red. Otherwise, we will paint them white.*

The probability measure, $\mathbb{P} \equiv (m, (P^{ij})_{i=1\ j=1}^{q\ r})$, *associated to the joint experiment satisfies the conditions given in definitions 1 and 3. Nevertheless, the probability measure that models the joint experiment (the probability measure* $Q = \sum_{i=1}^{3}\sum_{i=1}^{3} m(A_i \times B_j)P^{ij}$*) cannot be written as the product of its marginals. For instance, the probability of the result* (r, r) *is, approximately, 0.22. On the other hand* $Q(\{r\} \times \Omega_2) = 0.55$ *and* $Q(\Omega_1 \times \{r\}) \approx 0.37$. *Hence,* $Q(\{(r, r)\})$ *does not coincide with the product* $Q(\{r\} \times \Omega_2) \cdot Q(\Omega_1 \times \{r\})$.

Summarizing, each condition reflects a different aspect of the notion of independence. First condition (random set independence) reflects independence between the procedures used to select both balls from the urns. In last examples, this condition is satisfied, because each ball is selected from a different urn, in a stochastically independent way. Second condition reflects independence between the procedures to paint both balls, once they have been selected. Finally third condition reflects independence between the procedure used to select one ball from a urn and the procedure used to paint the other ball, once it has been selected.

In examples 1, 2 and 3 we show situations where some, but not all of these conditions are satisfied, and $P = \mathbb{P}_2$ cannot be written as a product. If $\mathbb{P} = (m, (P^{ij})_{i=1\ j=1}^{q\ r})$, satisfies conditions 1 to 3 then the probability measure $P = \mathbb{P}_2 = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \times B_j)P^{ij}$ can be factorized as $P = P_1 \otimes P_2$, as Fetz checks in [3]. Conversely, we easily check that every

product probability $P = P_1 \otimes P_2$ where $P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1})$ and $P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})$ can be written as $P = I\!P_2 = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \times B_j)P^{ij}$, where $I\!P$ satisfies conditions given in Definitions 1, 2 and 3. In next section we will make a further study about the connection between conditions 1 to 3 and independence in the selection.

4.2 Independence in the selection

As we pointed out in last subsection, any probability measure $P = P_1 \otimes P_2$ with $P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1})$, $P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})$ is associated to a probability measure $I\!P$ satisfying independence conditions given in last section. In other words, it can be written as a linear convex combination $P = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \times B_j)P^{ij}$, where $m = m_1 \odot m_2$ and $P^{ij} = P_1^i \otimes P_2^j,\ \forall\, i = 1 = 1, \ldots, q,\ j = 1, \ldots, r$. On the other hand, we can use different linear convex combinations and get the same probability measure, as we have checked in Remark 3. So we can ask ourselves whether we can find an alternative linear convex combination

$$P = \sum_{i=1}^{q}\sum_{j=1}^{r} m'(A_i \times B_j)Q^{ij},$$

where $I\!P \equiv (m', \{Q^{ij}\}_{i=1}^{q}{}_{j=1}^{r})$ does not satisfy the requirements considered in definitions 1, 2 and 3. In fact, it is possible, as we show below.

Example 4. *Suppose we have two urns, each one with 10 balls. The two of them have five red, and five unpainted balls. We select one ball from the first urn and then we select a similar ball (red or uncoloured) from the second urn. (There is stochastic dependence between both selections.) Once we have selected both balls, we use the following procedure to paint them in case they are uncoloured: we toss three coins, and check the number of heads:*

- *If the number is 3, we paint both balls with the colour red.*
- *If the number of heads is 2, we paint the first ball red, and the second one, white.*
- *If the number of heads is 1, we paint the first ball white, and the second one, red.*
- *Finally, if three tails are obtained, we paint white both of them.*

The probability measure that models this random experiment can be written as:

$$P = m(A_1 \times B_1)\, P^{11} + m(A_2 \times B_2)\, P^{22},$$

where $A_1 = B_1 = \{r\},\ A_2 = B_2 = \{r, w\}$,

$$m(A_1 \times B_1) = m(A_2 \times B_2) = 0.5 \text{ and}$$

$$P^{11} \equiv (1, 0, 0, 0) \text{ and } P^{22} \equiv (1/8, 3/8, 3/8, 1/8).$$

There does not exist m_1 and m_2 such that $m = m_1 \odot m_2$. On the other hand, each P^{ij} cannot be factorized as $P^{ij} = P_1^i \otimes P_2^j$. In other words, m and $\{P^{ij}\}_{i=1}^{2}{}_{j=1}^{2}$ do not satisfy the requirements from definitions 1 and 2. (It has no sense to check condition 3, since P_1^{12}, P_2^{12}, P_1^{21} and P_2^{22} can be arbitrarily defined.) Nevertheless, P coincides with the product of its marginals. In fact, $P(\{(r, r)\}) = 9/16$, $P(\{(r, w)\}) = P(\{(w, r)\}) = 3/16$, and $P(\{(w, w)\}) = 1/16$, and hence $P(A \times B) = P_1(A)\, P_2(B),\ \forall\, A, B \subseteq \{r, w\}$.

Since the probability measure that models last experiment can be written as a product, there must exists an alternative linear convex combination,

$$P = \sum_{i=1}^{2}\sum_{j=1}^{2} m_1(A_i)m_2(B_j)Q^{ij}, \tag{2}$$

where $Q^{ij} = Q_1^i \otimes Q_2^j, \forall\, i, j$. In fact, last experiment is equivalent to the following one: suppose we have two urns, each one with 10 balls. The two of them have five red, and five unpainted balls. We select one ball from each urn in a stochastically independent way. If some of the balls is uncoloured, we toss a coin to decide its colour (one coin for each ball). The probability measure associated to this new random experiment coincides with P and it can be written, in a natural way as in equation 2, where: $m_1(A_1) = m_1(A_2) = m_2(B_1) = m_2(B_2) = 0.5$, $Q_k^i(\{r\} = Q_k^i(\{w\}) = 0.5, i = 1, 2,\ k = 1, 2$.

In last example, we have built a product probability measure $P = P_1 \otimes P_2$ without having into account any of the requirements given in definitions 1 to 3. We can also get a product probability by using some or these rules, but not all of them. In next example, we will only take into account the requirement from definition 1, and we will get a product probability measure.

Example 5. *Consider a urn with 10 balls. Five of them are red, and the other five are unpainted. Suppose that a ball is drawn at random from the urn and replaced, and then a second ball is drawn at random, and the two drawings are stochastically independent. Once both balls are selected from the urn, we consider the following procedure to paint them:*

- *If both balls are red, we do not need to do anything.*
- *If the first ball is red and the second one is uncoloured, we paint it red with probability 5/8 and white, with probability 3/8.*

- *If the second ball is red and the first one is uncoloured, then we paint it red with probability 1/2 (and white, with the same probability).*
- *Finally, if both balls are unpainted, we assign them the pairs of colors (red, red), (red, white), (white, red), (white, white) with respective probabilities (1/8, 3/8, 1/4, 1/4).*

The probability measure, P, that models the joint experiment can be written as

$$P = \sum_{i=1}^{2}\sum_{j=1}^{2} m(A_i \times B_j)\, P^{ij}, \text{ where}$$

$$A_1 = B_1 = \{r\},\ A_2 = B_2 = \{r, w\},$$

$$m(A_1 \times B_1) = m(A_1 \times B_2) = m(A_2 \times B_1) =$$

$$m(A_2 \times B_2) = 0.25 \text{ and}$$

$$P^{11} \equiv (1,0,0,0) \quad P^{12} \equiv (\tfrac{5}{8}, \tfrac{3}{8}, 0, 0)$$
$$P^{21} \equiv (\tfrac{1}{2}, 0, \tfrac{1}{2}, 0) \quad P^{22} \equiv (\tfrac{1}{8}, \tfrac{3}{8}, \tfrac{1}{4}, \tfrac{1}{4}).$$

The probability measure $\mathbb{P} \equiv (m, (P^{ij})_{i=1}^{2}{}_{j=1}^{2})$ satisfies first condition of independence, but it does not satisfy the second and the third ones. On the other hand, the probability measure $P = \sum_{i=1}^{2}\sum_{i=1}^{2} m(A_i \times B_j)\, P^{ij}$ can be identified with the tuple

$$P \equiv \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right),$$

so it can be factorized as

$$P = P_1 \otimes P_2 \equiv (3/4, 1/4) \otimes (3/4, 1/4).$$

We can also build some $\mathbb{P}$ satisfying the requirements from definitions 2 and 3, but not the property from definition 1, and such the probability measure $P = \mathbb{P}_2$ can be written as the product of its marginals. Let us show it in next example:

Example 6. *Suppose that we have three urns. First one has 3 balls: one white, one red and one uncoloured. Second urn has two balls: one red and one white. Third urn has two unpainted balls. We select one ball from the first urn. If it is coloured, we select another ball from second urn. If, otherwise, it is uncoloured, we select a ball from the second urn. Once the balls have been selected, we drop two coins to decide their colour (if they are uncoloured), one coin for each ball.*

The probability measure that models this experiment can be written as:

$$P = \sum_{i=1}^{3}\sum_{j=1}^{3} m(A_i \times B_j) P_1^i \otimes P_2^j, \text{ where}$$

$$A_1 = B_1 = \{r\}, A_2 = B_2 = \{w\}, A_3 = B_3 = \{r, w\},$$

the mass assignment m is determined by:

	B_1	B_2	B_3
A_1	*1/6*	1/6	0
A_2	*1/6*	*1/6*	*0*
A_3	*0*	*0*	*1/3*

and the marginal probability measures defined on each focal are:

$$P_1^1 \equiv (1,0) \quad P_1^2 \equiv (0,1) \quad P_1^3 \equiv (0.5, 0.5)$$
$$P_2^1 \equiv (1,0) \quad P_2^2 \equiv (0,1) \quad P_2^3 \equiv (0.5, 0.5)$$

The mass assignment m cannot be written as the product of its marginals, i.e., $m \neq m_1 \odot m_2$. So, $\mathbb{P} = (m, \{P^{ij}\}_{i=1}^{3}{}_{j=1}^{3})$ does not satisfy the condition described in definition 1. But it satisfies the conditions described in definitions 2 and 3. (There is independence inside the focal elements, but not between focals.) On the other hand, we easily check that $P(\{(r,r)\}) = P(\{(r,w)\}) = P(\{(w,r)\}) = P(\{(w,w)\}) = 0.25$. So P can be factorized as the product of its marginals. In fact:

$$P \equiv (0.25, 0.25, 0.25, 0.25) =$$

$$(0.5, 0.5) \otimes (0.5, 0.5) = P_1 \otimes P_2.$$

4.3 Random set independence and independence in the selection

Let $m_1 : \wp(\Omega_1) \to [0,1]$, $m_2 : \wp(\Omega_2) \to [0,1]$ two arbitrary mass assignments and let $m : \wp(\Omega_1 \times \Omega_2) \to [0,1]$ satisfy $m(A \times \Omega_2) = m_1(A)$, $m(\Omega_1 \times B) = m_2(B), \forall\, A \subseteq \Omega_1, B \subseteq \Omega_2$. As we have pointed out in Section 4.1, the class of probability measures $P = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \times B_j) P^{ij}$, where $\mathbb{P} = (m, (P^{ij})_{i=1}^{q}{}_{j=1}^{r})$ safisties the three conditions considered in last definitions coincides with the family of product probability measures:

$$\{P_1 \otimes P_2 \,:\, P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1}), P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})\}.$$

On the other hand, we easily check that the class of probability measures $P = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \times B_j) P^{ij}$ where $\mathbb{P} = (m, P^{ij})_{i=1}^{q}{}_{j=1}^{r})$ satisties the first condition coincides with $\mathcal{P}(\mathrm{Pl}_{m_1 \odot m_2})$. Thus, the following inclusion holds:

$$\{P_1 \otimes P_2 \,:\, P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1}), P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})\}$$

$$\subseteq \mathcal{P}(\mathrm{Pl}_{m_1 \odot m_2}) \tag{3}$$

The left hand side is associated to type 1 independence. The right hand side is related to random set independence. We may ask ourselves whether the inclusion in equation 3 is strict or not, for any pair of mass asignments m_1, m_2. Let us notice that the probability measure $I\!P \equiv (m, (P^{ij})_{i=1}^{q}{}_{j=1}^{r})$ in example 5 satisfies the first condition of independence, but it does not satisfy the second and the third ones. Nevertheless, the probability measure $P = I\!P_2 = \sum_{i=1}^{q}\sum_{j=1}^{r} m(A_i \otimes B_j)P^{ij}$ can be factorized as $P = P_1 \otimes P_2$, and hence it belongs to the class $\{P_1 \otimes P_2 \,:\, P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1}), P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})\}$. So, we ask ourselves

> Does there exists some pair m_1, m_2 such that any
>
> $$P = \sum_{i=1}^{q}\sum_{j=1}^{r} m_1(A_i)\, m_2(B_j) P^{ij}$$
>
> can be written as the product of its marginals, $P = P_1 \otimes P_2$?

The answer is "no", except for the cases where m_1 and m_2 represent trivial situations. Let us show the following result:

Theorem 2. *Let us consider two finite universes Ω_1 and Ω_2 and two arbitrary mass assignments $m_1 : \wp(\Omega_1) \to [0,1]$ and $m_2 : \wp(\Omega_2) \to [0,1]$. Let m be the "product mass assignment", i.e. $m : \wp(\Omega_1 \times \Omega_2) \to [0,1]$ such that $m(A \times B) = m_1(A) \cdot m_2(B),\ \forall\, A, B$. Let us assume that $\mathcal{P}(\mathrm{Pl}_m)$ coincides with the family:*

$$\{P_1 \otimes P_2 \,:\, P_1 \in \mathcal{P}(\mathrm{Pl}_{m_1}),\ P_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})\}.$$

Then, some of the following conditions holds:

- Pl_{m_1} *and* Pl_{m_2} *are probability measures (they are additive).*
- Pl_{m_1} *or* Pl_{m_2} *is a degenerate probability measure (I.e., al least one of the families $\mathcal{F}_{m_1}$ or $\mathcal{F}_{m_2}$ has only one focal with only one element.)*

Proof: (Sketch) Let us assume that Pl_{m_2} is not a degenerate probability measure. Then there exists $B \subseteq \Omega_2$ and $Q_2 \in \mathcal{P}(\mathrm{Pl}_{m_2})$ such that $Q_2(B) \in (0,1)$. Let A be an arbitrary subset of Ω_1 and let $P_1, Q_1 \in \mathcal{P}(\mathrm{Pl}_{m_1})$ such that $P_1(A) = \mathrm{Pl}_{m_1}(A)$ and $Q_1(A) = \mathrm{Bel}_{m_1}(A)$. (The existence of such P_1, Q_1 and Q_2 is easily checked.) Let $\vec{\alpha}$, $\vec{\alpha}'$ and $\vec{\beta}$ be respectively associated to each one of them. Let $\vec{\gamma} = (\gamma_{ij})_{i=1}^{q}{}_{j=1}^{r}$ be defined as $\gamma_{ij}(x,y) = \alpha_i(x)\beta_j(y)\, I_B(y) + \alpha'_i(x)\beta_j(y)\, I_{B^c}(y)$. We can check that $\vec{\gamma}$ represents a probability measure, R, on $\Omega_1 \times \Omega_2$ such that (a) $R \in \mathcal{P}(\mathrm{Pl}_m)$, (b) $R_2 = Q_2$, $R_2(A \times B) = P_1(A)Q_2(B)$ and (c) $R_2(A \times B^c) = Q_1(A)Q_2(B)$. We easily derive that $\mathrm{Pl}_{m_1}(A) = P_1(A) = Q_1(A) = \mathrm{Bel}_{m_1}(A)$. Since A is an arbitrary set, we conclude that Pl_{m_1} is a additive.

5 Conclusion and open problems

We have considered three rules to build probability measures on product spaces in Evidence Theory framework. Each one of them reflects a particular aspect of independence, as we illustrate in Examples 1, 2 and 3. They are simple examples about drawing pairs of balls from urns. As we show there, first condition reflects that the selections of both balls are independent. Second condition means that there is independence between the procedures of painting the balls, for a particular selection of a pair of balls. Finally, third condition reflects independence between the selection of a ball and the procedure used to choose the colour to paint the other ball.

In a more general and applied context, first condition is related to the idea of independence between mechanisms of observation of variables. If we add second and third conditions, independence between the actual variables holds. But, as we have checked in Examples 4, 5 and 6, none of these conditions is strictly necessary to guarantee this independence. When there is no imprecision in the observations, second and third conditions do not apply (they are trivially satisfied when the focals are singletons). In that case, independence between the variables and between their observations are the same (perception and reality do coincide). But when imprecision appears, there is no an implication relationship between independence of the observations and independence of the variables.

All these ideas can be extended to non finite universes. In the general context, pairs of upper and lower probabilities associated to multi-valued mappings play the role of pairs of plausibility-belief functions. Furthermore, the probability measures induced by the selections of the multi-valued mapping are dominated by its upper probability. So, in the general context, the mass assignment $m : \wp(\Omega_1 \times \Omega_2) \to [0,1]$ will be replaced by a multi-valued mapping $\Gamma = \Gamma_1 \times \Gamma_2 : \Lambda \to \wp(\Omega_1 \times \Omega_2)$, such that $\Gamma(\lambda) = \Gamma_1(\lambda) \times \Gamma_2(\lambda)$. (The images of the multi-valued mapping play the role of the focal sets of the basic mass assignment.) Furthermore, each probability measure on $\Omega_1 \times \Omega_2$ induced by a selection (X_1, X_2) is dominated by the upper probability of Γ. Hence, the finite tuple of probability measures $(P^{ij})_{i=1}^{q}{}_{j=1}^{r}$ will be replaced by the conditional distribution of (X_1, X_2) given Γ. In this new setting, we will say that first condition of independence is sat-

isfied when Γ_1 and Γ_2 are stochastically independent (random set independence). Second condition will be satisfied when X_1 and X_2 are conditionally independent, given Γ. Finally, third condition will be satisfied when X_1 and Γ_2 are conditionally independent given Γ_1 and X_2 and Γ_1 are conditionally independent given Γ_2. In this general context, there is independence in the selection when X_1 and X_2 are stochastically independent. We intuitively observe that when the three conditions are satisfied, then X_1 are X_2 are stochastically independent. But the converse is not true. Furthermore, there is no implication relationship between the independence of Γ_1 and Γ_2 (random set independence) and the independence between X_1 and X_2 (independence in the selection), as it happens in the finite case.

Acknowledgements

I wish to thank the referees for their helpful comments and suggestions. This work has been supported by grant MTM2004-01269.

References

[1] L.M. de Campos, S. Moral, Independence concepts for convex sets of probabilities. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Ph. Besnard, S. Hanks (eds.) Morgan Kaufmann (San Mateo), 108-115.

[2] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities, Risk Decision and Policy **5** 165–181, 2000.

[3] T. Fetz, Sets of joint probability measures generated by weighted marginal focal sets Thomas Fetz. Proceedings of ISIPTA'01, 171–178, Ithaca, NY (USA), 2001.

[4] M. Grabisch, H.T. Nguyen, E.A. Walker, Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference, Kluwer Academic Publishers, 1995.

[5] G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976.

[6] P. Smets, What is Dempster-Shafer's model? Advances in the Dempster-Shafer Theory of Evidence, R.R. Yager, M. Fedrizzi and J. Kacprzyk (eds.), Wiley (1994) 5-34.

[7] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, 1991.

[8] B. Ben Yaghlane, P. Smets and K. Mellouli, Belief function independence: I. The marginal case, Int. J. of Approximate Reasoning **31** 31-75, 2002.

On various definitions of the variance of a fuzzy random variable

Inés Couso
Dep. Statistics and O.R.
Univ. of Oviedo, Spain
couso@uniovi.es

Didier Dubois
IRIT-CNRS
Univ. of Toulouse, France
dubois@irit.fr

Susana Montes
Dep. Statistics and O.R.
Univ. of Oviedo, Spain
montes@uniovi.es

Luciano Sánchez
Dep. Computer Sciences
Univ. of Oviedo, Spain
luciano@uniovi.es

Abstract

According to the current literature, there are two different approaches to the definition of the variance of a fuzzy random variable. In the first one, the variance is defined as a fuzzy interval, offering a gradual description of our incomplete knowledge about the variance of an underlying, imprecisely observed, classical random variable. In the second case, the variance of the fuzzy random variable is defined as a crisp number, that makes it easier to handle in further processing. In this work, we introduce yet another definition of the variance of a fuzzy random variable, in the context of the theory of imprecise probabilities. The new variance is not defined as a fuzzy or crisp number, but it is a real interval, which is a compromise between both previous definitions. Our main objectives are twofold: first, we show the interpretation of the new variance and, second, with the help of simple examples, we demonstrate the usefulness of all these definitions when applied to particular situations.

Keywords: Fuzzy random variable, random set, variance, second order possibility measure.

1 Introduction

The concept of fuzzy random variable, that extends the classical definition of random variable, was introduced by Féron [14] in 1976, and modified by other authors like Kwakernaak [22], Puri and Ralescu [31], Kruse and Meyer [21], or Diamond and Kloeden [9], among others. In [18], Krätschmer surveys all of these definitions and proposes an unified approach. In all of these works, a fuzzy random variable is defined as a function that assigns a fuzzy subset to each possible output of a random experiment. The different definitions in the literature disagree on the measurability conditions imposed to this mapping, and in the properties of the output space, but all of them intend to model situations that combine fuzziness and randomness. Since the introduction of this concept, many works have generalized different probabilistic concepts and classical results to the case in which all observations associated to the different results of the experiment are fuzzy sets.

Regarding the generalizations, in this context, of definitions of parameters associated to a probability distribution, we can divide them into two groups. On the one hand, some parameters have been defined as fuzzy values: the expectation [31], the distribution function in a point [5, 21], the variance [20] and the covariance[1] [26]. On the other hand, the expectation [23], the variance [13, 17, 24], the covariance[13] or the inequality index [25] have also been defined as crisp values. The introduction of these last definitions is guided by the interest of the authors in solving decision problems involving parameters with numerical, not fuzzy values.

In spite of the great amount of studies about fuzzy random variables, there are few works that study the different interpretations that could be given to their various definitions. The same can be said about the new concepts arising from them (for instance, some of the mentioned parameters.) It is well known that fuzzy sets admit of many different meanings (see, for example [12]) and each one of these meanings could lead to an interpretation of the concept of fuzzy random variable.

In this work, we shall observe that there are different extensions of the concept of variance to fuzzy random variables. We shall review different definitions of variance, found in the literature, and we shall propose an additional definition, that could be cast in a model of imprecise probabilities. We pay attention to the interpretation of each definition. Guided by simple examples, we shall observe the advantages and drawbacks of each definition in different contexts.

[1] We must remark that the concept of fuzzy random variable not only extends the concept of one-dimensional random variable, but also of random n-dimensional vector.

2 Fuzzy random variables

It was mentioned in the introduction that a fuzzy random variable is a function that assigns a fuzzy subset to each outcome of a random experiment. The different definitions of fuzzy random variable differ in the measurability conditions imposed to the random variables.

Kwakernaak [22] and Puri and Ralescu [31] rely on the α-cut mappings (the multivalued functions that map each element in the initial probability space to the respective α-cuts of the fuzzy set-valued image) to translate such condition. While Kwakernaak restricts himself to images that are fuzzy subsets of $\mathbb{R}$ and the boundaries of the α-cuts are measurable functions, Puri and Ralescu impose that the graph of the images itself be measurable (i.e. lies in the product σ-algebra.) On the other hand, Klement et al. [16] and Diamond and Kloeden [9] consider different metrics over the class of fuzzy sets of the output space and impose that the function is measurable with respect to the Borel σ-algebra induced by the corresponding metric. Krätschmer [18] reviews all the previous concepts and offers a unified vision when he considers a certain topology, defined over the class of fuzzy subsets of $\mathbb{R}^n$, with non empty compact α-cuts. In this work, we shall not deal with formal aspects of each particular definition, but with the interpretation of the various concepts of fuzzy random variable.

Fuzzy sets have been given different interpretations [12], therefore a fuzzy random variable admits of various meanings as well. In the remaining part of this section, we briefly review two existing interpretations of fuzzy random variables, and introduce a new one. For every interpretation, we shall describe the information provided by the fuzzy random variable by means of a specific underlying model, namely, a classical probability model, an order 2 imprecise probability model and an order 1 imprecise probability model, respectively.

2.1 Linguistic random variables

In [31], Puri and Ralescu consider that the observations of some random experiments do not consist of numerical outputs, but are represented by vague linguistic terms. According to this idea, some authors consider that a fuzzy random variable is a measurable function, in the classical sense, between a certain σ-algebra of events in the original space and a σ-algebra defined over a class of fuzzy subsets of $\mathbb{R}$. In this context, the probability distribution induced by the fuzzy random variable can be used to summarize the probabilistic information that the variable provides. If the fuzzy random variable has a finite number of images forming a linguistic term set, probability values can be assigned to the different linguistic labels. For example, the following model could be generated: the result is "high" with probability 0.5, "medium" with probability 0.25 and "low" with probability 0.25, where "high", "medium" and "low" are linguistic labels associated to fuzzy subsets of $\mathbb{R}$.

2.2 Ill-known classical random variables

On the contrary, Kruse and Meyer [21] choose a possibilistic interpretation of fuzzy sets. Each fuzzy set is viewed as modeling incomplete knowledge about an otherwise precise value. These authors then claim that the fuzzy random variable represents imprecise or vague knowledge about a *classical* random variable, $X_0 : \Omega \to \mathbb{R}$, they refer to as the "original random variable." Therefore, the membership degree of a point x to the fuzzy set $\tilde{X}(\omega)$ represents the possibility degree of the assertion "$X_0(\omega)$ is x", i.e., the image of element ω coincides with x. This way, the authors get all the elements needed to define a possibility measure over the set of all random variables. They define the "acceptability degree" of each random variable, $X : \Omega \to \mathbb{R}$, as the value: $\mathrm{acc}(X) = \inf_{\omega\in\Omega} \tilde{X}(\omega)(X(\omega))$. The function "acc" takes values in the unity interval. Therefore, it can be regarded as the possibility distribution associated to a possibility measure, $\Pi_{\tilde{X}}$, defined over the set of all random variables. $\mathrm{acc}(X)$ represents the possibility degree of X being the "true" random variable that models the studied experiment. If the fuzzy random variable were a random set (its images are crisp subsets of $\mathbb{R}$,) the acceptability function would assign the value 1 to random variables in a certain set, and the value 0 to the remaining ones. In the particular case when the fuzzy random variable is a classical random variable (all images are sets with only one element) the acceptability function would assign the value 1 to only one random variable, which is the true random variable that models the experiment. In this case, its observation is completely precise.

Under this framework, we can build (see [6]) a possibility measure over the set of all the probability distributions in $\mathbb{R}$. The possibility distribution, $\pi_{P_{\tilde{X}}}$, that characterizes such possibility measure is defined as follows:

$$\pi_{P_{\tilde{X}}}(Q) = \sup\{\mathrm{acc}(X) \mid P_X = Q\} =$$

$$\Pi_{\tilde{X}}(\{X : \Omega \to \mathbb{R} \text{ measurable} \mid P_X = Q\}).$$

$\pi_{P_{\tilde{X}}}(Q)$ represents the degree of possibility that the original random variable is one of those that induce the probability distribution Q in $\mathbb{R}$. The possibility measure $\Pi_{P_{\tilde{X}}}$ is a "second-order possibility" formally

equivalent to those considered in [8]. It is so called, because it is a possibility distribution defined over a set of probability measures.

A possibility measure on a set represents the same information as a family of probability measures on this set (the family of probability measures that are dominated by the possibility measure and dominate the dual necessity measure [11].) Therefore, a second-order possibility measure is associated to a set of (meta-) probability measures, each of them defined, in turn, over a set of probability measures. Thus, a second-order possibility would allow us to state assertions like "the subjective probability that the true probability of the value 7 is 0.5 is between 0. and 0.7."

2.3 Known random process with imprecisely perceived output

Also in accordance with the possibilistic interpretation of fuzzy sets, in this work we are going to proceed in a slightly different way, in order to describe the information provided by $\tilde{X}$. We follow the path started in [30] for the particular case of the random sets and continued in [1] and [4] for fuzzy random variables. Suppose we have partial information about the probability distribution that models a sequence of two random experiments whose sample spaces are Ω and $I\!R$, respectively. For instance, the first one describes some random phenomenon of interest and the second one accounts for a measurement process applied to outcomes of the first one. Let us suppose, on the one hand, that the probability distribution that models the first one, $P : \mathcal{A} \to [0, 1]$, is completely determined (in the preceding expression, $\mathcal{A}$ denotes a σ-algebra of events over Ω.) On the other hand, the other experiment is only known via a family of conditional possibility measures $\{\Pi(\cdot \mid \omega)\}_{\omega\in\Omega}$, each of them inducing the fuzzy set $\tilde{X}(\omega)$. This family of possibility measures models our knowledge about the relationship between the outcome of the first subexperiment an the possible outcomes of the second one. (If the result of the first experiment is ω, then the possibility degree of x occurring in the second one is $\tilde{X}(\omega)(x)$.) In other words, we know the probability measure that drives the primary random process but the measurement process of outcomes is tainted with uncertainty.

The combination, using natural extension techniques [32] of both sources of information, allows to describe the available information about the probability distribution on $\beta_{I\!R}$ (the probability distribution that rules the second sub-experiment) by means of an upper probability (a standard imprecise probability model, not an order-2 model, like the one described before.) Let the reader notice that the conditional possibility measure is coherent,under fairly general conditions on $\tilde{X}$. This way, we are able to state assertions like the following: "the probability of observing an outcome between 3 and 7 lies between 0.3 and 0.6."

3 Several definitions of variance

Each of the three models described in the preceding section leads a different understanding of the variance. In this section we consider the different definitions, according to each model, and emphasize their usefulness in different contexts. We shall restrict ourselves to the case where the images of the fuzzy random variable are fuzzy subsets of $I\!R$. In the last section of this work, we shall make some considerations about the generalization to the multi-dimensional case.

3.1 Classical model

Let us consider a probability space, $(\Omega, \mathcal{A}, P)$, and a metric, d, defined over the class of the fuzzy subsets of $I\!R$, $\tilde{\mathcal{P}}(I\!R)$, (or over a subclass) and let us suppose that $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(I\!R)$ is a function $\mathcal{A}-\beta(d)$-measurable (here, $\beta(d)$ represents the Borel σ-algebra induced by d.)

Definition 1 *We call* classical variance *of* $\tilde{X}$ *the quantity*

$$\mathrm{Var}_{\mathrm{Cl}}(\tilde{X}) = \int_\Omega d(X, E(\tilde{X}))^2 \, dP.$$

The different definitions of variance in the literature that fit this formulation differ in the used metric and in the definition of the expectation of a fuzzy random variable. With respect to this, we briefly comment some details about the definitions of Körner [17] and Lubiano et al. [24]. On the one hand, Körner considers Fréchet's definition of expectation [15] for measurable functions taking values in a metric space. It is noticeable that Fréchet defines the expectation of a measurable function Z, with values in a metric space (M, d) as a solution $a = E^{(d)}(Z)$, (not necessarily unique) of the problem $\min_{a\in M} E[d(Z, a))^2]$. Körner [17] checks that Puri and Ralescu's expectation [31] is the only Fréchet expectation for a certain family of metrics defined over the class of compact and normal fuzzy sets of $I\!R$, which they generically denote ρ_2. According to this, given a distance ρ_2, the variance of a fuzzy random variable $\tilde{X}$ is the amount

$$\mathrm{Var}_{\rho_2}(\tilde{X}) = \int_\Omega \rho_2(\tilde{X}, E_{PR}(\tilde{X}))^2 \, dP. \qquad (1)$$

With respect to the family of variances defined by Lubiano *et al.* in [24], the considered expectation is

also that of Puri-Ralescu, $E_{PR}(\tilde{X})$, and the class of distances is that defined by Bertoluzza et al. in [3], which in turn is a subclass of the family defined by Körner. In [17] and [24] we can find some interesting properties of the families of variances defined there. In this work we only comment some particular aspects of those, to show some of their advantages and also some drawbacks, if compared to other definitions of variance. Even though these definitions are stated for general fuzzy random variables in [17] and [24], in this work , it is sufficient to use their formulation in the particular case when $\tilde{X}$ is a multi-valued mapping (a function whose images are "crisp" subsets of the final space.)

In this case, we can easily check that the definitions of Körner and Lubiano et al. are of the form:

$$\mathrm{Var}(\tilde{X}) = \pi_1 \mathrm{Var}(X_1) + \pi_2 \mathrm{Cov}(X_1, X_2) + \pi_3 \mathrm{Var}(X_2),$$

where $\pi_1 = \lambda_1 + 0.25\lambda_2$, $\pi_2 = 0.5\lambda_2$, $\pi_3 = \lambda_3 + 0.25\lambda_2$, $\lambda_i \geq 0$, $i = 1, 2, 3$, $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and X_1, X_2 are the random variables defined over Ω as $X_1(\omega) = \inf \tilde{X}(\omega)$ and $X_2(\omega) = \sup \tilde{X}(\omega)$, $\forall\, \omega \in \Omega$, respectively. (Under the measurability conditions imposed to $\tilde{X}$ by the authors, the functions X_1 and X_2 are $\mathcal{A} - \beta_{I\!R}$ measurable.) Therefore, in the particular case in which π_2 is null (thus $\pi_3 = 1 - \pi_1$), the variance of the random set will be a convex linear combination of the variances of their boundaries[2]. Additionally, if $\pi_3 = (1 - \sqrt{\pi_1})^2$, the variance of $\tilde{X}$ coindices with the variance of the convex linear combination of X_1 and X_2 given by the expression $\sqrt{\pi_1} X_1 + (1 - \sqrt{\pi_1}) X_2$. In other words, in this case, for every element in the sample space, ω, we can choose a representative point, $\alpha X_1(\omega) + (1 - \alpha) X_2(\omega)$ (with $\alpha \in [0, 1]$), of the image of the fuzzy random variable, and then calculate the variance of the classical random variable that results. The idea of computing the scalar variance using a representative substitute point to each fuzzy observation is used by Baudrit et al. [2], as one piece of information to be extracted from the hybrid propagation of fuzzy and probabilistic information through a mathematical model. With these examples, we observe that families of variances so defined allow us to quantify the dispersion of the (fuzzy, or set-valued) images of $\tilde{X}$, regarded as a measurable function from a classical point of view, and that can be useful when the images of $\tilde{X}$ are linguistic labels. In the context of a linguistic variable, $Var(\tilde{X})$ thus evaluates the variation across the possible linguistic labels.

The following example illustrates the shortcomings of this "classical" variance when quantifying the information available about the variance of an underlying random variable, when a "possibilistic" view of the fuzzy random variables is used, instead of the above setting.

Example 1 *Let us consider first a unitary sample space (that models a deterministic experiment,) $\Omega_1 = \{\omega_1\}$, a probabilistic space with the only σ-algebra that can be defined over it, $\mathcal{A}_1 = \mathcal{P}(\Omega_1)$ and the only probability measure P_1 that is possible. Let us define a random set $\Gamma_1 : \Omega_1 \to \mathcal{P}(I\!R)$, as $\Gamma_1(\omega_1) = [-K, K]$. In this example, Γ_1 is instrumental to represent the output (imprecisely known) of a deterministic experiment. For example, it represents the amount of money that, with absolute confidence, we shall receive, if we only know that it lies between $-K$ and K.*

Then, let us also consider another probabilistic space $(\Omega_2, \mathcal{A}_2, P_2)$ that corresponds to the outcome of tossing a fair coin ($\Omega_2 = \{h, t\}$), and the random set $\Gamma_2 : \Omega_2 \to \mathcal{P}(I\!R)$ defined as $\Gamma_2(h) = \Gamma_2(t) = [-K, K]$). In turn, Γ_2 can be used to represent our gain after tossing a fair coin: the amount we are going to receive depends on the coin. The two outcomes are fixed before we perform the experiment, but we only know them in an imprecise manner, and actually we have the same knowledge $[-K, K]$ about these different values. The random sets Γ_1 and Γ_2, if regarded as classical measurable functions on the power set, induce the same possibility distribution (degenerated in the interval $[-K, K]$). Therefore, they both have the same Aumann expectation[3] (that coincides with their own image) and they have null "classical" variance, since they are constant set-valued functions. But, if we follow Kruse and Meyer's, interpretation, we suppose that each of the maps models the imprecise observation of a classical random variable. Let us recall that, when the fuzzy random variable is reduced to multi-valued mapping, Γ, the information that it provides us about the original random variable X_0, can be interpreted as follows: for every $\omega \in \Omega$, all we know about the image of ω, $X_0(\omega)$, is that it is in the set $\Gamma(\omega)$. So, returning to the example, in the case of Γ_1, we are certain that the variance of the original random variable is 0. In the second case (Γ_2) we only know that it is a value between 0 and K^2.

In practice, a fuzzy random variable can also be used to represent the imprecise observation of a certain property of the elements of a population Ω. To represent the information provided by the imprecise observations about the variance of the (classical) underlying random variable that models this property, we must resort to the variance defined by Kruse and Meyer.

[2] The definition given by Feng in [13] and cited in the introduction fits this formulation for $\pi_1 = \pi_3 = 0.5$.

[3] The Aumann expectation of a random set is defined as the union of the expectations of all its measurable selections.

3.2 Second-order imprecise model

In [20], Kruse defines the variance of a multi-valued mapping, $\Gamma : \Omega \to \mathcal{P}(\mathbb{R})$, as the set:

$$\mathrm{Var}_{\mathrm{Kr}}(\Gamma) = \{\mathrm{Var}(X) \mid X \in S(\Gamma)\},$$

where $S(\Gamma)$ represents the set of all measurable selections of the multi-valued mapping. The preceding definition can be easily extended to the case of fuzzy random variables as follows:

Definition 2 *Let us call* Kruse's variance *of the fuzzy random variable $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(\mathbb{R})$, the only fuzzy set determined by the nested family of sets:*

$$F(\alpha) := \mathrm{Var}_{\mathrm{Kr}}(\tilde{X}_\alpha), \forall\, \alpha,$$

where $\tilde{X}_\alpha$ is the multi-valued mapping α-cut of $\tilde{X}$.

We refer to the fuzzy set whose membership function is given by the expression

$$\pi(x) = \sup\{\alpha \in (0,1] \mid x \in \mathrm{Var}_{\mathrm{Kr}}(\tilde{X}_\alpha)\},\ \forall\, x \in \mathbb{R}$$

Let us notice that:

$$\{x \mid \pi(x) > \alpha\} \subseteq F(\alpha) \subseteq \{x \mid \pi(x) \geq \alpha\}, \forall\, \alpha \in (0,1).$$

Hence, it is easy to see that the following equality holds:

$$\pi(x) = \sup\{\mathrm{acc}(X) \mid \mathrm{Var}(X) = x\},\ \forall\, x \in \mathbb{R}.$$

From now on, we shall denote by $\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})$ the fuzzy set with membership function π. It is clear that this definition is compatible with the second-order possibility model shown in Section 2. Therefore, the membership degree of a value x to the fuzzy set $\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})$ represents the maximal possibility degree of the original random variable among those whose variance is equal to x. See [19] for the computation of the empirical set-valued variance of a finite set of set-valued realizations and [10] for the fuzzy case.

When the outputs of a random experiment are imprecisely observed, our knowledge about their dispersion is also imprecise. So, Kruse's variance can be called *potential variance*, since $\pi(x)$ is the degree of possibility that x is the variance (in case it exists) of the actual underlying random variable. $\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})$ reflects the imprecision pervading the observation of the outcome of a random experiment. Therefore, it produces a crisp set of potentially attainable variances (when the imprecise observations of the random variable are set-valued) or a fuzzy set (when it is represented by a fuzzy random variable). It does not produce a real value, like the "classical" observable variance of the previous section. Thus, when the random set (or the fuzzy random variable) represents the imprecise observation of a "classical" random variable, the description of the changes of the observed sets or fuzzy sets via a classical variance is not enough to inform about the variability of the underlying phenomenon. Let us show an illustrative example.

Example 2

(a) The set $\Omega = \{\omega_1, \ldots, \omega_4\}$ comprises four objects, whose actual weights are $X_0(\omega_1) = 10.2$, $X_0(\omega_2) = 10.0$, $X_0(\omega_3) = 10.4$, $X_0(\omega_4) = 9.7$. We sense the weights with a digital device that rounds the measure to the nearest integer, and displays the value '10' in all of these cases. Therefore, we get the constant random set $\Gamma(\omega_i) = [9.5, 10.5], \forall\, i = 1, \ldots, 4$. The true variance of the four measurements is 0.067. Since we only know the information provided by Γ, all we can say about the variance is that it is bounded by the values 0 and 0.25. This is the information that Kruse's variance gives us. Misleadingly, the classical variance of Γ returns the value 0.

(b) Case (a) is an example where the classical variance of the random set Γ is not an upper bound of the actual value of the variance of X_0. Neither is it, in general, a lower bound, as we are going to show. Let us suppose that four objects $\omega_1, \ldots, \omega_4$ weigh the same: $X_0(\omega_1) = X_0(\omega_2) = X_0(\omega_3) = X_0(\omega_4) = 9.8g$. Let us also suppose that, for some reason, the weight of the fourth object was imprecisely measured, and we only know that it is between the values 9.5 and 10.5. Our knowledge about the variable $\tilde{X}_0$ is given by the random set $\Gamma : \omega \to \mathcal{P}(\mathbb{R})$ defined as $\Gamma(\omega_1) = \Gamma(\omega_2) = \Gamma(\omega_3) = \{9.8\}$ and $\Gamma(\omega_4) = [9.5, 10.5]$. The true variance of X_0 is 0, but the "classical" variance assigns a strictly positive value to it. On the other hand, Kruse's variance produces the interval $[0, 0.092]$.

The last case suggests that the observed classical variance of a fuzzy random variable can be misleading. It may reflect the variance *of the imprecision* of the output (the knowledge of object ω_4 is more imprecise than the knowledge of the other objects), rather than the actual variability of the underlying phenomenon.

On the other hand, neither Kruse's variance is determined by the classical one, nor the converse holds. Let us illustrate these ideas with the aid of the following examples.

Example 3 *Let us consider now the random sets in Example 1. According to Kruse, their respective vari-*

ances represent the sets of possible values of the variance of the corresponding original random variable. Thus, in that example, the respective variances are, according to this definition, $\mathrm{Var}_{\mathrm{Kr}}(\Gamma_1) = \{0\}$ *and* $\mathrm{Var}_{\mathrm{Kr}}(\Gamma_2) = [0, K^2]$. *However, the classical variance assigns the value 0 to both random sets.*

We observe that Kruse's variance allows us to distinguish between two fuzzy random variables with the same "classical" probability distribution (the probability measure induced by the fuzzy random variable in the classical model) when they are used in this context. However, it does not always associates different values to two fuzzy random variables with different "classical" variance, as we shall see in the example that follows.

Example 4 *Let us consider the probability space* $(\Omega_2, \mathcal{A}_2, P_2)$ *of Example 1 and the constant random set* $\Gamma_2 : \Omega_2 \to \mathcal{P}(I\!R)$, *defined there. Let us also define the random set* $\Gamma_3 : \Omega_2 \to \mathcal{P}(I\!R)$ *as follows:* $\Gamma_3(h) = [-K, 0]$ *and* $\Gamma_3(t) = [0, K]$. *In both cases, Kruse's variance produces the interval* $[0, K^2]$. *But the classical variance would assign the value 0 to* Γ_2 *and a strictly positive value to* Γ_3.

The last example serves us to observe that Kruse's variance does not allow, generally speaking, to quantify the dispersion of the images of a fuzzy random variable, when it is considered as a classical measurable function.

In fact, the scalar variance of section 3.1 could be used in the context of an imprecisely observed random variable, but it could only account for an "observable variance", namely the part of the variance that can be measured, despite the imprecision of the observation. Indeed in Example 1, the fair die case leads to a zero observable variance, because the variability of the die is drowned into the imprecision of the observation. However, it seems that the scalar variance, when non-zero, may partially account for the variability of the underlying phenomenon: if the fuzzy random variable represents an imprecisely observed random variable with disjoint imprecise realizations, then it has a positive scalar variance that reveals the non-deterministic nature of the underlying process (even if only partially). On the other hand, as the above examples show, a zero scalar variance is not enough to conclude whether the observed phenomenon is random or not. Nor does a positive scalar variance reveal the actual randomness of the phenomenon if the realizations are nested fuzzy sets. It only points out the variability of the imprecision of the observed outcomes. In fact, one way of computing the observable variance as a scalar is to choose an appropriate distance between fuzzy sets instead of ρ_2 in the scalar variance (1), namely one that vanishes when the two fuzzy intervals overlap: consider two fuzzy intervals F and G, and let

$$d_{\min}(F_\alpha, G_\alpha) = \inf\{\mid x - y \mid, x \in F_\alpha, y \in G_\alpha\},$$

and (for instance) $d_{\min}(F, G) = \inf_{\alpha>0} d_{\min}(F_\alpha, G_\alpha)$.

We can check that this new scalar variance is less that the lower bound of Kruse's variance. In example 2(b), the above scalar variance is now 0, and so is it in example 4. In example 2(b), the Körner scalar variance essentially reflects the variability of the precision of the observation.

3.3 First-order imprecise model

In this section, we propose a model that also takes imprecision into account, although in a different manner. We consider here a first-order imprecise probability model, instead of a second-order one. Therefore, the new variance assigns a crisp set to every fuzzy random variable. With the help of easy examples, we shall show the similarities and differences between this new model and the present one.

The present definition of variance is based upon the first-order, imprecise probabilities model that was shown at the end of section 2. As we pointed out there, we consider, on the one hand, the probability measure P (defined over $\mathcal{A}$), that models a first sub-experiment, and, on the other hand, a family of conditional possibility measures, $\{\Pi(\cdot \mid \omega)\}_{\omega\in\Omega}$, defined as follows:

$$\Pi(A \mid \omega) = \Pi_{\tilde{X}(\omega)}(A) = \sup_{x\in A} \tilde{X}(\omega)(x),\ \forall\, A \in \beta_{I\!R},\ \forall\, \omega.$$

In the preceding formula, $\Pi_{\tilde{X}(\omega)}$ represents the possibility measure determined by the possibility distribution $\tilde{X}(\omega) : I\!R \to [0, 1]$. So, the value $\Pi(A|\omega)$ is an upper bound for the probability that the final outcome is in A, verifying the hypothesis that the outcome of the initial experiment is ω. This family of possibility measures represents our (imprecise) knowledge carried by $\tilde{X}$ about the relation that exists between the outcome of the first sub-experiment and the set of all the possible outcomes of the second one.

Therefore, the relationship between the two experiments is given by a **transition probability** $Q(\cdot|\cdot)$ on $\beta_{I\!R} \times \Omega$, i.e., a function such that:

1. $Q(\cdot|\omega)$ is a probability measure for all $\omega \in \Omega$.
2. $Q(A|\cdot)$ is $\mathcal{A} - \beta_{[0,1]}$-measurable for all $A \in \beta_{I\!R}$,

and the available knowledge about this transition probability is modelled by the conditional possibil-

ity measures $\{\Pi(\cdot|\omega)\}_{\omega\in\Omega}$, in the sense that $Q(\cdot|\omega) \leq \Pi(\cdot|\omega)$ for all $\omega \in \Omega$.

Within this context, all we know about the probability distribution that models the second experiment is that it is given by the formula:

$$Q_2(B) = \int_\Omega Q(B|\omega)\, dP(\omega),\ \forall\, B \in \beta_{I\!R},$$

where $Q(\cdot|\cdot)$ belongs to the class:

$$\mathcal{C} = \{Q(\cdot|\cdot)\ \ | \, Q(A|\omega) \leq \Pi(A|\omega)\ \forall\, A \in \beta_{I\!R},\ \omega \in \Omega\}.$$

In other words, all we know about Q_2 is that it is in the set

$$\mathcal{C}_2 = \{Q_2 : \beta_{I\!R} \to I\!R \mid \exists\, Q(\cdot|\cdot) \in \mathcal{C} \text{ where}$$

$$Q_2(B) = \int_\Omega Q(B|\omega)\, dP(\omega),\ \forall\, B \in \beta_{I\!R}\}, \qquad (2)$$

It is easily observed that this is a generalization of the concept of probability induced by a classical random variable. Let us suppose that the images of the fuzzy random variable $\tilde{X}$ are real values. In other words, let us suppose that for all $\omega \in \Omega$, $\Pi(\cdot|\omega)$ is, in particular, the degenerated probability measure in a point $X(\omega)$. In this case, we are admitting a complete confidence about the relationship between both sub-experiments (if the result of the first sub-experiment is ω, then we are absolutely certain the outcome of the second experiment is $X(\omega)$). It is easy to prove that the class $\mathcal{C}_2$ in equation (2) is reduced to the singleton $\{P_X\}$ (in this case, the probability induced by $X : \Omega \to I\!R$ in $\beta_{I\!R}$ is the only probability measure compatible with P and $\Pi(\cdot|\cdot)$). Besides, the variance of a classical random variable, $\mathrm{Var}(X) = \int_\Omega [X - E(X)]^2 dP$, can be alternatively expressed as the following Lebesgue integral with respect to P_X:

$$\mathrm{Var}(P_X) = \int_{I\!R} \left(\mathrm{id} - \int_{I\!R} \mathrm{id}\, dP_X\right)^2 dP_X,$$

where id:$I\!R \to I\!R$ is the identity function[4]. Therefore, in the proposed imprecise probabilities model, all we know about the variance of the output of the second sub-experiment is that it belongs to the set $\mathrm{Var}_{\mathrm{Im}\ 1}(\tilde{X})$ defined as follows:

Definition 3 *Consider a probability space $(\Omega, \mathcal{A}, P)$, and a fuzzy random variable defined over it, $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(I\!R)$. For each $\omega \in \Omega$, let $\Pi(\cdot|\omega)$ denote the possibility measure associated to the possibility distribution $\tilde{X}(\omega)$. We define the* first-order imprecise variance of *$\tilde{X}$ as the (crisp) set:*

$$\mathrm{Var}_{\mathrm{Im}-1}(\tilde{X}) = \{\mathrm{Var}(Q_2) \mid Q_2 \in \mathcal{C}_2\}$$

where

$$\mathcal{C}_2 = \{Q_2 : \beta_{I\!R} \to I\!R \mid \exists\, Q(\cdot|\cdot) \in \mathcal{C} \text{ s.t.}$$

$$Q_2(B) = \int_\Omega Q(B|\omega)\, dP(\omega),\ \forall\, B \in \beta_{I\!R}\},$$

and

$$\mathcal{C} = \{Q(\cdot|\cdot)\ \ | \, Q(A|\omega) \leq \Pi(A|\omega)\ \forall\, A \in \beta_{I\!R},\ \omega \in \Omega\}.$$

$\mathrm{Var}_{\mathrm{Im}-1}(\tilde{X})$ is the set of possible values of the variance of the second sub-experiment, according to the available information. We are going to compare, on an example, the information provided by $\mathrm{Var}_{\mathrm{Im}\,1}$ and $\mathrm{Var}_{\mathrm{Kr}}$ about the variance of the "original" probability distribution.

Example 5 *Let us consider the unit interval, $\Omega = [0, 1]$, equipped with the Lebesgue measure. Let us also consider the fuzzy random variable $\tilde{X} : \Omega \to \tilde{\mathcal{P}}(I\!R)$ constant in the fuzzy set $\tilde{A}$ determined by the α-cuts $\tilde{A}_\alpha = [-(1-\alpha), 1-\alpha]$. It can be easily checked that $[\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})]_\alpha = [0, (1-\alpha)^2],\ \forall\, \alpha > 0$. On the other hand, we can observe that $\mathrm{Var}_{\mathrm{Im}-1}(\tilde{X})$ is the interval [0,1/3][5]. It is clear that this interval is strictly contained in the support of $\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})$. Therefore, under the first-order model here described, the variance of the results of the experiment is known to be less than or equal that 1/3, while under the second-order probability model, a strictly positive possibility degree is also assigned to all variables between 1/3 and 1.*

Despite the fact that the two models considered in last example (orders 1 and 2 imprecise probability models) are associated to a possibilistic interpretation of fuzzy sets, the meaning of the two definitions of variance derived from them are quite different. In the second-order model, the fuzzy random variable, $\tilde{X}$, represents an imprecise observation of a particular (classical) random variable, $X_0 : \Omega \to I\!R$. For each possible result of the random experiment, $\omega \in \Omega$, the value $X_0(\omega)$ is fixed but we have imprecise knowledge about it. However, in the first-order model, the fuzzy random variable $\tilde{X}$ represents our (imprecise) knowledge about the link between two steps of a random experiment. Thus, the same result ω in the first step can be associated to different outcomes of the second step. Under the first-order model assumptions, we must combine the probability measure associated

[4]Since the variance of a classical random variable is a function of its induced probability distribution,we shall commit a small abuse of the language from now on and we shall express it as the variance of such probability distribution.

[5]It is actually equal to $\frac{1}{2}\int_0^1 (\inf A_\alpha - \sup A_\alpha)^2 d\alpha$. See Dubois et al.[10].

to the first step with the probability measure that relates the first step with the second one. As our knowledge about the latter conditional probability measure is given by a pair of upper-lower probability measures, so is our knowledge about the probability measure that governs the whole process.

Let us examine now the relation between both models in the particular case where $\tilde{X}$ is a random set. ($\tilde{X}(\omega)$ is a crisp set, $\forall\,\omega \in \Omega$.) In this case, Kruse's variance is defined as:

$$\mathrm{Var}_{\mathrm{Kr}}(\tilde{X}) = \{\mathrm{Var}(P_X) \mid X \in S(\tilde{X})\}$$

$$= \{\mathrm{Var}(Q) \mid Q \in \mathcal{P}(\tilde{X})\},$$

where $\mathcal{P}(\tilde{X})$ is the set of probability measures associated to the measurable selections of $\tilde{X}$,

$$\mathcal{P}(\tilde{X}) = \{P_X | X \in S(\tilde{X})\}.$$

On the other hand, the first-order imprecise variance is given by the formula:

$$\mathrm{Var}_{\mathrm{Im}\,1}(\tilde{X}) = \{\mathrm{Var}(Q_2) \mid Q_2 \in \mathcal{C}_2\}, \text{ where}$$

$$\mathcal{C}_2 = \{Q_2 \mid Q_2 \text{ marginal of } P \times Q(\cdot|\cdot), Q(\cdot|\cdot) \in \mathcal{C}\},$$

and $\mathcal{C}$ is the set of transition probability measures:

$$\mathcal{C} = \{Q(\cdot|\cdot) \mid Q(A|\omega) \leq \Pi(A|\omega)\ \forall\, A \in \beta_{I\!R},\ \omega \in \Omega\}.$$

In the above formula, $\Pi(\cdot|\omega)$ is the Boolean possibility measure associated to the (crisp) set $\tilde{X}(\omega)$. For an arbitrary measurable selection of $\tilde{X}$, $X \in S(\tilde{X})$, and a fixed $\omega \in \Omega$, let us consider the probability measure degenerated on the point $X(\omega)$, $\delta_{X(\omega)}$. Let us construct the function $Q(\cdot|\cdot) : \beta_{I\!R} \times \Omega \to [0,1]$ as $Q(\cdot|\omega) = \delta_{X(\omega)}$, $\forall\,\omega \in \Omega$. It is easy to see that $Q(\cdot|\cdot)$ is a transition probability measure and it belongs to the set $\mathcal{C}$. So the probability measure $P_X : \beta_{I\!R} \to [0,1]$ belongs to $\mathcal{C}_2$. Thus, we observe that the set $\mathcal{P}(\tilde{X})$ is included in $\mathcal{C}_2$ and so $\mathrm{Var}_{\mathrm{Kr}}(\tilde{X})$ is contained in $\mathrm{Var}_{\mathrm{Im}\,1}(\tilde{X})$. Furthermore $\mathcal{C}_2$ is a convex set of probability measures, but $\mathcal{P}(\tilde{X})$ is not convex in general. (The properties of $\mathcal{P}(\tilde{X})$ are studied in detail in [6, 7, 27, 28, 29].) These differences can influence the calculation of the variances, as shown in the following example.

Example 6 *Consider again the random sets used in example 1. According to the model described in that section, in the first case the first sub-experiment is deterministic, and the relationship between both sub-experiments is determined by* Γ_1. *This random set represents an "empty" conditional probability distribution over* $[-K, K]$. *Therefore, the set of conditional probability measures* $Q(\cdot|\omega_1)$, *that are compatible with them is the set of all measures that assign probability 1 to the set* $[-K, K]$. *This way, the following information is given: once the first experiment is performed, a random number between* $-K$ *and* K *is chosen, and not a number selected beforehand. This is the difference between the second-order model described before and the current model. In the second-order model, the number was selected beforehand, but it was unknown.*

Now, in the case of Γ_2, *the first sub-experiment consists in tossing a coin. Once the result has been observed, it is chosen, whatever the result is, a random number between* $-K$ *and* K. *Therefore, it is intuitively clear in this example that, regarding the outcome of the second sub-experiment, we could obviate tossing the coin (we could not in the second order model) and then* Γ_1 *and* Γ_2 *show, according to the interpretation of the first-order model the same information. Thus we observe that* $\mathrm{Var}_{\mathrm{Im}-1}(\Gamma_1) = \mathrm{Var}_{\mathrm{Im}-1}(\Gamma_2) = [0, K^2]$.

Let us comment on some relationships that exist between the variance of this imprecise, first-order model, and the classical variance of section 3.1.

We easily observe that none of them can be calculated as a function of the other one:

Example 7 *Let us consider, on the one hand, the random set* Γ_1 *defined in Example 1 and, on the other, the random set* Γ_5, *defined over the same space, of the form* $\Gamma_5(\omega_1) = \{0\}$. *The "classical" variance assigns value 0 to both random sets, while the imprecise variance assigns the set of values* $[0, K^2]$ *to the first problem, and the singleton* $\{0\}$ *to the second.*

In a similar manner, we can check that the classical variance can not either be expressed as a function of the variance that is considered in this section. It is enough to observe the random sets of Example 4.

4 Concluding remarks

In this work we have studied different proposals to generalize the concept of variance of a real random variable to fuzzy random variables. In Körner 's work [17] is stated a more general definition, valid when the final space is $I\!R^n$, with arbitrary $n \in I\!N$. In that work, the variance of $\tilde{X}$ is defined as the expectation of the squares of the distances of their images to their Fréchet expectation. In the particular case where $\tilde{X}$ is a classical random vector and the chosen distance is Euclidean, the result of this calculation is the *moment of inertia*. This way, Körner's procedure generalizes, in the n-dimensional case, a concept that may be useful to measure the dispersion of the images of the fuzzy random variable, but not directly related

to the concept of variance-covariance matrix.

If, on the contrary, the aim is to generalize the latter concept, Kruse's procedure can be applied without too many changes. Using similar reasoning methods as those of this author, a fuzzy set over the class of square matrices can be obtained. It associates, to each particular matrix, a degree of possibility. This fuzzy set models the imprecise knowledge available about the variance-covariance matrix of the "original" random vector. In [26], Meyer proposes a definition of covariance following a path similar to Kruse's. According to our intuition, the combination of the information provided separately about the variance of every component and about the covariance between them is more imprecise that the straight information about the variance-covariance matrix.

With respect to the different definitions of variance considered in this work, we think that none of them is, in general terms, preferable to the others, but they either serve different purposes or reflect different models of the observed phenomenon, as well as different observation settings. Therefore, according to the problem under concern, it should be decided whether the dispersion needs to be measured as a number, a fuzzy set or a crisp set. If the fuzzy random variable is interpreted as a classical measurable function, the most appropriate decision would involve Feng, Körner or Lubiano et al.'s definitions. It measures the variability of the observed membership function, not the variability of the quantity it possibly describes. Such classical definitions do not take into account any kind of imprecision, but they merely quantify the dispersion of the (fuzzy) images of the fuzzy random variable.

Some of these classical definitions are equivalent to considering first a representative (numerical) element of every image of the fuzzy random variable (the center point of the 0-cut, for instance) and then calculate the dispersion of these numerical values. Part of the actual variability can be observed and measured by means of a scalar if the fuzzy outcomes are precise enough and often disjoint. On the other hand, the average precision of the fuzzy random variable, and the variance of the precision are other useful evaluations.

If the fuzzy random variable represents an imprecise measurement of a certain characteristic of the elements of the sample space, one of the two non-scalar definitions must be used. For example: let us suppose we intend to calculate the dispersion of the weights of a bunch of apples, and we use an imprecise scale. Let us suppose that, for every confidence level $1-\alpha$ we know that the real weight is at most at d_α from the value produced by the scale. In this case, every α-cut of Kruse's variance represents our knowledge about the true dispersion of the weights of the apples, for every confidence level $1-\alpha$. On the other hand, the variance proposed in section 3.3 represents the set of all possible values for the dispersion of the weights, if we combine the initial randomness (tied to the random experiment "choose an apple") with the randomness originated in the degrees of confidence associated to the scale accuracy. Therefore, if the fuzzy random variable represents the knowledge about the relationship between the two sub-experiments ("if we choose the apple ω, the degree of possibility of its weight x is $\tilde{X}(\omega)(x)$"), then the definition proposed in section 3.3 should be used.

Acknowledgements

This work has been supported by grants MTM2004-01269 and TIN2005-08036-C05-05. Both grants partially participate to FEDER funds.

References

[1] C. Baudrit, I. Couso, D. Dubois (2007) Joint propagation of probability and possibility in risk analysis: Towards a formal framework, *Int. J. of Approximate Reasoning*, 45, 82-105.

[2] C. Baudrit, D. Dubois, D. Guyonnet, H. Fargier (2006) Joint treatment of imprecision and randomness in uncertainty propagation. In : *Modern Information Processing: From Theory to Applications*. B. Bouchon-Meunier, G. Coletti, R.R. Yager (Eds.), Elsevier, p. 37-47.

[3] C. Bertoluzza, A. Salas, N. Corral (1995) On a new class of distances between fuzzy numbers. *Mathware and Soft Computing* **2** 71-84.

[4] I. Couso, E. Miranda, G. de Cooman (2004) A possibilistic interpretation of the expectation of a fuzzy random variable. In *Soft methodology and random information systems* (eds: M. López-Díaz, M. A. Gil; P. Grzegorzewski, O. Hyrniewicz, and J. Lawry), 133-140. Springer, Heidelberg.

[5] I. Couso, S. Montes, P. Gil (1998) Función de distribución y mediana de variables aleatorias difusas, *Proceedings of the Conference ESTYLF'98*. Pamplona. (In Spanish)

[6] I. Couso, S. Montes, P. Gil (2002) Second-order possibility measure induced by a fuzzy random variable. In *Statistical modeling, analysis and management of fuzzy data* (eds: C. Bertoluzza,

M. A. Gil and D. A. Ralescu), 127-144. Physica-Verlag, Heidelberg.

[7] I. Couso, L. Sánchez, P. Gil (2004) Imprecise distribution function associated to a random set, *Information Sciences* **159** 109-123.

[8] G. de Cooman, P. Walley (2002) An imprecise hierarchical model for behaviour under uncertainty, *Theory and Decision* **52** 327-374.

[9] P. Diamond, P. Kloeden (1994) *Metric Spaces of Fuzzy Sets*, World Scientific, Singapur.

[10] D. Dubois, H. Fargier, J. Fortin (2005) The empirical variance of a set of fuzzy intervals. *Proc. IEEE Int. Conf. on Fuzzy Systems*, Reno, Nevada, IEEE Press, p. 885-890.

[11] D. Dubois, H. Prade (1987) The mean value of a fuzzy number, *Fuzzy Sets and Systems* **24** 279-300.

[12] D. Dubois, H. Prade (1997) The three semantics of fuzzy sets, *Fuzzy Sets and Systems* **90** 141-150.

[13] Y. Feng, L. Hu, H. Shu (2001) The variance and covariance of fuzzy random variables and their applications, *Fuzzy Sets and Systems* **120** 487-497.

[14] R. Féron (1976) Ensembles aléatoires flous, *C.R. Acad. Sci. Paris Ser. A* **282** 903-906.

[15] M. Fréchet (1948) Les éléments aleatoires de natures quelconque dans un espace distancié, *Ann. Inst. H. Poincaré* **10** 215-310.

[16] E.P. Klement, M.L. Puri, D.A. Ralescu (1986) Limit theorems for fuzzy random variables *Proc. Roy. Soc. London A* **407** 171-182.

[17] R. Körner (1997) On the variance of fuzzy random variables, *Fuzzy Sets and Systems* **92** 83-93.

[18] V. Krätschmer (2001) A unified approach to fuzzy random variables, *Fuzzy Sets and Systems* **123** 1-9.

[19] V. Kreinovich, G. Xiang and S. Ferson (2006) Computing mean and variance under Dempster-Shafer uncertainty: Towards faster algorithms *Int.J. of Approximate Reasoning*, 42(3), 212-227.

[20] R. Kruse (1987) On the variance of random sets, *J. Math. Anal. Appl.* **122** 469-473.

[21] R. Kruse, K.D. Meyer (1987) *Statistics with vague data* D. Reidel Publishing Company.

[22] Kwakernaak (1989) Fuzzy random variables. Definition and theorems. *Inform. Sci.* **15** 1-29.

[23] Y.K. Liu, B. Liu (2003) A class of fuzzy random optimization: expected valued models, *Information Sciences* **155** 89-102.

[24] M.A. Lubiano, M.A. Gil, M. López-Díaz, M.T. López-García (2000), The $\vec{\lambda}$-mean squared dispersion associated with a fuzzy random variable, *Fuzzy Sets and Systems* **111** 307-317.

[25] M.A. Lubiano, M.A. Gil (2002) f-inequality indices for fuzzy random variables, In *Statistical modeling, analysis and management of fuzzy data* C. Bertoluzza, M. A. Gil and D. A. Ralescu (Eds.), 43-63. Physica-Verlag, Heidelberg.

[26] K.D. Meyer, R. Kruse (1990) On calculating the covariance in the presence of vague data. *In: Progress in Fuzzy Sets and Systems.* W.H. Janko, M.Roubens and H.J. Zimmermann (Eds.) Kluwer Academic Publishers, Dordrecht.

[27] E. Miranda, I. Couso, P. Gil (2002) Upper probabilities and selectors of random sets. *En: Soft methods in probability, statistics and data analysis* P. Gzegorzewski, O. Hryniewicz, M.A. Gil (Eds.). Physica-Verlag, Heidelberg, Alemania.

[28] E. Miranda, I. Couso, P. Gil (2003) Study of the probabilistic information of a random set, *Proceedings of the 3rd ISIPTA Conference*. Lugano, Suiza.

[29] E. Miranda, I. Couso, P. Gil (2005) Random sets as imprecise random variables. *Journal of Mathematical Analysis and Applications* **307** 32-47.

[30] E. Miranda, G. de Cooman, I. Couso (2005) Imprecise probabilities induced by multi-valued mappings, *J. Stat. Plann. Inference.* **133** 173-197.

[31] M.L. Puri, D. Ralescu (1986) Fuzzy Random Variables, *J. Math. Anal. Appl.* **114** 409-422.

[32] P. Walley (1991) *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

Inference in Credal Networks Through Integer Programming

Cassio Polpo de Campos
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo, Brazil
cassiopc@usp.br

Fabio Gagliardi Cozman
Escola Politécnica
Universidade de São Paulo, Brazil
fgcozman@usp.br

Abstract

A credal network associates a directed acyclic graph with a collection of sets of probability measures; it offers a compact representation for sets of multivariate distributions. In this paper we present a new algorithm for inference in credal networks based on an integer programming reformulation. We are concerned with computation of lower/upper probabilities for a variable in a given credal network. Experiments reported in this paper indicate that this new algorithm has better performance than existing ones for some important classes of networks.

Keywords. Credal networks, Integer programming.

1 Introduction

This paper presents novel techniques for marginal inference in *credal networks*. The goal is to provide an algorithm that can handle graphical models for precise and imprecise probabilistic assessments based on integer programming.

Credal networks represent a set of joint probability measures through a directed acyclic graph and a collection of local sets of probability measures [3, 5, 13]. The structure of the graph indicates relations of independence between variables; the "size" of the sets of probabilities encodes the imprecision in the probability values. Section 2 reviews basic properties of credal networks and Section 3 addresses the inference problem we are interested in. Basically, a belief updating inference in the context of credal networks is a computation of upper/lower probability for some conjunction of events, given observations. Most existing algorithms for belief updating cannot handle large networks; when they can, they suffer from numerical instability [9].

This paper aims at enlarging the class of networks that can be successfully processed exactly. We focus on developing a reformulation that is particularly efficient for polytree-shaped networks. Section 4 reviews a multilinear reformulation and presents a new inference algorithm based on bilinear and integer programming. Section 5 shows through experiments that the algorithm can process large polytree networks, surpassing existing algorithms [8, 9]. Section 6 concludes the paper.

2 Credal sets and credal networks

A few preliminary definitions are important. A convex set of probability distributions is called a *credal set* [18]. A credal set for X is denoted by $K(X)$; we assume that every random variable is categorical and that every credal set has a finite number of vertices. A conditional credal set is a set of conditional distributions, obtained by applying Bayes rule to each distribution in a credal set of joint distributions. The theory of sets of probability distributions adopted in this paper can be placed in the framework of coherent behavior by selecting axioms advocated by several authors, for instance by Walley [23]. We emphasize that our setting is restricted to categorical variables, thus we can brush away subtle but crucial differences between proposed frameworks concerning issues of conglomerability and countable additivity.

The sets $K(X|Y)$ are *separately specified* when there is no constraint on the conditional set $K(X|Y = y_1)$ that is based on the properties of $K(X|Y = y_2)$, for any $y_2 \neq y_1$ — that is, the conditional sets bear no relationship to each other. In this paper we assume that local credal sets are always separately specified; justifications for this separability assumption can be found in [7]. Given a number of marginal and conditional credal sets, an *extension* of these sets is a joint credal set with the given marginal and conditional credal sets. In this paper we are exclusively concerned with the largest possible extension for any collection of marginal and conditional credal sets.

Given a credal set $K(X)$ and an event A, the *upper* and *lower* probability of A are respectively $\overline{P}(A) = \max_{p(X)\in K(X)} P(A)$ and $\underline{P}(A) = \min_{p(X)\in K(X)} P(A)$.

A *credal network* $N = (G, \mathbb{X}, \mathbb{K})$ is composed by a directed acyclic graph $G = (V, E)$ where each node of V is associated with a random variable $X_i \in \mathbb{X}$ and with a collection of conditional credal sets $K(X_i|\text{pa}(X_i)) \in \mathbb{K}$, where $\text{pa}(X_i)$ denotes the parents of the node associated to X_i in the graph. In the remainder of this paper, we refer to X_i and its associated node interchangeably. Note that we have a conditional credal set related to X_i for each instantiation of $\text{pa}(X_i)$. A root node is associated with a single marginal credal set. We take that in a credal network every random variable is independent of its nondescendants nonparents given its parents; this is the *Markov condition* on the network. In this paper we adopt the concept of *strong independence*[1]: two random variables X and Y are strongly independent when every extreme point of $K(X, Y)$ satisfies standard stochastic independence of X and Y (that is, $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$) [5]. Strong independence is the most commonly adopted concept of independence for credal sets, probably due to its obvious connection with standard stochastic independence. There are concepts of independence that are less precise in the sense that they admit distributions that do not factorize; an example is epistemic independence [11, 23].

Given a credal network, an *extension* of the network is any joint credal set that satisfies all constraints encoded in the network. The *strong extension* of a credal network is the largest joint credal set such that every variable is strongly independent of its nondescendants nonparents given its parents. The strong extension of a credal network is the joint credal set that contains every possible combination of vertices for all credal sets in the network [6]; that is, each vertex of a strong extension factorizes as follows:

$$p(X_1, \ldots, X_n) = \prod_i p(X_i|\text{pa}(X_i)) \,. \qquad (1)$$

3 Inference with strong extensions

A *marginal inference* in a credal network is the computation of lower/upper probabilities in an extension of the network. If X_q is a *query* variable and $\mathbf{X}_E$ represents a set of *observed* variables, then an inference is the computation of tight bounds for $p(X_q|\mathbf{X}_E)$ for one or more values of X_q. For inferences in strong extensions, it is known that the distributions that minimize/maximize $p(X_q|\mathbf{X}_E)$ belong to the set of vertices of the extension [13].

[1] We note that other concepts of independence are found in the literature [4, 12].

An inference can be produced by combinatorial optimization, as we must find a vertex for each local credal set $K(X_i|\text{pa}(X_i))$ so that Expression (1) leads to a maximum/minimum of $p(X_q|\mathbf{X}_E)$. In general, inference offers tremendous computational challenges — consider the following example, taken from Rocha et al. [8]. Take a network with three nodes, $X \to Y \leftarrow Z$, where X, Y and Z have four categories each, and where all credal sets have four vertices each. There are 4^{18} different joint distributions factorizing as Expression (1), where local distributions are vertices of local credal sets. Rocha et al. [8] discuss branch-and-bound procedures that can handle situations such as this, but that still have difficulties in large networks. The only known polynomial algorithm for strong extensions is the 2U algorithm, which only processes polytrees with *binary* variables [13]. Other exact inference algorithms based on enumeration examine all potential vertices of the strong extension to produce the required lower/upper values [2, 3, 5, 7]; these algorithms face serious difficulties in large networks.

A different way to look at the computation of inferences is to recognize that a lower/upper value for $p(X_q|\mathbf{X}_E)$ is obtained by minimization/maximization of a fraction containing polynomials in probability values. This is in fact the strategy discussed in Section 4; our results suggest that this is the most profitable strategy to take for exact inference with strong extensions.

4 Inference as a multilinear programming problem

A marginal inference for a strong extension can be formulated as a multilinear programming problem. The goal is to minimize/maximize the expression

$$\sum_{X_i \setminus X_q} \prod_i p(X_i|\text{pa}(X_i)) \qquad (2)$$

subject to constraints on the local probabilities $p(X_i|\text{pa}(X_i))$. For a query with evidence, we may use the constraint $p(X_q|\mathbf{X}_E) = \frac{p(X_q, \mathbf{X}_E)}{p(\mathbf{X}_E)}$, that can be turned into a multilinear constraint.[2] In this problem we must deal with a large number of terms in the multilinear objective function (the number of terms is exponential on the size of the network), as shown in Example 1.

Example 1 *Take the network presented in Figure 1. Suppose that random variables are binary and we want*

[2] We assume that the probability of evidence is strictly greater than zero, leaving for future work the important case where lower probabilities equal to zero may happen.

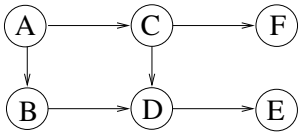


Figure 1: Simple multi-connected network.

to evaluate the maximum possible value for the probability of $(E = e) \wedge (F = f)$*; this is obtained by solving:*

$$\max \sum_{A,B,C,D} p(f|C) \cdot p(e|D) \cdot p(D|B,C) \cdot \quad (3)$$
$$\cdot p(B|A) \cdot p(C|A) \cdot p(A)\,,$$

subject to linear constraints (from local credal sets). We have a multilinear objective function with 16 nonlinear terms of degree six. The probability functions p are seen as optimization variables in the multilinear program.

As presented by Campos and Cozman [9], we can run a symbolic variable elimination algorithm to obtain a simpler objective function.

Example 2 *Take the network and specifications of Example 1. Instead of optimizing Expression (3), the symbolic variable elimination procedure transforms it into a problem with simpler multilinear functions of degree at most three, by grouping terms and introducing new optimization variables. The result is a multilinear program with 22 nonlinear terms:*

$$\max \sum_{D} p(e|D)\, p(D,f) \quad \text{subject to}$$
$$p(B,C) = \sum_{A} p(B|A)\, p(C|A)\, p(A)\,, \ \textit{for all } B, C$$
$$p(C,D) = \sum_{B} p(D|B,C)\, p(B,C)\,, \ \textit{for all } C, D$$
$$p(D,f) = \sum_{C} p(f|C)\, p(C,D)\,, \ \textit{for all } D$$

plus the linear constraints.

Note that the reformulated problem presented in Example 2 contains terms of smaller degree than the original problem (Example 1). This is important in multilinear programming, as it is difficult to handle problems with high degree. Besides that, the transformation usually leads a smaller number of nonlinear terms than in the direct version given by Expression (1), although that was not the case in Example 2.

4.1 A bilinear transformation

The new multilinear terms obtained with the symbolic variable elimination procedure just described have smaller degree than the original ones, but the maximum degree is at least as large as the tree-width of network's moral graph (even if we know an optimal elimination ordering for the variables) [9].

We present in this section a new transformation procedure, which naturally produces multilinear programming problems with maximum degree of two (that is, bilinear problems) regardless of network's topology, and where the following property holds: each bilinear term has at least one variable that is defined (even though by a credal set) in the input (that is, each bilinear term has at most one auxiliary variable). This property will be essential for obtaining an integer program in Section 4.2, and it is specially efficient for polytree networks, which are defined by graph without any cycles (directed or not).

Prior to the algorithm itself, we must present some useful definitions.

Definition 3 *An ordering for the network variables is said a* precedence ordering *if, for each variable in the ordering, all its ancestors in the network's graph appear before it in the ordering.*

Definition 4 (Robertson and Seymour [20, 21]). *Given a graph* $G = (V, E)$*, a sequence* $V_1, \ldots, V_r$ *of subsets of* V *is a* path-decomposition *of* G *if the following conditions are satisfied:*

- $\bigcup_i V_i = V$.
- *For every edge* $e \in E$*, some* V_i *contains both endpoints of* e.
- *For* $1 \leq i \leq j \leq k \leq r$*,* $V_i \cap V_k \subseteq V_j$.

Definition 5 (Robertson and Seymour [20, 21]). *The* path-width *of* G*, denoted by* $pw(G)$*, is the minimum value* $h \geq 0$ *such that* G *has a path-decomposition* $V_1, \ldots, V_r$ *with* $|V_i| \leq h + 1$ *for* $i = 1, \ldots, r$.

Definition 6 *The path-width of a credal network* $N = (G, \mathbb{X}, \mathbb{K})$*, or just* $pw(N)$*, is the path-width of its graph* G.

The idea of `Bilinear-Transformation` algorithm is to process the network variables top-down, using a *precedence ordering*. At each step we construct a constraint that defines the relationship between the query and the current variable being processed. A variable may be processed only if all its ancestors have already been processed. The active nodes at each step form a path-decomposition of the network's graph. Note that we cannot use other decompositions such as joint trees, because we would get multilinear terms with

more than one auxiliary variable, that is, the result would not be a bilinear programming problem with that described property. We proceed with the idea of the transformation using an example.

Example 7 *Suppose we want to query the probability of e, f in the network presented in Figure 1. The first step of the* `Bilinear-Transformation` *algorithm is to choose a precedence ordering for the network variables (when there is evidence, all the process must be repeated for the queries and for the observed variables). We will use the ordering A, C, B, D, E, F. The first variable to be processed is A (it is the only variable without parents). We have the queries e, f and will write their joint probability using $p(A)$ (which is defined in the network specification) and inserting A in the conditional part. So we create the constraint*

$$p(e, f) = \sum_{A \in \{a, \overline{a}\}} p(A) \cdot p(e, f|A).$$

Functions $p(e, f|A)$ are auxiliary (they do not appear in the network), and we must create constraints to define them (for all possible instantiations of A). The current variable to be processed is C. Thus, for all $A \in \{a, \overline{a}\}$:

$$p(e, f|A) = \sum_{C \in \{c, \overline{c}\}} p(C|A) \cdot p(e, f|A, C).$$

At this stage, our queries are conditioned on A and C. Following the idea, we process B, obtaining

$$p(e, f|A, C) = \sum_{B \in \{b, \overline{b}\}} p(B|A) \cdot p(e, f|B, C),$$

which must be written for all $A \in \{a, \overline{a}\}, C \in \{c, \overline{c}\}$. Note that at this point A disappeared from the conditioning side, because B and C together separate the query variables from A. Now the current variable to be treated is D, and our queries are conditioned on B, C, that is, we must define how to evaluate $p(e, f|B, C)$. We have, for all $B \in \{b, \overline{b}\}, C \in \{c, \overline{c}\}$, that

$$p(e, f|B, C) = \sum_{D \in \{d, \overline{d}\}} p(D|B, C) \cdot p(e, f|C, D).$$

At this moment, e, f are conditioned on C, D (again, B is not present anymore as C, D separate the queries from B). Now we will process E, but because the only two remaining variables are E and F and they are not parent of each other, their order in fact does not matter. Thus,

$$p(e, f|C, D) = p(e|D) \cdot p(f|C),$$

for all $C \in \{c, \overline{c}\}, D \in \{d, \overline{d}\}$. Note that, as $p(f|C)$ is specified in the network, we can stop. We have completed the procedure as both $p(e|D)$ and $p(f|C)$ appear in the network. Note that, if we had chosen another ordering such as A, B, C, D, E, F or A, B, C, F, D, E, more constraints might be needed.

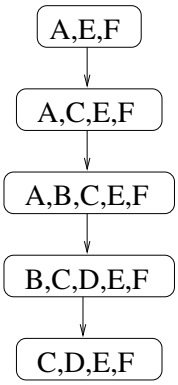


Figure 2: Path-decomposition of Example 7.

Figure 2 shows the path-decomposition induced by the ordering of Example 7. Note that there is an one-to-one relation between decomposition components and constraints in the example. The elements in a given component appear together in some constraint of the reformulation, either in the conditioned (including queries) or in the conditioning sides.

The algorithm is presented using pseudo-code in Figure 3. Functions g appearing in line 24 of the algorithm are just conditioned probability functions. We use the letter g instead of p because of another transformation presented in next section, where functions g have special meaning. If we just want a bilinear transformation, g should be simply replaced by p, even though we note that names of optimization variables are not an issue (they just need to be coherent among each other). Regarding the complexity of the algorithm, we define the path-width of a precedence ordering as the width of the path-decomposition induced by that ordering, and present the following theorem.

Theorem 8 *Let $N = (G, \mathbb{X}, \mathbb{K})$ be a credal network where the maximum number of categories of a random variable is $O(|V|)$. Suppose o' is a precedence ordering for the variables in $\mathbb{X}$. Then* `Bilinear-Transformation` *runs in time $O(|V|^{pw(o')+k})$, for k constant.*

Proof: All non-loop lines of the algorithm can clearly be executed in polynomial time in the number of nodes, that is, $O(|V|^K)$, for K constant.

The loop of line 4 is executed twice if we have evidence, and only once if we do not have evidence. Line 12 loop is executed $|U|$ times, that is, $O(|V|)$. The loop of line 26 is executed $O(c)$ times, where c is the

BILINEAR-TRANSFORMATION$(N, \mathbb{Q}, \mathbb{E})$

$N = (G, \mathbb{X}, \mathbb{K})$ is the network, $G = (V, E)$ its graph, $\mathbb{X}$ its variables and $\mathbb{K}$ its local credal sets.
$\mathbb{Q}$ is an instantiation for a set of queried variables.
$\mathbb{E}$ is an instantiation for a set of observed variables.
The result of this procedure is a bilinear programming problem
(bilinear objective function of line 1 and a set of bilinear constraints from lines 2 and 24).

1 ▷ The program will maximize or minimize t, which will be the objective function.
2 Insert the following constraint into the bilinear program:
$p(\mathbb{Q}, \mathbb{E}) = t \cdot p(\mathbb{E})$
3 ▷ Now we create constraints to evaluate $p(\mathbb{Q}, \mathbb{E})$ and $p(\mathbb{E})$.
4 **for** $W = \mathbb{Q} \cup \mathbb{E}$ and $W = \mathbb{E}$
5 **do**
6 ▷ U is the set of all relevant variables.
7 $U \leftarrow \{X \in V \setminus W$ such that X is an ancestor of some $w \in W\}$.
8 ▷ There are many ways to choose a precedence ordering. Do it polynomially.
9 Rename the variables of U as $X_1, X_2, \ldots, X_{|U|}$ according to a precedence ordering.
10 ▷ Initially, functions are not conditioned. L is a list of sets of conditioning variables.
11 Let L be an empty queue of sets. Insert $\emptyset$ in the end of L.
12 **for** $i \leftarrow 1$ **to** $|U|$
13 **do**
14 ▷ L' will have the conditioning sets to be considered on the next loop step.
15 Let L' be an empty queue of sets.
16 ▷ Conditioning sets from the previous step are processed.
17 **while** L is not empty
18 **do**
19 $S \leftarrow$ first element of L (remove S from L).
20 ▷ S' are separated variables (with respect to the query variables) when inserting X_i in the conditioning part. The separation is based on the graph structure (known as d-separation [19]).
21 $S' \leftarrow \{s \in S$ such that $\{X_i\} \cup S \setminus \{s\}$ separates W from $s\}$.
22 ▷ The variables in S' are no more relevant.
23 $R \leftarrow S \setminus S'$.
24 Depending on W, insert the following constraint into the bilinear program:
if $W = \mathbb{Q} \cup \mathbb{E}$ **then** $p(W|S) = \sum_{x_{ij}} p(x_{ij}|\text{pa}(X_i)) \cdot p(W|R, x_{ij})$.
if $W = \mathbb{E}$ **then** $g(W|S) = \sum_{x_{ij}} p(x_{ij}|\text{pa}(X_i)) \cdot g(W|R, x_{ij})$.
▷ x_{ij} is a category of X_i.
▷ $\text{pa}(X_i)$ is an instantiation complying with W, R and S.
25 ▷ Now we insert into L' the conditioning sets for the next step.
26 **for** each x_{ij} of X_i
27 **do**
28 $R' \leftarrow R \cup \{x_{ij}\}$.
29 Insert R' in the end of L'.
30 ▷ End of **for**
31 ▷ End of **while**
32 $L \leftarrow L'$
33 ▷ End of **for**
34 ▷ End of **for**

Figure 3: Reformulation algorithm for inferences in credal networks.

maximum number of categories of a random variable.

Line 17 loop executes $|L|$ times. Let u be the maximum size of a set stored in L. Then the number of elements in L at each round is $O(c^u)$, because at most c^u unequal instantiations for u variables are possible. Note that the sizes of sets stored in L are exactly the sizes of components in the path-decomposition induced by o'. So the bottleneck is the size of the elements in L, which is equal (in worst case) to $pw(o')$. Thus the complexity of `Bilinear-Transformation` is ($k = K + 2$):

$$2 \cdot O(|V|) \cdot O(|V|^K) \cdot O(c) \cdot O(c^{pw(o')}) = O(|V|^{pw(o')+k}).$$

Note that, with recent results of Feige et al. [14], it is possible to approximate the optimum path-width of the network by $\log |V| \sqrt{\log pw(N)}$. So, the algorithm runs in time $O(|V|^{\log c \cdot pw(N) \cdot \sqrt{\log pw(N)}+k})$. □

Corollary 9 *When restricted to polytrees, the algorithm* `Bilinear-Transformation` *runs in polynomial time in the size of input.*

Proof: In a polytree, each variable separates its parents from its descendants. The path-width is bounded by d, the maximum degree of network's graph, and it is easy to find an ordering with such width (a greedy algorithm will succeed). So the complexity is $O(|V|^d \cdot |V|^k)$, k constant. As the input size needed to specify the local credal sets of the network is already exponential on d, the corollary follows. □

4.2 An integer programming version

We show in this section how to obtain an integer program from that bilinear program generated by `Bilinear-Transformation`. We must note some useful properties:

1. A multiplication of a rational optimization variable $x \in [0, 1]$ by a boolean variable $b \in \{0, 1\}$ can be encoded by linear constraints: replace the nonlinear term $x \cdot b$ by a new variable y_{xb} and insert the constraints:

$$\begin{aligned} 0 \leq \quad & y_{xb} \quad \leq b \\ x - 1 + b \leq \quad & y_{xb} \quad \leq x \end{aligned}$$

2. We can represent each local credal set as a combination of its vertices. Suppose X is a network variable with parents $Y_1, \ldots, Y_r$, and that vertices $\alpha_1, \ldots, \alpha_s$ define the credal set for $p(X|y_1, \ldots, y_r)$ (the dimension of each α_i equals the number of categories of X). For each instantiation of $X, Y_1, \ldots, Y_r$ we have

$$p(x|y_1, \ldots, y_r) = \sum_{i=1}^{s} \alpha_i(x) \cdot b^{(i)}_{y_1,\ldots,y_r}, \qquad (4)$$

where $\alpha_i(x)$ are known values (specified in the network) and $b^{(i)}_{y_1,\ldots,y_r}$ are boolean variables such that

$$\sum_{i=1}^{s} b^{(i)}_{y_1,\ldots,y_r} = 1,$$

that is, only one of these $b^{(i)}_{y_1,\ldots,y_r}$ variables is one, thus selecting a vertex.

3. Each nonlinear term appearing in the constraints created by `Bilinear-Transformation` is a multiplication of a rational variable and a variable appearing in the network specification (defined by the local credal sets).

These observations lead us to the following procedure to replace each product $r \cdot p(x|y_1, \ldots, y_r)$ of each constraint created by `Bilinear-Transformation`:

$$\begin{aligned} r \cdot p(x|y_1, \ldots, y_r) \quad &:= \quad \sum_{i=1}^{s} \alpha_i(x) \cdot y^{(i)}_{rb_{y_1,\ldots,y_r}}, \\ y^{(i)}_{rb_{y_1,\ldots,y_r}} \quad &\geq \quad 0, \\ y^{(i)}_{rb_{y_1,\ldots,y_r}} \quad &\leq \quad b^{(i)}_{y_1,\ldots,y_r}, \\ y^{(i)}_{rb_{y_1,\ldots,y_r}} \quad &\geq \quad r - 1 + b^{(i)}_{y_1,\ldots,y_r}, \\ y^{(i)}_{rb_{y_1,\ldots,y_r}} \quad &\leq \quad r, \\ \sum_{i=1}^{s} b^{(i)}_{y_1,\ldots,y_r} \quad &= \quad 1, \end{aligned}$$

where $A := B$ means to replace A by B. Although we need to work with all vertices of credal sets and it may be hard to enumerate all of them, many important models can easily be translated into lists of vertices. For example, capacities of infinite order (also known as belief functions) can be expressed by mass assignments that are attached to sets of categories; vertices are simply obtained by combining the ways in which mass assignments are to be distributed [22, 23].

There is still a problem to address to get an integer version. The constraint inserted during line 2 of algorithm `Bilinear-Transformation` is nonlinear and the variables involved in its product are not in the network specification and thus cannot be directly replaced by some linear constraints and integer variables. We solve this problem by calling the loop of line 4 twice: in the first time, we evaluate $p(\mathbb{Q}, \mathbb{E})$ (using functions named p); in the second time, we evaluate $g(\mathbb{E}) = t \cdot p(\mathbb{E})$, that is, functions g do not mean probability functions but t times probability functions.

For example, the constraint in line 2 becomes simply $p(\mathbb{Q}|\mathbb{E}) = g(\mathbb{E})$. Each constraint inserted in line 24 of the algorithm

$$g(W|S) = \sum_{x_{ij}} p(x_{ij}|\mathrm{pa}(X_i)) \cdot g(W|R, x_{ij})$$

in fact means

$$t \cdot p(W|S) = \sum_{x_{ij}} p(x_{ij}|\mathrm{pa}(X_i)) \cdot t \cdot p(W|R, x_{ij}),$$

with the t variable hidden inside the g functions, whose are seen as optimization variables. So, on the last step of the inner loop, the constraint $g(W|S) = t \cdot p(W|S)$ must be included to pull t out of g, transforming it into p again. Because the $p(W|S)$ of the last step is certainly specified in the credal network input, this product can now be linearized using those ideas described on items from 1 to 3.

Example 10 *Suppose we want to evaluate $p(a|d)$ in the network of Figure 1. First, we symbolically evaluate $p(a, d)$ using p functions:*

$$\begin{aligned}
p(a,d) &= p(a) \cdot p(d|a) \\
p(d|a) &= p(b|a) \cdot p(d|a,b) + p(\overline{b}|a) \cdot p(d|a,\overline{b}) \\
p(d|a,b) &= p(c|a) \cdot p(d|b,c) + p(\overline{c}|a) \cdot p(d|b,\overline{c}) \\
p(d|a,\overline{b}) &= p(c|a) \cdot p(d|\overline{b},c) + p(\overline{c}|a) \cdot p(d|\overline{b},\overline{c})
\end{aligned}$$

Note that we have both p functions defined in the network and auxiliary p functions. Now we evaluate $g(d)$, using g functions that hide t until the last step:

$$\begin{aligned}
g(d) &= p(a) \cdot g(d|a) + p(\overline{a}) \cdot g(d|\overline{a}) \\
g(d|a) &= p(b|a) \cdot g(d|a,b) + p(\overline{b}|a) \cdot g(d|a,\overline{b}) \\
g(d|\overline{a}) &= p(b|\overline{a}) \cdot g(d|\overline{a},b) + p(\overline{b}|\overline{a}) \cdot g(d|\overline{a},\overline{b}) \\
g(d|a,b) &= p(c|a) \cdot g(d|b,c) + p(\overline{c}|a) \cdot g(d|b,\overline{c}) \\
g(d|a,\overline{b}) &= p(c|a) \cdot g(d|\overline{b},c) + p(\overline{c}|a) \cdot g(d|\overline{b},\overline{c}) \\
g(d|\overline{a},b) &= p(c|\overline{a}) \cdot g(d|b,c) + p(\overline{c}|\overline{a}) \cdot g(d|b,\overline{c}) \\
g(d|\overline{a},\overline{b}) &= p(c|\overline{a}) \cdot g(d|\overline{b},c) + p(\overline{c}|\overline{a}) \cdot g(d|\overline{b},\overline{c}) \\
g(d|b,c) &= t \cdot p(d|b,c) \\
g(d|b,\overline{c}) &= t \cdot p(d|b,\overline{c}) \\
g(d|\overline{b},c) &= t \cdot p(d|\overline{b},c) \\
g(d|\overline{b},\overline{c}) &= t \cdot p(d|\overline{b},\overline{c})
\end{aligned}$$

To force t as the variable to maximize/minimize, we impose that $p(a, d) = g(d)$ (remember that $g(d)$ means $t \cdot p(d)$). Now take the last constraint ($g(d|\overline{b},\overline{c}) = t \cdot p(d|\overline{b},\overline{c})$) to illustrate the linearization of a product (the same idea must be applied to all products in all constraints). Suppose that $p(d|\overline{b},\overline{c}) \in [l, u]$, with l and u known. The constraint becomes

$$g(d|\overline{b},\overline{c}) = l \cdot y^{(1)}_{tp(d|\overline{b},\overline{c})} + u \cdot y^{(2)}_{tp(d|\overline{b},\overline{c})}$$

and we include

$$\begin{aligned}
y^{(1)}_{tp(d|\overline{b},\overline{c})} &\geq 0, \\
y^{(1)}_{tp(d|\overline{b},\overline{c})} &\leq b^{(1)}_{d|\overline{b},\overline{c}}, \\
y^{(1)}_{tp(d|\overline{b},\overline{c})} &\geq t - 1 + b^{(1)}_{d|\overline{b},\overline{c}}, \\
y^{(1)}_{tp(d|\overline{b},\overline{c})} &\leq t, \\
y^{(2)}_{tp(d|\overline{b},\overline{c})} &\geq 0, \\
y^{(2)}_{tp(d|\overline{b},\overline{c})} &\leq b^{(2)}_{d|\overline{b},\overline{c}}, \\
y^{(2)}_{tp(d|\overline{b},\overline{c})} &\geq t - 1 + b^{(2)}_{d|\overline{b},\overline{c}}, \\
y^{(2)}_{tp(d|\overline{b},\overline{c})} &\leq t, \\
b^{(1)}_{d|\overline{b},\overline{c}} + b^{(2)}_{d|\overline{b},\overline{c}} &= 1,
\end{aligned}$$

where the new created boolean variables $b^{(i)}_{d|\overline{b},\overline{c}}$ indicate which vertex to use: l or u. The variable $p(d|\overline{b},\overline{c})$ has disappeared (its possible values l and u still remain), and t and new variables $b^{(i)}_{d|\overline{b},\overline{c}}$ appear linearly in the constraints.

Because only one vertex of each local credal set will be chosen, we can go further in the reformulation, obtaining a smaller number of boolean optimization variables. According to the linearization just described, we represent each local credal set as a combination of its vertices and create one boolean optimization variable for each vertex of each local credal set. Instead of this transformation, we can use another idea, interpreting the boolean optimization variables as the binary representation of a vertex index. Suppose $\alpha_0, \ldots, \alpha_{s-1}$ are the vertices and that s is a power of two (we do not loose generality because, if s is not a power of two, we can always repeat several times one of the already existent vertices to reach the next power of two; these additional vertices do not change the result as they are equal to some old vertex). Now let $1 \leq j \leq \log_2 s$ be an integer indexing a bit of the number i, and for each instantiation of variable X with parents $Y_1, \ldots, Y_r$ we define

$$\begin{aligned}
p(x|y_1, \ldots, y_r) = \sum_{i=0}^{s-1} \alpha_i(x) \times \prod_{\text{bit } j \text{ of } i} b^j_{y_1,\ldots,y_r} \times \\
\prod_{\text{not bit } j \text{ of } i} (1 - b^j_{y_1,\ldots,y_r}), \quad (5)
\end{aligned}$$

where (not) bit j of i means that the jth bit of i is (not) one. That is, instead of a boolean variable that indicates (with a zero or one) if a given vertex should be used (and only one of them actually should), we multiply a collection of boolean variables according

to the binary representation of i (the vertex index). This product guarantees that the result is one if and only if all b variables of its binary representation are set to one.

Example 11 *Let X be a random variable with three categories (x_0, x_1, x_2) and one parent, named Z. Let Z have two categories $(z, \overline{z})$. Suppose the credal set for $p(X|z)$ has four vertices $(\alpha_0, \ldots, \alpha_3)$ with three dimensions each. Then we define the boolean optimization variables $b_z^{(1)}, b_z^{(2)}$ and the constraints:*

$$\begin{aligned}
p(x_0|z) &= \alpha_0(x_0)\cdot(1-b_z^{(1)})\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_1(x_0)\cdot b_z^{(1)}\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_2(x_0)\cdot(1-b_z^{(1)})\cdot b_z^{(2)} + \\
&\quad \alpha_3(x_0)\cdot b_z^{(1)}\cdot b_z^{(2)} \\
p(x_1|z) &= \alpha_0(x_1)\cdot(1-b_z^{(1)})\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_1(x_1)\cdot b_z^{(1)}\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_2(x_1)\cdot(1-b_z^{(1)})\cdot b_z^{(2)} + \\
&\quad \alpha_3(x_1)\cdot b_z^{(1)}\cdot b_z^{(2)} \\
p(x_2|z) &= \alpha_0(x_2)\cdot(1-b_z^{(1)})\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_1(x_2)\cdot b_z^{(1)}\cdot(1-b_z^{(2)}) + \\
&\quad \alpha_2(x_2)\cdot(1-b_z^{(1)})\cdot b_z^{(2)} + \\
&\quad \alpha_3(x_2)\cdot b_z^{(1)}\cdot b_z^{(2)}
\end{aligned}$$

After some simple algebraic manipulation of Equations (5), we still have to deal with products of boolean variables. The procedure is straightforward: If $b_1, b_2, \ldots, b_r$ are boolean variables, then the product $\prod_i b_i$ can be replaced by the continuous variable y, with additional constraints:

$$\begin{aligned}
0 \leq\ & y \leq 1 \\
& y \leq b_i, \text{ for all } i \qquad (6) \\
\sum_i b_i - r + 1 \leq\ & y
\end{aligned}$$

The number of boolean optimization variables in the integer programming version is $O(\log_2 \prod_{X \in V} c_X)$, where c_X is the number of categories of the random variable associated to node X. Thus, the reformulation to an integer programming problem is performed by running the `Bilinear-Transformation` algorithm together with the linearization step. The linearization inserts a logarithmic number of new constraints for each constraint generated by `Bilinear-Transformation` (when using the ideas of Expressions (5) and (6)). The number of new boolean optimization variables is small and does not increase the overall complexity of the reformulation. For polytrees, we still have a polynomial time procedure.

5 Computational results

To illustrate the behavior of our methods, we present two sets of experiments. First we deal with test sets containing multi-connected networks (randomly generated using the BNGenerator software [17] or using the topology of the Alarm network [1]). Latter we treat randomly generated polytrees. In each network we perform a belief updating inference with a predefined variable (we have chosen the most challenging variables).

Table 1 shows results of `Bilinear-Transformation` followed by the linearization step for four different network. Rows present type of the network, total number of nodes, number of nodes involved in the inference, number of vertices in the credal sets, resulting continuous optimization variables, resulting boolean variables and resulting optimization constraints. All the tests were done by transforming inferences in multi-connected credal networks into integer programming problems. The chosen inferences represent the most challenging inference for each network. We processed networks with different variables (binary and ternary), and different sizes of credal sets per node of the network. Note that the size of resulting problems (specially the number of boolean variables to optimize) is large. Existing exact optimization solvers usually can not handle such large number of boolean variables, but approximation ideas are still possible. As we can see, the number of integer variables is too high for processing such networks.

Restricting our attention to polytrees, Table 2 presents twenty polytree-shaped credal networks. They have ternary variables and at most three vertices by locally and separately specified credal set. Rows present name of network, total number of nodes, number of nodes involved in the inference, generated continuous optimization variables, generated boolean variables, generated constraints, time for solving the integer programming problem and number of branch-and-bound nodes evaluated by the solver.

Analyzing Table 1 (multi-connected networks) and Table 2 (polytree networks), we see that the algorithm could generate much smaller problems in the latter case. That happens because of the relationship between tree-width and path-width of a polytrees: they are almost the same.

Because the integer programming reformulation is usually less dependant on some convergence criteria and numerical problems than nonlinear programming techniques (such as multilinear programming [8, 9]), the integer programming reformulation achieves good performance together with reliable results. All tests

Network topology	Nodes	Active Nodes	Vertices by credal set	Continuous variables	Boolean variables	Constraints
Dense binary	10	10	2	1523	120	1523
Dense ternary	10	10	3	3954	202	3954
Alarm network	37	24	2	34537	161	34351
Alarm network	37	24	4	51293	322	65075

Table 1: Size of integer programming problems generated from some multi-connected credal networks.

were performed on a Intel Xeon 2.8Ghz (4MB of L2 cache) with 4GB of RAM memory. The integer programming problems were solved using the AMPL modeling language [15, 16] and the CPLEX solver.

6 Conclusion

We have discussed in this paper a new idea for inferences in credal networks. The main contribution is the use of bilinear and integer programming techniques. Although many authors have suggested and worked with multilinear programming as a possible approach to inference, as far as we know no investigation or implementation of bilinear and/or integer programming reformulations have been conducted.

Results produced in our experiments seem promising and surpass existing exact algorithms for inference in polytree-shaped credal networks with respect to the number of network variables that could be dealt [5, 9]. Although there is a polynomial time algorithm for inferences in binary polytrees [13], the problem is `NP-Complete` in general polytrees [10]. So our reformulation is a new idea to address this hard problem, with good empirical performance. Furthermore, integer programming produces outer bounds based on linear programming relaxations, which can be used for approximate procedures. Known approximate techniques, such as cutting planes, can be applied.

Other multilinear programming techniques could certainly be investigated in future work, perhaps combining some ideas from multilinear programming with integer programming. As it happens with multilinear programming, integer programming will fail for large networks; in this case approximate inference is the natural solution. The reformulations presented in this paper also contribute in that direction, as approximations for bilinear and integer programming problems are well studied in the literature.

Acknowledgements

This work has been supported by FAPESP grant 2004/09568-0; the second author is partially supported by CNPq grant 3000183/98-4.

References

[1] I. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *European Conf. on Artificial Intelligence in Medicine*, pages 247–256, Berlin, 1989. Springer-Verlag.

[2] A. Cano, J. E. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 978–983, Paris, France, 1994.

[3] J. Cano, M. Delgado, and S. Moral. An axiomatic framework for propagating uncertainty in directed acyclic networks. *International Journal of Approximate Reasoning*, 8:253–280, 1993.

[4] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk, Decision and Policy*, 5:165–181, 2000.

[5] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[6] F. G. Cozman. Separation properties of sets of probabilities. In In *Conference on Uncertainty in Artificial Intelligence*, pages 107–115, San Francisco, 2000. Morgan Kaufmann.

[7] J. C. F. da Rocha and F. G. Cozman. Inference with separately specified sets of probabilities in credal networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 430–437, San Francisco, 2002. Morgan Kaufmann.

[8] J. C. F. da Rocha, F. G. Cozman, and C. P. de Campos. Inference in polytrees with sets of probabilities. In *Conference on Uncertainty in Artificial Intelligence*, pages 217–224, New York, 2003. Morgan Kaufmann.

[9] C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Second Starting AI Researcher Symposium*, pages 50–61, Valencia, 2004. IOS Press.

Network name	Network nodes	Active nodes	Continuous variables	Boolean variables	Constraints	Solution time (sec)	B&B nodes
poly50v0	50	16	2101	40	2076	6.976	791
poly50v1	50	17	3775	84	3967	7.965	258
poly50v2	50	15	511	50	532	0.187	23
poly50v3	50	15	3244	42	2911	60.224	2743
poly50v4	50	25	1779	33	1996	3.988	977
poly50v5	50	18	1533	38	1582	2.429	935
poly50v6	50	17	1049	41	1169	2.598	1456
poly50v7	50	19	768	20	912	2.332	1307
poly50v8	50	15	2459	53	2569	12.386	708
poly50v9	50	16	775	37	829	0.549	314
poly100v0	100	22	3435	68	3324	48.772	3191
poly100v1	100	28	4101	95	4511	1095.353	89560
poly100v2	100	23	1179	74	1255	628.776	333809
poly100v3	100	21	2540	40	2745	296.844	41915
poly100v4	100	24	1787	63	1837	5.751	3695
poly100v5	100	25	1067	39	1150	8.006	3428
poly100v6	100	25	3340	51	3474	372.833	21352
poly100v7	100	26	672	60	725	0.458	232
poly100v8	100	28	773	47	897	25.146	28860
poly100v9	100	23	10635	57	9853	576.663	2880

Table 2: Tests with random polytree networks.

[10] C. P. de Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. In *International Joint Conference on Artificial Intelligence*, pages 1313–1318, 2005.

[11] C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44:244–260, 2007.

[12] L. de Campos and S. Moral. Independence concepts for convex sets of probabilities. In *Conference on Uncertainty in Artificial Intelligence*, pages 108–115, San Francisco, 1995. Morgan Kaufmann.

[13] E. Fagiuoli and M. Zaffalon. 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[14] U. Feige, M. Hajiaghayi, and J. R. Lee. Improved approximation algorithms for minimum-weight vertex separators. In *ACM Symposium on Theory of computing*, pages 563–572, New York, NY, USA, 2005. ACM Press.

[15] R. Fourer, D. M. Gay, and Brian W. Kernighan. A modeling language for mathematical programming. *Management Science*, 36:519–554, 1990.

[16] R. Fourer, D. M. Gay, and Brian W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Duxbury Press, Brooks/Cole Publishing Company, 2002.

[17] J. S. Ide and F. G. Cozman. Random generation of Bayesian networks. In *Brazilian Symposium on Artificial Intelligence*, pages 366–375, Recife, 2002. Springer-Verlag.

[18] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

[20] N. Robertson and P. D. Seymour. Graph minors – II. Algorithmic aspects of tree-width. *J. Algorithms*, 7:309–322, 1986.

[21] N. Robertson and P.D. Seymour. Graph minors – I. Excluding a forest. *J. Combinatorial Theory Series B*, 35:39–61, 1983.

[22] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.

[23] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Relating practical representations of imprecise probabilities

S. Destercke
Institut de Radioprotection et de sûreté nucléaire (IRSN), Cadarache, France
sebastien.destercke@irsn.fr

D. Dubois
Institut de Recherche en Informatique de Toulouse (IRIT)
Toulouse, France
dubois@irit.fr

E. Chojnacki
Institut de Radioprotection et de sûreté nucléaire (IRSN), Cadarache, France
eric.chojnacki@irsn.fr

Abstract

There exist many practical representations of probability families that make them easier to handle. Among them are random sets, possibility distributions, probability intervals, Ferson's p-boxes and Neumaier's clouds. Both for theoretical and practical considerations, it is important to know whether one representation has the same expressive power than other ones, or can be approximated by other ones. In this paper, we mainly study the relationships between the two latter representations and the three other ones.

Keywords. Random Sets, possibility distributions, probability intervals, p-boxes, clouds.

1 Introduction

There are many representations of uncertainty. The theory of imprecise probabilities (including lower/upper previsions) [27] is the most general framework. It formally encompasses all the representations proposed by other uncertainty theories, regardless of their possible different interpretations.

The more general the theory, the more expressive it can be, and, usually, the more expensive it is from a computational standpoint. Simpler (but less flexible) representations can be useful if judged sufficiently expressive. They are mathematically and computationally easier to handle, and using them can greatly increase efficiency in applications.

Among these simpler representations are random sets [7], possibility distributions [28], probability intervals [2], p-boxes [15] and, more recently, clouds [21, 22]. With such a diversity of simplified representations, it is then natural to compare them from the standpoint of their expressive power. Building formal links between such representations also facilitates a unified handling of uncertainty, especially in propagation techniques exploiting uncertain data modeled by means of such representations. This is the purpose of the present study. It extends some results by Baudrit and Dubois [1] concerning the relationships between p-boxes and possibility measures.

The paper is structured as follows: the first section briefly recalls the formalism of random sets, possibility distributions and probability intervals, as well as some existing results. Section 3 then focuses on p-boxes, first generalizing the notion of p-boxes to arbitrary finite spaces before studying the relationships of these generalized p-boxes with the three former representations. Finally, section 4 studies the relationships between clouds and the preceding representations. For the reader convenience, longer proofs are put in the appendix.

2 Preliminaries

In this paper, we consider that uncertainty is modeled by a family $\mathcal{P}$ of probability assignments, defined over a finite referential $X = \{x_1, \ldots, x_n\}$. We also restrict ourselves to families that can be represented by their lower and upper probability bounds, defined as follows:

$$\underline{P}(A) = \inf_{P \in \mathcal{P}} P(A) \text{ and } \overline{P}(A) = \sup_{P \in \mathcal{P}} P(A)$$

Let $\mathcal{P}_{\underline{P},\overline{P}} = \{P | \forall A \subseteq X, \underline{P}(A) \leq P(A) \leq \overline{P}(A)\}$. In general, we have $\mathcal{P} \subset \mathcal{P}_{\underline{P},\overline{P}}$, since $\mathcal{P}_{\underline{P},\overline{P}}$ can be seen as a projection of $\mathcal{P}$ on events. Although they are already restrictions from more general cases, dealing with families $\mathcal{P}_{\underline{P},\overline{P}}$ often remains difficult.

2.1 Random Sets

Formally, a random set [20] is a mapping Γ from a probability space to the power set $\wp(X)$ of another space X, also called a multi-valued mapping. This mapping induces lower and upper probabilities on X [7]. In the continuous case, the probability space is often $[0, 1]$ equipped with Lebesgue measure, and Γ is a point-to-interval mapping.

In the finite case, these lower and upper probabilities are respectively called belief and plausibility measures, and it can be shown that the belief measure is a ∞-monotone capacity [4]. An alternative (and useful) representation of the random set consists of a normalized assignment of positive masses m over the power set $\wp(X)$ s.t. $\sum_{E\subseteq X} m(E) = 1$ and $m(\emptyset) = 0$ [25]. A set E that receives strictly positive mass is said to be focal. Belief and plausibility functions are then defined as follows:

$$Bel(A)=\sum_{E,E\subseteq A} m(E)$$

$$Pl(A)=1-Bel(A^c)=\sum_{E,E\cap A\neq\emptyset} m(E).$$

The set

$$\mathcal{P}_{Bel} = \{P|\forall A \subseteq X, Bel(A) \leq P(A) \leq Pl(A)\}$$

is the probability family induced by the belief measure.

Although $2^{|X|}$ values are still needed to fully specify a general random set, the fact that they can be seen as probability assignments over subsets of X allows for simulation by means of some sampling process.

2.2 Possibility distributions

A possibility distribution π [12] is a mapping from X to the unit interval such that $\pi(x) = 1$ for some $x \in X$. Formally, a possibility distribution is the membership function of a fuzzy set. Several set-functions can be defined from a distribution π [11]:

- $\Pi(A) = \sup_{x\in A} \pi(x)$ (possibility measures);
- $N(A) = 1 - \Pi(A^c)$ (necessity measures);
- $\Delta(A) = \inf_{x\in A} \pi(x)$ (sufficiency measures).

Possibility degrees express the extent to which an event is plausible, i.e., consistent with a possible state of the world, necessity degrees express the certainty of events and sufficiency (also called guaranteed possibility) measures express the extent to which all states of the world where A occurs are plausible. They apply to so-called guaranteed possibility distributions [11] generally denoted by δ.

A possibility degree can be viewed as an upper bound of a probability degree [13]. Let

$$\mathcal{P}_\pi = \{P, \forall A \subseteq X, N(A) \leq P(A) \leq \Pi(A)\}$$

be the set of probability measures encoded by a possibility distribution π. A possibility distribution is also equivalent to a random set whose realizations are nested.

From a practical standpoint, possibility distributions are the simplest representation of imprecise probabilities (as for precise probabilities, only $|X|$ values are needed to specify them). Another important point is their interpretation in term of collection of confidence intervals [10], which facilitates their elicitation and makes them natural candidate for vague probability assessments (see [5]).

2.3 Probability intervals

Probability intervals are defined as lower and upper probability bounds restricted to singletons x_i. They can be seen as a collection of intervals $L = \{[l_i, u_i], i = 1, \ldots, n\}$ defining a probability family:

$$\mathcal{P}_L = \{P|l_i \leq p(x_i) \leq u_i \ \forall x_i \in X\}.$$

Such families have been extensively studied in [2] by De Campos et al.

In this paper, we consider non-empty families (i.e. $\mathcal{P}_L \neq \emptyset$) that are reachable (i.e. each lower or upper bound on singletons can be reached by at least one probability assignment of the family $\mathcal{P}_L$). Conditions of non-emptiness and reachability respectively correspond to avoiding sure loss and achieving coherence in Walley's behavioural theory.

Given intervals L, lower and upper probabilities $\underline{P}(A), \overline{P}(A)$ are calculated by the following expressions

$$\begin{aligned} \underline{P}(A) &= \max(\textstyle\sum_{x_i\in A} l_i, 1 - \sum_{x_i\notin A} u_i) \\ \overline{P}(A) &= \min(\textstyle\sum_{x_i\in A} u_i, 1 - \sum_{x_i\notin A} l_i) \end{aligned} \qquad (1)$$

De Campos et al. have shown that these bounds are Choquet capacities of order 2 ($\underline{P}$ is a convex capacity).

The problem of approximating $\mathcal{P}_L$ by a random set has been treated in [17] and [8]. While in [17], Lemmer and Kyburg find a random set m_1 that is an inner approximation of $\mathcal{P}_L$ s.t. $Bel_1(x_i) = l_i$ and $Pl_1(x_i) = u_i$, Denoeux [8] extensively studies methods to build a random set that is an outer approximation of $\mathcal{P}_L$. The problem of finding a possibility distribution approximating $\mathcal{P}_L$ is treated by Masson and Denoeux in [19].

Two common cases where probability intervals can be encountered as models of uncertainty are confidence intervals on parameters of multinomial distributions built from sample data, and expert opinions providing such intervals.

3 P-boxes

We first recall some usual notions on the real line that will be generalized in the sequel.

Let Pr be a probability function on the real line with density p. The *cumulative* distribution of Pr is denoted F^p and is defined by $F^p(x) = \Pr((-\infty, x])$.

Let $F_1(x)$ and $F_2(x)$ be two *cumulative* distributions. Then, $F_1(x)$ is said to stochastically dominate $F_2(x)$ iff $F_1(x) \leq F_2(x)\ \forall x$.

A P-box [15] is defined by a pair of cumulative distributions $\underline{F} \leq \overline{F}$ ($\underline{F}$ stochastically dominates $\overline{F}$) on the real line. It brackets the cumulative distribution of an imprecisely known probability function with density p s.t. $\underline{F}(x) \leq F^p(x) \leq \overline{F}(x) \quad \forall x \in \Re$.

3.1 Generalized Cumulative Distributions

Interestingly, the notion of cumulative distribution is based on the existence of the natural ordering of numbers. Consider a probability assignment (probability vector) $\lambda = (\lambda_1 \dots \lambda_n)$ defined over the finite space X; λ_i denotes the probability $\Pr(x_i)$ of the i-th element x_i, and $\sum_{j=1}^{n} \lambda_j = 1$. In this case, no natural notion of cumulative distribution exists. In order to make sense of this notion over X, one must equip it with a complete preordering $\leq_R$, which is a reflexive, complete and transitive relation. An R-downset is of the form $\{x_i : x_i \leq_R x\}$, and denoted $(x]_R$.

Definition 1. *[9] The generalized R-cumulative distribution of a probability assignment λ on a finite, completely preordered set $(X, \leq_R)$ is the function $F_R^\lambda : X \to [0,1]$ defined by $F_R^\lambda(x) = \Pr((x]_R)$.*

The usual notion of stochastic dominance can also be defined for generalized cumulative distributions. Consider another probability assignment $\kappa = (\kappa_1 \dots \kappa_n)$ on X. The corresponding R-dominance relation of λ over κ can be defined by the pointwise inequality $F_R^\lambda < F_R^\kappa$. Clearly, a generalized cumulative distribution can always be considered as a simple one, up to a reordering of elements.

Any generalized cumulative distribution F_R^λ with respect to a complete preorder $\leq_R$ on X, of a probability measure Pr, with assignment λ on X, can also be used as a possibility distribution π_R whose associated measure dominates Pr, i.e. $\max_{x \in A} F_R^\lambda(x) \geq \Pr(A), \forall A \subseteq X$. This is because a (generalized) cumulative distribution is constructed by computing the probabilities of events $\Pr(A)$ in a nested sequence of downsets $(x_i]_R$. [10].

3.2 Generalized p-box

Using the generalizations of the notions of cumulative distributions and of stochastic dominance described in section 3.1, we define a generalized p-box as follows

Definition 2. *A R-P-box on a finite, completely preordered set $(X, \leq_R)$ is a pair of R-cumulative distributions $F_R^\lambda(x)$ and $F_R^\kappa(x)$, s.t. $F_R^\lambda(x) \leq F_R^\kappa(x)$ (i.e. κ is a probability assignment R-dominated by λ)*

The probability family induced by a R-P-box is

$$\mathcal{P}_{p-box} = \{P | \forall x,\ F_R^\lambda(x) \leq F_R(x) \leq F_R^\kappa(x).\}$$

If we choose a relation R with $x_i \leq_R x_j$ iff $i < j$, and, $\forall x_i \in X$, consider the sets $A_i = (x_i]_R$, it comes down to a family of nested confidence sets $\emptyset \subseteq A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subset X$. The family $\mathcal{P}_{p-box}$ can then be represented by the following restrictions on probability measures [9]:

$$\alpha_i \leq P(A_i) \leq \beta_i \qquad i = 1, \dots, n \tag{2}$$

with $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n \leq 1$ and $\beta_1 \leq \beta_2 \leq \dots \leq \beta_n \leq 1$. Choosing $X = \Re$ and $A_i = (-\infty, x_i]$, it is easy to see that we find back the usual definition of P-boxes.

A generalized cumulative distribution being fully specified by $|X|$ values, it follows that $2|X|$ values must be given to completely determine a generalized p-box. Moreover, we can interpret p-boxes as a collection of nested confidence intervals with upper and lower probability bounds (which could come, for example, from expert elicitation). In order to make notation simpler, the upper and lower cumulative distributions will respectively be noted F^*, F_* in the sequel and, unless stated otherwise, we will consider (without loss of generality) the order R s.t. $x_i \leq_R x_j$ iff $i < j$ with the associated nested sets A_i.

3.3 Generalized P-boxes in the setting of possibility theory

Given that sets A_i can be interpreted as nested confidence intervals with upper and lower bounds, it is natural to search a connection with possibility theory, since possibility distributions can be interpreted as a collection of nested confidence intervals (a natural way of expressing expert knowledge). We thus have the following proposition

Proposition 1. *A family $\mathcal{P}_{p-box}$ described by a generalized P-box can be encoded by a pair of possibility distributions π_1, π_2 s.t. $\mathcal{P}_{p-box} = \mathcal{P}_{\pi_1} \cap \mathcal{P}_{\pi_2}$ with $\pi_1(x) = F^*(x)$ and $\pi_2(x) = 1 - F_*(x)$*

Proof of proposition 1. Consider the definition of a generalized p-box and the fact that a generalized cumulative distribution can be used as a possibility distribution π_R dominating the probability distribution Pr (see section 3.1). Then, the set of constraints $(P(A_i) \geq \alpha_i)_{i=1,n}$ from equation (2) generates a possibility distribution π_1 and the set of constraints $(P(A_i^c) \geq 1 - \beta_i)_{i=1,n}$ generates a possibility distribution π_2. Clearly $\mathcal{P}_{p-box} = \mathcal{P}_{\pi_1} \cap \mathcal{P}_{\pi_2}$. □

3.4 Generalized P-boxes are special case of random sets

The following proposition was proved in [9]

Proposition 2. *A family $\mathcal{P}_{p-box}$ described by a generalized P-box can be encoded by a random set s.t. $\mathcal{P}_{p-box} = \mathcal{P}_{Bel}$.*

Algorithm 1: R-P-box → random set

Input: Nested sets $\emptyset, A_1, \ldots, A_n, X$ and bounds α_i, β_i
Output: Equivalent random set
for $k = 1, \ldots, n+1$ **do**
  Build partition $F_i = A_i \setminus A_{i-1}$
Rank α_i, β_i increasingly
for $k = 0, \ldots, 2n+1$ **do**
  Rename α_i, β_i by γ_l s.t.
  $\alpha_0 = \gamma_0 = 0 \leq \gamma_1 \leq \ldots \leq \gamma_l \leq \ldots \leq \gamma_{2n} \leq 1 = \gamma_{2n+1} = \beta_{n+1}$
Define focal set $E_0 = \emptyset$
for $k = 1, \ldots, 2n+1$ **do**
  if $\gamma_{k-1} = \alpha_i$ **then**
    $E_k = E_{k-1} \cup F_{i+1}$
  if $\gamma_{k-1} = \beta_i$ **then**
    $E_k = E_{k-1} \setminus F_i$
  Set $m(E_k) = \gamma_k - \gamma_{k-1}$

Algorithm 1 provides an easy way to build the random set encoding a generalized p-box. It is similar to algorithms given in [16, 24], and extends them to more general spaces. The main idea of the algorithm is to use the fact that a generalized p-box can be seen as a random set whose focal elements are unions of adjacent sets in a partition. Thanks to the nested nature of sets A_i, we can build a partition of X made of $F_i = A_i \setminus A_{i-1}$, and then add or substract consecutive elements of this partition to build the focal sets (of the form $\bigcup_{j \leq i \leq k} F_i$) of the random set equivalent to the generalized p-box. The following example illustrates both the notion of generalized p-box and algorithm 1.

Example 1. *Consider a space X made of six elements $\{x_1, \ldots, x_6\}$ (These elements could be, for instance, successive components on a production line). For various reasons, one can only observe whether the event $A_1 = \{x_1, x_2\}$, $A_2 = \{x_1, x_2, x_3\}$, $A_3 = \{x_1, x_2, x_3, x_4, x_5\}$ or the whole X happens. Suppose an expert must evaluate the likelihood of these events, and only gives us probability intervals :*

$$P(A_1) \in [0, 0.3] \quad P(A_2) \in [0.2, 0.7] \quad P(A_3) \in [0.5, 0.9]$$

So we have a generalized p-box, the order of the elements being determined by the possible observations (notice that we are indifferent to the order of x_1, x_2 and of x_4, x_5). Applying algorithm 1, we have :

$$F_1 = \{x_1, x_2\} \quad F_2 = \{x_3\} \quad F_3 = \{x_4, x_5\} \quad F_4 = \{x_6\}$$

and

$$0(\alpha_0) \leq 0(\alpha_1) \leq 0.2(\alpha_2) \leq 0.3(\beta_1) \leq 0.5(\alpha_3) \leq 0.7(\beta_2) \leq 0.9(\beta_3) \leq 1$$

which gives us the following corresponding random set

$$m(E_1)=m(\{x_1,x_2\})=0 \quad m(E_2)=m(\{x_1,x_2,x_3\})=0.2$$
$$m(E_3)=m(\{x_1,x_2,x_3,x_4,x_5\})=0.1 \quad m(E_4)=m(\{x_3,x_4,x_5\})=0.2$$
$$m(E_5)=m(\{x_3,x_4,x_5,x_6\})=0.2 \quad m(E_6)=m(\{x_4,x_5,x_6\})=0.2$$
$$m(E_7)=m(\{x_6\})=0.1$$

which makes the imprecision of the available information more visible.

3.5 Generalized P-boxes and probability intervals

Provided an order R has been defined on elements x_i, a method to build a p-box from probability intervals L can be easily derived from equations (1). Lower an upper generalized cumulative distributions can be computed as follows

$$F_*(x_i) = \underline{P}(A_i) = \max\Big(\sum_{x_i \in A_i} l_j, 1 - \sum_{x_i \notin A_i} u_j\Big)$$
$$F^*(x_i) = \overline{P}(A_i) = \min\Big(\sum_{x_i \in A_i} u_i, 1 - \sum_{x_i \notin A_i} l_i\Big) \qquad (3)$$

Transforming a p-box into probability intervals is also an easy task. First, let us assume that each element F_i of the partition used in algorithm 1 is reduced to a singleton x_i. Corresponding probability intervals are then given by the two following formulas:

$$\underline{P}(F_i) = \underline{P}(x_i) = l_i = \max(0, \alpha_i - \beta_{i-1})$$
$$\overline{P}(F_i) = \overline{P}(x_i) = u_i = \beta_i - \alpha_{i-1}$$

if a set F_i is made of n elements $x_{i1}, \ldots, x_{in}$, it is easy to see that $l(x_{ij}) = 0$ and that $u(x_{ij}) = \overline{P}(F_i)$, since $x_{ij} \in F_i$.

Let us note that transforming probability intervals into p-boxes (and conversely) generally loses information, except in the degenerated cases of precise probability assignment and of total ignorance. If no obvious

order relation R between elements x_i is to be privileged, and if one wants to transform probability intervals into generalized p-boxes, we think that a good choice for the order R is the one s.t.

$$\sum_{i=1}^{n} F^*(x_i) - F_*(x_i)$$

is minimized, so that a minimal amount of information is lost in the process.

Another interesting fact to pinpoint is that both cumulative distributions given by equations (3) can be interpreted as possibility distributions dominating the family $\mathcal{P}_L$ (for F_*, the associated possibility distribution is $1 - F_*$). Thus, computing either F^* or F_* is a method to find a possibility distribution approximating $\mathcal{P}_L$, which is different from the one proposed by Masson and Denoeux [19].

4 Clouds

We begin this section by recalling basic definitions and results due to Neumaier [21], cast in the terminology of fuzzy sets and possibility theory. A *cloud* is an Interval-Valued Fuzzy Set F such that $(0,1) \subseteq \cup_{x \in X} F(x) \subseteq [0,1]$, where $F(x)$ is an interval $[\delta(x), \pi(x)]$. In the following, it is either defined on a finite space X, or it is a continuous interval-valued fuzzy interval (IVFI) on the real line (a "cloudy" interval). In the latter case each fuzzy set has cuts that are closed intervals. When the upper membership function coincides with the lower one, ($\delta = \pi$) the cloud is called *thin*, and when the lower membership function is identically 0, the cloud is called *fuzzy* by Neumaier. Let us note that these names are somewhat counter-intuitive, since a *thin* cloud correspond to a fuzzy set with precise membership function, while a fuzzy cloud is equivalent to a probability family modeled by a possibility distribution.

A random variable x with values in X is said to belong to a cloud F if and only if $\forall \alpha \in [0,1]$:

$$P(\delta(x) \geq \alpha) \leq 1 - \alpha \leq P(\pi(x) > \alpha) \qquad (4)$$

under all suitable measurability assumptions.

If X is a finite space of cardinality n, a *cloud* can be defined by the following restrictions :

$$P(B_i) \leq 1 - \alpha_i \leq P(A_i) \text{ and } B_i \subseteq A_i, \qquad (5)$$

where $1 = \alpha_0 > \alpha_1 > \alpha_2 > \ldots > \alpha_n > \alpha_{n+1} = 0$ and $\emptyset = A_0 \subset A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq A_{n+1} = X; \emptyset = B_0 \subseteq B_1 \subseteq B_2 \subseteq \ldots \subseteq B_n \subseteq B_{n+1} = X$. The confidence sets A_i and B_i are respectively the strong and regular α-cut of fuzzy sets π and δ ($A_i = \{x_i, \pi(x_i) > \alpha_{i+1}\}$ and $B_i = \{x_i, \delta(x_i) \geq \alpha_{i+1}\}$).

As for probability intervals and p-boxes, eliciting a cloud requires $2|X|$ values.

4.1 Clouds in the setting of possibility theory

Let us first recall the following result regarding possibility measures (see [10]):

Proposition 3. *$P \in \mathcal{P}_\pi$ if and only if $1 - \alpha \leq P(\pi(x) > \alpha), \forall \alpha \in (0,1]$*

The following proposition directly follows

Proposition 4. *A probability family $\mathcal{P}_{\delta,\pi}$ described by the cloud (δ, π) is equivalent to the family $\mathcal{P}_\pi \cap \mathcal{P}_{1-\delta}$ described by the two possibility distributions π and $1 - \delta$.*

Proof of proposition 4. Consider a cloud (δ, π), and define $\overline{\pi} = 1 - \delta$. Note that $P(\delta(x) \geq \alpha) \leq 1 - \alpha$ is equivalent to $P(\overline{\pi} > \beta) \geq 1 - \beta$, letting $\beta = 1 - \alpha$. So it is clear from equation (4) that probability measure P is in the cloud (δ, π) if and only if it is in $\mathcal{P}_\pi \cap \mathcal{P}_{\overline{\pi}}$. So a cloud is a family of probabilities dominated by two possibility distributions (see [14]) . □

This property is common to generalized p-boxes and clouds: they define probability families upper bounded by two possibility measures. It is then natural to investigate their relationships.

4.2 Finding clouds that are generalized p-boxes

Proposition 5. *A cloud is a generalized p-box iff $\{A_i, B_i, i = 1, \ldots, n\}$ form a nested sequence of sets (i.e. there is a linear preordering with respect to inclusion)*

Proof of proposition 5. Assume the sets A_i and B_j form a globally nested sequence whose current element is C_k. Then the set of constraints defining a cloud can be rewritten in the form $\gamma_k \leq P(C_k) \leq \beta_k$, where $\gamma_k = 1 - \alpha_i$ and $\beta_k = \min\{1 - \alpha_j : A_i \subseteq B_j\}$ if $C_k = A_i$; $\beta_k = 1 - \alpha_i$ and $\gamma_k = \max\{1 - \alpha_j : A_j \subseteq B_i\}$ if $C_k = B_i$.

Since $1 = \alpha_0 > \alpha_1 > \ldots > \alpha_n < \alpha_{n+1} = 0$, these constraints are equivalent to those of a generalized p-box. But if $\exists\ B_j, A_i$ with $j > i$ s.t. $B_j \not\subset A_i$ and $A_i \not\subset B_j$, then the cloud is not equivalent to a p-box, since confidence sets would no more form a complete preordering with respect to inclusion. □

In term of pairs of possibility distributions, it is now easy to see that a cloud (δ, π) is a generalized p-box

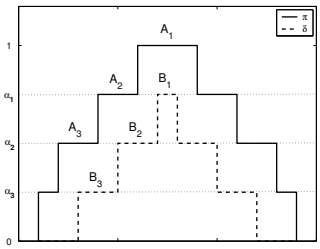


Figure 1: Comonotonic cloud

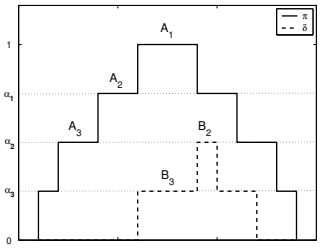


Figure 2: Non-Comonotonic cloud

if and only if π and δ are comonotonic. We will thus call such clouds comonotonic clouds. If a cloud is comonotonic, we can thus directly adapt the various results obtained for generalized p-boxes. In particular, because comonotonic clouds are generalized p-boxes, algorithm 1 can be used to get the corresponding random set. Notions of comonotonic and non-comonotonic clouds are respectively illustrated by figures 1 and 2

4.3 Characterizing and approximating non-comonotonic clouds

The following proposition characterizes probability families represented by most non-comonotonic clouds, showing that the distinction between comonotonic and non-comonotonic clouds makes sense (since the latter cannot be represented by random sets).

Proposition 6. *If (δ, π) is a non-comonotonic cloud for which there are two overlapping sets A_i, B_j that are not nested (i.e. $A_i \cap B_j \neq \{A_i, B_j, \emptyset\}$), then the lower probability of the induced family $\mathcal{P}_{\delta,\pi}$ is not even 2-monotone.*

The proof can be found in the appendix.

Remark 1. *The case for which we have $B_j \cap A_i \in \{A_i, B_j\}$ for all pairs A_i, B_j is the case of comonotonic clouds. Now, if a cloud is such that for all pairs A_i, B_j : $B_j \cap A_i \in \{A_i, B_j, \emptyset\}$ with at least one empty intersection, then it is still a random set, but no longer a generalized p-box. Let us note that this special case can only occur for discrete clouds.*

Since it can be computationally difficult to work with capacities that are not 2-monotone, one could wish to work either with outer or inner approximations. We propose two such approximations, which are easy to compute and respectively correspond to necessity (possibility) measures and belief (plausibility) measures.

Proposition 7. *If $\mathcal{P}_{\delta,\pi}$ is the probability family described by the cloud (δ, π) on a referential X, then, the following bounds provide an outer approximation of the range of $P(A)$:*

$$\max(N_\pi(A), N_{1-\delta}(A)) \leq P(A) \leq \min(\Pi_\pi(A), \Pi_{1-\delta}(A)) \; \forall A \subset X \quad (6)$$

Proof of proposition 7. Since we have that $\mathcal{P}_{\delta,\pi} = \mathcal{P}_{1-\delta} \cap \mathcal{P}_\pi$, and given the bounds defined by each possibility distributions, it is clear that equation 6 give bounds of $P(A)$. □

We can check that the bounds given by equation (6) are the one considered by Neumaier in [21]. Since these bounds are, in general, not the infinimum and supremum of $P(A)$ on $\mathcal{P}_{\delta,\pi}$, Neumaier's claim that clouds are only vaguely related to Walley's previsions or random sets is not surprising. Nevertheless, if we consider the relationship between clouds and possibility distributions, taking this outer approximation, that is very easy to compute, seems very natural.

Nevertheless, these bounds are not, in general, the infinimum and the supremum of $P(A)$ over $\mathcal{P}_{\delta,\pi}$. To see this, consider a discrete cloud made of four non-empty elements A_1, A_2, B_1, B_2. It can be checked that

$$\begin{aligned} \pi(x) &= 1 \text{ if } x \in A_1; \\ &= \alpha_1 \text{ if } x \in A_2 \setminus A_1; \\ &= \alpha_2 \text{ if } x \notin A_2. \\ \delta(x) &= \alpha_1 \text{ if } x \in B_1; \\ &= \alpha_2 \text{ if } x \in B_2 \setminus B_1; \\ &= 0 \text{ if } x \notin B_2. \end{aligned}$$

Since $P(A_2) \geq 1 - \alpha_2$ and $P(B_1) \leq 1 - \alpha_1$, from (5), we can easily check that $\underline{P}(A_2 \setminus B_1) = \underline{P}(A_2 \cap B_1^c) = \alpha_1 - \alpha_2$. Now, $N_\pi(A_2 \cap B_1^c) = \min(N_\pi(A_2), N_\pi(B_1^c)) = 0$ since $\Pi_\pi(B_1) = 1$ and $B_1 \subseteq A_1$. Considering distribution δ, we can have $N_{1-\delta}(A_2 \cap B_1^c) = \min(N_{1-\delta}(A_2), N_{1-\delta}(B_1^c)) = 0$ since $N_{1-\delta}(A_2) = \Delta_\delta(A_2^c) = 0$ since $B_2 \subseteq A_2$. Equation (6) can thus result in a trivial lower bound, different from $\underline{P}(A_2 \setminus B_1)$.

The next proposition provides an inner approximation of $\mathcal{P}_{\delta,\pi}$

Proposition 8. *Given the sets $\{B_i, A_i, i = 1, \ldots, n\}$ inducing the distributions (δ, π) of a cloud and the*

corresponding α_i, the belief and plausibility measures of the random set s.t. $m(A_i \setminus B_{i-1}) = \alpha_{i-1} - \alpha_i$ are inner approximations of $\mathcal{P}_{\delta,\pi}$.

It is easy to see that this random set can always be defined. We can see that it is always an inner approximation by using the contingency matrix advocated in the proof of proposition 6 (see appendix). In this matrix, the random set defined above comes down to concentrating weights on diagonal elements. This inner approximation is exact in case of comonotonicity or when we have $A_i \cap B_j \in \{A_i, B_j, \emptyset\}$ for any pair of sets A_i, B_j defining the clouds.

4.4 A note on thin and continuous clouds

Thin clouds ($\delta = \pi$) constitute an interesting special case of clouds. In this latter case, conditions defining clouds are reduced to

$$P(\pi(x) \geq \alpha) = P(\pi(x) > \alpha) = 1 - \alpha, \forall \alpha \in (0, 1).$$

On finite sets these constraints are generally contradictory, because $P(\pi(x) \geq \alpha) > P(\pi(x) > \alpha)$ for some α, hence the following theorem:

Proposition 9. *If X is finite, then $\mathcal{P}(\pi) \cap \mathcal{P}(1 - \pi)$ is empty.*

which is proved in [14], where it is also shown that this emptiness is due to finiteness. A simple shift of indices solves the difficulty. Let $\pi(u_i) = \alpha_i$ such that $\alpha_1 = 1 > \ldots > \alpha_n > \alpha_{n+1} = 0$. Consider $\delta(u_i) = \alpha_{i+1} < \pi_1(u_i)$. Then $\mathcal{P}(\pi) \cap \mathcal{P}(1 - \delta)$ contains the unique probability measure P such that the probability weight attached to u_i is $p_i = \alpha_i - \alpha_{i+1}, \forall i = 1 \ldots n$. To see it, refer to equation (5), and note that in this case $A_i = B_i$.

In the continuous case, a thin cloud is non-trivial. The inclusions $[\delta(x) \geq \alpha] \subseteq [\pi(x) > \alpha]$ (corresponding to $B_i \subseteq A_i$) again do not work but we may have $P(\pi(x) \geq \alpha) = P(\pi(x) > \alpha) = 1 - \alpha, \forall \alpha \in (0, 1)$. For instance, a cumulative distribution function, viewed as a tight p-box, defines a thin cloud containing the only random variable having this cumulative distribution (the "right" side of the cloud is rejected to ∞). In fact, it was suggested in [14] that a thin cloud contains in general an infinity of probability distributions.

Insofar as Proposition 5 can be extended to the reals (this could be shown, for instance, by proving the convergence of some finite outer and inner approximations of the continuous model, or by using the notion of directed set [5] to prove the complete monotonicity of the model), then a thin cloud can be viewed as a generalized p-box and is thus a (continuous) belief function with uniform mass density, whose focal sets are doubletons of the form $\{x(\alpha), y(\alpha)\}$ where $\{x : \pi(x) \geq \alpha\} = [x(\alpha), y(\alpha)]$. It is defined by the Lebesgue measure on the unit interval and the multimapping $\alpha \longrightarrow \{x(\alpha), y(\alpha)\}$. This result gives us a nice way to characterize the infinite set of random variables contained in a thin cloud. In particular, concentrating the mass density on elements $x(\alpha)$ or on elements $y(\alpha)$ would respectively give the upper and lower cumulative distributions that would have been associated to the possibility distribution π alone (let us note that every convex mixture of those two cumulative distributions would also be in the thin cloud). It is also clear that $Bel(\pi(x) \geq \alpha) = 1 - \alpha$. More generally, if Proposition 5 holds in the continuous case, a comonotonic cloud can be characterized by a continuous belief function [26] with uniform mass density, whose focal sets would be unions of disjoint intervals of the form $[x(\alpha), u(\alpha)] \cup [v(\alpha), y(\alpha)]$ where $\{x : \pi(x) \geq \alpha\} = [x(\alpha), y(\alpha)]$ and $\{x : \delta(x) \geq \alpha\} = [u(\alpha), v(\alpha)]$.

4.5 Clouds and probability intervals

Since probability intervals are 2-monotone capacities, while clouds are either ∞-monotone capacities or not even 2-monotone capacities, there is no direct correspondence between probability intervals and clouds. Nevertheless, given previous results, we can easily build a cloud approximating a family $\mathcal{P}_L$ defined by a set L of probability intervals (but perhaps not the most "specific" one): indeed, any generalized p-box built from the probability intervals is a comonotonic cloud encompassing the family $\mathcal{P}_L$.

Finding the "best" (i.e. keeping as much information as possible, given some information measure) method to transform probability intervals into cloud is an open problem. Any such transformation in the finite case should follow some basic requirements such as:

1. Since clouds can model precise probability assignments, the method should insure that a precise probability assignment will be transformed into the corresponding (almost thin) cloud.

2. Given a set L of probability intervals, the transformed cloud $[\delta, \pi]$ should contain $\mathcal{P}_L$ (i.e. $\mathcal{P}_{\delta,\pi} \subset \mathcal{P}_L$) while being as close to it as possible.

Let us note that using the transformation proposed in section 3.5 for generalized p-boxes satisfies these two requirements. Another solution is to extend Masson and Denoeux's [19] method that builds a possibility distribution covering a set of probability intervals, completing it by a lower distribution δ (due to lack of space, we do not explore this alternative here).

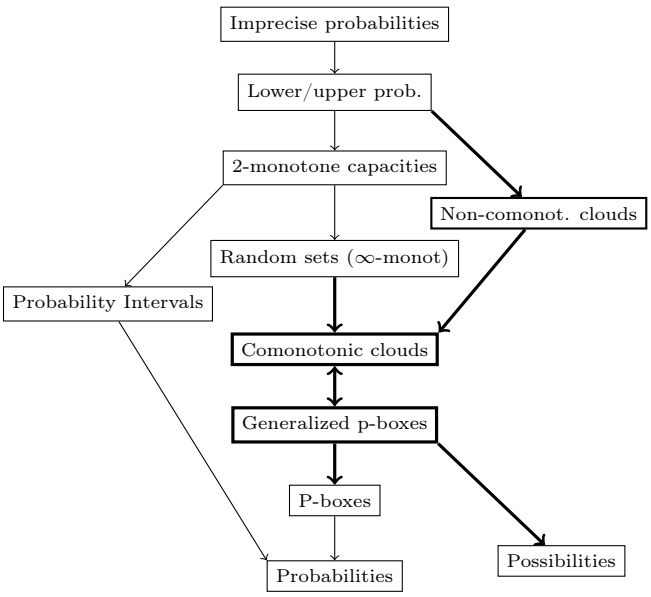


Figure 3: Representations relationships. $A \longrightarrow B$: B is a special case of A

5 Conclusions

Figure 3 summarizes our results cast in a more general framework of imprecise probability representations (our main contributions in boldface).

In this paper, we have considered many practical representations of imprecise probabilities, which are easier to handle than general probability families. They often require less data to be fully specified and they allow many mathematical simplifications, which may prove to increase computational efficiency (except, perhaps, for non-comonotonic clouds).

Some clarifications are provided concerning the situation of the cloud formalism. The fact that non-comonotonic clouds are not even 2-monotone capacities tends to indicate that, from a computational standpoint, they may be more difficult to exploit than the other formalisms. Nevertheless, as far as we know, they are the only simple model generating capacities that are not 2-monotone.

A work that remains to be done to a large extent is to evaluate the validity and the usefulness of these representations, particularly from a psychological standpoint (even if some of it has already been done [23, 18]). Another issue is to extend presented results to continuous spaces or to general lower/upper previsions (by using results from, for example [26, 6]). Finally, a natural continuation to this work is to explore various aspects of each formalisms in a manner similar to the one of De campos et al. [2]. What becomes of random sets, possibility distributions, generalized p-boxes and clouds after fusion, marginalization, conditioning or propagation? Do they preserve the representation? and under which assumptions ? To what extent are these representations informative ? Can they easily be elicited or integrated ? If many results already exist for random sets and possibility distributions, there are fewer results for generalized p-boxes or clouds, due to their novelty.

Acknowledgements

This paper has been supported by a grant from the Institut de Radioprotection et de Sûreté Nucléaire (IRSN). Scientific responsibility rests with the authors.

References

[1] C. Baudrit and D. Dubois. Practical representations of incomplete probabilistic knowledge. *Computational Statistics and Data Analysis*, 51(1):86–108, 2006.

[2] L. de Campos, J. Huete, and S. Moral. Probability intervals : a tool for uncertain reasoning. *I. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.

[3] A. Chateauneuf. Combination of compatible belief functions and relation of specificity. In *Advances in the Dempster-Shafer theory of evidence*, pages 97–114. John Wiley & Sons, Inc, New York, NY, USA, 1994.

[4] G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[5] G. de Cooman. A behavioural model for vague probability assessments. *Fuzzy sets and systems*, 154:305–358, 2005.

[6] G. de Cooman, M. Troffaes, and E. Miranda. n-monotone lower previsions and lower integrals. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *Proc. 4th International Symposium on Imprecise Probabilities and Their Applications*, 2005.

[7] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[8] T. Denoeux. Constructing belief functions from sample data using multinomial confidence regions. *I. J. of Approximate Reasoning*, 42, 2006.

[9] S. Destercke and D. Dubois. A unified view of some representations of imprecise probabilities.

In J. Lawry, E. Miranda, A. Bugarin, and S. Li, editors, *Int. Conf. on Soft Methods in Probability and Statistics (SMPS)*, Advances in Soft Computing, pages 249–257, Bristol, 2006. Springer.

[10] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing*, 10:273–297, 2004.

[11] D. Dubois, P. Hajek, and H. Prade. Knowledge-driven versus data-driven logics. *Journal of logic, Language and information*, 9:65–89, 2000.

[12] D. Dubois and H. Prade. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press, 1988.

[13] D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

[14] D. Dubois and H. Prade. Interval-valued fuzzy sets, possibility theory and imprecise probability. In *Proceedings of International Conference in Fuzzy Logic and Technology (EUSFLAT'05)*, Barcelona, September 2005.

[15] S. Ferson, L. Ginzburg, V. Kreinovich, D. Myers, and K. Sentz. Construction probability boxes and dempster-shafer structures. Technical report, Sandia National Laboratories, 2003.

[16] E. Kriegler and H. Held. Utilizing belief functions for the estimation of future climate change. *I. J. of Approximate Reasoning*, 39:185–209, 2005.

[17] J. Lemmer and H. Kyburg. Conditions for the existence of belief functions corresponding to intervals of belief. In *Proc. 9th National Conference on A.I.*, pages 488–493, 1991.

[18] G. N. Linz and F. C. de Souza. A protocol for the elicitation of imprecise probabilities. In *Proceedings 4th International Symposium on Imprecise Probabilities and their Applications*, Pittsburgh, 2005.

[19] M. Masson and T. Denoeux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, february 2006.

[20] I. Molchanov. *Theory of Random Sets*. Springer, 2005.

[21] A. Neumaier. Clouds, fuzzy sets and probability intervals. *Reliable Computing*, 10:249–272, 2004.

[22] A. Neumaier. On the structure of clouds. available on www.mat.univie.ac.at/∼neum, 2004.

[23] E. Raufaste, R. Neves, and C. Mariné. Testing the descriptive validity of possibility theory in human judgments of uncertainty. *Artificial Intelligence*, 148:197–218, 2003.

[24] H. Regan, S. Ferson, and D. Berleant. Equivalence of methods for uncertainty propagation of real-valued random variables. *I. J. of Approximate Reasoning*, 36:1–30, 2004.

[25] G. Shafer. *A mathematical Theory of Evidence*. Princeton University Press, 1976.

[26] P. Smets. Belief functions on real numbers. *I. J. of Approximate Reasoning*, 40:181–223, 2005.

[27] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, 1991.

[28] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1:3–28, 1978.

Appendix

Proof of proposition 6 *(sketch).* Our proof uses the following result by Chateauneuf [3]: Let m_1,m_2 be two random sets with focal sets $\mathcal{F}_1, \mathcal{F}_2$, each of them respectively defining a probability family $\mathcal{P}_{Bel_1}, \mathcal{P}_{Bel_2}$. Here, we assume that those families are "compatible" (i.e. $\mathcal{P}_{Bel_1} \cap \mathcal{P}_{Bel_2} \neq \emptyset$).

Then, the result from Chateauneuf states the following : the lower probability $\underline{P}(E)$ of the event E on $\mathcal{P}_{Bel_1} \cap \mathcal{P}_{Bel_2}$ is equal to the least belief measure $Bel(E)$ that can be computed on the set of joint normalized random sets with marginals m_1,m_2. More formally, let us consider a set $\mathcal{Q}$ s.t. $Q \in \mathcal{Q}$ iff

- $Q(A,B) > 0 \Rightarrow A \times B \in \mathcal{F}_1 \times \mathcal{F}_2$ (masses over the cartesian product of focal sets)
- $A \cap B = \emptyset \Rightarrow Q(A,B) = 0$ (normalization constraints)
- $m_1(A) = \sum_{B \in \mathcal{F}_2} Q(A,B)$ and $m_2(B) = \sum_{A \in \mathcal{F}_1} Q(A,B)$ (marginal constraints)

and the lower probability $\underline{P}(E)$ is given by the following equation

$$\underline{P}(E) = \min_{Q \in \mathcal{Q}} \sum_{(A \cap B) \subseteq E} Q(A,B) \qquad (7)$$

where $\mathcal{Q}$ is the set of joint normalized random sets. This result can be applied to clouds, since the family described by a cloud is the intersection of two families modeled by possibility distributions.

To illustrate the general proof, we will restrict ourselves to a 4-set cloud (the most simple non-trivial cloud that can be found). We thus consider four sets A_1, A_2, B_1, B_2 s.t. $A_1 \subset A_2$,$B_1 \subset B_2$,$B_i \subset A_i$ together with two values α_1, α_2 s.t. $1(=\alpha_0) > \alpha_1 > \alpha_2 > 0(=\alpha_3)$ and the cloud is defined by enforcing the inequalities $P(B_i) \leq 1-\alpha_i \leq P(A_i)$ $i = 1, 2$. The random sets equivalent to the possibility distributions $\pi, 1-\delta$ are summarized in the following table:

π	$1-\delta$
$m(A_1) = 1-\alpha_1$	$m(B_0^c = X) = 1-\alpha_1$
$m(A_2) = \alpha_1 - \alpha_2$	$m(B_1^c) = \alpha_1 - \alpha_2$
$m(A_3 = X) = \alpha_2$	$m(B_2^c) = \alpha_2$

Furthermore, we add the constraint $A_1 \cap B_2 \neq \{A_1, B_2, \emptyset\}$, related to the non-monotonicity of the cloud. We then have the following contingency matrix, where the mass m_{ij} is assigned to the intersection of the corresponding sets at the beginning of line i and the top of column j:

	$B_0^c = X$	B_1^c	B_2^c	$\sum$
A_1	m_{11}	m_{12}	m_{13}	$1-\alpha_1$
A_2	m_{21}	m_{22}	m_{23}	$\alpha_1 - \alpha_2$
$A_3 = X$	m_{31}	m_{32}	m_{33}	α_2
$\sum$	$1-\alpha_1$	$\alpha_1 - \alpha_2$	α_2	1

We now consider the four events $A_1, B_2^c, A_1 \cap B_2^c, A_1 \cup B_2^c$. Given the above contingency matrix, we immediately have $\underline{P}(A_1) = 1-\alpha_1$ and $\underline{P}(B_2^c) = \alpha_2$, since A_1 only includes the (joint) focal sets in the first line and B_2^c in the third column.

It is also easy to see that $\underline{P}(A_1 \cap B_2^c) = 0$, by considering the mass assignment $m_{ii} = \alpha_{i-1} - \alpha_i$ (we then have $m_{13} = 0$, which is the mass of the only joint focal set included in $A_1 \cap B_2^c$).

Now, concerning $\underline{P}(A_1 \cup B_2^c)$, let us consider the following mass assignment:

$$\begin{aligned} A_2 \cap B_1^c: &\quad m_{22} = \alpha_1 - \alpha_2 \\ A_3 \cap B_0^c: &\quad m_{31} = \min(1-\alpha_1, \alpha_2) \\ A_1 \cap B_0^c: &\quad m_{11} = 1 - \alpha_1 - m_{31} \\ A_3 \cap B_2^c: &\quad m_{33} = \alpha_2 - m_{31} \\ A_1 \cap B_2^c: &\quad m_{13} = m_{31} \end{aligned}$$

it can be checked that this mass assignment satisfies the constraints of the contingency matrix, and that the only joint focal sets included in $A_1 \cup B_2^c$ are those with masses m_{11}, m_{33}, m_{13}. Summing these masses, we have $\underline{P}(A_1 \cup B_2^c) = \max(\alpha_2, 1-\alpha_1)$. Hence:

$$\begin{aligned} \underline{P}(A_1 \cup B_2^c) + \underline{P}(A_1 \cap B_2^c) &< \underline{P}(B_2^c) + \underline{P}(A_1) \\ \max(\alpha_2, 1-\alpha_1) &< 1 - \alpha_1 + \alpha_2 \end{aligned}$$

an inequality that clearly violates the 2-monotonicity property. We have thus shown that in the 4-set case, 2-monotonocity never holds for families modeled by non-comonotonic clouds.

Now, in the general case, we have the following contingency matrix

	B_0^c	$\cdot$	B_j^c	$\cdot$	B_n^c	$\sum$
A_1	m_{11}					$1-\alpha_1$
$\cdot$		$\cdot$				
A_i			$m_{i(j+1)}$			$\alpha_{i-1} - \alpha_i$
$\cdot$				$\cdot$		
A_{n+1}					$m_{(n+1)(n+1)}$	α_n
$\sum$	$1-$		α_j-		α_n	
	α_1		α_{j+1}			

Under the hypothesis of proposition 6, there are two sets A_i, B_j s.t. $A_i \cap B_j \neq \{A_i, B_j, \emptyset\}$. Due to the inclusion relationships between the sets, and similarly to what was done in the 4-set case, we have

$$\begin{aligned} \underline{P}(A_i) &= 1 - \alpha_i \\ \underline{P}(B_j^c) &= \alpha_j \\ \underline{P}(A_i \cap B_j^c) &= 0 \end{aligned}$$

Next, let us concentrate on event $A_i \cup B_j^c$ (which is different from X by hypothesis). Let us suppose that $m_{kk} = \alpha_{k-1} - \alpha_k$, except for masses $m_{(j+1)i}, m_{ii}, m_{i(j+1)}, m_{(j+1)(j+1)}$. This is similar to the 4-set case with masses $m_{(j+1)i}, m_{ii}, m_{i(j+1)}, m_{(j+1)(j+1)}$ and we define the following assignment

$$\begin{aligned} A_i \cap B_j^c: &\quad m_{i(j+1)} = \min(\alpha_{i-1} - \alpha_i, \alpha_j - \alpha_{j+1}) \\ A_i \cap B_{i-1}^c: &\quad m_{ii} = \alpha_{i-1} - \alpha_i - m_{i(j+1)} \\ A_{j+1} \cap B_j^c: &\quad m_{(j+1)(j+1)} = \alpha_j - \alpha_{j+1} - m_{i(j+1)} \\ A_{j+1} \cap B_{i-1}^c: &\quad m_{(j+1)i} = \min(\alpha_{i-1} - \alpha_i, \alpha_j - \alpha_{j+1}) \end{aligned}$$

Given this specific mass assignment (which is always inside the set $\mathcal{Q}$), and summing assignments given to subsets of $A_i \cup B_j^c$, the following inequality results:

$$\underline{P}(A_i \cup B_j^c) \leq \max(\alpha_{j+1} + 1 - \alpha_i, \alpha_j + 1 - \alpha_{i-1})$$

so,

$$\underline{P}(A_i \cup B_j^c) + \underline{P}(A_i \cap B_j^c) < \underline{P}(A_i) + \underline{P}(B_j^c),$$

which clearly violates the 2-monotonicity property.

□

Coherence and Fuzzy Reasoning

Serena Doria
Department of Geotechnologies,
University G.D'Annunzio
Chieti, Italy
s.doria@dst.unich.it

Abstract

Upper and lower conditional previsions are defined by the Choquet integral with respect to the Hausdorff outer and inner measures when the conditioning events have positive and finite Hausdorff outer or inner measures in their dimension; otherwise, when conditioning events have infinite or zero Hausdorff outer or inner measures in their dimension, they are defined by a 0-1 valued finitely, but not countably additive probability. It is proven that, if we consider the restriction of the (outer) Haudorff measures to the Borel σ-field, these (upper) conditional and unconditional previsions satisfy the disintegration property in the sense of Dubins with respect to all countable partitions of Ω. This result is obtained as a consequence of the fact that non-disintegrability characterizes finitely as opposed to countably additive probability. Moreover upper and lower conditional previsions are proven to be coherent, in the sense of Walley, with the unconditional previsions.
Properties related to the coherence of upper conditional probabilities are extended to the case where information is represented by fuzzy sets. In particular, given an infinite set Ω, a conditioning rule for possibility distribution is proposed so that it is coherent and it is coherent with the unconditional possibility distribution.
Through this conditional possibility distribution, a conditional possibility measure with respect to the partition of all singletons of [0,1] is defined. It is proved it satisfies the conglomerative principle of de Finetti.

Keywords. Upper and lower conditional previsions, Hausdorff outer and inner measures, disintegration property, fuzzy reasoning, conditional possibility distribution.

1 Introduction

Fuzzy reasoning has been introduced as a tool to handle vague and ambiguous information about linguistic or numerical variables. In [2],[16],[17] probabilistic and fuzzy reasoning are compared. The common aim is to extend the conditions of coherence, which characterize upper and lower conditional previsions to uncertainty measures used to manage vague and ambiguous information, represented by fuzzy sets.
In this paper two different problems are considered; firstly we continue the research about the possibility to define coherent upper and lower conditional probabilities by a class of Hausdorff outer and inner measures. In particular, when the conditioning event has positive and finite Hausdorf outer (inner) measure in its dimension, upper (lower) conditional previsions are defined by the Choquet integral ([5]) with respect to the outer (inner) Hausdorff measures, which are particular examples of monotone set functions. Otherwise, when the conditioning event has Hausdorff outer (inner) measure in its dimension equal to zero or infinity, upper (lower) conditional previsions are defined by a 0-1 valued finitely but not countably additive probability.
Moreover when we consider the restriction of the (outer) Haudorff measures to the Borel σ-field (upper) conditional and unconditional previsions are proven to satisfy the disintegration property in the sense of Dubins with respect to all countable partitions of Ω and to be coherent in the sense of Walley.
The second problem analysed in this paper is a comparison between probabilistic and fuzzy reasoning.
If upper and lower conditional previsions are defined with respect to outer and inner Hausdorff measures some properties are assured. We focus the attention on the property (P1 Section 2), which assures that coherent conditional probability is an uncertainty measure able to manage precise information represented by the singletons and on the disintegration property.
If we represent information by fuzzy sets and partial knowledge by conditional possibility measures do we lose these properties assured by the coherence?
In Section 5 of this paper we define a conditional possibility distribution on an infinite set Ω that is coherent and coherent with respect to the unconditional

possibility distribution. Moreover, through this conditional possibility distribution we obtain a possibility conditional measure with respect to the partition of all singletons that is coherent and that satisfies the (weak) conglomerative principle of de Finetti.

2 Upper and Lower Conditional Previsions Separately Coherent and Coherent with respect to the Unconditional Prevision.

In the approach of Walley ([15]) coherent conditional previsions are required to be separately coherent and coherent with respect to a given unconditional lower prevision $\underline{P}$. Given a non empty set Ω, a gamble X is a bounded function from Ω to R (the set of real numbers) and let **L** be the set of all gambles on Ω. When **K** is a linear space of gambles a coherent lower prevision $\underline{P}$ is a real function defined on **K**, such that the following conditions hold for every X and Y in **K**:
1)$\underline{P}(X) \geq \inf(X)$
2)$\underline{P}(\lambda X) = \lambda \underline{P}(X)$ for each positive constant λ
3)$\underline{P}(X+Y) \geq \underline{P}(X) + \underline{P}(Y)$

Lower previsions have a behavioural interpretation. If the gambles X in **K** are regarded as uncertain rewards, the lower prevision $\underline{P}(X)$ can be regarded as a supremum buying price for the gamble X.
Suppose that $\underline{P}$ is a lower prevision defined on a linear space **K**, its conjugate upper prevision $\overline{P}$ is defined on the same domain **K** by $\overline{P}(X) = -\underline{P}(-X)$.
If **K** contains only gambles that are indicator functions of events then a coherent lower (upper) prevision $\underline{P}$ defined on **K** is a coherent lower (upper) probability. So in this note we use the same symbol for (conditional) probability measure and (conditional) prevision.
Let **B** denote a partition of Ω, which is a non-empty, pair wise-disjoint subsets whose union is Ω. For B in **B** let **H**(B) be the set of gambles defined on B which includes the gamble B (we denote with the same symbol the set that represents an event and the indicator function of the event). A lower conditional prevision $\underline{P}(X|B)$ is a real function defined on **H**(B). Lower coherent conditional previsions $\underline{P}(X|B)$, defined for B in **B** and X in **H**(B) are required to be *separately coherent,* that is for every conditioning event B $\underline{P}(\cdot|B)$ is a coherent lower prevision on the domain **H**(B) and $\underline{P}(B|B) = 1$.
$\underline{P}(X|B)$ can be interpreted as the supremum buying price for X after we make the observation of a set B, that is we learn that the true state ω is in B. (This interpretation amounts to one of several possible "conditionalization" principles).
A gamble X is **B**-*measurable* when it is constant on each set B in **B**. Given a σ-field **G** of subsets of Ω, a gamble X is **G-**measurable if for every Borel set C of R the sets $\{\omega \in \Omega : \omega \in X^{-1}(C)\}$ belong to **G**.

Measurability with respect to a partition is, in general, a stronger condition than the measurability with respect to a σ-field. In fact, given two σ-fields **F** and **G** with **G** properly contained in **F** and generated by the partition **B**, fix A in **F**-**G.** We have that the indicator function of A is **B**-measurable, but not **G**-measurable.
Let **G**(**B**) be the class of **B**-measurable gambles. We denote by $\underline{P}(X|\mathbf{B})$ the function from **H** into **G**(**B**) whose image is the collection of coherent lower previsions $\{\underline{P}(\cdot|B) : B \text{ in } \mathbf{B}\}$. $\underline{P}(X|\mathbf{B})$ is separately coherent if all the lower conditional previsions are separately coherent. Let $\overline{P}(\cdot \mid B)$ be the conjugate upper conditional prevision. If $\underline{P}(\cdot|\mathbf{B})$ are linear previsions, that is $\underline{P}(\cdot \mid B) = \overline{P}(\cdot \mid B)$ for every B in **B,** then a *linear conditional prevision* $P(X|\mathbf{B})$ is defined by $P(X \mid B) = \underline{P}(X \mid B) = \overline{P}(X \mid B)$ for every B in **B**.

Given a non-empty set Ω, a partition **B** of Ω and a **B**-measurable gamble X if upper and lower conditional previsions are separately coherent then we have that $\underline{P}(X \mid B) = \overline{P}(X \mid B) = X$ for every B in **B** (Walley [15] pag. 292).
In particular if **B** is the partition of Ω that consists of all singletons and $X = I_A(\omega)$ is the indicator function of an event A then the previous property implies that
(P1) $P(A|\{\omega\}) = I_A(\omega)$ for every $\omega \in \Omega$ and for every $A \subseteq \Omega$.
The intuitive meaning of property (P1) is that coherent conditional probability is an uncertainty measure able to manage "precise" information, which is represented by the singletons of Ω.
This basic property is not always satisfied, in the continuous case, by the axiomatic definition of conditional probability given by the Radon-Nykodim; in fact if the conditioning σ-field is not countably generated, we may have that the regular conditional distributions, given that σ-field, are maximally improper (Seidenfeld et al. [13]) and therefore it does not verify the property (P1). It implies that conditional probability defined by the Radon-Nikodym derivative cannot always be used to represent uncertainty (see Example 33.11 of Billingsley [1]).

Walley ([15], 6.3) discusses the conditions in which an unconditional lower prevision $\underline{P}$ is coherent with $\underline{P}(\cdot|\mathbf{B})$. Given a class **D** of gambles, we say that **D** is a class of desirable gambles if, for each X in **D** and positive δ, we are disposed to accept the gamble X+δ. X is almost-desirable if we are not necessarily disposed to accept X itself.
The link between unconditional and conditional previsions can be expressed in terms of desirability, by the conglomerative principle (Walley [15], 6.3.3):

*If a gamble X is B-desirable, i.e we intend to accept X provided we observe only the event B, for every set B in the partition **B**, then X is desirable.*

Definition 1. Let $\underline{P}$ be an unconditional lower prevision defined on **K** and $\underline{P}(\cdot|\mathbf{B})$ be a conditional lower prevision on the domain **H** separately coherent on **H**. Assume that **H** and **K** are linear spaces containing all constant gambles and denote by $G(X) = X - \underline{P}(X)$, $G(Y|\mathbf{B}) = Y - \underline{P}(Y|\mathbf{B})$ and $G(W|B) = B(W - \underline{P}(W|B))$. Say that $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ are coherent if

a) $\sup[G(X) + G(Y|\mathbf{B}) - G(Z)] \geq 0$
and
b) $\sup[G(X) + G(Y|\mathbf{B}) - G(W|B)] \geq 0$ if $X,Z \in \mathbf{K}$, $Y,W \in H$ and $B \in \mathbf{B}$.

The previous definition quantifies over infinitely many unconditional previsions and called-off previsions since the superior operation appears. It is an important difference with respect to de Finetti's criterion of coherence that permits only finitely many unconditional and called-off previsions to enter into an assessment of coherence. For this reason, in de Finetti's theory, coherence does not entail that $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ are coherent.
Conditions a) and b) automatically hold when either domain **K** contains only constant gambles or **H** contains only **B**-measurable gambles.
In the first case we have that G(X) and G(Z) are equal to zero and by the separate coherence of $\underline{P}(\cdot|\mathbf{B})$ we have that conditions a) and b) are satisfied.
In the second case we have that $G(Y|B) = 0$ for every **B**-measurable gamble Y and by the coherence of $\underline{P}$ we have that conditions a) and b) are satisfied.
The gamble $G(X|\mathbf{B})$, in which we pay the uncertain price $\underline{P}(X|\mathbf{B})$ for X can be regarded as a two-stage gamble: firstly we observe B and pay price $\underline{P}(X|B)$, then we observe ω in B and receive $X(\omega)$.

A general characterization of coherence of the unconditional lower prevision with respect to the lower conditional prevision can be given by two axioms (Walley [15] 6.5.1).
Let $\overline{P}_B(Y|\mathbf{B})$ denote the gamble $B\,\underline{P}(Y|B) + B^c\,\overline{P}(Y|\mathbf{B})$.
Then $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ are coherent if and only if they satisfy the two axioms:
1) If $X \in \mathbf{K}$, $Y \in \mathbf{H}$ and $X \geq Y$ then $\underline{P}(X) \geq \inf \underline{P}(\cdot|\mathbf{B})$
2) If $B \in \mathbf{B}$, $X \in \mathbf{K}$, $Y \in \mathbf{H}$ and $X \leq Y$ then $\underline{P}(X) \leq \sup \overline{P}_B(Y|\mathbf{B})$.

The axioms simplify further when one of the domains contains the other.
In particular if **H** contains **K** $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ are coherent if and only if
3) $\underline{P}(X) \geq \inf \underline{P}(X|\mathbf{B})$ whenever $X \in \mathbf{K}$
4) $\underline{P}(X) \leq \sup \overline{P}_B(X|\mathbf{B})$ whenever $X \in \mathbf{K}$ and $B \in \mathbf{B}$.

If $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ are respectively linear unconditional and conditional previsions their coherence can be characterized by simpler conditions. In particular in Walley ([15] Section 6.5.3 and section 6.5.7) the following result has been proven:

Proposition 1. *Given P defined on* **K** *and P(X|***B***) defined on* **H** *such that they are respectively linear unconditional and conditional previsions with* **H** *contained in* **K** *and P(X|***B***) separately coherent, then P and P(X|***B***) are coherent if and only if the following conglomerative property is satisfied*

$$P(X) = P(P(X|\mathbf{B})).$$

Given a partition **B** of Ω, the unconditional probability P(X) is **B**-conglomerable if it satisfies the conglomerative property in the partition **B**.
When the unconditional prevision P(X) is **B**-conglomerable for every partition **B** of Ω then it is called *fully conglomerable* (Walley [15] 6.8.1).
In the paper of Dubins ([7]) the following definitions are introduced.
Given a partition **B** of Ω, a linear prevision P(X) is *disintegrable* with respect to linear conditional previsions $P(X|\mathbf{B})$ if the equality $P(X) = P(P(X|B))$ is satisfied for every bounded variable X on Ω and for every B in **B**.
A linear prevision P(X) is defined to be *conglomerative* with respect to a partition **B** of Ω if the following condition is satisfied: for every bounded variable X and for every B in **B** we have that $P(X|B) \geq 0$ implies $P(X) \geq 0$.
It has been proven (Theorem 1 of [7]) that a prevision is disintegrable with respect to a partition if and only if it is conglomerative with respect to the same partition.
When **H** and **K** are equal to the set **L** of all bounded gambles on Ω then the conglomerative property of Walley is equivalent to the notion of *disintegrability* of a prevision P(X) with respect to a partition of Ω, introduced by Dubins ([7]). The author calls *strategies* linear conditional previsions that are separately coherent and defined on the set of all bounded gambles on Ω.
So if linear conditional previsions $P(X|\mathbf{B})$ and linear unconditional prevision P(X), defined on the class of all bounded gambles, are such that they satisfy the disintegration property with respect to a given partition **B** of Ω, then they are coherent.
The notion of conglomerability given by Dubins can be seen as a generalization to the class of all bounded variables of the *conglomerative principle*, introduced by de Finetti ([4] pp.99) for probabilities:

Given a partition **B** *of Ω we say that the probability P is conglomerable with respect to the partition* **B** *if for every*

event A and for every B *in* **B** *we have that* $a \leq P(A|B) \leq b$ *implies* $a \leq P(A) \leq b$.

Generally the disintegrability in the sense of Dubins is stronger than the conglomerability in the sense of de Finetti.
In fact if the conglomerative principle is satisfied it does not imply that the disintegration property is satisfied; but when the domain of the conditional and unconditional linear previsions is a linear space then the notion of conglomerability in the sense of de Finetti is equivalent to the notion of disintegrability in the sense of Dubins.

An important aspect, analysed in literature is the relationship between conglomerability and countable additivity.
In Schervish et. al. [11] it has been proven that when a probability P is defined on a σ-field, it takes infinitely many values and it is countably additive then it is disintegrable (conglomerable) in the sense of Dubins in every countable partition of Ω.
In particular if P is defined on the class of all subsets of Ω and it takes infinitely many different values then it is fully conglomerable if and only if it is countably additive on every partition of Ω.
We have that for non-countable partitions countable additivity of the unconditional prevision is not a sufficient condition to assure that it is coherent with the conditional previsions (Kadane, Schervish, Seidenfeld [9] Example 6.1).
The previous results imply that there is no fully conglomerable linear prevision P defined on the set of all bounded gambles **L** that takes many different values on events and satisfies $P(\{\omega\}) = 0$ for all $\omega \in \Omega$. For example there is no fully conglomerable linear extension of Lebesgue measure to all bounded gambles on the unit interval. Otherwise the Lebesgue lower prevision on **L**, which is the natural extension of the Lebesgue (inner) measure to all bounded gambles is fully conglomerable (Walley [15], 6.9.6), that is there is a lower conditional prevision with respect to every partition **B** coherent with the Lebesgue lower previsions.

3 Hausdorff Outer and Inner Measures

In this section we recall some preliminaries about Hausdorff measures, that we use to define conditional previsions P(X|B) when the conditioning events B have finite and positive Hausdorff measure in their dimension. For more details about Hausdorff measures see for example Falconer ([8]).
Let (Ω,d) be the Euclidean metric space with $\Omega = [0,1]$. The diameter of a nonempty set U of Ω is defined as $|U| = \sup\{|x-y|: x,y \in U\}$ and if a subset A of Ω is such that $A \subset \bigcup_i U_i$ and $0 < |U_i| < \delta$ for each i, the class $\{U_i\}$ is called a δ-cover of A. Let s be a non-negative number. For $\delta > 0$ we define $h^s_\delta(A) = \inf \sum_{i=1}^{\infty} |U_i|^S$, where the infimum is over all δ-covers $\{U_i\}$. The Hausdorff s-dimensional outer measure of A, denoted by $h^s(A)$, is defined as $h^s(A) = \lim_{\delta \to 0} h^S\delta(A)$. This limit exists, but may be infinite, since $h^s_\delta(A)$ increases as δ decreases.
The *Hausdorff dimension* of a set A, $\dim_H(A)$, is defined as the unique value, such that

$$h^s(A) = \begin{cases} \infty & \text{if} \quad 0 \leq s \leq \dim_H(A) \\ 0 & \text{if} \quad \dim_H(A) < s < \infty \end{cases}$$

We can observe that if $0 < h^s(A) < \infty$ then $\dim_H(A) = s$, but the converse is not true. We assume that the Hausdorff dimension of the empty set is equal to –1 so no event has Hausdorff dimension equal to the empty set. If an event A is such that $\dim_H(A) = s < 1$, then the Hausdorff dimension of the complementary set A^c is equal to 1 since the following relation holds:

$$\dim_H(A \cup B) = \max\{\dim_H(A); \dim_H(B)\}.$$

A subset A of Ω is called measurable with respect to the outer measure h^s if it decomposes every subset of Ω additively, that is if $h^s(E) = h^s(AE) + h^s(E-A)$ for all sets $E \subset \Omega$.

The restriction of h^s to the σ-field of h^s-measurable sets, containing the σ-field of the Borel sets, is called *Hausdorff s-dimensional measure*. In particular the Hausdorff 0-dimensional measure is the counting measure and the Hausdorff 1-dimensional measure is the Lebesgue measure.

4 Upper and Lower Conditional Previsions defined by the Hausdorff Outer and Inner Measures

In [6] upper and lower conditional probabilities are obtained as *natural extensions* (Theorem 3.1.5 [15]) of a finitely additive conditional probability in the sense of Dubins, assigned by a class of Hausdorff measures. They are proven to be separately coherent and so they satisfy the necessary condition for the coherence (P1).
In this Section upper and lower conditional previsions are defined as extensions of the previous upper and lower conditional probabilities. In particular, when the conditioning event has positive and finite Hausdorff outer (inner) measure in its dimension, they are defined

by the Choquet integral ([5]) with respect to outer (inner) Hausdorff measures, which are particular examples of monotone set functions. Otherwise, when the conditioning event has Hausdorff outer (inner) measure in its dimension equal to zero or infinity, they are defined by a 0-1 valued finitely but not countably additive probability.
In Theorem 2 and 3 of this Section we prove that when Hausdorff measures are defined on the Borel σ-field and the class of all Borel-measurable gambles is considered, then linear conditional and unconditional previsions defined with respect to Hausdorff measures satisfy the Dubin's disintegration property with respect to every countable partition of Ω.
In Theorem 2 we consider Ω equal to [0,1] and in Theorem 3 we consider the general case in which Ω is an infinite set with Hausdorff measure equal to 1 in its dimension.
Moreover linear conditional previsions are coherent with the unconditional previsions in the sense of Walley, since in this case coherence in the sense of Walley is equivalent to the disintegration property of Dubins (see Proposition 1 of Section 2).
The role of Hausdorff measures in the previous results is crucial.
In fact it is important to observe that if we define conditional and unconditional previsions with respect to a coherent finitely but not countably additive probability we cannot obtain the same results.
In fact from Theorem 3.1 of ([11]) we have that for each finitely but not countable additive probability P defined on a σ-field there is a partition (in that σ-field) where P is not disintegrable in the sense of Dubins.
This implies that linear conditional and unconditional previsions defined with respect to a merely finitely additive probability cannot be disintegrable on every countable partition of Ω.

We recall some results given in ([6]).
Let Ω be a non empty set and let **F** and **G** be two fields of subsets of Ω, with $\mathbf{G} \subseteq \mathbf{F}$ or with **G** an additive subclass of **F**, P* is a *finitely additive conditional probability* ([7]) defined on **(F,G)** if it is a real function defined on $\mathbf{F}\times\mathbf{G}^0$, where $\mathbf{G}^0=\mathbf{G}-\emptyset$, such that the following conditions hold:

I) given any $H\in\mathbf{G}^0$ and $A_1,...,A_n\in\mathbf{F}$ with $A_iA_j=\emptyset$ for $i\neq j$, the function $P^*(\cdot|H)$ defined on **F** is such that

$$I)\ P^*(A|H) \geq 0, \qquad P^*\Big(\bigcup_{k=1}^{n} A_k \,|\, H\Big) = \sum_{k=1}^{n} P^*(A_k \,|\, H),$$

$$P^*(\Omega|H) = 1$$

II) $P^*(H|H) = 1$ if $H\in \mathbf{FG}^0$

III) given $E \in \mathbf{F}$, $H\in\mathbf{F}$ $EH \in \mathbf{F}$ with $A\in\mathbf{G}^0$ and $EA \in\mathbf{G}^0$ then $P^*(EH|A) = P^*(E|A)P^*(H|EA)$.

From conditions I) and II) we have

II') $P^*(A|H) = 1$ if $A\in \mathbf{F}$, $H\in\mathbf{G}^0$ and $H\subset A$.

Such approach to conditional probability allows to give probability assessments on arbitrary finite family of conditional events through the notion of *coherence* as proposed by de Finetti ([3], [4]). In fact, if **F** and **G** are arbitrary finite families of subsets of Ω, then the real function P, defined on $\mathbf{F}\times\mathbf{G}^0$ is *coherent* if and only if it is the restriction of a finitely additive conditional probability defined on $\mathbf{D}\times\mathbf{D}^0$, where **D** is the field generated by the sets of **F** and **G**.
In [6] a finitely additive conditional probability in the sense of Dubins is defined by a class of Hausdorff dimensional measures. Moreover, upper (lower) conditional probability is given by Hausdorff s-dimensional outer (inner) measures if the conditioning event has positive and finite Hausdorff s-dimensional outer (inner) measure in its dimension; otherwise upper conditional probability is defined by a 0-1 finitely additive (but not countable additive) probability so that condition III) of a finitely additive conditional probability in the sense of Dubins is satisfied. They are proven to be separately coherent in the sense of Walley. The unconditional probability is obtained as particular case when the conditioning event is Ω.

Theorem 1. *Let Ω=[0,1], let **F** be the σ-field of all subsets of Ω and let **G** be an additive sub-class of **F**. Let us denote by h^s the Hausdorff s-dimensional outer measure and let us define on $\mathbf{C}=\mathbf{F}\times\mathbf{G}^0$ the function $\overline{P}$ by*

$$\overline{P}(A|H) = \begin{cases} \dfrac{h^s(AH)}{h^s(H)} & \text{if} \quad 0 < h^s(H) < \infty \\ \\ m(AH) & \text{if} \quad h^s(H) = 0, \infty \end{cases}$$

where m is a 0-1 valued finitely additive (but not countably additive) probability measure. Then the function $\overline{P}$ is an upper conditional probability.

The existence of the measure m is a consequence of the prime ideal theorem.
The conjugate lower conditional probability $\underline{P}$ can be defined as in Theorem 1 if h^s denotes the Hausdorff s-dimensional inner measure.

When the family of the conditioning events is a partition of Ω the conditional probabilities can be defined in a similar way.

Definition 2. *Let Ω = [0,1], let **F** be the σ-field of all subsets of Ω and let **B** be a partition of Ω. Let us denote*

by s the Hausdorff dimension of the conditioning event B belonging to **B** *and by* h^s *the outer Hausdorff s-dimensional measure. Let us define an upper conditional probability on* $\mathbf{F}\times\mathbf{B}$ *by the function*

$$\overline{P}(A|B)=\begin{cases}\dfrac{h^s(AB)}{h^s(B)} & \text{if} \quad 0<h^s(B)<\infty \\ m(AB) & \text{if} \quad h^s(B)=0,\infty\end{cases}$$

where m is a 0-1 valued finitely additive (but not countably additive) probability measure.

The two definitions of upper conditional probabilities can be compared when **G** is the σ-field generated by the partition **B**. In particular, given a probability space (Ω,**F**,P), let **G** be equal to or contained in the σ-field generated by a countable class **C** of subsets of **F** and let **B** be the partition generated by the class the **C.** Denote by Ω' = **B** and $\psi_{\mathbf{B}}$ the function from Ω to Ω' that associates to every $\omega\in\Omega$ the atom B of the partition **B** that contains ω; then we have that $\overline{P}(\cdot|\mathbf{G}) = \overline{P}(\cdot|\mathbf{B})\circ\psi_{\mathbf{B}}$ (See Koch [10] p. 262).

Upper (lower) conditional prevision is obtained as extension of upper (lower) conditional probability assigned by a class of outer Hausdorff measures.
It is defined by the Choquet integral ([5]) with respect to outer (inner) Hausdorff measures, which are particular examples of monotone set functions.

Definition 3. Let *Ω = [0,1], let* ***L*** *be the class of all bounded gambles on Ω and let* **B** *be a partition of Ω . Let us denote by s the Hausdorff dimension of the conditioning event B belonging to* **B** *and by* h^s *the Hausdorff s-dimensional outer measure. Let us define an upper conditional prevision on* $\boldsymbol{L}\times\mathbf{B}$ *by the function*

$$\overline{P}(X\mid B)=\begin{cases}\dfrac{1}{h^s(B)}\displaystyle\int_B X dh^s & \text{if} \quad 0<h^s(B)<\infty \\ m(XB) & \text{if} \quad h^s(B)=0,\infty\end{cases}$$

where m is a 0-1 valued finitely additive (but not countably additive) probability measure.

From the definition it follows that upper conditional previsions are separately coherent for every partition **B** of Ω.
We prove that when the (outer) Hausdorff measures are defined on the Borel σ-field and **L** is the class of all Borel-measurable gambles, then linear conditional and unconditional previsions defined with respect to Hausdorff measures satisfy the Dubin's disintegration property with respect to every countable partition of Ω and they are coherent in the sense of Walley.
The following results can be obtained as a consequence of the fact that non-disintegrability characterizes finitely as opposed to countably additive probability as proven in [11]. Each arbitrary finitely additive probability P can be decomposed uniquely into a convex combination of a countably additive probability P_c and a purely finitely additive probability P_D, that is

$$P=\alpha P_c+\beta P_D \qquad \text{with } \alpha+\beta=1,\ \alpha,\beta\geq 0.$$

In [11] the coefficient β has been proven to be an upper bound for failures of conglomerability in all denumerable partitions.
In Theorem 3.1 of [11] it has been proven that if $\beta\neq 0$, if the range of P is not limited to finitely many distinct values and if P is defined on a σ-field of event, then the upper bound on the failure of conglomerability, β, must be approached.

Theorem 2. Let *Ω = [0,1], let* ***F*** *be the Borel σ-field of subsets of Ω and let* **L** *be the class of all Borel-measurable gambles on Ω. If* **B** *is a countable partition of Ω, consisting of sets belonging to* ***F****, then the linear conditional prevision defined on* $\boldsymbol{L}\times\mathbf{B}$ *by Definition 3, is coherent with the unconditional prevision P(·|Ω).*

Proof. Since Ω is equal to [0,1] then the linear unconditional prevision P(·|Ω) is defined with respect to the Hausdorff measure of order 1, h^1, that is the Lebesgue measure. It is defined on the Borel σ-field, it takes infinitely many different values and it is countably additive. As shown in [11] this is equivalent to the disintegrability of h^1 in the sense of Dubins with respect to all countable partitions of Ω.
Since for every s, the σ-field of h^s-measurable sets contains the Borel σ-field and **L** is the class of all Borel–measurable gambles, we also have that the conditional previsions are linear.
So the unconditional and conditional previsions are coherent in the sense of Walley; in fact from Proposition 1 of Section 2, disintegrability in the sense of Dubins is equivalent to the coherence of linear conditional previsions with respect to the linear unconditional prevision.□

The previous result can be generalized to the case where Ω is an infinite set with Hausdorff measure in its dimension equal to 1.

Theorem 3. Let *Ω be an infinite set with Hausdorff measure equal to 1 in its dimension, let* ***F*** *be the Borel σ-field of subsets of Ω and let* **L** *be the class of all Borel-measurable gambles on Ω. If* **B** *is a countable partition*

of Ω, consisting of sets belonging to ***F,*** *then the conditional prevision defined on* $\mathbf{L}\times\mathbf{B}$ *as Theorem 2, is coherent with the unconditional prevision* $P(\cdot|\Omega)$.

Proof. Denoted by s the Hausdorff dimension of Ω, then the unconditional prevision $P(\cdot|\Omega)$ is defined with respect to the s-dimensional Hausdorff measure h^s, which is a probability since $h^s(\Omega) = 1$, it is defined on the Borel σ-field, it takes infinitely many different values and it is countably additive since for every s, the σ-field of h^s-measurable sets contains the Borel σ-field.
Then the result can be obtained in a similar way of Theorem 2.□

Remark 1. It is important to note the crucial role of the Hausdorff measures in the previous theorems. In fact if the unconditional prevision is defined with respect to the s-dimensional Hausdorff measure, where s is the Hausdorff dimension of Ω and **F** is the Borel σ-field, then in Theorem 2 and in Theorem 3 the unconditional prevision is defined with respect to a countably additive probability. This implies ([11]) that the disintegration property in the sense of Dubins is satisfied on every countable partition of Ω.
Otherwise if we define the unconditional prevision with respect to a coherent finitely but not countably additive probabilility P, defined on a σ-field then there is (Theorem 3.1 of [11]) a countably partition where P fails disintegrability in the sense of Dubins.

Example 1. We recall the definition of the Cantor set, which is the most familiar set of real numbers of non-integer Hausdorff dimension.

Let $E_0 = [0,1]$, $E_1 = [0,1/3] \cup [2/3,1]$, $E_2 = [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1]$, etc., where E_{j+1} is obtained by removing the open middle third of each interval in E_j. The Cantor's set is the perfect set $E = \bigcap_{j=0}^{\infty} E_j$. The Hausdorff dimension of the Cantor set is s = log2/log3 and $h^s(E) = 1$ (see [8] Theorem 1.14).

Let Ω be equal to the Cantor set, let **F** be the Borel σ-field of subsets of Ω and let **L** be the class of all Borel-measurable gambles on Ω. If **B** is a countable partition of Ω, consisting of sets belonging to **F,** then the conditional prevision defined on $\mathbf{L}\times\mathbf{B}$ as in Theorem 2, is coherent with the unconditional prevision $P(\cdot|\Omega)$.

5 Coherence of Conditional Possibility Distribution

A first criterion to decide if an uncertainty measure is a good tool to handle imprecise and vague information about a linguistic or numerical variable is to verify if it is, first of all, able to manage "precise" information, which is represented by the singletons of Ω.
In the theory of imprecise probabilities this property is formalised by property (P1) as recalled in Section 2.
(P1) for every x belonging to Ω $P(A|\{x\})$ is equal to 1 if x belongs to A and it is equal to 0 if x does not belong to A.
We analyze the possibility to extend the properties assured by the coherence to uncertainty measures used when information is represented by fuzzy sets.
In this Section a conditional possibility distribution on an infinite set Ω that is coherent and coherent with respect to the unconditional possibility distribution.
The conditional possibility distribution satisfies the property (P1). Moreover, through this conditional possibility distribution we obtain a possibility conditional measure with respect to the partition of all singletons that is coherent and such that the (weak) conglomerative principle of de Finetti is verified.

In ([2], [16]) possibility measures are proven to be an important special class of upper probabilities; moreover in [17] a necessary and sufficient condition for the coherence of rules for defining conditional possibility distributions is given when the possibility space Ω is finite. In the quoted paper conditioning on variables rather then events is considered. Given two variables X and Y whose sets of possible values are finite the problem of examining whether a conditioning rule produces conditional distribution π(x|y) that is coherent with the joint possibility distribution π(x,y) has been investigated; moreover it has been investigated when they generate possibility measures (or equivalently upper probability measures) Π and Π(·|y) that are coherent.
Given an infinite set Ω, in this section we consider conditioning on the class of fuzzy sets of Ω, and we investigate the problem to define conditional possibility distribution π(x|y) coherent with the unconditional possibility distribution π. Moreover the coherence of the conditional possibility measures Π(·|y) with the unconditional possibility measure Π is analyzed.
Several conditioning rules are proposed in literature for defining conditional possibility distributions or measures from unconditional ones.
The approach followed in this section is quite different: firstly we define the conditional possibility distribution π(x|y) and the conditional possibility measure Π(·|y) such that they satisfy the condition (P1) for the coherence, then we consider their coherence with the unconditional possibility distribution and the unconditional possibility measure.

Given a non-empty set Ω a *fuzzy set* A is defined by a membership function that associates to each element x of Ω a real number A(x) between 0 and 1, which represents the degree to which x belongs to A.

If the membership function is equal to the indicator function then A is a *crisp* set.
The *support* of a fuzzy set is the crisp set where the membership function of the fuzzy set is greater than zero; the *core* of a fuzzy set is the crisp set where the membership function is equal to one. A fuzzy singleton is a fuzzy set whose core is a singleton.
Given two fuzzy sets A(x) and B(x) their union is defined by $A(x) \cup B(x) = \max\{A(x),B(x)\}$ for every x in Ω.
Given an infinite set Ω we denote by $P(\Omega)$ the class of the fuzzy sets of Ω; a fuzzy measure over $P(\Omega)$ is a function m: $P(\Omega) \rightarrow [0,1]$ such that $m(\varnothing) = 0$ and $m(\Omega) = 1$, $E \subset F \Rightarrow m(E) \leq m(F)$.
A measure of possibility is a fuzzy measure Π such that $\Pi\left(\bigcup_{j\in J} A_j\right) = \sup_{j\in J} \Pi(A_j)$.
A possibility distribution over Ω is a function π: $\Omega \rightarrow [0,1]$ such that $\pi(\omega) = \Pi(\{x\})$
Using a possibility distribution π over Ω, it is possible to construct a possibility measure Π over $P(\Omega)$ by the formula
$\Pi(A) = \sup_{\omega\in\Omega} \{\min(\pi(\omega), A(\omega))\}$;
A possibility distribution and a possibility measure are normalized if $\Pi(\Omega) = \sup\{\pi(\omega) : \omega \in \Omega\} = 1$.
In this paper we assume they are normalized.
As recalled in Walley ([16] p. 35) the information represented by a fuzzy set, for example "Mary is young" can be modeled by a possibility distribution defined on the set of possible ages. The number $\pi(\omega)$ lies between zero and one and it represents " the degree to which it is possible that Mary has a precise age ω, given she is young"
In the same way we can interpret a conditional possibility distribution $\pi(x|y)$ as "the degree to which it is possible that Mary has a precise age x, given she is y years old"
So if we want the conditional distribution to satisfy the condition (P1) necessary for the coherence of an upper conditional probability we have to define

$$\pi(x|y) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \text{if} \quad x \neq y \end{cases} \qquad (1)$$

that is the conditional distribution $\pi(x|y)$ is equivalent to the indicator function of the (fuzzy) singleton; so for every y in Ω $\pi(x|y)$ is concentrated on the singleton y.

In order to find conditions that assure the coherence of the conditional distribution $\pi(x|y)$ with the unconditional distribution π it is important to determine relations between them.

Given two fuzzy sets A and B we introduce a joint possibility distribution $\pi(x,y)$ for all $x \in A$ and $y \in B$.
According to Hisdial a conditional possibility distribution $\pi(x|y)$ is implicitly defined as

$$\pi(x,y) = \min(\pi(y), \pi(x|y)) \qquad (2)$$

If we define conditional possibility distribution by (1) and we require that also (2) is satisfied we obtain

$$\pi(x, y) = \begin{cases} \pi(y) & \text{if} \quad x = y \\ 0 & \text{if} \quad x \neq y \end{cases}$$

At this point the question is:" how do we have to choose $\pi(y)$ so that the conditional possibility distribution $\pi(x|y)$ and the unconditional possibility distribution are coherent?"

We observe that the definition of conditional possibility distribution proposed in this section is similar to the one proposed by Ramer [14]. This conditioning rule consists in picking one x_0 such that $\pi(x_0,y) = \pi(y)$, letting $\pi(x_0,y) = 1$ and $\pi(x|y) = \pi(x,y)$. It produces normal $\pi(\cdot|y)$, but it has the disadvantage of requiring an arbitrary choice whenever there is more than one x that maximizes $\pi(\cdot|y)$. Moreover, Ramer's rule produces conditional possibility distributions which are incoherent with joint distribution if $0 < \pi(x,y) < \pi(y) < 1$ as pointed out in [17].
The definition of conditional possibility distribution given in this section by (1) avoids this problem since the only value, which maximizes $\pi(\cdot|y)$ is $x_0 = y$
Given $y \in \Omega$ and for every $x \in \Omega$ we define $\pi(x,y)$ equal to

$$\pi(x, y) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \text{if} \quad x \neq y \end{cases}$$

So that $\pi(\cdot)$ is the 0-1 valued finitely additive measure concentrated on the singleton $\{y\}$.

When Ω is a finite set a necessary and sufficient condition for the coherence of $\pi(\cdot,\cdot)$ and $\{\pi(\cdot \mid y) : y \in \Omega\}$ has been proposed in [17]; it claims that $\pi(\cdot,\cdot)$ and $\{\pi(\cdot \mid y) : y \in \Omega\}$ are coherent if and only if the conditional possibility distribution $\pi(x|y)$ is greater or equal to the conditional possibility distribution defined by the Dempster's rule ($\pi_{DE}(x, y)$) and it is less or equal to the conditional possibility distribution defined by the natural extension ($\pi_{NE}(x, y)$) if $\pi(y) > 0$.

The definition of conditional possibility distribution $\pi(x|y)$ given by (1) satisfies the previous condition in fact we have that

$\pi(x|y) = \pi(x,y) = \pi(y) = 1 = \pi_{DE}(x, y) = \pi_{NE}(x, y)$.

If Ω is a countable set the unconditional possibility distribution $\pi(\cdot)$ and the conditional possibility distribution $\pi(\cdot|y)$ defined by (1) are equal to the measure concentrated on the singleton $\{y\}$.

They are coherent as proven in Seidenfeld et al. ([12] Lemma 1).

Definition 4. Let us define a conditional possibility measure by

$$\Pi(A \mid y) = \sup_{x \in A} \{\min(\pi(x \mid y), A(y))\}$$

then we obtain

$$\Pi(A \mid y) = A(y)$$

Remark 2. If A is a crisp set, so that its membership function A(x) is equal to the indicator function of A, then $\Pi(A \mid y)$ is the indicator function of A and so property (P1) necessary for the coherence is satisfied; moreover from the definition of possibility measure we have that $\Pi(A \cup B \mid y) = \max\{\Pi(A \mid y), \Pi(B \mid y)\}$.

So, if A is a crisp set, the conditional possibility measure $\Pi(A \mid y)$ is the 0-1 finitely additive measure concentrated on the singleton $\{y\}$, which is a particular kind of (upper) conditional probability

Example 2. Let Ω be the set of natural numbers N, then the conditional possibility measure $\Pi(A \mid y)$ is the 0-1 finitely additive measure concentrated on the singleton $\{y\}$ and we have that

$$\Pi(N \mid y) = \sup_{x \in N} \{\min(\pi(x \mid y), N(y))\} = 1$$

We have defined the conditional possibility distribution $\pi(\cdot)$ equal to the 0-1 valued finitely additive measure concentrated on the singleton $\{y\}$. So we obtain that the conditional possibility measure is equal to

$$\Pi(A) = \sup_{y \in \Omega} \{\min(\pi(y), A(y))\} = \sup_{y \in \Omega} \{A(y)\}$$

In particular if A is a fuzzy singleton $A = \{x\}$ we have that $\Pi(\{x\}) = \pi(x) = 1$.

The next result shows that for every fuzzy set A the normalized possibility measure Π and the normalized conditional possibility measure $\Pi(A \mid y)$ satisfy the conglomerative principle of de Finetti with respect to the partition of all singletons of an infinite set Ω.

Theorem 4. *Let Ω be an infinite set and let Π be a normalized possibility measure over $P(\Omega)$ the class of all fuzzy sets of Ω defined by* $\Pi(A) = \sup_{y \in \Omega} \{\min(\pi(y), A(y))\} = \sup_{y \in \Omega} \{A(y)\}$. *Moreover let* $\Pi(A \mid y)$ *be the conditional possibility measure defined by* $\Pi(A \mid y) = A(y)$. *Then for every y belonging to Ω, we have that*

$a \le \Pi(A|y) \le b$ implies $a \le \Pi(A) \le b$.

Proof. Since we have that

$$\Pi(A \mid y) = \sup_{x \in A} \{\min(\pi(x \mid y), A(y))\}$$

from the definition of conditional possibility distribution $\pi(x|y)$ given by (1) we obtain that $\Pi(A \mid y) = A(y)$.

So, if for every y in Ω we have that

$$a \le \Pi(A \mid y) = A(y) \le b$$

it implies that

$$a \le \sup_{y \in \Omega} \{A(y)\} \le b$$

that is $a \le \Pi(A) \le b$.□

6 Summary and Conclusions

A new model of upper and lower conditional previsions is proposed in this paper.

When the conditioning event has positive and finite Hausdorf outer (inner) measure in its dimension, upper (lower) conditional previsions are defined by the Choquet integral ([5]) with respect to the outer (inner) Hausdorff measures, which are particular examples of monotone set functions. Otherwise, when the conditioning event has Hausdorff outer (inner) measure in its dimension equal to zero or infinity, upper (lower) conditional previsions are defined by a 0-1 valued finitely but not countably additive probability.

These upper and lower conditional previsions are proven to be separately coherent for every partition **B** of Ω.

Moreover when we consider the restriction of the (outer) Haudorff measures to the Borel σ-field (upper) conditional and unconditional previsions are proven to satisfy the disintegration property in the sense of Dubins with respect to all countable partitions of Ω and to be coherent in the sense of Walley.

Another problem analyzed in this paper is the extension of upper conditional probability properties assigned by a class of Hausdorff outer measures when information is represented by fuzzy sets.

A conditional possibility distribution on an infinite set Ω that is coherent and coherent with respect to the unconditional possibility distribution is defined.

Moreover through this conditional possibility distribution we obtain a possibility conditional measure with respect to the partition of all singletons that is coherent and that satisfies the conglomerative principle of de Finetti.

Acknowledgements

I wish to thank the reviewers for pointing out trivial aspects in the first version of the paper and for their valuable remarks.

References

[1] P.Billingsley, *Probability and measure*, John Wiley, New York, 1985.

[2] G. de Cooman, A behavioural model for vague probability assessments, *Fuzzy sets and Systems*, 154, 305-358, 2005.

[3] B. de Finetti, *Teoria della Probabilità*, Einaudi Editore, Torino, 1970.

[4] B. de Finetti, *Probability, Induction and Statistics*. New York, Wiley, 1972.

[5] D. Denneberg, *Non-additive measure and integral*, Kluwer Academic Publishers, 1994.

[6] S. Doria, Probabilistic independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures, accepted for publication in *International Journal of Approximate Reasoning*, 2007.

[7] L. Dubins Finitely additive conditional probabilities, conglomerability and disintegrations, *The Annals of Probability*, Vol.3, No1, 89-99, 1975.

[8] K.J. Falconer, *The geometry of fractal sets*, Cambridge University Press, 1986.

[9] J.B. Kadane, M. Schervish, T. Seidenfeld, Statistical implications of finitely additive probability. In *Bayesian Inference and Decision Techniques With Applications* (P.Goel and A. Zellner, eds.) 59-76. North-Holland, Amsterdam, 1986.

[10] G. Koch, *La matematica del probabile*, Aracne Editrice, 1997.

[11] M. Schervish, T. Seidenfeld, J.B. Kadane, The extent of non-conglomerability of finitely additive probabilities. *Z. Warsch.Verw.Gebiete* 66, 205-226, 1984.

[12] T. Seidenfeld, M. Schervish, J.B. Kadane, Non-conglomerability for finite-valued, finitely additive probability, *The Indian Journal of Statistics*, Special issue on Bayesian Analysis, Vol.60, Series A, 476-491, 1998.

[13] T. Seidenfeld, M. Schervish, J.B. Kadane, Improper regular conditional distributions, *The Annals of Probability*, Vol.29, No 4, 1612-1624, 2001

[14] A.Ramer Conditional possibility measures. *Cybernetics and Systems*, 20:233-247, 1989.

[15] P.Walley *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, 1991.

[16] P.Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83: 1.-58, 1996.

[17] P.Walley, G.de Cooman, Coherence of rules for defining conditional possibility, *International Journal of Approximate Reasoning*, vol. 21, 63-107, 1999.

Distributions over Expected Utilities in Decision Analysis

Love Ekenberg
Dept. of Computer and Systems Sciences
Stockholm University and Royal Institute of Technology
lovek@dsv.su.se
Address: Forum 100
SE-164 40 Kista, Sweden

Mikael Andersson
Dept. of Mathematical Statistics
Stockholm University
mikaela@math.su.se

Mats Danielson
Dept. of Computer and Systems Sciences
Stockholm University
mad@dsv.su.se

Aron Larsson
Dept. of Information Technology and Media
Mid Sweden University
aron.larsson@miun.se

Abstract

It is often recognised that in real-life decision situations, classical utility theory puts too strong requirements on the decision-maker. Various interval approaches for decision making have therefore been developed and these have been reasonably successful. However, a problem that sometimes appears in real-life situations is that the result of an evaluation still has an uncertainty about which alternative is to prefer. This is due to expected utility overlaps rendering discrimination more difficult. In this article we discuss how adding second-order information may increase a decision-maker's understanding of a decision situation when handling aggregations of imprecise representations, as is the case in decision trees or influence diagrams.

Keywords. Decision analysis, Imprecise probabilities, Imprecise utilities, Hierarchical models.

1 Introduction

In classical types of utility theories, a widespread opinion is that utility theory captures the concept of rationality. However, the shortcomings of this standpoint are sometimes severe. Among other things, the question has been raised whether people are capable of providing the inputs that utility theory requires, when, for instance, most people cannot clearly distinguish between probabilities ranging over substantial intervals. Similar problems arise in the case of artificial decision-makers, since utility-based artificial agents usually base their reasoning on human assessments, for instance in the form of induced preference functions. Furthermore, even if a decision-maker is able to discriminate between different probabilities, very often complete, adequate, and precise information is missing.

Thus, the requirement to provide numerically precise information in such models has often been considered unrealistic for real-life decision situations and after quite intense activities in the area, particularly during recent years, a number of models with representations allowing imprecise probability statements have been suggested. Such models include possibility theory [4], capacities (of order 1 and 2) [3], [13], [5], evidence theory and belief functions [19], various kinds of logic [22], upper and lower probabilities [7], hierarchical models [21], [10], and sets of probability measures [15]. Some general approaches to evaluating imprecise decision situations include probabilities and utilities. [16] is an early example and more recently some other interesting approaches have been suggested, e.g., [17], [14], [1], [6], and [2].

2 Decision Trees

In this paper, we let an *information frame* represent a decision problem. The idea with such a frame is to collect all information necessary for the model into one structure. The representational issues are of two kinds, structure (trees) and constraints (statements).

Decisions under risk (probabilistic decisions) are often given a tree representation, cf. [18]. One of the building blocks of a frame is a decision tree. Formally, a decision tree is a graph.

Definition 1. *A graph is a structure $\langle V, E\rangle$ where V is a set of nodes and E is a set of node pairs (edges).*

A general graph structure is, however, too permissive for representing a decision tree. Hence, we will restrict the possible degrees of freedom of expression in the decision tree.

Definition 2. *A tree is a connected graph without cycles. A decision tree is a tree containing a finite set of nodes and that has a dedicated node at level 0. The adjacent nodes, except for the nodes at level $i-1$, to a node at level i is at level $i+1$. A node at level $i+1$ that is adjacent to a node at level i is a child of the latter. A node at level 1 is an* alter-

native. *A node at level i is a leaf or* consequence *if it has no adjacent nodes at level $i+1$. A node that is at level 2 or more and has children is an* event *(an intermediary node). The depth of a rooted tree is* $\max(n|$*there exists a node at level* $n)$.

Thus, a decision tree is a way of modelling a decision situation where the alternatives are nodes at level 1 and the set of final consequences are the set of nodes without children. Intermediary nodes are called events. For convenience we can, for instance, use the notation that the n children of a node c_i are denoted $c_{i1}, c_{i2}, \ldots, c_{in}$ and the m children of the node c_{ij} are denoted $c_{ij1}, c_{ij2}, \ldots, c_{ijm}$, etc.

Figure 1 shows a decision tree. Over the sets of events and consequences, different functions can be defined, such as probability measures and utility functions.

3 Intervals in Decision Making

For numerically imprecise decision situations, one option is to define probability and utility functions in the classical way. Another, more elaborate option is to define sets of candidates of possible probability and utility functions. For instance, in [7] such an approach is suggested. The possible functions are expressed as vectors in polytopes that are solution sets to, so called, *probability* and *utility bases* (see below).

For instance, the probability (or utility) of c_{ij} being between the numbers a_k and b_k is expressed as $p_{ij} \in [a_k, b_k]$ ($u_{ij} \in [a_k, b_k]$). This approach also includes relations: a measure (or function) of c_{ij} is greater than a measure (or function) of c_{kl} is expressed as $p_{ij} \geq p_{kl}$ and analogously $u_{ij} \geq u_{kl}$. Each statement can thus be represented by one or more constraints.

Definition 3. *Given a decision tree D, a utility base is a set of linear constraints of the types $u_{ij} \in [a_k, b_k]$, $u_{ij} \geq u_{kl}$ and, for all consequences $\{c_{ij}\}$ in D, $u_{ij} \in [0,1]$. A probability base has the same structure, but, for all nodes N (except the root node) in D, also includes $\sum_{j=1}^{m_i} p_{ij} = 1$ for the children $\{c_{ij}\}_{j=1,\ldots,m_i}$ of N.*

Since a vector in the polytope can be considered to represent a distribution, a probability base $\mathcal{P}$ can be interpreted as constraints defining the set of all possible probability measures over the consequences. Similarly, a utility base $\mathcal{U}$ consists of constraints defining the set of all possible utility functions over the consequences. The bases $\mathcal{P}$ and $\mathcal{U}$ together with the decision tree constitute the *information frame*.

Primary evaluation rules of a decision tree model are based on the expected utility. Since neither probabilities nor utilities are fixed numbers, the evaluation of the expected utility yields multi-linear expressions.

Definition 4. *Given a decision tree T and an alternative $A_i \in A$ the expression*

$$\mathrm{E}(A_i) = \sum_{i_1=1}^{n_{i_0}} p_{ii_1} \sum_{i_2=1}^{n_{i_1}} p_{ii_1i_2} \cdots \sum_{i_{m-1}=1}^{n_{i_{m-2}}} p_{ii_1i_2\ldots i_{m-2}i_{m-1}} \sum_{i_m=1}^{n_{i_{m-1}}} p_{ii_1i_2\ldots i_{m-2}i_{m-1}i_m} u_{ii_1i_2\ldots i_{m-2}i_{m-1}i_m}$$

where m is the depth of the tree corresponding to A_i, n_{i_k} is the number of possible outcomes following the event with probability p_{i_k}, $p_{\ldots i_j \ldots}$, $j \in [1, \ldots, m]$, denote probability variables and $u_{\ldots i_j \ldots}$ denote utility variables as above, is the expected utility of alternative A_i in T.

Maximisation of such non-linear objective functions subject to linear constraint sets (statements on probability and utility variables) are computationally demanding problems to solve for an interactive decision tool in the general case, using techniques from the area of non-linear programming. In, e.g., [7], [8], and [6], there are discussions about computational procedures reducing the evaluation of non-linear decision problems to systems with linear objective functions, solvable with ordinary linear programming methods. The approach taken is to model probability and utility intervals as constraint sets, containing statements on upper and lower bounds. Furthermore, normalisation constraints for the probabilities are added (representing that the consequences from a parent node are exhaustive and pairwise disjoint). Such constraints are always on the form $\sum_{j=1}^{n} p_{ij} = 1$.

The solution sets to probability and utility constraint sets are polytopes. The evaluation procedures then yield first-order interval estimates of the evaluations, i.e. upper and lower bounds for the expected utilities of the alternatives.

An advantage of approaches using upper and lower probabilities is that they do not require taking particular probability distributions into consideration. On the other hand, the expected utility range resulting from an evaluation is also an interval. To our experience, in real-life decision situations, it is then sometimes hard to discriminate between the alternatives. In effect, an interval based decision procedure keeps all alternatives with overlapping expected utility intervals, even if the overlap is quite small. Therefore, it is interesting to extend the representation of the decision situation using more information, such as distributions over classes of probability and utility measures, in pursuit of more discriminative power.

4 Including Second-Order Information

Basically, distributions have been used for expressing various beliefs over multi-dimensional spaces where each dimension corresponds to, for instance, possible probabilities or utilities of consequences. The distributions can consequently be used to express strengths of beliefs in different vectors in the polytopes.

Beliefs of such kinds are expressed using higher-order distributions (hierarchical models). Approaches for extending the interval representation using distributions over classes of probability and value measures have been developed into various hierarchical models, such as second-order probability theory. A quite early approach was suggested in [11] and [12]. A more recent example is [20] that provides a model for one-level trees similar to [9].

In the following, we will pursue the idea of adding more information and discuss some interesting properties that appear when evaluating second-order models as well as the effects of aggregating such distributions over expected utilities. The main conclusion here is that the actual deep and breadth of the decision tree under consideration is of large importance for the interpretation of the result. We will also see that the detailed shapes of the distributions are not utterly important compared with this and approximates are sufficient.

4.1 Distributions over Information Frames

Interval estimates can be considered as special cases of representations based on distributions over polytopes. For instance, a distribution can be defined to have a positive support only for $x_i \leq x_j$. More formally, the solution set to a probability or utility constraint set is a subset of a unit cube since both variable sets have $[0, 1]$ as their ranges. This subset can be represented by the support of a distribution over the cube.

Definition 5. *Let a unit cube be represented by $B = (b_1, \ldots, b_n)$. The b_i can be explicitly written out to make the labelling of the dimensions clearer. (More rigorously, the unit cube should be represented by all the tuples $(x_1, \ldots, x_n)$ in $[0, 1]^n$.)*

Definition 6. *By a second-order distribution over B, we denote a positive distribution F defined on the unit cube B such that*

$$\int_B F(x)\, dV_B(x) = 1\,,$$

where V_B is the n-dimensional Lebesque measure on B. The set of all second-order distributions over B is denoted by $BD(B)$.

For our purposes here, second-order *probabilities* are an important sub-class of these distributions and will be used below as a measure of belief, i.e. a second-order joint probability distribution. Marginal distributions are obtained from the joint ones in the usual way.

Definition 7. *Let a unit cube $B = (b_1, \ldots, b_n)$ and $F \in BD(B)$ be given. Furthermore, let $B_i^- = (b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n)$. Then*

$$f_i(x_i) = \int_{B_i^-} F(x)\, dV_{B_i^-}(x)$$

is a marginal distribution over the axis b_i.

Such distributions can then straightforwardly be defined over the information frames. However, regardless of the actual shapes of the distributions involved, constraints such as $\sum_{i=1}^n x_i = 1$ must be satisfied since it is not reasonable to believe in an inconsistent point such as $(0.15, 0.25, 0.4, 0.3)$ if the vector is supposed to represent a probability distribution over four mutually exclusive outcomes. Therefore, a convenient and general way of modelling random weights in $[0, 1]$ is the *Dirichlet distribution*.

Definition 8. *Let the notation be as above. Then the probability density function of the Dirichlet distribution is defined as*

$$f_{Dir}(p, \alpha) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots p_n^{\alpha_n - 1}$$

on a set $\{p = (p_1, \ldots, p_n) \mid p_1, p_2, \ldots, p_n \geq 0, \sum p_i = 1\}$, where $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a parameter vector in which each α_i is a positive parameter and $\Gamma(\alpha_i)$ is the Gamma function.

This distribution is particularly popular among Bayesian statisticians because it is conjugate with respect to the multinomial distribution, i.e. if we choose the prior to be the Dirichlet distribution then the posterior will also become Dirichlet. It is also convenient in the sense that it is not hard to choose parameters to reflect our prior knowledge about the weights $p_1, p_2, \ldots, p_n$. If we choose large values for $\alpha_1, \alpha_2, \ldots, \alpha_n$ we obtain small variances, which reflect a large measure of certainty about the probabilities involved.

Formally, this probability density function does not fulfil our requirement for a belief distribution, but as demonstrated in [9], the issue with the dimension loss can be solved using the *the Dirac distribution*, $\delta_p(x)$, with pole at the point p.

Definition 9. Let A be a subset of a unit cube B, and let f be a belief distribution in A. The natural extension $\tilde{f}_A(x)$ of f with respect to A is defined by

$$\tilde{f}_A(x) = \begin{cases} f(x) & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Definition 10. Let A be a subset of B. A distribution g_A over B is called a characteristic distribution for A in B if

$$f(p) = \int_B \delta_p(x)\tilde{f}_A(x)g_A(x)\,dV_B(x)$$

for every belief distribution f over A, and for every point p in A.

Now let $A = \{(p_1, \ldots, p_n) \mid \sum_{i=1}^n p_i = 1\}$ and let g_A be a Dirichlet distribution. From distribution theory follows that for every measurable subset A in a unit cube B, there exists a characteristic distribution for A in B. It also follows that $\tilde{f}_A(x) \cdot g_A(x)$ is a belief distribution over B and equals 0 outside A.

4.2 Marginal Distributions

A marginal distribution of a Dirichlet distribution is a beta distribution. For instance, if the distribution is uniform, the resulting marginal distribution (over an axis) is a polynomial of degree $n-2$, where n is the dimension of a cube B: let $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1$. Then the Dirichlet distribution is uniform and the marginal distribution is

$$f(x_i) = \int\limits_{B_i^-} dV_{B_i^-}(x) = (n-1)(1-x_i)^{n-2}\,.$$

Example 1. *The marginal distribution $f(x_i)$ of the uniform Dirichlet distribution in a 4-dimensional cube is*

$$f(x_i) = \int_0^{1-x_i}\int_0^{1-y-x_i} 6\,dz\,dy = 3(1-2x_i+x_i^2)$$

$$= 3(1-x_i)^2\,.$$

This tendency is the result of a general phenomenon that becomes more emphasised as the dimension increases. As it will be discussed in the next section, this observation of marginal probabilities is important for the analysis of expected values in decision trees and similar structures.

4.3 The Expected Value and its Variance

Consider a decision tree with only one level of events and n alternatives. Let p_i denote probabilities and u_i utilities of the consequences of an alternative A_j. We assume that $u_1, u_2, \ldots, u_n$ can be considered as independent random variables and we denote the mean and the variance of u_i by μ_i and σ_i^2, respectively. We also assume that $p_1, p_2, \ldots, p_n$ are random variables in the interval $[0,1]$ satisfying the condition $\sum_i p_i = 1$.

Using the Dirichlet distribution, the expected value of $\sum_{i=1}^n p_i u_i$ can be calculated straightforwardly. Let y below represent the (uncertain) expected utility of the alternative A_j such that $y = \sum_{i=1}^n p_i u_i$. Then

$$\mathrm{E}(y) = \mathrm{E}\left(\sum_{i=1}^n p_i u_i\right) = \sum_{i=1}^n \mathrm{E}(p_i)\mathrm{E}(u_i) = \sum_{i=1}^n \frac{\alpha_i}{\alpha}\mu_i$$

When calculating the variance, we have to take the dependence of the p_i-variables into account.

We use the convenient formula

$$\mathrm{Var}(y) = \mathrm{E}(y^2) - \mathrm{E}(y)^2$$

where

$$\begin{aligned}
\mathrm{E}(y^2) &= \mathrm{E}\left(\left(\sum_{i=1}^n p_i u_i\right)^2\right) \\
&= \mathrm{E}\left(\sum_{i=1}^n p_i^2 u_i^2\right) + 2\mathrm{E}\left(\sum_{i<j} p_i p_j u_i u_j\right) \\
&= \sum_{i=1}^n \mathrm{E}(p_i^2)\mathrm{E}(u_i^2) + 2\sum_{i<j}\mathrm{E}(p_i p_j)\mathrm{E}(u_i)\mathrm{E}(u_j) \\
&= \sum_{i=1}^n (\mathrm{E}(p_i)^2 + \mathrm{Var}(p_i))(\mathrm{E}(u_i)^2 + \mathrm{Var}(u_i)) \\
&+2\sum_{i<j}(\mathrm{E}(p_i)\mathrm{E}(p_j) + \mathrm{Cov}(p_i,p_j))\mu_i\mu_j \\
&= \sum_{i=1}^n \left(\frac{\alpha_i^2}{\alpha^2} + \frac{\alpha_i(\alpha-\alpha_i)}{\alpha^2(\alpha+1)}\right)(\mu_i^2+\sigma_i^2) \\
&+2\sum_{i<j}\left(\frac{\alpha_i\alpha_j}{\alpha^2} - \frac{\alpha_i\alpha_j}{\alpha^2(\alpha+1)}\right)\mu_i\mu_j \\
&= \sum_{i=1}^n \frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)}(\mu_i^2+\sigma_i^2) + 2\sum_{i<j}\frac{\alpha_i\alpha_j}{\alpha(\alpha+1)}\mu_i\mu_j
\end{aligned}$$

where $\alpha = \sum_i \alpha_i$, and

$$\mathrm{E}(y)^2 = \left(\sum_{i=1}^n \frac{\alpha_i}{\alpha}\mu_i\right)^2 = \sum_{i=1}^n \frac{\alpha_i^2}{\alpha^2}\mu_i^2 + 2\sum_{i<j}\frac{\alpha_i\alpha_j}{\alpha^2}\mu_i\mu_j$$

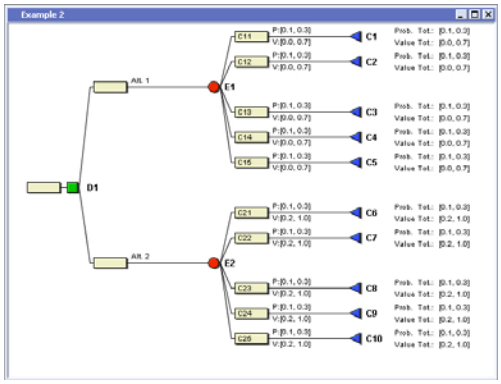


Figure 1: The decision tree in Example 2.

Combining these results yields the variance

$$\mathrm{Var}(y) = \frac{1}{\alpha^2(\alpha+1)}\left(\sum_{i=1}^{n}\alpha_i((\alpha-\alpha_i)\mu_i^2+\alpha(\alpha_i+1)\sigma_i^2)-2\sum_{i<j}\alpha_i\alpha_j\mu_i\mu_j\right)$$

For the uniform case, we obtain

$$\mathrm{E}(y) = \sum_{i=1}^{n}\frac{1}{n}\mu_i = \bar{\mu}$$

and

$$\mathrm{Var}(y) = \frac{1}{n^2(n+1)}\left(\sum_{i=1}^{n}((n-1)\mu_i^2+2n\sigma_i^2)-2\sum_{i<j}\mu_i\mu_j\right)$$

Example 2. *Let an information frame contain a decision tree with two alternatives* A_1 *and* A_2*. Assume that each have five consequences* $C_{i1},\ldots,C_{i5}$ *with probabilities* $p_{ij}\in[0.1,0.3], j=1,\ldots,5, i=1,2$ *and with utilities* $u_{1j}\in[0,0.7], j=1,\ldots,5, u_{2j}\in[0.2,1], j=1,\ldots,5$*. This tree is shown in Figure 1. An interval analysis yields* $\mathrm{E}(A_1)\in[0,0.7]$ *and* $\mathrm{E}(A_2)\in[0.2,1]$*. The major overlap between the two alternatives' expected utility intervals,* $[0.2,0.7]$*, makes it difficult to supply the decision-maker with any advice. If, for example, the distributions over the information frame are uniform, we can see that the distribution of mass over the expected utility clearly discriminates the alternatives. The expected values are* 0.35 *and* 0.6 *and the variances are around* 0.015*. Furthermore, in Figure 2 and Figure 3 the alternatives are entirely separated already for* 75% *of the belief mass (the darker areas). A comparison of the two alternatives is further demonstrated in Figure 4, showing the distribution over the difference* $\mathrm{E}(A_1)-\mathrm{E}(A_2)$*.*

If we do not know any specifics of the underlying distributions, we can utilise *Chebyshev's inequality* which

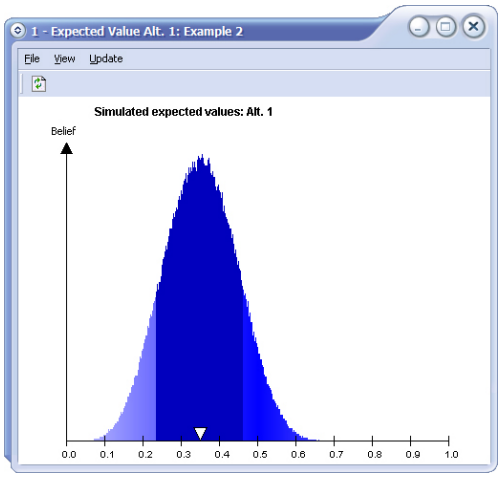


Figure 2: Distribution over $\mathrm{E}(A_1)$ in Example 2.

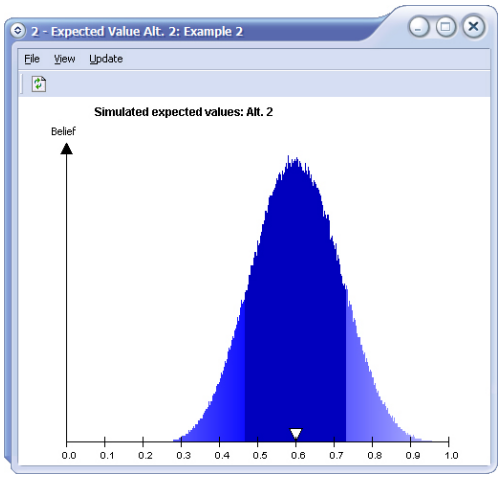


Figure 3: Distribution over $\mathrm{E}(A_2)$ in Example 2.

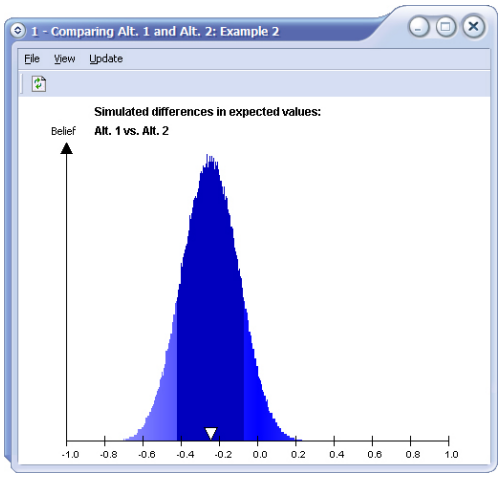


Figure 4: Distribution over $\mathrm{E}(A_1)-\mathrm{E}(A_2)$ in Example 2.

can be formulated in a number of different ways depending on the application. The most common and useful version is

$$\mathrm{P}(|X-\mu| > c\sigma) \leq \frac{1}{c^2}$$

where X is a random variable with mean μ and standard deviation σ and c is an arbitrary constant. For instance, if we want to determine a symmetric 95 % interval around μ, we choose $c = \sqrt{20} = 4.47$. For many classical distributions, this approximation is unfortunately quite rough, even if it is possible to find distributions where equality is attained. For instance, the normal distribution satisfies $\mathrm{P}(|X-\mu| > 1.96\sigma) = 0.05$, which yields an interval being less than half as wide as the Chebyshev approximation.

In any case, it should be noted that we can add information to the decision tree by utilising second-order information. Moreover, the distributions resulting from multiplications generally have shapes very different from their marginal components and we will further investigate this effect below. As will be seen, this has some implications for trees deeper than one level.

4.4 Aggregations

The characteristic of a decision tree is that the marginal (or conditional) probabilities of the event nodes are multiplied in order to obtain the joint probability of a combined event, i.e. of a path from the root to a leaf. In the evaluation of a decision tree the operations involved are multiplications and additions. There are therefore two effects present at the same time when calculating expected utilities in decision trees. Those are additive effects (for joint probabilities aggregated together with the utilities at the leaf nodes) and multiplicative effects (for intermediate probabilities).

One important effect is that multiplied distributions become considerably warped compared to the corresponding component distributions. Such multiplications occur in obtaining the expected utility in decision trees and probabilistic networks, enabling discrimination while still allowing overlap. Properties of additions of components follow from ordinary convolution, i.e. there is a strong tendency towards the middle.

We will now investigate the combined effect and consider how to put second-order information into use to further discriminate between alternatives. The main idea is not to require a total lack of overlap but rather allowing overlap by interval parts carrying little belief mass, i.e. representing a very small part of the decision-maker's belief. Then, the non-overlapping parts can be thought of as being the core of the decision-maker's appreciation of the decision situation, thus allowing discrimination. In addition, effects from varying belief (i.e. differing forms of belief distribution) should be taken into account.

Evaluations of expected utilities in trees lead to multiplication of probabilities using a type of "multiplicative convolution" of two densities.

Let G be a distribution over the two cubes A and B. Assume that G has a positive support on the feasible probability distributions at level i in a decision tree, i.e. is representing these (the support of G in cube A), as well as on the feasible probability distributions of the children of a node x_{ij}, i.e. $x_{ij1}, x_{ij2}, \ldots, x_{ijm}$ (the support of G in cube B). Let $f(x)$ and $g(y)$ be the marginal distributions of $G(z)$ on A and B, respectively.

Definition 11. *The cumulative distribution of the two belief distributions $f(x)$ and $g(y)$ is*

$$H(z) = \iint\limits_{\Gamma_z} f(x)g(y)\,dx\,dy = \int\limits_0^1 \int\limits_0^{z/x} f(x)g(y)\,dy\,dx = \int\limits_0^1 f(x)G(z/x)\,dx = \int\limits_z^1 f(x)G(z/x)\,dx\,,$$

where G is a primitive function to g, $\Gamma_z = \{(x,y) \mid x \cdot y \leq z\}$, and $0 \leq z \leq 1$.

Let $h(z)$ be the corresponding density function. Then

$$h(z) = \frac{d}{dz}\int\limits_z^1 f(x)G(z/x)\,dx = \int\limits_z^1 \frac{f(x)g(z/x)}{x}\,dx\,.$$

The addition of such products is analogous to the product rule for standard probabilities and we can use the ordinary convolution of two densities restricted to the cubes. The distribution h on a sum $z = x + y$ of two independent variables associated with belief distributions $f(x)$ and $g(y)$ is therefore given by

$$h(z) = \int\limits_0^z f(x)g(z-x)\,dx\,.$$

Example 3. *Consider an information frame containing an alternative A_1 with depth 3 and with 3 consequences at each event node. Let $p_{1i} \in [0,1], p_{1ij} \in$*

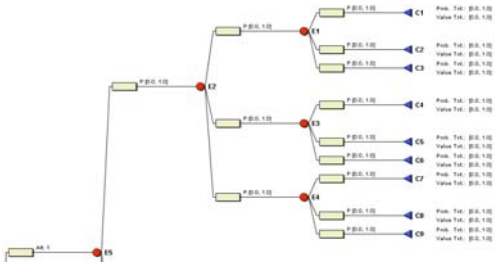


Figure 5: The upper one third of the decision tree in Example 3.

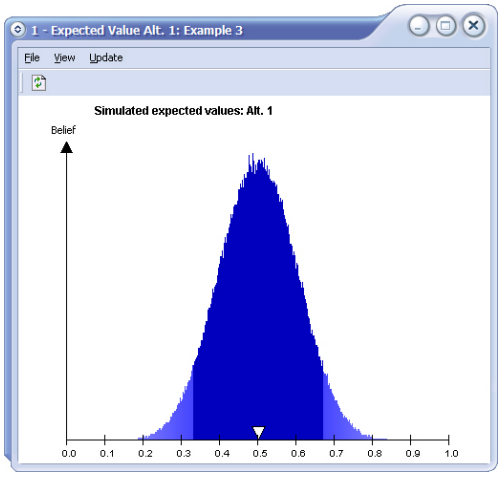


Figure 6: Distribution over $\mathrm{E}(A_1)$ in Example 3.

$[0,1], p_{1ijk} \in [0,1], u_{1ijk} \in [0,1]$, *and* $i,j,k \in \{1,2,3\}$. *This means that no numerical information, except for the trivial constraints, is provided. Part of the tree is shown in Figure 5. Looking at the upper and lower bounds for the expected utility, we find that* $\mathrm{E}(A_1) \in [0,1]$. *If, for example, the second-order distributions over the information frame are uniform, we find that the resulting distribution from each path is* $-4(-12+12z-6\ln(z)-6z\ln(z)-\ln(z)^2+z\ln(z)^2)$ *and that, e.g., 90% of the mass is over the interval* $[0.33, 0.67]$, *see Figure 6.*

As can be seen, second-order data, for instance in terms of Dirichlet distributions, may provide important information in decision evaluation. The example above is taking a particular distribution into account, but as in the previous discussion, these results apply for all types of distributions.

5 Summary and Conclusions

In classic decision theory it is assumed that a decision-maker can assign precise numerical values corresponding to the true value of each consequence, as well as precise numerical probabilities for their occurrences. In attempting to address real-life problems, where uncertainty in the input data prevails, some kind of representation of imprecise information is important and several have been proposed. In particular, representations such as sets of probability measures, upper and lower probabilities, and interval probabilities and utilities of various kinds have been perceived as enabling a better representation of the input sentences for a subsequent decision analysis. However, higher-order analysis can sometimes add important information to the analysis, enabling further discrimination between alternatives.

In this paper, we have discussed the effects of employing second-order information in decision trees. As was seen from Definition 11, the multiplicative effects on probabilities in decision trees increase with tree depth. We have also shown that the multiplicative and additive effects strongly influence the resulting distribution over the expected values.

These effects combined yield a method that sometimes can offer more discriminative power in selecting alternatives in decision trees. The main idea of the method is to allow a small overlap where the belief mass is kept under control. While the discussion focuses on probabilistic decision trees, the results also apply to other formalisms involving products of probabilities, such as probabilistic networks, and to formalisms dealing with other products of interval entities such as interval weight trees in hierarchical multi-criteria decision models.

References

[1] T. Augustin. On Decision Making under Ambiguous Prior and Sampling Information. *Proceedings of ISIPTA'01*, 2001.

[2] T. Augustin. On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View - A Cautionary Note on Updating Imprecise Priors. *Proceedings of ISIPTA'03*, 2001.

[3] G. Choquet. Theory of Capacities. *Ann. Inst. Fourier*, 5: 131–295, 1953/54.

[4] G. de Cooman. Possibility Theory. *International Journal of General Systems*, 25(4): 291–371, 1997.

[5] V. Cutello and J. Montero. Fuzzy Rationality Measures. *Fuzzy Sets and Systems*, 62: 39–54, 1994.

[6] M. Danielson and L. Ekenberg. Computing Upper and Lower Bounds in Interval Decision Trees. *Eu-*

ropean Journal of Operational Research, 181(2): 808–816, 2007.

[7] M. Danielson and L. Ekenberg. A Framework for Analysing Decisions under Risk. *European Journal of Operational Research*, 104(3): 474–484, 1998.

[8] X. S. Ding, M. Danielson and L. Ekenberg. Nonlinear Programming Solvers for Decision Analysis Support Systems. *Operations Research Proceedings 2003 - Selected Papers of the International Conference on Operations Research (OR 2003)*, 475–482, 2004.

[9] L. Ekenberg and J. Thorbiörnson. Second-order Decision Analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(1): 13–38, 2001.

[10] L. Ekenberg, J. Thorbiörnson and T. Baidya. Value Differences using Second Order Distributions. *International Journal of Approximate Reasoning*, 38(1): 81–97, 2005.

[11] P. Gärdenfors and N.-E. Sahlin. Unreliable Probabilities, Risk Taking, and Decision Making. *Synthese*, 53: 361–386, 1982.

[12] P. Gärdenfors and N.-E. Sahlin. Decision Making with Unreliable Probabilities. *British Journal of Mathematical and Statistical Psychology* 36: 240–251, 1983.

[13] P. J. Huber. The Case of Choquet Capacities in Statistics. *Bulletin of the International Statistical Institute*, 45: 181–188, 1973.

[14] J.-Y. Jaffray. Rational Decision Making with Imprecise Probabilities. *Proceedings of ISIPTA'99*, 1999.

[15] I. Levi. *The Enterprise of Knowledge*, MIT Press, 1980.

[16] I. Levi. On Indeterminate Probabilities. *The Journal of Philosophy*, 71: 391–418, 1974.

[17] R. F. Nau. The Aggregation of Imprecise Probabilities. *Journal of Statistical Planning and Inference*, 105: 265–282, 2002.

[18] H. Raiffa. *Decision Analysis*, Addison Wesley, 1968.

[19] P. Smets. Practical Uses of Belief Functions. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[20] L. Utkin and T. Augustin. Decision Making with Imprecise Second-order Probabilities. *Proceedings of ISIPTA'03*, 547–561, 2003.

[21] L. Utkin. Imprecise Second-order Hierarchical Uncertainty Model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(3): 301–317. 2003.

[22] N. Wilson. A Logic of Extended Probability. *Proceedings of ISIPTA'99*, 1999.

Multiparameter models: Probability distributions parameterized by random sets.

Thomas Fetz
Institut für Grundlagen der Bauingenieurwissenschaften,
Arbeitsbereich Technische Mathematik
Universität Innsbruck, Austria
Thomas.Fetz@uibk.ac.at

Abstract

This paper is devoted to the construction of sets of joint probability measures for the case that the marginal sets of probability measures are generated by probability measures with uncertain parameters where the uncertainty of these parameters is modelled by random sets. Further we show how different conditions on the choice of the weights of the joint focal sets and on the probability measures associated to these sets lead to different sets of joint probability measures including the cases of strong independence, random set independence and unknown interaction.

Keywords. Random sets, lower and upper probabilities, sets of probability measures, parameterized probability measures, sets of joint probability measures, strong independence, random set independence, unknown interaction.

1 Introduction

Let a mapping

$$g : D \subseteq \mathbb{R}^m \longrightarrow \mathbb{R} : \ (x_1, \ldots, x_m) \mapsto g(x_1, \ldots, x_m)$$

be given. The variables x_k are assumed to be uncertain where the uncertainty is modelled by sets of probability measures for each variable separately. What we want to know is the lower and upper probabilities if the value $g(x)$, $x = (x_1, \ldots, x_m)$, is lower (or greater) than a certain value. Therefore we had to propagate the uncertainty of the variables x_k through this multivariate model g, c.f. [1].

As a short motivation we want to mention a few applications where this problem of propagating uncertain variables is arising:

Reliability analyis: In this case the above mapping g is the so called *failure function* where $g(x) \leq 0$ means failure and $g(x) > 0$ means no failure of buildings like bridges and tunnels in civil engineering; or of slopes and dams in geotechnical engineering. The aim is to describe the risk of failure, that means we want to have the upper probability $\overline{P}(\{g(x) \leq 0\})$ of failure. The variables x_k are parameters as elastic modulus E, angle of friction ϕ or flood heights.

Construction management: Here the values of $g(x)$ are costs or durations which should not exceed a certain bound a where the variables x_k are costs, durations or similar parameters as above. Then we want to have the upper probability $\overline{P}(\{g(x) \geq a\})$.

In most cases all these variables are not precisely known, especially parameters arsing in geotechnical engineering are only very vaguely known. In engineering there are several approaches used to describe the uncertainty of these variables: wellknown ones as *probability distributions* or *intervals* and more modern ones as *fuzzy sets* and *random sets*. The uncertainty of the variables is given separately and often modelled by different ways. So a unifying approach is needed to combine and propagate the different models of uncertainty trough the function g. This is provided by the concept of *sets of probability measures* where these sets are generated by random sets which are including the other three approaches (probability distributions, intervals and fuzzy sets).

In some applications the type of probability distributions to discribe the uncertainty of a variable is known, e.g. gaussian distributions, exponential distributions in queueing theory, etreme value distributions in flood risk analysis, but the parameters of these probability distributions are often only vaguely known. In these cases we have to model the uncertainty of the parameters of these distributions. So we introduce here the concept of sets of probability measures which are generated by *parameterized probability measures* where these parameters are uncertain and the uncertainty is described by random sets. All models of uncertainty mentioned before are special cases of this concept.

Since the uncertainty of the variables is given separately, we have to model the joint uncertainty, that means to construct the set of joint probability measures. There are certain ways to generate such sets, e.g. according to strong independence [2, 11] if we assume stochastically independence of the variables, or according to unknown interaction [2] if we do not know how the variables interact, or according to random set independence [3] since random sets are involved. These cases are already studied for sets of probability measures generated by random sets in [5, 6, 7, 8]. Here in this paper we extend this to sets of probability measures generated by parameterized probability measures with uncertain parameters.

To propagate the uncertainty through a multivariate model in a computational efficient way it is essential to make use of the structure of the random sets. We show how different conditions on the parts of this structure (on the choice of the weights of the joint focal sets and on the probability measures associated to these sets) lead to different sets of joint probability measures. But our goal is not to create new artificial types of sets of joint probability measures, but to get sets according to strong independence or unknown interaction by using the random set structure.

The plan of this paper is as follows:

Section 2 is devoted to random sets and the parameterization of probability measures by random sets in the univariate case. In Section 3 we construct sets of joint probability measures which are generated by probability measures which are parameterized by ordinary sets as preliminary work for Sec. 5. In Section 4 we recall from [5, 7] the general formulation for constructing sets of joint probability measures for the case where random sets are involved and list different conditions on choosing the weights of the joint focal sets and the probability measures associated to these sets. In Section 5 we show that some of these cases lead to strong independence, random set independence and unknown interaction.

2 Sets of probability measures generated by probability measures parameterized by random sets

We want to model the uncertainty about the value of a variable x by a convex set $\mathcal{K}$ of probability measures in the univariate case. Here in this paper we generate such sets $\mathcal{K}$ by a parameterized probability measure $p^{\,\theta}$ where $\theta = (\theta^1, \theta^2, \dots)$ are the parameters of the probability measure. These parameters are assumed to be uncertain. The uncertainty of θ is modelled by random sets. So we have to recall the concept of random sets and sets of probability measures generated by random sets. Further we need two different measurable spaces: $(\Omega, \mathcal{A})$ for the uncertain variable itself and $(\Theta, \mathfrak{A})$ for the uncertain parameters of the probability measures.

2.1 Random sets

First we want to model the uncertainty of a variable x by random sets. Let a measurable space $(\Omega, \mathcal{A})$ be given. A random set $(\mathcal{F}, m)$ [3, 4] consists of a finite class

$$\mathcal{F} = \{F^1, F^2, \dots, F^n\} \subseteq \mathcal{A}$$

of focal sets and of a weight function

$$m : \mathcal{F} \longrightarrow [0,1] : F \mapsto m(F)$$

with $\sum_{i=1}^{|\mathcal{F}|} m(F^i) = 1$ where $|\mathcal{F}|$ is the number of focal sets. Then the plausibility measure Pl or upper probability $\overline{P}$ of a set $A \in \mathcal{A}$ is defined by

$$\overline{P}(A) = \mathrm{Pl}(A) = \sum_{F^i \cap A \neq \varnothing} m(F^i)$$

and the belief measure Bel or lower probability $\underline{P}$ by

$$\underline{P}(A) = \mathrm{Bel}(A) = \sum_{F^i \subseteq A} m(F^i).$$

2.2 Sets of probability measures generated by random sets

The focal set F^i has the weight $m(F^i)$, but we do not know how this weight is distributed on the elements of the focal set which reflects the uncertainty modelled by a random set. Let

$$\mathcal{K}(F^i) := \{P : \ P(F^i) = 1\} \tag{1}$$

be the set of all probability measures "on" the focal set F^i. Then $m(F^i)\mathcal{K}(F^i)$ is the set of all possible distributions of the weight on the focal set. A convex set of probability measures $\mathcal{K}$ is generated by the random set $(\mathcal{F}, m)$ as follows [5]:

$$\mathcal{K} := \mathcal{K}(\mathcal{F}, m) := \sum_{i=1}^{|\mathcal{F}|} m(F^i)\, \mathcal{K}(F^i) := \tag{2}$$

$$= \left\{ P : \ P = \sum_{i=1}^{|\mathcal{F}|} m(F^i) P^i, \ P^i \in \mathcal{K}(F^i) \right\}.$$

This set $\mathcal{K}(\mathcal{F}, m)$ coincides with the set of probability measures defined by

$$\{P : \ \forall A \in \mathcal{A} : \mathrm{Bel}(A) \leq P(A) \leq \mathrm{Pl}(A)\},$$

c.f. [3, 4, 10].

Remark: There is a second approach of defining random sets: using multivalued mappings and measurable selections [9, 10]. This approach leads to a set $\mathcal{M}$ of probability measures which is a subset of $\mathcal{K}$ and which is not convex in general. The set of probability measures associated to the measurable selctions is $P_{\overline{\Omega}}(\Gamma) = \{P_X : \ X \in S(\Gamma)\}$, where $\Gamma : \overline{\Omega} \to \mathcal{A}$ is a multivalued mapping defined on a probability space $(\overline{\Omega}, \overline{\mathcal{A}}, P_{\overline{\Omega}})$. $S(\Gamma)$ is the set of measurable selections of Γ, that means the class of random variables $X : \overline{\Omega} \to \Omega$ with $X(\overline{\omega}) \in \Gamma(\overline{\omega})$.

Now let $P_X \in P(\Gamma)$, $X \in S(\Gamma)$, be given. Then

$$P_X(A) = P_{\overline{\Omega}}(X^{-1}(A)) = \sum_{i=1}^{|\overline{\Omega}|} P_{\overline{\Omega}}(\{\overline{\omega}^i\})\chi_A(X(\overline{\omega}^i))$$
$$= \sum_{i=1}^{|\mathcal{F}|} m(F^i)\chi_A(\omega^i) = \sum_{i=1}^{|\mathcal{F}|} m(F^i)\delta_{\omega^i}(A)$$

with $\omega^i = X(\overline{\omega}^i) \in \Gamma(\overline{\omega}^i) = F^i$ and $P_{\overline{\Omega}}(\{\overline{\omega}^i\}) = m(F^i)$ where χ_A is the indicator function of A.

So in our above notation the set $\mathcal{M}$ would be generated by

$$\mathcal{M} := \mathcal{M}(\mathcal{F}, m) := \sum_{i=1}^{|\mathcal{F}|} m(F^i)\, \mathcal{M}(F^i)$$

with

$$\mathcal{M}(F^i) = \{\delta_\omega : \delta_\omega(F^i) = 1\} = \{\delta_\omega : \omega \in F^i\} \subseteq \mathcal{K}(F^i) \tag{3}$$

where δ_ω is the Dirac measure at $\omega \in \Omega$ corresponding to the selections. The connections between $\mathcal{M}$ and $\mathcal{K}$ are discussed in [9, 10].

2.3 Sets of parameterized probability measures

Now we generate the set $\mathcal{K}$ of probability measures by a probability measure $p^{\,\theta}$ on $(\Omega, \mathcal{A})$ which is parameterized by an uncertain θ. For modelling the uncertainty of the parameter θ we need the following: A measurable space $(\Theta, \mathfrak{A})$ where Θ is the universal set for θ, $\mathfrak{A}$ a σ-Algebra and $\mathfrak{K}$ a set of probability measures μ on $(\Theta, \mathfrak{A})$. The σ-Algebra $\mathfrak{A}$ has to be chosen in a way that for all $A \in \mathcal{A}$ the mapping

$$\theta \mapsto p^{\,\theta}(A)$$

is $\mathfrak{A}$-measurable.

The set $\mathcal{K}$ is defined by

$$\mathcal{K} := \mathcal{K}(\mathfrak{K}, p^{\,\theta}) := \left\{ P = \int_\Theta p^{\,\theta}(\cdot)\mu(\mathrm{d}\theta) : \ \mu \in \mathfrak{K} \right\}. \tag{4}$$

Then the upper and lower probabilities for a set $A \in \mathcal{A}$ is computed as follows:

$$\overline{P}(A) = \sup\{P(A) : \ P \in \mathcal{K}\} = \sup_{\mu \in \mathfrak{K}} \int_\Theta p^{\,\theta}(A)\ \mu(\mathrm{d}\theta),$$
$$\underline{P}(A) = \inf\{P(A) : \ P \in \mathcal{K}\} = \inf_{\mu \in \mathfrak{K}} \int_\Theta p^{\,\theta}(A)\ \mu(\mathrm{d}\theta).$$

In the following the set $\mathfrak{K}$ is either a set of probability measures generated by ordinary sets or by random sets. The usage and meaning of the symbols $\mathcal{K}$ and $\mathfrak{K}$ is summarized in the following table:

notation	set of probability measures
$\mathcal{K}(F)$	on $(\Omega, \mathcal{A})$ generated by a set F
$\mathfrak{K}(F)$	on $(\Theta, \mathfrak{A})$ generated by a set F
$\mathcal{K}(\mathcal{F}, m)$	on $(\Omega, \mathcal{A})$ gen. by a random set $(\mathcal{F}, m)$
$\mathfrak{K}(\mathcal{F}, m)$	on $(\Theta, \mathfrak{A})$ gen. by a random set $(\mathcal{F}, m)$
$\mathcal{K}(\mathfrak{K}, p^{\,\theta})$	on $(\Omega, \mathcal{A})$ gen. by $\mathfrak{K}$ and $p^{\,\theta}$ as in (4) and where $\mathfrak{K}$ is either a $\mathfrak{K}(F)$ or $\mathfrak{K}(\mathcal{F}, m)$

So $\mathcal{K}$ is always a set of probability measures on $(\Omega, \mathcal{A})$ and $\mathfrak{K}$ a set of probability measures on the parameter space of θ, namely Θ.

2.4 Generation of $\mathfrak{K}$ by probability measures μ on ordinary sets F, $\mathfrak{K} := \mathfrak{K}(F)$

We take the set $\mathfrak{K} := \mathfrak{K}(F)$ of probability measures on $F \in \mathfrak{A}$ and $\mathcal{K} := \mathcal{K}(\mathfrak{K}(F), p^{\,\theta})$ for the set $\mathcal{K}$ of probability measures which are generated by $\mathfrak{K}(F)$ and the parameterized probability measure $p^{\,\theta}$. Then the upper and lower probability are given by

$$\overline{P}(A) = \sup_{\mu \in \mathfrak{K}(F)} \int_\Theta p^{\,\theta}(A)\ \mu(\mathrm{d}\theta) = \tag{5}$$
$$= \sup_{\theta_0 \in F} \int_\Theta p^{\,\theta}(A)\ \delta_{\theta_0}(\mathrm{d}\theta) = \sup_{\theta_0 \in F} p^{\,\theta_0}(A)$$

and $\underline{P}(A) = \inf_{\theta_0 \in F} p^{\,\theta_0}(A)$. Further we have for the special case $(\Theta, \mathfrak{A}) := (\Omega, \mathcal{A})$ and $p^{\,\omega} := \delta_\omega$:

$$\mathcal{K}(F) = \mathcal{K}(\mathfrak{K}(F), \delta_\omega), \tag{6}$$

because

$$\mathcal{K}(\mathfrak{K}(F), \delta_\omega) = \left\{ P = \int_\Omega \delta_\omega(\cdot)\mu(\mathrm{d}\omega) : \ \mu \in \mathfrak{K}(F) \right\} =$$
$$= \{\mu \in \mathfrak{K}(F)\} = \mathfrak{K}(F) = \mathcal{K}(F)$$

and $\omega \mapsto p^{\,\omega}(A) = \delta_\omega(A) = \chi_A(\omega)$ is $\mathcal{A}$-measurable for all $A \in \mathcal{A}$. So the set of probability measures generated by an ordinary set is integrated into the new concept.

2.5 Generation of $\mathfrak{K}$ by random sets, $\mathfrak{K} := \mathfrak{K}(\mathcal{F}, m)$

Here we take $\mathfrak{K} := \mathfrak{K}(\mathcal{F}, m)$ and $\mathcal{K} := \mathcal{K}(\mathfrak{K}(\mathcal{F}, m), p^{\,\theta})$. A probability measure $P \in \mathcal{K}$ is written as follows:

$$\begin{aligned} P &= \int_\Theta p^{\,\theta}(\cdot)\, \mu(\mathrm{d}\theta) = \\ &= \int_\Theta p^{\,\theta}(\cdot) \left(\sum_{i=1}^{|\mathcal{F}|} m(F^i)\, \mu^i(\mathrm{d}\theta) \right) = \\ &= \sum_{i=1}^{|\mathcal{F}|} m(F^i) \int_\Theta p^{\,\theta}(\cdot)\, \mu^i(\mathrm{d}\theta) = \sum_{i=1}^{|\mathcal{F}|} m(F^i) P^i \end{aligned}$$

where $\mu \in \mathfrak{K}(\mathcal{F}, m)$. $\mu = \sum_{i=1}^{|\mathcal{F}|} m(F^i)\mu^i$ is a decomposition of μ according to the focal sets and $P^i = \int_\Theta p^{\,\theta}(\cdot)\mu^i(\mathrm{d}\theta)$ is a probability measure in $\mathcal{K}(\mathfrak{K}(F^i), p^{\,\theta})$. So for the set $\mathcal{K}(\mathfrak{K}(\mathcal{F}, m), p^{\,\theta})$ we also can write

$$\mathcal{K}(\mathfrak{K}(\mathcal{F}, m), p^{\,\theta}) = \sum_{i=1}^{|\mathcal{F}|} m(F^i)\, \mathcal{K}(\mathfrak{K}(F^i), p^{\,\theta}) \qquad (7)$$

which is formula Eq. (2) but with $\mathcal{K}(F^i)$ replaced by $\mathcal{K}(\mathfrak{K}(F^i), p^{\,\theta})$. The set $\mathcal{K}(F^i)$ used in Eq. (2) is a set of probability measures on F^i, but the probability measures in the set $\mathcal{K}(\mathfrak{K}(F^i), p^{\,\theta})$ are only associated to F^i via the parameter θ.

Similar to the section above we have for the upper and lower probability:

$$\begin{aligned} \overline{P}(A) &= \sum_{i=1}^{|\mathcal{F}|} m(F^i) \sup_{\mu^i \in \mathfrak{K}(F^i)} \int_\Theta p^{\,\theta}(A) \mu^i(\mathrm{d}\theta) = \\ &= \sum_{i=1}^{|\mathcal{F}|} m(F^i) \sup_{\theta_0 \in F^i} p^{\,\theta_0}(A) = \sum_{i=1}^{|\mathcal{F}|} m(F^i) \overline{P}^i(A) \end{aligned}$$

and

$$\underline{P}(A) = \sum_{i=1}^{|\mathcal{F}|} m(F^i) \inf_{\theta_0 \in F^i} p^{\,\theta_0}(A) = \sum_{i=1}^{|\mathcal{F}|} m(F^i) \underline{P}^i(A).$$

2.6 An example for $p^{\,\theta}$, gaussian distribution

$(\Omega, \mathcal{A}) := (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $(\Theta, \mathfrak{A}) = (\mathbb{R} \times \mathbb{R}^{>0}, \mathcal{B}(\mathbb{R} \times \mathbb{R}^{>0}))$, $\theta := (\mu, \sigma^2)$. The function

$$(\mu, \sigma^2) \mapsto p^{(\mu,\sigma^2)}(A) := \int_{\mathbb{R}} \chi_A(x) \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x$$

is continuous and therefore $\mathfrak{A}$-measurable. We compute the upper and lower probability of the set $A = [a, \infty)$ using a set $\mathfrak{K}(\mathcal{F}, m)$ to describe the uncertainty of μ and σ^2 with $F^i = [\underline{\mu}_i, \overline{\mu}_i] \times [\underline{\sigma}_i^2, \overline{\sigma}_i^2]$, $i = 1, \ldots, n$, as follows:

$$\overline{P}^i([a, \infty)) = \begin{cases} p^{(\overline{\mu}_i, \underline{\sigma}_i^2)}([a, \infty)) & \overline{\mu}_i \geq a, \\ p^{(\overline{\mu}_i, \overline{\sigma}_i^2)}([a, \infty)) & \text{otherwise,} \end{cases}$$

and

$$\underline{P}^i([a, \infty)) = \begin{cases} p^{(\underline{\mu}_i, \underline{\sigma}_i^2)}([a, \infty)) & \underline{\mu}_i \leq a, \\ p^{(\underline{\mu}_i, \overline{\sigma}_i^2)}([a, \infty)) & \text{otherwise.} \end{cases}$$

Then

$$\overline{P}([a, \infty)) = \sum_{i=1}^{n} m(F^i) \overline{P}^i([a, \infty))$$

and

$$\underline{P}([a, \infty)) = \sum_{i=1}^{n} m(F^i) \underline{P}^i([a, \infty)).$$

3 Sets of joint probability measures generated by ordinary sets

3.1 Preliminaries

In this paper we restrict ourselves to the combination of only two sets of probability measures. In the following we always need the measurable spaces $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$ and $(\Omega, \mathcal{A})$ with $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$ for the uncertain variables and $(\Theta_1, \mathfrak{A}_1)$, $(\Theta_2, \mathfrak{A}_2)$ and $(\Theta, \mathfrak{A})$ with $\Theta = \Theta_1 \times \Theta_2$ for the uncertain parameters of the probabilty measures with σ-Algebras such that mappings like $\theta \mapsto p^{\,\theta}(A)$ are measurable. A set A will be always in $\mathcal{A}$.

The generation of a set of joint probability measures by two marginal sets $\mathcal{K}_1$ and $\mathcal{K}_2$ of probability measures will be written as $\mathcal{K}(\mathcal{K}_1, \mathcal{K}_2)$. First we recall two general ways of combining sets $\mathcal{K}_1$ and $\mathcal{K}_2$ of probability measures.

Unknown interaction: The set of joint probability measures according to unknown interaction [2] is generated by

$$\mathcal{K}_{\mathrm{U}} := \{P :\ P(\cdot \times \Omega_2) \in \mathcal{K}_1,\ P(\Omega_1 \times \cdot) \in \mathcal{K}_2\}. \quad (\mathrm{U})$$

Strong independence: The set of joint probability measures according to strong independence [2, 11] is generated by

$$\mathcal{K}_{\mathrm{S}} := \{P_1 \otimes P_2 :\ P_1 \in \mathcal{K}_1,\ P_2 \in \mathcal{K}_2\} \subseteq \mathcal{K}_{\mathrm{U}}. \quad (\mathrm{S})$$

Notation for the corresponding probabilities:

$$\begin{aligned}
\overline{P}_{\mathrm{S}}(A) &:= \sup\{P_{\mathrm{S}}(A):\ P_{\mathrm{S}} \in \mathcal{K}_{\mathrm{S}}\},\\
\underline{P}_{\mathrm{S}}(A) &:= \inf\{P_{\mathrm{S}}(A):\ P_{\mathrm{S}} \in \mathcal{K}_{\mathrm{S}}\},\\
\overline{P}_{\mathrm{U}}(A) &:= \sup\{P_{\mathrm{U}}(A):\ P_{\mathrm{U}} \in \mathcal{K}_{\mathrm{U}}\},\\
\underline{P}_{\mathrm{U}}(A) &:= \inf\{P_{\mathrm{U}}(A):\ P_{\mathrm{U}} \in \mathcal{K}_{\mathrm{U}}\}.
\end{aligned}$$

In the following we are analyzing very special cases of sets of joint probability measures which is a preliminary work for Sec. 4 and 5 for dealing with the joint focals sets in these sections.

3.2 $\mathcal{K}_k := \mathcal{K}(F_k) = \mathcal{K}(\mathfrak{K}(F_k), \delta_{\omega_k})$

Given subsets $F_k \in \mathcal{A}_k$, $k = 1, 2$, we generate the sets $\mathcal{K}_{\mathrm{U}}$ and $\mathcal{K}_{\mathrm{S}}$ of joint probability measures by the sets $\mathcal{K}(F_1)$ and $\mathcal{K}(F_2)$. $\mathcal{K}_{\mathrm{U}}(\mathcal{K}(F_1), \mathcal{K}(F_2))$ is the set of joint probability measures generated by the two sets $\mathcal{K}(F_1)$ and $\mathcal{K}(F_2)$ of probability measures according to (U). Since the marginals of all probability measures on $F_1 \times F_2$ are in the sets $\mathcal{K}(F_1)$ and $\mathcal{K}(F_2)$, respectively, we have $\mathcal{K}_{\mathrm{U}}(\mathcal{K}(F_1), \mathcal{K}(F_2)) = \mathcal{K}(F_1 \times F_2)$. To get the upper and lower probability $\overline{P}_{\mathrm{U}}(A)$ and $\underline{P}_{\mathrm{U}}(A)$ it is sufficient to put a Dirac measure at the appropriate place. Since a Dirac measure is a product measure we get

$$\overline{P}_{\mathrm{S}}(A) = \overline{P}_{\mathrm{U}}(A) \text{ and } \underline{P}_{\mathrm{U}}(A) = \underline{P}_{\mathrm{S}}(A).$$

Now we make a first step towards sets of joint probability measures generated by parameterized probabilities doing the same for $\mathcal{K}\big(\mathfrak{K}(F_k), \delta_{\omega_k}\big)$ in the more general notation. We already know that $\mathcal{K}(F_k) = \mathcal{K}\big(\mathfrak{K}(F_k), \delta_{\omega_k}\big)$ and therefore

$$\begin{aligned}
\mathcal{K}_{\mathrm{U}} &= \mathcal{K}_{\mathrm{U}}\big(\mathcal{K}(F_1), \mathcal{K}(F_2)\big) = \mathcal{K}(F_1 \times F_2) =\\
&= \mathcal{K}\big(\mathfrak{K}(F_1 \times F_2), \delta_{\omega_1} \otimes \delta_{\omega_2}\big).
\end{aligned}$$

Further we have for strong independence

$$\begin{aligned}
\mathcal{K}_{\mathrm{S}} &= \mathcal{K}_{\mathrm{S}}\big(\mathcal{K}(F_1), \mathcal{K}(F_2)\big) =\\
&= \mathcal{K}_{\mathrm{S}}\big(\mathcal{K}(\mathfrak{K}(F_1), \delta_{\omega_1}), \mathcal{K}(\mathfrak{K}(F_2), \delta_{\omega_2})\big) =\\
&= \mathcal{K}\big(\mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1), \mathfrak{K}(F_2)), \delta_{\omega_1} \otimes \delta_{\omega_2}\big),
\end{aligned}$$

because

$$\begin{aligned}
P_{\mathrm{S}}(A) &= \Big(P_1 \otimes P_2\Big)(A) = \int\limits_{\Omega_1} P_2(A_{\omega_1}) P_1(\mathrm{d}\omega_1) = \quad (8)\\
&= \int\limits_{\Omega_1} \Big(\int\limits_{\Omega_2}\int\limits_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2) \delta_{\omega_2'}(\mathrm{d}\omega_2) \mu_2(\mathrm{d}\omega_2') \Big) P_1(\mathrm{d}\omega_1) =\\
&= \int\limits_{\Omega_1}\int\limits_{\Omega_1} \Big(\int\limits_{\Omega_2}\int\limits_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2) \delta_{\omega_2'}(\mathrm{d}\omega_2) \mu_2(\mathrm{d}\omega_2') \Big) \cdot\\
&\qquad \cdot \delta_{\omega_1'}(\mathrm{d}\omega_1) \mu_1(\mathrm{d}\omega_1') =\\
&= \int\limits_{\Omega_1}\int\limits_{\Omega_1} \Big(\int\limits_{\Omega_2}\int\limits_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2)\ \delta_{\omega_2'}(\mathrm{d}\omega_2) \delta_{\omega_1'}(\mathrm{d}\omega_1) \Big) \cdot\\
&\qquad \cdot \mu_2(\mathrm{d}\omega_2') \mu_1(\mathrm{d}\omega_1') =\\
&= \int\limits_{\Omega_1}\int\limits_{\Omega_2} \Big[\Big(\delta_{\omega_1'} \otimes \delta_{\omega_2'} \Big)(A) \Big]\ \mu_2(\mathrm{d}\omega_2') \mu_1(\mathrm{d}\omega_1') =\\
&= \int\limits_{\Omega_1 \times \Omega_2} \Big[\Big(\delta_{\omega_1'} \otimes \delta_{\omega_2'} \Big)(A) \Big]\ \mu(\mathrm{d}(\omega_1', \omega_2'))
\end{aligned}$$

with $\mu_1 \in \mathfrak{K}(F_1)$, $\mu_2 \in \mathfrak{K}(F_2)$, $\mu \in \mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1), \mathfrak{K}(F_2))$ and $A_{\omega_1} = \{\omega_2 \in \Omega_2 :\ (\omega_1, \omega_2) \in A\}$.

3.3 $\mathcal{K}_k := \mathcal{K}(\mathfrak{K}(F_k), p_k^{\theta_k})$

Now we replace the Diracs by parameterized probability measures $p_k^{\theta_k}$ and analyze the cases of strong independence and unknown interaction.

3.3.1 Strong independence

Similar to Eq. (8) it holds:

$$\begin{aligned}
P_{\mathrm{S}}(A) &= \int\limits_{\Theta_1}\int\limits_{\Omega_1} \Bigg(\int\limits_{\Theta_2}\int\limits_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2) p_2^{\theta_2}(\mathrm{d}\omega_2) \mu_2(\mathrm{d}\theta_2) \Bigg) \cdot\\
&\qquad \cdot p_1^{\theta_1}(\mathrm{d}\omega_1) \mu_1(\mathrm{d}\theta_1) =\\
&= \int\limits_{\Theta_1}\int\limits_{\Theta_2} \Bigg(\int\limits_{\Omega_1}\int\limits_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2)\ p_2^{\theta_2}(\mathrm{d}\omega_2) p_1^{\theta_1}(\mathrm{d}\omega_1) \Bigg) \cdot\\
&\qquad \cdot \mu_2(\mathrm{d}\theta_2) \mu_1(\mathrm{d}\theta_1) =\\
&= \int\limits_{\Theta_1}\int\limits_{\Theta_2} \Big[\Big(p_1^{\theta_1} \otimes p_2^{\theta_2} \Big)(A) \Big]\ \mu_2(\mathrm{d}\theta_2) \mu_1(\mathrm{d}\theta_1).
\end{aligned}$$

So we get

$$\mathcal{K}_{\mathrm{S}} = \mathcal{K}\big(\mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1), \mathfrak{K}(F_2)), p_1^{\theta_1} \otimes p_2^{\theta_2}\big).$$

3.3.2 Unknown interaction

For strong independence the joint probability measure generated by $p_1^{\theta_1}$ and $p_2^{\theta_2}$ was $p_1^{\theta_1} \otimes p_2^{\theta_2}$, a single probability measure. In case of unknown interaction we would need the whole set of all possible joint probability measures on $(\Omega, \mathcal{A})$. Maybe on the other hand we have more information how the joint probability measure, say p^θ, is generated by $p_1^{\theta_1}$ and $p_2^{\theta_2}$ than how the parameters of the joint probability measure interact. So we introduce the sets $\mathcal{K}_{(\mathrm{US})}$ and $\mathcal{K}_{(\mathrm{Up}^\theta)}$ of joint probability measures for which the choice of μ is according to (U) and the choice of the joint parameterized probability measure is according to (S) or defined by p^θ.

Then it holds:

$$\begin{aligned}\mathcal{K}_{\mathrm{S}} :&= \mathcal{K}\big(\mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1),\mathfrak{K}(F_2)), p_1^{\theta_1}\otimes p_2^{\theta_2}\big) \subseteq \\ &\subseteq \mathcal{K}\big(\mathfrak{K}_{\mathrm{U}}(\mathfrak{K}(F_1),\mathfrak{K}(F_2)), p_1^{\theta_1}\otimes p_2^{\theta_2}\big) = \\ &= \mathcal{K}\big(\mathfrak{K}(F_1\times F_2), p_1^{\theta_1}\otimes p_2^{\theta_2}\big) =: \mathcal{K}_{(\mathrm{US})}.\end{aligned}$$

For the upper and lower probabilities we have

$$\overline{P}_{\mathrm{S}}(A) = \overline{P}_{(\mathrm{US})}(A) \text{ und } \underline{P}_{\mathrm{S}}(A) = \underline{P}_{(\mathrm{US})}(A),$$

because we can obtain the upper and lower probabilities from $\mathcal{K}_{(\mathrm{US})}$ by means of Dirac measures in $\mathfrak{K}(F_1\times F_2)$ which are also in $\mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1),\mathfrak{K}(F_2))$.

4 General formulation of the generation of sets of joint probability measures by random sets

Let random sets $(\mathcal{F}_k, m_k)$, $k=1,2$, be given for modelling the uncertainty of the variables x_1 and x_2. As a consequence of Dempster's rule of combination [3, 4] the joint random set $(\mathcal{F}, m)$ is defined by

$$\mathcal{F} = \{F^{ij} :\ i=1,\dots,n_1;\ j=1,\dots,n_2\}$$

where

$$F^{ij} := F_1^i \times F_2^j$$

and

$$m(F_1^i\times F_2^j) := m_1(F_1^i)m_2(F_2^j) \tag{9}$$

which is the case of random set independence (RS-independence).

For our more general approach we start with the multivariate analogon of Eq. (2):

$$\mathcal{K}_? = \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i\times F_2^j)\mathcal{K}_?(\mathcal{K}_1^i,\mathcal{K}_2^j)$$

where the question mark in $\mathcal{K}_?(\mathcal{K}_1^i,\mathcal{K}_2^j)$ indicates the possibility of different choices in combining the sets of probability measures $\mathcal{K}_1^i$ and $\mathcal{K}_2^j$ associated with the marginal focal sets F_1^i and F_2^j. Further we have to define how the joint weights $m(F_1^i\times F_2^j)$ are computed (perhaps not in the way of Eq. (9)) and to think about possible interactions between probability measures in the different sets $\mathcal{K}_?(\mathcal{K}_1^i,\mathcal{K}_2^j)$.

The consequences of these different choices are different sets of joint probability measures $\mathcal{K}_?$ and the goal is to generate sets according to strong independence, unknown interaction and RS-independence. In the following we describe the different choices we have for the above formula and discuss their consequences for the set of joint probability measures.

4.1 The choice of the joint weights $m(F_1^i\times F_2^j)$

The weights m_1 and m_2 are discrete probability measures on the sets of focal sets $\{F_1^1,\dots,F_1^{n_1}\}$, $\{F_2^1,\dots,F_2^{n_2}\}$ respectively. So if we want to choose the joint focal sets in a stochastically independent way, then $m = m_1\otimes m_2$ which means $m(F_1^i\times F_2^j) = m_1(F_1^i)m_2(F_2^j)$ for all i, j. If we do not know how m_1 and m_2 interact, we allow all possible combinations, that means unknown interaction.

Case (U−−): Unknown interaction, m must satisfy the following conditions:

$$\begin{aligned}m_1(F_1^i) &= \sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i\times F_2^j), \quad i=1,\dots,|\mathcal{F}_1|,\\ m_2(F_2^j) &= \sum_{i=1}^{|\mathcal{F}_1|} m(F_1^i\times F_2^j), \quad j=1,\dots,|\mathcal{F}_2|.\end{aligned}$$

In this case m is not uniquely defined and is determined later on by solving an optimization problem for the lower or upper probabilities.

Case (S−−): Stochastic independence:

$$m(F_1^i\times F_2^j) := m_1(F_1^i)m_2(F_2^j).$$

4.2 The choice of P^{ij}, $\mathcal{K}^{ij}$, respectively

$P^{ij}\in\mathcal{K}^{ij}$ is a probability measure associated to the joint focal set $F_1^i\times F_2^j$. How a P^{ij} looks like depends on how $\mathcal{K}^{ij}$ is constructed from $\mathcal{K}_1^i$ and $\mathcal{K}_2^j$.

Case (−U−): $\mathcal{K}_{\mathrm{U}}^{ij} := \mathcal{K}_{\mathrm{U}}(\mathcal{K}_1^i,\mathcal{K}_2^j))$ which is the set of all joint probability measures generated by the sets $\mathcal{K}_1^i$ and $\mathcal{K}_2^j$ according to condition (U).

Case (−S−): $\mathcal{K}_{\mathrm{S}}^{ij} := \mathcal{K}_{\mathrm{S}}(\mathcal{K}_1^i,\mathcal{K}_2^j)$ which is the set generated according to strong independence (S).

4.3 The choice of interactions between the P^{ij}

Case (−−1): Row- and columnwise equality conditions on the marginals of the probability measures on the joint focal sets:

$$\begin{aligned}P_1^i &:= P_1^{i,i1} = \dots = P_i^{i,in_2},\ i=1,\dots,n_1,\\ P_2^j &:= P_2^{j,1j} = \dots = P_i^{j,n_1j},\ j=1,\dots,n_2\end{aligned}$$

where

$$P_1^{i,ik} = P_1^{ik}(\cdot\times\Omega_2) \text{ and } P_2^{j,kj} = P_2^{kj}(\Omega_1\times\cdot).$$

This condition seems to be very artificial, but we need this to get results according to strong independence later on.

Case (−−0): No interactions, this means that we can choose a $P^{ij} \in \mathcal{K}^{ij}$ on $F_1^i \times F_2^j$ irrespective of the probability measures chosen on other joint focal sets.

Remark: It is clear that it should hold that the convex sum

$$\sum_k \frac{1}{m_1(F_1^i)} m(F_1^i \times F_2^j) P_1^{i,ik}$$

is in $\mathcal{K}_1^i$. This is always true for convex sets $\mathcal{K}_1^i$ of probability measures, but for sets which are generated by measurable selections (see Eq. (3)) it is not true in general. In this case one should introduce a more restrictive condition than (−−1).

4.4 The choice of the joint marginals

We emphasize that the choice of the Cartesian products $F_1^i \times F_2^j$ as joint focals is no restriction of generality. Joint focal sets $V \subseteq F_1^i \times F_2^j$ of arbitrary shape can be subsumed in our approach by restricting sets of joint probability measures on $F_1^i \times F_2^j$ to those whose support lies in V. Such subsets would describe specific types of dependence or interaction between the marginal focal sets F_1^i and F_2^j. But such interactions are not investigated in this paper.

5 The different cases

Now we will discuss combinations of the above cases which lead to random set independence, unknown interaction, strong independence. The cases are indicated by indices of the form (ABC) where for example (SU0) means m according (S−−), P^{ij} according to (−U−) and no interaction between the P^{ij}.

We want to stress that again, that it is not our goal to introduce a number of eight (all possible combinations) new types of joint probability measures, but to identify the combinations which leads to the desired types of sets joint probability measures. We do this for RS-independence, unknown interaction and strong independence. In this very technical part we first recall for each of these types the case where "pure random sets" are used, that means the case where no parameterized probabilities are involved. Then we generalize the results to the case of parameterized probabilities. So the sets $\mathcal{K}_k^i$ are first sets of probability measures $\mathcal{K}(F_k^i)$ and then in a second part replaced by sets $\mathcal{K}(\mathfrak{K}(F_k^i), p^{\,\theta})$ associated with the marginal focal set F_k^i.

5.1 (SU0), (SS0) and RS-independence

5.1.1 General formulation

The sets $\mathcal{K}_{\mathrm{SU0}}$ and $\mathcal{K}_{\mathrm{SS0}}$ of joint probability measures are generated by

$$\mathcal{K}_{\mathrm{SU0}} = \sum_{i=1}^{|\mathcal{F}_1|} \sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i) m_2(F_2^j) \mathcal{K}_{\mathrm{U}}(\mathcal{K}_1^i, \mathcal{K}_2^j)$$

$$\mathcal{K}_{\mathrm{SS0}} = \sum_{i=1}^{|\mathcal{F}_1|} \sum_{j=1}^{|\mathcal{F}_2|} m_2(F_2^j) m_2(F_2^j) \mathcal{K}_{\mathrm{S}}(\mathcal{K}_1^i, \mathcal{K}_2^j).$$

5.1.2 $\mathcal{K}_1^i := \mathcal{K}(F_1^i)$, $\mathcal{K}_2^j := \mathcal{K}(F_2^j)$

We obtain the upper probability $\overline{P}_{\mathrm{SU0}}(A)$ for a set $A \in \mathcal{A}$ by

$$\overline{P}_{\mathrm{SU0}}(A) = \sum_{i=1}^{|\mathcal{F}_1|} \sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i) m_2(F_2^j) \overline{P}_{\mathrm{U}}^{ij}(A)$$

where

$$\overline{P}_{\mathrm{U}}^{ij}(A) = \sup\left\{P_{\mathrm{U}}^{ij}(A) :\ P_{\mathrm{U}}^{ij} \in \mathcal{K}_{\mathrm{U}}(\mathcal{K}(F_1^i), \mathcal{K}(F_2^j))\right\}$$

and

$$\mathcal{K}_{\mathrm{U}}(\mathcal{K}(F_1^i), \mathcal{K}(F_2^j)) = \mathcal{K}(F_1^i \times F_2^j).$$

So $\overline{P}_{\mathrm{U}}^{ij}(A)$ is computed very easily by

$$\overline{P}_{\mathrm{U}}^{ij}(A) = \sup\{\delta_\omega(A) :\ \omega \in F_1^i \times F_2^j\} = \\ = \begin{cases} 1 & \exists\, \omega \in A \cap F_1^i \times F_2^j, \\ 0 & \text{else} \end{cases}$$

which leads to the formula for the joint plausibility measure

$$\overline{P}_{\mathrm{R}}(A) := P_{\mathrm{SU0}}(A) = \mathrm{Pl}(A) = \sum_{i,j:\ F_1^i \times F_2^j \cap A \neq \varnothing} m_1(F_1^i) m_2(F_2^j)$$

which is the joint upper probability in the case of RS-independence indicated by the index R. Further we have $\overline{P}_{\mathrm{SU0}} = \overline{P}_{\mathrm{SS0}}$ because

$$\delta_\omega = \delta_{(\omega_1, \omega_2)} = \delta_{\omega_1} \otimes \delta_{\omega_2}.$$

is a product measure (case (−S−)). Similar to the upper probability we get for the lower probability

$$\underline{P}_{\mathrm{R}} := \mathrm{Bel} = \underline{P}_{\mathrm{SU0}} = \underline{P}_{\mathrm{SS0}}.$$

Contrary to the above equalities we have for the corresponding sets of joint probability measures only

$$\mathcal{K}_{\mathrm{R}} := \mathcal{K}_{\mathrm{SU0}} \supseteq \mathcal{K}_{\mathrm{SS0}}.$$

5.1.3 $\mathcal{K}_1^i := \mathcal{K}(\mathfrak{K}(F_1^i), p_1^{\theta_1})$, $\mathcal{K}_2^j := \mathcal{K}(\mathfrak{K}(F_2^j), p_2^{\theta_2})$

An idea would be to define $\mathcal{K}_{\mathrm{R}}$ by $\mathcal{K}_{\mathrm{SU0}}$ as before [6], but then we have the same problem as in Sec. 3.3.2. So another possibility would be to define $\mathcal{K}_{\mathrm{R}} := \mathcal{K}_{\mathrm{S(US)0}}$ or $\mathcal{K}_{\mathrm{R}} := \mathcal{K}_{\mathrm{S(UP^\theta)0}}$.

We start with the case of (SS0) and get

$$\begin{aligned}\mathcal{K}_{\mathrm{SS0}} &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\ \mathcal{K}_{\mathrm{S}}^{ij} \\ &\subseteq \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\mathcal{K}_{(\mathrm{US})}^{ij} = \\ &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\cdot \\ &\qquad \cdot \mathcal{K}\big(\mathfrak{K}(F_1^i \times F_2^j), p_1^{\theta_1} \otimes p_2^{\theta_2}\big) = \\ &= \mathcal{K}\big(\mathfrak{K}(\mathcal{F}, m), p_1^{\theta_1} \otimes p_2^{\theta_2}\big) =: \mathcal{K}_{\mathrm{S(US)0}} =: \mathcal{K}_{\mathrm{R}},\end{aligned}$$

with $\mathcal{K}_{\mathrm{S}}^{ij} := \mathcal{K}_{\mathrm{S}}\big(\mathcal{K}(\mathfrak{K}(F_1^i), p_1^{\theta_1}), \mathcal{K}(\mathfrak{K}(F_2^j), p_2^{\theta_2})\big)$ and $\mathcal{K}_{(\mathrm{US})}^{ij} := \mathcal{K}\big(\mathfrak{K}(F_1^i \times F_2^j), p_1^{\theta_1} \otimes p_2^{\theta_2}\big)$ and Eq. (7). $(\mathcal{F}, m)$ is the joint random set according to RS-independence. $\mathcal{K}_{\mathrm{S(SU)0}}$ is the set of probability measures where the parameterized probability measure is the product measure, but the uncertainty of the parameters of this poduct measure is discribed by the set $\mathfrak{K}(\mathcal{F}, m)$ of joint probability measures which are generated by the random set describing the uncertainty of θ_1 and θ_2.

For the upper and lower probabilities we have $\overline{P}_{\mathrm{SS0}} = \overline{P}_{\mathrm{S(US)0}}$ and $\underline{P}_{\mathrm{SS0}} = \underline{P}_{\mathrm{S(US)0}}$ by the same arguments as in Sec. 3.3.2.

5.2 (UU0), (US0) and unknown interaction

5.2.1 $\mathcal{K}_1^i := \mathcal{K}(F_1^i)$, $\mathcal{K}_2^j := \mathcal{K}(F_2^j)$

Let $\mathcal{K}_{\mathrm{UU0}}$ be the set of probability measures generated according to case (UU0). A computational method for $\overline{P}_{\mathrm{UU0}}(A)$ is obtained in the following way:

$$\begin{aligned}\overline{P}_{\mathrm{UU0}}(A) &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m^*(F_1^i \times F_2^j)\,\overline{P}_{\mathrm{U}}^{ij}(A) = \\ &= \sum_{\substack{i,j:\\ F_1^i \times F_2^j \cap A \neq \varnothing}} m^*(F_1^i \times F_2^j),\end{aligned}$$

where $\overline{P}_{\mathrm{U}}^{ij}(A)$ is computed by the same Dirac measures as for $\overline{P}_{\mathrm{R}}$ and the weights m^* by solving the following linear optimization problem:

$$\sum_{\substack{i,j:\\ F_1^i \times F_2^j \cap A \neq \varnothing}} m(F_1^i \times F_2^j) = \max!$$

subject to condition (U−−). Minimization instead of maximization leads to lower probability $\underline{P}_{\mathrm{UU0}}(A)$.

The set $\mathcal{K}_{\mathrm{UU0}}$ is just the set of probability measures which is generated by the least restrictive conditions on m and P^{ij}. It is proven in [5, 6] that $\mathcal{K}_{\mathrm{U}} = \mathcal{K}_{\mathrm{UU0}}$.

By the same arguments as in the previous cases we get $\overline{P}_{\mathrm{U}} = \overline{P}_{\mathrm{UU0}} = \overline{P}_{\mathrm{US0}}$ and $\underline{P}_{\mathrm{U}} = \underline{P}_{\mathrm{UU0}} = \underline{P}_{\mathrm{US0}}$.

5.2.2 $\mathcal{K}_1^i := \mathcal{K}(\mathfrak{K}(F_1^i), p_1^{\theta_1})$, $\mathcal{K}_2^j := \mathcal{K}(\mathfrak{K}(F_2^j), p_2^{\theta_2})$

Similar to Sec. 5.1.3 we can define sets

$$\begin{aligned}\mathcal{K}_{\mathrm{UU0}} &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i \times F_2^j)\cdot \\ &\qquad \cdot \mathcal{K}_{\mathrm{U}}\big(\mathcal{K}(\mathfrak{K}(F_1^i), p_1^{\theta_1}), \mathcal{K}(\mathfrak{K}(F_2^j), p_2^{\theta_2})\big) \\ \mathcal{K}_{\mathrm{US0}} &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i \times F_2^j)\cdot \\ &\qquad \cdot \mathcal{K}\big(\mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(F_1^i), \mathfrak{K}(F_2^j)), p_1^{\theta_1} \otimes p_2^{\theta_2}\big) \\ \mathcal{K}_{\mathrm{U(US)0}} &= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i \times F_2^j)\cdot \\ &\qquad \cdot \mathcal{K}\big(\mathfrak{K}(F_1^i \times F_2^j), p_1^{\theta_1} \otimes p_2^{\theta_2}\big),\end{aligned}$$

where in addition the joint weights can be choosen according to (U−−); and it also holds $\overline{P}_{\mathrm{US0}}(A) = \overline{P}_{\mathrm{U(US)0}}$, $\underline{P}_{\mathrm{US0}}(A) = \underline{P}_{\mathrm{U(US)0}}$ and $\mathcal{K}_{\mathrm{UU0}} \supseteq \mathcal{K}_{\mathrm{U(US)0}} \supseteq \mathcal{K}_{\mathrm{US0}}$.

But unfortunately we do not have $\mathcal{K}_{\mathrm{U}} = \mathcal{K}_{\mathrm{UU0}}$ in general what we show in the following example.

Example:

The sets $\mathcal{K}_1$ and $\mathcal{K}_2$ of probability measures are given by

$$\mathcal{K}_1 = \mathcal{K}(\mathcal{F}_1, m_1) = \mathcal{K}(\mathfrak{K}(\mathcal{F}_1, m_1), \delta_\omega)$$

and

$$\mathcal{K}_2 = \mathcal{K}(\mathfrak{K}(\mathcal{F}_2, m_2), p_2^{\theta_2})$$

where $p_2^{\theta_2}$ is defined by $p_2^{\theta_2}(\{0\}) = \theta_2$ and $p_2^{\theta_2}(\{1\}) = 1 - \theta_2$ and where

$\Omega_1 = \Omega_2 = \{0, 1\}$, $\Omega = \Omega_1 \times \Omega_2$
$\mathcal{F}_1 = \{\{0\}, \{1\}\}$, $m_1(\{0\}) = m_1(\{1\}) = \frac{1}{2}$,
$\mathcal{F}_2 = \{\{\frac{1}{2}\}\}$, $m_2(\{\frac{1}{2}\}) = 1$.

In this very special example both marginal sets of probability measures have only one element, namely the discrete uniform distribution on $\{0, 1\}$:

$\mathcal{K}_1 = \{P_1^1\}$, $\mathcal{K}_2 = \{P_2^1\}$, $P_1 = P_2$ and

$P_1(\{0\}) = P_1(\{1\}) = \frac{1}{2}$.

But this uniform distribution is "generated" by two different ways:

1. As a degenerated random set where the two focal sets are singletons.

2. As a realization of the parameterized probability measure $p_2^{\theta_2}$ with a parameterization by a random set with only one focal set.

The sets of probability measures associated with the marginal focal sets are given by

$\mathcal{K}_1^1 = \{P_1^1\},\ P_1^1(\{0\}) = 1,$

$\mathcal{K}_1^2 = \{P_1^2\},\ P_1^2(\{1\}) = 1,$

$\mathcal{K}_2^1 = \{P_2^1\} = \{P_2\}.$

Now we determine the joint focal sets and weights:

$\mathcal{F} = \{F^{11}, F^{21}\}$ with $F^{11} = \{(0, \frac{1}{2})\},\ F^{21} = \{(1, \frac{1}{2})\},$

$m(F^{11}) = m(F^{21}) = \frac{1}{2}.$

Since $|\mathcal{F}_2| = 1$ the joint weights are uniquely determined independent of (S−−) or (U−−).

The sets of probability measures associated with the joint focal sets:

$\mathcal{K}_{\mathrm{U}}^{11} = \mathcal{K}_{\mathrm{U}}(\mathcal{K}_1^1, \mathcal{K}_2^1) = \mathcal{K}_{\mathrm{U}}(P_1^1, P_2^1) = \{P_{\mathrm{U}}^{11}\}$ with

$P_{\mathrm{U}}^{11}(\{(0,0)\}) = P_{\mathrm{U}}^{11}(\{(0,1)\}) = \frac{1}{2}$

and

$\mathcal{K}_{\mathrm{U}}^{21} = \mathcal{K}_{\mathrm{U}}(\mathcal{K}_1^2, \mathcal{K}_2^1) = \mathcal{K}_{\mathrm{U}}(P_1^2, P_2^1) = \{P_{\mathrm{U}}^{21}\}$ with

$P_{\mathrm{U}}^{21}(\{(1,0)\}) = P_{\mathrm{U}}^{21}(\{(1,1)\}) = \frac{1}{2}.$

Let $A = \{(0,0), (1,1)\}$. Then

$\overline{P}_{\mathrm{UU0}}(A) = m(F^{11})P_{\mathrm{U}}^{11}(A) + m(F^{21})P_{\mathrm{U}}^{21}(A) =$

$= \frac{1}{2}P_{\mathrm{U}}^{11}(\{(0,0)\}) + \frac{1}{2}P_{\mathrm{U}}^{21}(\{(1,1)\}) =$

$= \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}.$

But it is clear that

$\overline{P}_{\mathrm{U}}(A) = \sup\{P_{\mathrm{U}}(A) :\ P_{\mathrm{U}} \in \mathcal{K}_{\mathrm{U}}(\mathcal{K}_1, \mathcal{K}_2)\} = 1$ for P_{U} defined by $P_{\mathrm{U}}(\{(0,0)\}) = P_{\mathrm{U}}(\{(1,1)\}) = \frac{1}{2}$.

5.3 The case (SS1), strong independence

5.3.1 $\mathcal{K}_1^i := \mathcal{K}(F_1^i)$, $\mathcal{K}_2^j := \mathcal{K}(F_2^j)$

We write a probability measure $P_{\mathrm{SS1}} \in \mathcal{K}_{\mathrm{SS1}}$ in the following way:

$$P_{\mathrm{SS1}}(A) = \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\,(P_1^i \otimes P_2^j)(A) =$$
$$= \left(\sum_{i=1}^{|\mathcal{F}_1|} m_1(F_1^i)\,P_1^i\right) \otimes \left(\sum_{j=1}^{|\mathcal{F}_2|} m_2(F_2^j)\,P_2^j\right)(A) =$$
$$= (P_1 \otimes P_2)(A) = P_{\mathrm{S}}(A)$$

with $P_1 \in \mathcal{K}(\mathcal{F}_1, m_1)$ and $P_2 \in \mathcal{K}(\mathcal{F}_2, m_2)$. This leads to

$$\mathcal{K}_{\mathrm{SS1}} = \mathcal{K}_{\mathrm{S}} =$$
$$= \big\{P_1 \otimes P_2 : P_1 \in \mathcal{K}(\mathcal{F}_1, m_1), P_2 \in \mathcal{K}(\mathcal{F}_2, m_2)\big\}$$

which is the case of strong independence.

Computational method:

Theorem 1. *The upper probability $\overline{P}_{\mathrm{S}}(A)$ is the solution of the following global optimization problem:*

$$\sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m(F_1^i \times F_2^j)\,\chi_A(\omega_1^i, \omega_2^j) = \max!$$

subject to

$$\omega_1^i \in F_1^i,\ i = 1, \ldots, |\mathcal{F}_1|,$$
$$\omega_2^j \in F_2^j,\ j = 1, \ldots, |\mathcal{F}_2|,$$

where χ_A is the indicator function of the set A. The lower probability $\underline{P}_{\mathrm{S}}(A)$ is obtained by minimization.

Proof: see [5, 8].

In general it is very hard to solve the above optimization problem because there may be many local maxima (or minima) and because the objective function is not continuous. Criteria when we have $\overline{P}_{\mathrm{S}} = \overline{P}_{\mathrm{R}}$ are given in [6]. In this case we automatically get $\overline{P}_{\mathrm{S}}$ by using the computationally cheaper $\overline{P}_{\mathrm{R}}$.

5.3.2 $\mathcal{K}_1^i := \mathcal{K}(\mathfrak{K}(F_1^i), p_1^{\theta_1})$, $\mathcal{K}_2^j := \mathcal{K}(\mathfrak{K}(F_2^j), p_2^{\theta_2})$

It holds

$$P_{\mathrm{S}}(A) = P_{\mathrm{SS1}}(A) = \big(P_1 \otimes P_2\big)(A) =$$
$$= \left[\left(\sum_{i=1}^{|\mathcal{F}_1|} m_1(F_1^i)P_1^i\right) \otimes \left(\sum_{j=1}^{|\mathcal{F}_2|} m_2(F_2^j)P_2^j\right)\right](A) =$$
$$= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\Big(P_1^i \otimes P_2^j\Big)(A) =$$
$$= \sum_{i=1}^{|\mathcal{F}_1|}\sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i)m_2(F_2^j)\cdot$$
$$\cdot \int_{\Theta_1}\int_{\Theta_2}\left(\int_{\Omega_1}\int_{\Omega_2} \chi_{A_{\omega_1}}(\omega_2)\; p_2^{\theta_2}(\mathrm{d}\omega_2)p_1^{\theta_1}(\mathrm{d}\omega_1)\right)\cdot$$
$$\cdot\, \mu_2^j(\mathrm{d}\theta_2)\mu_1^i(\mathrm{d}\theta_1) =$$

$$= \sum_{i=1}^{|\mathcal{F}_1|} \sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i) m_2(F_2^j) \cdot$$
$$\cdot \int_{\Theta_1} \int_{\Theta_2} \left[\left(p_1^{\theta_1} \otimes p_2^{\theta_2} \right)(A) \right] \mu_2^j(\mathrm{d}\theta_2) \mu_1^i(\mathrm{d}\theta_1) =$$
$$= \int_{\Theta_1} \int_{\Theta_2} \left[\left(p_1^{\theta_1} \otimes p_2^{\theta_2} \right)(A) \right] \mu_2(\mathrm{d}\theta_2) \mu_1(\mathrm{d}\theta_1) =$$
$$= \int_{\Theta_1 \times \Theta_2} \left[\left(p_1^{\theta_1} \otimes p_2^{\theta_2} \right)(A) \right] \mu(\mathrm{d}(\theta_1, \theta_2))$$

with $\mu \in \mathfrak{K}_{\mathrm{S}}(\mathfrak{K}(\mathcal{F}_1, m_1), \mathfrak{K}(\mathcal{F}_2, m_2))$, $\mu_1^i \in \mathfrak{K}(F_1^i)$, $\mu_2^j \in \mathfrak{K}(F_2^j)$, $\mu_1 \in \mathfrak{K}(\mathcal{F}_1, m_1)$ and $\mu_2 \in \mathfrak{K}(\mathcal{F}_2, m_2)$.

Computational method:

We get the following optimization problem for the computation of $\overline{P}_{\mathrm{S}}(A)$ and $\underline{P}_{\mathrm{S}}(A)$, respectively:

$$\sum_{i=1}^{|\mathcal{F}_1|} \sum_{j=1}^{|\mathcal{F}_2|} m_1(F_1^i) m_2(F_2^j) \left(p_1^{\theta_1^i} \otimes p_2^{\theta_2^j} \right)(A) = \sup! \quad (\inf!)$$

subject to $\theta_1^i \in F_1^i$ and $\theta_2^j \in F_2^j$. Proof: see [6].

6 Summary

We summarize the results where parameterized probabilities are involved: Fig. 1 depicts the relations between the sets of joint probability measures. For the upper probabilities see Fig. 2. There are three differences to the results for "pure random sets" in [5].

1. $\mathcal{K}_{\mathrm{UU0}}$ is only a subset of $\mathcal{K}_{\mathrm{U}}$ in general.
2. New cases induced by $(-(\mathrm{US})-)$ which coincide with $(-\mathrm{U}-)$ for "pure random sets" because of the Dirac measures.
3. Generalization of the computational method for $\overline{P}_{\mathrm{S}}$ and $\underline{P}_{\mathrm{S}}$.

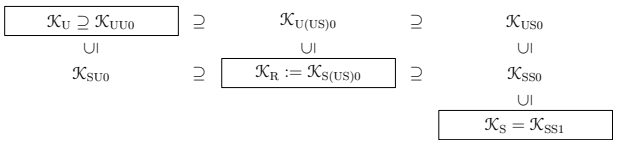


Figure 1: Relations between the sets of probability measures.

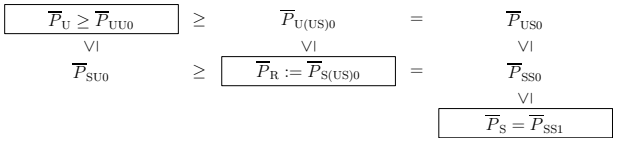


Figure 2: Relations between the upper probabilities.

References

[1] C. Baudrit and D. Dubois. Comparing methods for joint objective and subjective uncertainty propagation with an example in risk assessment. In *Proceedings of the 4th ISIPTA Conference*, Pittsburgh, 2005.

[2] I. Couso, S. Moral, and P. Walley. Examples of independence for imprecise probabilities. In *Proceedings of the 1st ISIPTA Conference*, pages 121–130, Ghent, 1999.

[3] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.*, 38:325–339, 1967.

[4] A. P. Dempster. Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statistics*, 39:957–966, 1968.

[5] Th. Fetz. Sets of joint probability measures generated by weighted marginal focal sets. In *Proceedings of the 2nd ISIPTA Conference*, pages 171–178, Ithaca (NY), 2001.

[6] Th. Fetz. *Mengen von gemeinsamen Wahrscheinlichkeitsmaßen erzeugt von zufälligen Mengen*. PhD thesis, Universität Innsbruck, 2003.

[7] Th. Fetz. Multi-parameter models: rules and computational methods for combing uncertainty. In Oberguggenberger,Vieider, Fellin, Lessmann, editors, *Analyzing Uncertainty in Civil Engineering*, Berlin, 2005. Springer.

[8] Th. Fetz and M. Oberguggenberger. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety*, 85(1-3):73 – 87, 2004.

[9] E. Miranda, I. Couso, and P. Gil. Random intervals as a model for imprecise information. *Fuzzy Sets and Systems*, 154(3):386 – 412, 2005.

[10] E. Miranda, I. Couso, and P. Gil. Random sets as imprecise random variables. *Journal of Mathematical Analysis and Applications*, 307:32–47, 2005.

[11] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, 1991.

An extension of chaotic probability models to real-valued variables

P.I.Fierens
Department of Physics and Mathematics,
Instituto Tecnológico de Buenos Aires (ITBA),
Argentina
pfierens@itba.edu.ar

Abstract

Previous works ([5][6][8])) have presented a frequentist interpretation of sets of measures as probabilistic models which have denominated *chaotic models*. Those models, however, dealt only with sets of probability measures on finite algebras, that is, probability measures which can be related to variables with a finite number of possible values. In this paper, an extension of chaotic models is proposed in order to deal with the more general case of real-valued variables.

Keywords. Imprecise probabilities, foundations of probability, chaotic probability models, frequentist interpretation.

1 Introduction

In a series of papers ([5][6]), we presented the first steps towards a frequentist interpretation of sets of measures as probability models, which we have called *chaotic probability models* in order to distinguish them from other plausible interpretations. This work was coherently presented in [4] and extended by Rêgo and Fine in [8]. In our previous work, we presented chaotic models as simply sets of probability measures whose domain is a finite set of events. In this sense, we may associate chaotic probability models to discrete "random"[1] variables with finite range (e.g., the outcome of the flipping of a coin or the tossing of a die). In this paper, we present a simple approach to the extension of chaotic probability models to real-valued variables (e.g., tomorrow's minimum temperature).

The paper is organized as follows. Section 2 presents some concepts of the previous work which are needed for this paper. In Section 3, we provide the basic motivation behind the model which is described in Section 4. In the latter Section, we also show that such a model is plausible. Section 5 is devoted to present extensions for this framework of the concepts of visibility and temporal homogeneity defined in previous works. Finally, in Section 6 we discuss the results presented in this paper and suggest future lines of work.

[1] We use quotation marks to denote the difference between these *chaotic* variables and the usual understanding of random variables.

2 Variables with finite range

We need to recall the interpretation of chaotic probability models for variables with finite range ([6][4][8]).

2.1 An Instrumental Description of the Model

The instrumental (that is, without commitment to reality) description of chaotic probability models presented in earlier works is basically preserved in this paper. Let $\mathbf{X}$ be a finite sample space. We denote by $\mathbf{X}^*$ the set of all finite sequences of elements taken in $\mathbf{X}$. A particular sequence of n samples from $\mathbf{X}$ is denoted by $x^n = \{x_1, x_2, \cdots, x_n\}$. $\mathbf{P}$ denotes the set of all measures on the power set of $\mathbf{X}$. A chaotic probability model $\mathbf{M}$ is a subset of $\mathbf{P}$ and models the "marginals" of some process generating sequences in $\mathbf{X}^*$.

Given any $n \in \mathbb{N}$, consider the generation of a sequence x^n of length n by the following algorithm[2]:

`FOR k = 1 TO k = n`

1. `Choose` $\nu = F(x^{k-1}) \in \mathbf{M}$.
2. `Generate` x_k `according to` ν.

where $F : \mathbf{X}^* \to \mathbf{M}$ is a function corresponding to the decisions causally made by the algorithm at each step. Let $\nu_k = F(x^{k-1})$. For any $k \leq n$, F determines the probability distribution of the *potential* kth outcome X_k of the sequence,

$$(\forall \mathbf{A} \subseteq \mathbf{X}) \;\; P\left(X_k \in \mathbf{A} | X^{k-1} = x^{k-1}\right) = \nu_k(X_k \in \mathbf{A}).$$

[2] We denote the empty string by x^0.

The probability of a particular realization x^n of a sequence of random variables X^n is given by

$$P\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{k=1}^{n} \nu_k(X_k = x_k).$$

We denote by $\mathbf{M}^*$ the family of all such process measures P, one for each possible function F. From the analysis of data, we do not expect in general to be able to pinpoint a single $P \in \mathbf{M}^*$ or even a small subset of $\mathbf{M}^*$, what we call a **fine-grained picture** of the source. On the contrary, we expect our knowable **operational quantities to be (large) subsets of $\mathbf{M}^*$** which provide an appropriate **coarse-grained** description of the source.

2.2 Data analysis and estimation

We begin the study of a sequence $x^n \in \mathbf{X}^*$ by analyzing it into several subsequences. These subsequences are selected by rules that satisfy the following

Definition 1. *A computable function $\psi : \mathbf{X}^* \to \{0, 1\}$ is a* **causal subsequence selection rule** *(also known as a Church place selection rule) if for any $x^n \in \mathbf{X}^*$, x_k is the j-th term in the generated subsequence $x^{\psi,n}$, of length $\lambda_{\psi,n}$, whenever*

$$\psi(x^{k-1}) = 1,\ \sum_{i=1}^{k} \psi(x^{i-1}) = j,\ \lambda_{\psi,n} = \sum_{k=1}^{n} \psi(x^{k-1}).$$

Let $\Psi = \{\psi_\alpha\}$ be a set of causal subsequence selection rules. For each $\psi \in \Psi$, we study the behavior of the relative frequency of marginal events along the chosen subsequence. That is, given x^n and a selection rule $\psi \in \Psi$ we determine the **frequentist empirical (relative frequency) measure** $\bar{\mu}_{\psi,n}$ along the subsequence $x^{\psi,n}$ through

$$(\forall \mathbf{A} \subset \mathbf{X})\ \bar{\mu}_{\psi,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} I_{\mathbf{A}}(x_k)\psi(x^{k-1}),$$

where $I_{\mathbf{A}}(\cdot)$ is the indicator function of the event $\mathbf{A}$.

A family of subsequence selection rules Ψ is key to our understanding of a chaotic probability model as given by a set of measures $\mathbf{M}$. It has been proved that:

- so long as we restrict to a family of causal selection rules of moderate size, we can with high probability avoid extracting arbitrary patterns through some of the selected subsequences and instead exhibit the patterns that have inductive validity (see [6] and [4]).
- chaotic probability models $\mathbf{M}$ can be estimated from the empirical relative frequency measures if the appropriate family of subsequence selection rules is chosen (see [6] and [4]). Rêgo and Fine [8] showed how to choose a universal family of place selection rules to make the model visible.
- the visibility (possibility of estimation) of a chaotic probability model $\mathbf{M}$ depends strongly on the choice of the subsequence selection family, i.e., there are cases where $\mathbf{M}$ can be estimated by a family Ψ_0 while another family Ψ_1 only "sees" one measure in $ch(\mathbf{M})$ (see [6] and [4]).

3 Motivation

In what follows, we shall assume that $(\mathbf{X}, \mathcal{X})$ is a measurable space and that $\mathbf{P}$ is the set of all probability measures on $\mathcal{X}$. A chaotic probability model is represented mainly by a set $\mathbf{M} \subset \mathbf{P}$.

The instrumental description of chaotic probability models summarized in Section 2.1 can be extended to variables with infinite (even uncountable) range without changes. Therefore, the problem of the extension of chaotic probability models to more general spaces lies on the task of making such models "visible" (in an intuitive sense) when they are represented as in Section 2.1.

One possibility is to allow, as in the finite case, for the estimation of the measures in $\mathbf{M}$ by means of the empirical relative frequencies:

$$\bar{\mu}_{\psi,n}(\mathbf{A}) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} I_{\mathbf{A}}(x_k)\psi(x^{k-1}).$$

The difficulty then becomes the choice of the sets $\mathbf{A} \subset \mathbf{X}$ that should be used. For, in general, it is impossible to compute $\bar{\mu}_{\psi,n}(\mathbf{A})$ for all $\mathbf{A}$ in a σ-field. Furthermore, it may make no sense at all to try to assess such a fine-grained model.

We may also charge the statistician with the responsibility of choosing a collection of subsets $\mathbf{A} \subset \mathbf{X}$ adequate for the problem at hand. If we follow this path, we may as well allow for greater generality by letting the practitioner to choose a suitable finite family $\mathbf{F}$ of real-valued bounded measurable *test functions* $f : \mathbf{X} \to \mathbb{R}$ and proceeding to the estimate by means of the empirical relative frequencies

$$\bar{\mu}_{\psi,n}(f) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} f(x_k)\psi(x^{k-1}).$$

We may conceive these functions as those which are of actual interest for the problem at hand. We may also understand the family of functions $\mathbf{F}$ together with the family of subsequence selection rules Ψ as a representation of the discernment power of the observer

or, at least, of the coarse-grainedness appropriate for the model. In particular note that if we restrict ourselves to bounded test functions, the family $\mathbf{F}$ may contain indicator functions of the type $I_{\mathbf{A}}$, $\mathbf{A} \subset \mathbf{X}$.

From a technical viewpoint, the trick is simple: we substitute the *finite* algebra of events related to a discrete variable by a *finite* set of test functions applied to a real variable. With this idea in mind, all previous results (e.g., those in [6]) can be easily extended, as it is shown in Section 5.

3.1 Test functions as gambles

From a behavioral stand, we may also consider $\mathbf{F}$ as a collection of gambles in the sense of Peter Walley:

> *"A gamble is a bounded real-valued function on* Ω [the sample space] *which is interpreted as a reward."* ([11], Chapter 2)

Therefore, from this point of view, the estimates $\bar{\mu}_{\psi,n}(f)$ can be understood as estimates of the set of linear previsions dominating a coherent lower prevision based on the gambles $\mathbf{F}$ (see [11], especially Chapters 2 and 3).

There is, however, a key difference with some of the work of Peter Walley in [11] in the sense that we are not interested in pursuing an equivalent to the natural extension in our framework. We assume that a finite set of test functions (or gambles) $\mathbf{F}$ is enough for the purposes of a given problem. In other words, we do not feel compelled to make probabilistic assessment over anything else than $\mathbf{F}$.

If we take the discussion in the last paragraph one step further, we may allow to specify a chaotic probability model by, not a precise set of probability measures $\mathbf{M}$, but by a collection of "previsions" $\mathbf{P}_r$ defined as

$$(\forall f \in \mathbf{F})\, \mathbf{P}_r(f) \subset \mathbb{R},\ \mathbf{P}_r = \{\mathbf{P}_r(f) :\ f \in \mathbf{F}\},$$

and where, intuitively,

$$\mathbf{P}_r(f) = \{\mu(f) :\ \mu \in \mathbf{M}\}$$

for some unspecified set $\mathbf{M}$.

There is another difference with the work of Peter Walley: we consider only countably additive linear previsions $\mathbf{P}$. Theorem 3.6.1 of Walley [11] shows that there is a one to one correspondence between the coherent lower previsions on the domain of the bounded gambles and the non-empty weak*-compact convex subsets $\mathbf{M}$ of $\mathbf{P}_l$, the set of all linear previsions. Although $\mathbf{P}_l$ is the weak*-closure of $\mathbf{P}$, the latter is strictly contained in the former if the sample space $\mathbf{X}$ is infinite (see [11], Appendix D and Section 3.6.8).

Since we do not desire to follow closely any behavioral interpretation and in order to avoid confusion, we continue to refer to the elements of $\mathbf{F}$ as test functions rather than gambles.

3.1.1 Gambling with nature

Shafer and Vovk ([10]) present a game-theoretic interpretation of probabilistic reasoning where there are three players:

1. `Nature`: It determines what may happen in the world, that is, the outcomes of a given game.
2. `Modeler`: `Modeler` suggests a theory about how `Nature` behaves. Based on this theory, `Modeler` proposes a game to Skeptic.
3. `Skeptic`: `Skeptic` tries to show that `Modeler`'s theory is wrong by betting on the proposed game. If `Modeler`'s theory is "correct", `Skeptic` should not be able to make a high profit (with "not-too-low probability").

It is easy to see a relationship between the work presented here and the theory of Shafer and Vovk. `Modeler` has an instrumental understanding of how `Nature` works (see Section 2.1): There is a certain set of measures $\mathbf{M} \subset \mathbf{P}$ such that

`FOR k = 1 TO k = n`

1. `Nature chooses` $\nu \in \mathbf{M}$ `based on` x^{k-1}.
2. `Nature generates` x_k `according to` ν.

Note that `Nature` is *causal, but not necessarily markovian.* Also note that `Modeler` does not need to know the set $\mathbf{M}$, but `Modeler` does know $\{\mu(f)\}_{\mu \in \mathbf{M}}$ for all $f \in \mathbf{F}$. Based on this understanding of `Nature`'s behavior, `Modeler` proposes any of the following gambles to `Skeptic`:

`Skeptic's initial capital is` $K_0 = 0$.

`Skeptic chooses` $f \in \mathbf{F}$.

`FOR k = 1 TO k = n`

1. `Skeptic chooses` $\theta_k \in \{0, 1\}$ `based on` x^{k-1}.
2. `Nature generates` x_k.
3. `Skeptic's capital is now`

$$K_k = K_{k-1} + \theta_k \left(f(x_k) - \sup_{\mu \in \mathbf{M}} \mu(f) \right).$$

Note that `Skeptic`'s sequence of bets $\{\theta_k\}$ can be associated with a causal subsequence selection rule ψ (see Def. 1), if `Skeptic` can only use computable strategies. Note also that, if `Skeptic` needs to keep track of his capital, he must be able to compute his earnings. One necessary (but not sufficient) requirement for this is that the test functions be computable in a reasonable sense (see Section 3.2). We shall come back to this game in Section 4.

3.2 Computability of test functions and place selection rules

One reasonable restriction on test functions is to ask them to be computable, i.e., that their value can be calculated. Also, place selection rules must be able to output values in $\{0, 1\}$ when having tuples of real values as inputs. The problem then becomes to find a reasonable definition of computable real-valued functions with real-valued input variables. To our knowledge, there are mainly two broad approaches to such a definition in the area of *computational analysis* (see, e.g., [2] [1]). On one hand, there are the traditional approach and its many variations and extensions which are based, loosely speaking, on the following ideas:

- Given a finite alphabet, say $\mathbf{A}$, and an adequate program M, a *description* of an element y of some space $\mathbf{Y}$ is a finite string $\underline{a} = a_1 a_2 \cdots a_k$, $a_i \in \mathbf{A}$, such that $y = M(\underline{a})$.
- It may be the case that not all the elements of some space have a description. However, an element y is considered to be computable if there is an *approximating sequence of descriptions* $\{\underline{a}_i\}$, i.e., the strings $\underline{a}_i$ are such that the outputs $y_i = M(\underline{a}_i)$ get increasingly closer to y.
- A function $f : \mathbf{Y} \to \mathbf{Z}$ is computable if for each computable number $y \in \mathbf{Y}$, with approximating sequence $\{\underline{a}_i\}$, there is a program P such that $\{P(\underline{a}_i)\}$ is an approximating sequence of descriptions for $f(y) \in \mathbf{Z}$.

This approach models well scientific computations. Moreover, most "calculator" functions (polynomials, $\log(x)$, $\sqrt{x}$, etc.) are computable under this approach.

On the other hand, there is the Blum-Shub-Smale (BSS) approach which is based on computing machines that can deal with elements of any field R (e.g, $R = \mathbb{R}$) and that are allowed to perform the field operations ($+$, $-$, $\times$ and $\%$) on R and can branch on comparisons ($<$, $>$, $\leq$) between elements of R if it is ordered. The fact that the BSS approach is very useful in numerical modelling should not come as a surprise.

Since we are focused on calculations that can be made on any personal computer, we shall take the first approach to computability of real-valued functions.

3.2.1 Computable functions of real variable

The material in this section is taken from Weihrauch [12] (see also [7]). There are other approaches which are equivalent and quite powerful, for example, that based on domain theory (see, e.g., Edalat [3]), but less intuitive.

Let $\mathbf{A}$ be any finite alphabet. The finite strings of elements of $\mathbf{A}$ will be denoted by $\mathbf{A}^*$ and the infinite sequences of elements of $\mathbf{A}$ will be denoted by $\mathbf{A}^\infty$.

Definition 2. (Computability by Type 2 machines)

1. *A* **Type 2 machine** *M is defined by two components:*
   (a) *a Turing machine with k one-way input tapes ($k \geq 0$), a single one-way output tape and finitely many work tapes,*
   (b) *a type specification $(\mathbf{Y}_1, \cdots, \mathbf{Y}_k, \mathbf{Y}_0)$ with $\{\mathbf{Y}_0, \cdots, \mathbf{Y}_k\} \subseteq \{\mathbf{A}^*, \mathbf{A}^\infty\}$.*
2. *The function $\rho_M : (\mathbf{Y}_1 \times \cdots \times \mathbf{Y}_k) \to \mathbf{Y}_0$ computed by the Type 2 machine M (the semantics of M) is defined as follows:*
   (a) *Case $\mathbf{Y}_0 = \mathbf{A}^*$ (finite output): $\rho_M(y_1, \cdots, y_k) = w$ iff M with input $(y_1, \cdots, y_k)$ halts with result w on the output tape.*
   (b) *Case $\mathbf{Y}_0 = \mathbf{A}^\infty$ (infinite output): $\rho_M(y_1, \cdots, y_k) = p$ iff M with input $(y_1, \cdots, y_k)$ computes forever writing the sequence p on the output tape.*
3. *We say that a function $\rho : (\mathbf{Y}_1 \times \cdots \times \mathbf{Y}_k) \to \mathbf{Y}_0$ is* **computable** *iff $\rho = \rho_M$ for some Type 2 machine M. A sequence y is a* **computable element of $\mathbf{Y}_0$** *iff the 0-place function $\rho : \{()\} \to \mathbf{Y}_0$ with $\rho() = y$ is computable.*

Type 2 machines can be considered as a certain kind of oracle Turing machines and computability with respect to them is entirely classical. In order to extend the concept of computability to functions, e.g., over the reals, we need the concept of a naming system. Indeed, objects like real numbers can be represented (named) by finite or infinite sequences of finite alphabets. For example, we can represent a real number

in $[0,1]$ by its (probably infinite) representation by a binary sequence. These ideas are formalized in the following

Definition 3. (Naming System. Reducibility)

1. *A* **notation** *of a set* $\mathbf{X}$ *is a surjective function* $\rho : \mathbf{A}^* \to \mathbf{X}$ *(naming by finite strings).*
2. *A* **representation** *of a set* $\mathbf{X}$ *is a surjective function* $\rho : \mathbf{A}^\infty \to \mathbf{X}$ *(naming by infinite sequences).*
3. *A* **naming system** *of a set* $\mathbf{X}$ *is a notation or a representation of* $\mathbf{X}$*.*
4. *For functions* $\gamma : \mathbf{Y} \to \mathbf{X}$ *and* $\gamma' : \mathbf{Y}' \to \mathbf{X}'$ *with* $\mathbf{Y}, \mathbf{Y}' \subseteq \{\mathbf{A}^*, \mathbf{A}^\infty\}$*, we call* γ **reducible** *to* γ'*,* $\gamma \preceq \gamma'$*, iff there exists a computable function* $\rho : \mathbf{Y} \to \mathbf{Y}'$ *such that* $(\forall y \in dom(\gamma))\ \gamma(y) = \gamma'(\rho(y))$*. We say that* γ *and* γ' *are* **equivalent***,* $\gamma \equiv \gamma'$*, iff* $\gamma \preceq \gamma'$ *and* $\gamma' \preceq \gamma$*.*

In order to clarify ideas, we present some common naming systems:

- *Binary representation of* $\mathbb{N}$: $\rho_{bin} : \{0,1\}^* \to \mathbb{N}$, $\rho_{bin}(a_0 a_1 \cdots a_k) = \sum_{i=0}^{k} a_i 2^i$.
- *Rational numbers*: $\rho_{\mathbb{Q}} : \{+,-\} \times \{0,1\}^* \times \{0,1\}^* \to \mathbb{Q}$, $\rho_{\mathbb{Q}}(s, b_n, b_d) = s\frac{\rho_{bin}(b_n)}{\rho_{bin}(b_d)}$.
- *Interval Representation of* $\mathbb{R}$: Let $\mathbf{S}_{\mathbb{Q}}$ be the set of all infinite sequences of triples (s,n,d) taken from $\{+,-\}\times\{0,1\}^*\times\{0,1\}^*$. Then define $\rho_{int} : \mathbf{S}_{\mathbb{Q}} \times \mathbf{S}_{\mathbb{Q}} \to \mathbb{R}$ by

$$\rho_{int}(a_0 a_1 a_2 \cdots, b_0 b_1 b_2 \cdots) = x \Leftrightarrow$$
$$\Leftrightarrow \lim_{n\to\infty} \rho_{\mathbb{Q}}(a_n) = \lim_{n\to\infty} \rho_{\mathbb{Q}}(b_n) = x$$

and

$$\rho_{\mathbb{Q}}(a_0) < \rho_{\mathbb{Q}}(a_1) < \cdots < x < $$
$$< \cdots < \rho_{\mathbb{Q}}(b_1) < \rho_{\mathbb{Q}}(b_0).$$

The latter naming system leads to the following

Definition 4. (Computable Real Numbers) $x \in \mathbb{R}$ *is computable if it is* ρ_{int}*-computable.*

The definition of naming systems leads to the extension of the definition of computable functions that we need for this paper:

Definition 5. (Relative Computability)

1. *For* $i = 0, 1, \cdots, k$*, let* $\gamma_i : \mathbf{Y}_i \to \mathbf{Z}_i$ *be naming systems. A function* $\delta : \mathbf{Z}_1 \times \cdots \times \mathbf{Z}_k \to \mathbf{Z}_0$ *is* $(\gamma_1, \cdots, \gamma_k, \gamma_0)$*-computable iff there is a Type 2-computable function (in the sense of Def. 2)* $\rho : \mathbf{Y}_1 \times \cdots \times \mathbf{Y}_k \to \mathbf{Y}_0$ *such that*

$$\delta(\gamma_1(y_1), \gamma_2(y_2), \cdots, \gamma_k(y_k)) = $$
$$= \gamma_0(\rho(y_1, y_2, \cdots, y_k)),$$

whenever $\delta(\gamma_1(y_1), \gamma_2(y_2), \cdots, \gamma_k(y_k))$ *exists.*

2. *We say that a real-valued function of a real variable is computable if it is* (ρ_{int}, ρ_{int})*-computable.*

One important consequence of the definition of computability is that all computable functions are continuous (see [12]).

We shall require all *admissible test functions* to be computable. Some examples of real-valued computable functions are: $+$, $-$, $\times$, $1/x$, exp, log, sin, cos, $\sqrt{}$, min, max, etc.

We shall also require *place selection rules* to be computable functions of tuples of real variables, in the sense of Definition 5, which take only values in $\{0,1\}$. In other words, we shall require of a place selection rule ψ to be $(\rho_{int}, \cdots, \rho_{int})$-computable, where ρ_{int} appears $k+1$ times, for each $k \geq 0$.

We shall also need the following

Definition 6. (Computable Probability Mass Function) *Let* $(\mathbf{X}, \mathcal{X})$ *be a measurable space, with* $\mathcal{X}$ *containing the singleton sets. Then, we say that a probability mass function on* $(\mathbf{X}, \mathcal{X})$ *is computable if each of the probability values is computable in the sense of Def. 4.*

4 Chaotic probability model

Let $\Psi = \{\psi_\alpha\}$ be a set of causal subsequence selection rules and $\mathbf{F} = \{f_\beta\}$ a collection of bounded real-valued test functions. For each $\psi \in \Psi$, we study the behavior of the relative frequency of (only) f_β along the chosen subsequence. That is, given x^n and a selection rule $\psi \in \Psi$ we determine the **frequentist empirical (relative frequency) measure** $\bar{\mu}_{\psi,n}$ along the subsequence $x^{\psi,n}$ through

$$(\forall f \in \mathbf{F})\ \bar{\mu}_{\psi,n}(f) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} f(x_k)\psi(x^{k-1}).$$

In a similar manner, for all such rules ψ, we define the **time average conditional measure** $\bar{\nu}_{\psi,n}$ $(\forall f \in \mathbf{F})$

$$\bar{\nu}_{\psi,n}(f) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} \mathrm{E}\left[f(X_k)|X^{x-1} = x^{k-1}\right] \psi(x^{k-1}).$$

Rewritten in terms of our instrumental understanding of the measure selection function F,

$$\bar{\nu}_{\psi,n}(f) = \frac{1}{\lambda_{\psi,n}} \sum_{k=1}^{n} \nu_k(f)\psi(x^{k-1}),$$

where $\nu_k = F(x^{k-1})$. Note that, since we assume F to be unknown, the time average conditional measure $\bar{\nu}_{\psi,n}$ is also unknown. Since we want to expose some of the structure of the chaotic probability model $\mathbf{M}$ by means of the rules in Ψ, we are interested in how good $\bar{\mu}_{\psi,n}$ is as an estimator of $\bar{\nu}_{\psi,n}$.

Define the metric $d_{\mathbf{F}}$ on $\mathbf{P}$ by

$$d_{\mathbf{F}}(\nu,\mu) = \max_{f\in\mathbf{F}} |\mu(f) - \nu(f)|, \ (\forall \mu,\nu \in \mathbf{P}).$$

We call **F-causally faithful** a set of rules Ψ such that any $\psi \in \Psi$ yields a small value of $d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \bar{\mu}_{\psi,n})$ with high probability. The existence of such a set of rules is stated by

Theorem 1. *Let $m \leq n$ and fix Ψ and $\mathbf{F}$ of finite cardinality, denoted by $\|\Psi\|$ and $\|\mathbf{F}\|$ respectively. Then* $(\forall P \in \mathbf{M}^*)$

$$P\left(\max_{\psi\in\Psi}\{d_{\mathbf{F}}(\bar{\mu}_{\psi,n},\bar{\nu}_{\psi,n}) : \ \lambda_{\psi,n} \geq m\} \geq \varepsilon\right) \leq \\ \leq 2\|\mathbf{F}\|\|\Psi\|\mathrm{e}^{-\frac{\varepsilon^2 m^2}{8\beta^2 n}},$$

where

$$\beta = \max_{f\in\mathbf{F}} \sup_{x\in\mathbf{X}} |f(x)|.$$

The proof of the theorem is completely analog to that of Theorem 1 in [6] (see also the appendix to Chapter 4 in [4]). The consequence of this theorem is that, as long as we restrict to small-sized families of causal selection rules we can with high probability avoid extracting arbitrary patterns through some of the selected subsequences.

Recall the game in Section 3.1.1 proposed by `Modeler` to `Skeptic`. If `Modeler` is right, the probability that `Skeptic` becomes rich is very low. This is exactly what the following result shows.

Lemma 1. *Consider the game played by* `Modeler` *and* `Skeptic`. *Then* $(\forall \varepsilon > 0)$ $(\forall m \leq n)$

$$P(K_n \geq m\varepsilon) \leq 2\mathrm{e}^{-\frac{\varepsilon^2 m^2}{8\beta^2 n}},$$

where

$$\beta = \max_{f\in\mathbf{F}} \sup_{x\in\mathbf{X}} |f(x)|.$$

The proof of this lemma follows along the same lines as the proof of Theorem 1.

4.1 Collection of expected values as a model

In Section 3.1, we suggested the idea of taking the collection of expected values as the actual model, defining implicitly the set of probability measures $\mathbf{M}$. The following Lemma shows that $\mathbf{M}$ defined in this way has a particularly simple structure.

Lemma 2. *Let $(\mathbf{X},\mathcal{X})$ be a measurable space and $\mathbf{P}$ the set of all probability measures on it. Assume that $\mathcal{X}$ contains the singletons. Let $\mathbf{F} = \{f_1,\cdots,f_N\}$ be a finite collection of real-valued bounded functions. Let the set*

$$\mathbf{P}_r \subset \left[\inf_{x\in\mathbf{X}} f_1(x), \sup_{x\in\mathbf{X}} f_1(x)\right] \times \cdots \\ \cdots \times \left[\inf_{x\in\mathbf{X}} f_N(x), \sup_{x\in\mathbf{X}} f_N(x)\right] \subset \mathbb{R}^{\|\mathbf{F}\|}$$

be given. Define a set of measures by

$$\mathbf{M}_{\mathbf{P}_r} = \{\mu \in \mathbf{P} : \ (\mu(f_1),\cdots,\mu(f_N)) \in \mathbf{P}_r\}.$$

Then, the measures in $\mathbf{M}_{\mathbf{P}_r}$ are ε-indistinguishable from measures with finite support in the sense that for each $\varepsilon > 0$ there are points $x_1,\cdots,x_{L(\varepsilon)}$ in $\mathbf{X}$ such that $(\forall \mu \in \mathbf{M}_{\mathbf{P}_r})(\exists \nu \in \mathbf{M}_{\mathbf{P}_r})$ such that

$$d_{\mathbf{F}}(\mu,\nu) \leq \varepsilon, \ \textit{and} \ \sum_{i=1}^{L(\varepsilon)} \nu(\{x_i\}) = 1.$$

In other words, Lemma 2 tells us that, as long as we restrict ourselves to a *finite* set of test functions, there is no substantial difference (what concerns the test functions) between the behavior of a given *chaotic* real variable and that of a particular *chaotic* discrete variable with finite range. This fact not only opens up the door to the reuse of previous results which were originally conceived for discrete variables, but it also shows the way in which chaotic real variables can be simulated. Indeed, the simulation of chaotic real variable is not different from that of an adequate chaotic discrete variable according to Lemma 2, and the simulation of the latter type of variables was already explained in [6] (see also the proof of Theorem 3 in the Appendix).

Hence, using a collection of expected values of a finite set of test functions $\mathbf{P}_r$ as a model, gives us only a coarse-grained, blurred view of how a real variable behaves. This model may be as precise as we are capable of (or willing to) build it. However, the model is so fuzzy, our view so blurred, that we cannot distinguish with certainty whether we observe a real-valued variable or just a simple discrete variable which takes only a few values. By the way, this should not be

very surprising for, that who observes a finite number of outcomes of a uniformly distributed random variable in $[0,1]$, how can he be certain that he was dealing with a real random variable or just a complex discrete random variable.

5 Visibility and Temporal Homogeneity

In this section, we present extensions to those concepts of visibility and temporal homogeneity which were defined in [6]. The proofs of the results that follow are also analog to the proofs of the results in [6] thanks to the finiteness of the set of bounded test functions $\mathbf{F}$ and Lemma 2.

The possibility of exposing all of $\mathbf{M}$ by means of the rules in Ψ is expressed in the following

Definition 7. (**Visibility**)

(a) $\mathbf{M}$ *is* **made F-visible** (Ψ,θ,δ,m,n) *by* $P\in\mathbf{M}^*$ *if*

$$P\left(\bigcap_{\mu\in\mathbf{M}}\bigcup_{\psi\in\Psi}\mathbf{C}_\psi\right)\geq 1-\delta,$$

where

$$\mathbf{C}_\psi=\{X^n:\lambda_{\psi,n}(X^n)\geq m, d_{\mathbf{F}}(\bar{\mu}_{\psi,n},\mu)\leq\theta\}.$$

(b) *A subset* $\mathbf{M}'$ *of* $\mathbf{M}^*$ **renders M uniformly F-visible** (Ψ,θ,δ,m,n) *if* $\mathbf{M}$ *is made* $\mathbf{F}$*-visible* (Ψ,θ,δ,m,n) *by each* $P\in\mathbf{M}'$*. The maximal such subset is denoted* $\mathbf{M}_V(\Psi)$ *and* $\mathbf{M}_V(\Psi)$ *may be empty.*

The non-triviality of Definition 7**(a)**, and, hence, of Definition 7**(b)**, is asserted in

Theorem 2. *Let* $\mathbf{M}$ *be a set of probability measures and* $\mathbf{F}$ *a finite family of real-valued bounded functions on* $\mathbf{X}$*. Given* $0<2\varepsilon<\theta$*, for large* n*, there exists a process measure* P *and a family* Ψ *of size* N_ε *such that* $\mathbf{M}$ *is made* $\mathbf{F}$*-visible* (Ψ,θ,δ,m,n) *by* P *with*

$$\delta=2(\|\mathbf{F}\|+1)N_\varepsilon \mathrm{e}^{-\frac{(\theta-2\varepsilon)^2m^2}{8\beta^2n}},$$

where

$$\beta=\max_{f\in\mathbf{F}}\sup_{x\in\mathbf{X}}|f(x)|,$$

$$N_\varepsilon\leq\left\lceil\frac{2\beta}{\varepsilon}\right\rceil^{\|\mathbf{F}\|}.$$

The fact that not every set of rules Ψ can expose all of $\mathbf{M}$ is expressed by the concept of temporal homogeneity defined as follows.

Definition 8. (**Temporal Homogeneity**)

(a) $P\in\mathbf{M}^*$ *is* **F-temporally homogeneous** (Ψ,θ,δ,m,n) *if*

$$P\left(\Delta_\Psi\leq\theta\right)\geq 1-\delta,$$

where

$$\Delta_\Psi=$$
$$=\max_{\psi_1,\psi_2\in\Psi}\left\{d_{\mathbf{F}}(\bar{\mu}_{\psi_1,n},\bar{\mu}_{\psi_2,n}):\lambda_{\psi_1,n},\lambda_{\psi_2,n}\geq m\right\}.$$

(b) *A subset* $\mathbf{M}'$ *of the set of all possible process measures* $\mathbf{M}^*$ *is* **uniformly F-temporally homogeneous** (Ψ,θ,δ,m,n) *if each of the elements of* $\mathbf{M}'$ *is temporally homogeneous* (Ψ,θ,δ,m,n)*. The maximal such subset is denoted* $\mathbf{M}_T(\Psi)$*.*

As it was the case with chaotic variables with finite range, a model $\mathbf{M}$ may be visible under a certain family of subsequence selection rules and temporal homogeneous under another, as the following result shows.

Theorem 3. *Let* $\mathbf{F}$*,* $\mathbf{P}_r$ *and* $\mathbf{M}_{\mathbf{P}_r}$ *be as in Lemma 2. Let* $\varepsilon>\frac{\beta}{m}$*, where*

$$\beta=\max_{f\in\mathbf{F}}\sup_{x\in\mathbf{X}}|f(x)|.$$

Let Ψ_0 *be a set of (causal deterministic) place selection rules. Then, there are a process measure* P *and a family* Ψ_1 *such that, for large enough* n*,* P *will both render* $\mathbf{M}_{\mathbf{P}_r}$ $\mathbf{F}$*-visible* $(\Psi_1,3\varepsilon,\delta,m,n)$ *and ensure* $\mathbf{F}$*-temporal homogeneity* $(\Psi_0,6\varepsilon,\delta,m,n)$ *with*

$$\delta=2\|\mathbf{F}\|\max\{\|\Psi_0\|,\|\Psi_1\|\}\,\mathrm{e}^{-\frac{\varepsilon^2m^2}{8\beta^2n}}.$$

Although the proof of this theorem is very similar to that of Theorem 4 in [6], we include a sketch in the appendix because it shows clearly how the concepts of computability of real-valued functions, the finiteness of the set of test functions $\mathbf{F}$ and Lemma 2 are applied in order to reuse previous results under the current framework.

6 Conclusions and future work

The extension of chaotic probability models proposed in this paper does not carry in itself any technical novelties with respect to previous works, except perhaps for Lemma 2. Although this may seem disappointing, we believe it is the best feature of the current presentation, i.e., that it allows a smooth and simple extension of chaotic models to real-valued variables.

Besides extending previous works on chaotic models, a different viewpoint on them is offered in Section

4.1, where we suggest to get rid of the set of measures and work directly with the assessment of "expected" values of the test functions. Although this idea is not novel in itself, it is in the framework of chaotic models.

The relation between gambles and test functions sketched in Section 3.1 may allow to those pursuing behavioral interpretations of probability to deal with chaotic models without any sense of guilt.

Lemma 2 shows that the finiteness of our discernment is implicitly embedded in the finite number of test functions.

There are several matters which were left out of this paper. For example, it is easy to see that the same ideas can be applied to tuples of variables. Then, the question becomes what the relation is between chaotic models on tuples of variables and the "marginal" chaotic models and how independence can be characterized. The problem of marginalizing chaotic models on tuples is difficult because the corresponding test functions must also be marginalized.

Acknowledgements

This paper would not have existed had it not been for the encouragement from T.L. Fine and Leandro Rêgo. The author is most grateful to T.L. Fine for he found a mistake in an earlier version of Lemma 2. The author would also like to thank the unknown reviewers for their helpful and useful comments.

This work was partially supported by anonymous contributors through the project "Prevention and early detection of forest fires by means of sensor networks" which is being developed at the Instituto Tecnológico de Buenos Aires (ITBA).

Appendix A: Proof of Lemma 2

In order to prove Lemma 2, we need the following preliminary result.

Proposition 1. *Let $\varepsilon > 0$ be given and let $\beta > 0$ be defined as*

$$\beta = \max_{f \in \mathbf{F}} \sup_{x \in \mathbf{X}} \|f(x)\|.$$

Then, there is a finite set $\mathbf{M}_\varepsilon = \{\nu_1, \nu_2, \cdots, \nu_{N_\varepsilon}\} \subset \mathbf{M}_{\mathbf{P}_r}$ such that

$$N_\varepsilon \leq \left\lceil \frac{2\beta}{\varepsilon} \right\rceil^{\|\mathbf{F}\|},$$

and

$$\sup_{\mu \in \mathbf{M}} \min_{1 \leq i \leq N_\varepsilon} d_{\mathbf{F}}(\mu, \nu_i) \leq \varepsilon.$$

Proof. Let $N = \|\mathbf{F}\|$ and consider the set in $\mathbb{R}^N$

$$\mathbf{A} = \{(\mu(f_1), \cdots, \mu(f_N)) : \mu \in \mathbf{M}_{\mathbf{P}_r}\}.$$

Then, it is clear that $\mathbf{A}$ is included in the closed hypercube $[-\beta, +\beta]^N$. Moreover, this hypercube can be covered by a set of $\left\lceil \frac{2\beta}{\varepsilon} \right\rceil^N$ smaller hypercubes of side ε. □

We also need the following result from Rudin [9] (see Lemma after Theorem 3.25 in Rudin [9], page 73).

Lemma 3. *If y lies in the convex hull of a set $\mathbf{E} \subset \mathbb{R}^N$, then y lies in the convex hull of a subset of $\mathbf{E}$ which contains at most $N + 1$ points.*

Now, we are ready for the proof of Lemma 2:

Proof. Let $\underline{f}_x = (f_1(x), \cdots, f_N(x))$ for all $x \in \mathbf{X}$, and $\underline{\mu} = (\mu(f_1), \cdots, \mu(f_N))$ for all $\mu \in \mathbf{M}_{\mathbf{P}_r}$. Consider the following set:

$$\mathbf{E} = \left\{ \underline{f}_x : x \in \mathbf{X} \right\} \subset \mathbb{R}^N.$$

It is clear that $\mathbf{P}_r \subseteq ch(\mathbf{E})$, where $ch(\mathbf{E})$ is the convex hull of $\mathbf{E}$. By Lemma 3, for each $\nu_k \in \mathbf{M}_\varepsilon$, where $\mathbf{M}_\varepsilon$ is as in Prop. 1, there are at most $N + 1$ points $\underline{f}_{x_1^{\{k\}}}, \cdots, \underline{f}_{x_L^{\{k\}}}$ in $\mathbf{E}$ such that

$$\underline{\nu}_k = \sum_{i=1}^{L} p_i^{\{k\}} \underline{f}_{x_i^{\{k\}}},$$

where

$$p_i^{\{k\}} \geq 0, \sum_{i=1}^{L} p_i^{\{k\}} = 1.$$

Define the probability measures $\{\nu'_k\}$ on $(\mathbf{X}, \mathcal{X})$ by

$$\nu'_k(\{x\}) = \begin{cases} p_i^{\{k\}} & \text{if } x = x_i^{\{k\}}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{A} = \cup_k \left\{ x_1^{\{k\}}, \cdots, x_L^{\{k\}} \right\}$. Clearly, the cardinality of $\mathbf{A}$ is

$$L(\varepsilon) = \|\mathbf{A}\| \leq (N+1) \times \left\lceil \frac{2\beta}{\varepsilon} \right\rceil^{\|\mathbf{F}\|}.$$

Then the measures ν'_k are as required by the Lemma. □

Appendix B: Proof of Theorem 3

By Lemma 2 and Prop. 1, there is a set of measures with finite support $\mathbf{M}_\varepsilon = \left\{ \nu'_1, \nu'_2, \cdots, \nu'_{N_\varepsilon} \right\}$ such that

$$\sup_{\mu \in \mathbf{M}} \min_{1 \leq i \leq N_\varepsilon} d_{\mathbf{F}}(\mu, \nu'_i) \leq \frac{\varepsilon}{2},$$

where the supremum is over all measures in $\mathbf{M}_{\mathbf{P}_r}$ and

$$N_\varepsilon \leq \left\lceil \frac{4\beta}{\varepsilon} \right\rceil^{\|\mathbf{F}\|},$$

where β is defined as in Prop. 1. Since the mass on each of the supporting atoms can be approximated as closely as desired by rational numbers and the fact that rational numbers are computable in the sense of Section 3.2, then it is easy to see that for each $\nu'_k \in \mathbf{M}_\varepsilon$ there is a computable probability mass function (in the sense of Def. 6) μ_k such that $d_{\mathbf{F}}(\nu'_k, \mu_k) \leq \frac{\varepsilon}{2}$ and, hence,

$$\sup_{\mu \in \mathbf{M}} \min_{1 \leq i \leq N_\varepsilon} d_{\mathbf{F}}(\mu, \mu_i) \leq \varepsilon,$$

where the supremum is over all measures in $\mathbf{M}_{\mathbf{P}_r}$. Note that the measures $\{\mu_i\}$ are not necessarily in $\mathbf{M}_{\mathbf{P}_r}$[3]. We may assume that ε is a computable number w.l.o.g.. Consider the following construction of P:

- Choose any measure $\nu_0 \in \mathbf{M}_{\mathbf{P}_r}$. Let μ_0 be any computable probability mass function such that $d_{\mathbf{F}}(\nu_0, \mu_0) < \varepsilon$.
- Define N_ε counters $i(1), \cdots, i(N_\varepsilon)$ and set them to 0.
- For each $k > 0$ define,
  - $(\forall \psi \in \Psi_0)$, $\bar{\nu}_{\psi,k-1} = \frac{1}{\lambda_{\psi,k-1}} \sum_{l=1}^{k-1} \psi(x^{l-1})\nu_l$ if $\lambda_{\psi,k-1} > 0$, and $\bar{\nu}_{\psi,k-1} = \mu_0$ otherwise.
  - $\alpha_k = 0$ if $(\forall \psi \in \Psi_0)$ $\psi(x^{k-1}) = 0$, and $\alpha_k = \max_{\psi \in \Psi_0} \left\{ d_{\mathbf{F}}(\bar{\nu}_{\psi,k-1}, \mu_0) : \psi(x^{k-1}) = 1 \right\}$ otherwise.
  - $j_k = \operatorname{argmin} i(j)$.

  Note that α_k depends only on x^{k-1}.
- If $\alpha_k > \varepsilon$, let $\nu_k = \mu_0$. Otherwise, let ν_k be the computable probability measure μ_{j_k} and increment $i(j_k)$ by 1.
- Generate x_k according to ν_k.

Note that all the steps in the construction are computable, with the exception of the generation of the outcomes.

Proposition 2. *For $\varepsilon > \beta/m$ and large enough n, P is $\mathbf{F}$-temporally homogeneous $(\Psi, 6\varepsilon, \delta, m, n)$, with*

$$\delta = 2\|\mathbf{F}\|\|\Psi_0\| \mathrm{e}^{-\frac{\varepsilon^2 m^2}{8\beta^2 n}}.$$

[3] Although we think that this restriction can easily be removed, it does not pose any problem to the proof of the theorem.

Proof. Suppose that there is some $\psi \in \Psi_0$ such that $d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \mu_0) > \varepsilon$ and $\lambda_{\psi,n} \geq m$. Let

$$\delta(\mu_0) = \max_{\nu \in \mathbf{M}_\varepsilon} d_{\mathbf{F}}(\mu_0, \nu).$$

Since, by construction, as soon as $d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \mu_0) > \varepsilon$ outcomes start to be generated according to μ_0, then we must have

$$d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \mu_0) < \frac{(\lambda_{\psi,n} - 1)\varepsilon + \delta(\mu_0)}{\lambda_{\psi,n}} \leq \varepsilon + \frac{\beta}{m} \leq 2\varepsilon.$$

Since by Theorem 1 we have

$$P\left(\max_{\psi \in \Psi_0} \{ d_{\mathbf{F}}(\bar{\mu}_{\psi,n}, \bar{\nu}_{\psi,n}) : \lambda_{\psi,n} \geq m \} \geq \varepsilon \right) \leq \leq 2\|\mathbf{F}\|\|\Psi_0\| \mathrm{e}^{-\frac{\varepsilon^2 m^2}{8\beta^2 n}},$$

the proposition is proved. □

Proposition 3. *Let*

$$n \geq \frac{\delta(\mu_0) N_\varepsilon m}{\varepsilon} \|\Psi_0\| + N_\varepsilon m - 1.$$

Then

$$\sum_{j=1}^{N_\varepsilon} i(j) \geq N_\varepsilon m, \tag{1}$$

and, hence,

$$\min_{1 \leq j \leq N_\varepsilon} i(j) \geq m. \tag{2}$$

Proof. We call k an **exceeding time** when

$$\alpha_k = \\ = \max \left\{ d_{\mathbf{F}}(\bar{\nu}_{\psi,k-1}, \mu_0) : \psi \in \Psi_0,\ \psi(x^{k-1}) = 1 \right\} > \\ > \varepsilon.$$

By the construction of P, it is clear that Eqn. 2 follows immediately from Eqn. 1. Suppose that Eqn. 1 does not hold. This means that there have been at least $(n - N_\varepsilon m + 1)$ exceeding times. Since by hypothesis

$$\frac{\delta(\mu_0) N_\varepsilon m}{\varepsilon} \|\Psi_0\| \leq n - N_\varepsilon m + 1,$$

there must be a $\psi \in \Psi_0$ such that, for its corresponding subsequence, $d_{\mathbf{F}}(\bar{\nu}_{\psi,k}, \mu_0)$ has been greater than ε at least $\frac{\delta(\mu_0) N_\varepsilon m}{\varepsilon}$ times. Note that, for each exceeding time

$$\varepsilon < d_{\mathbf{F}}(\bar{\nu}_{\psi,k}, \mu_0) \leq \\ \leq \frac{(\lambda_{\psi,k} - \lambda_{\psi,k,\mathbf{M}}) \times 0 + \lambda_{\psi,k,\mathbf{M}} \delta(\mu_0)}{\lambda_{\psi,k}} = \\ = \frac{\lambda_{\psi,k,\mathbf{M}}}{\lambda_{\psi,k}} \delta(\mu_0),$$

where $\lambda_{\psi,k,\mathbf{M}}$ is the number of times, along the subsequence selected by ψ, such that $\alpha_k \leq \varepsilon$. From the last inequality, it follows that

$$\lambda_{\psi,k,\mathbf{M}} > \frac{\varepsilon}{\delta(\mu_0)}\lambda_{\psi,k}.$$

Therefore, for ψ's last exceeding time we have

$$\lambda_{\psi,k,\mathbf{M}} > \frac{\varepsilon}{\delta(\mu_0)}\lambda_{\psi,k} \geq \frac{\varepsilon}{\delta(\mu_0)}\frac{\delta(\mu_0)N_\varepsilon m}{\varepsilon} = N_\varepsilon m.$$

However, this contradicts our initial assumption that there were less than $N_\varepsilon m$ exceeding times along the entire sequence. Thus, we must conclude that Eqn. 1 holds. □

Let $\Psi_1 = \{\psi_1, \psi_2, \cdots, \psi_{N_\varepsilon}\}$ be a set of N_ε place selection rules such that, for $1 \leq l \leq N_\varepsilon$, ψ_l selects the subsequence where the measure μ_l has been used. The fact that such a family Ψ_1 of computable place selection rules exists follows from the construction of P. Note that Proposition 3 implies that the subsequences selected by the rules in Ψ_1 have length larger than or equal to m.

Proposition 4. $\mathbf{M}$ *is* $\mathbf{F}$*-visible* $(\Psi_1, 3\varepsilon, \delta, m, n)$*, where*

$$\delta = 2\|\mathbf{F}\|\|\Psi_1\|\mathrm{e}^{-\frac{\varepsilon^2 m^2}{8\beta^2 n}}.$$

Proof. It is clear that, by construction, for all $\mu \in \mathbf{M}$ there is a measure $\mu_i \in \mathbf{M}_\varepsilon$ and a rule $\psi \in \Psi_1$ such that

$$d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \mu) \leq d_{\mathbf{F}}(\bar{\nu}_{\psi,n}, \mu_i) + d_{\mathbf{F}}(\mu_i, \mu) \leq \varepsilon + \varepsilon \leq 2\varepsilon.$$

Then the proposition follows from Theorem 1 and the fact that Proposition 3 implies that $(\forall \psi \in \Psi_1)\ \lambda_{\psi,n} \geq m$. □

The **proof of Theorem 3** follows from Propositions 2-4.

References

[1] Mark Braverman. On the complexity of real functions. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*. Technical Committee on Mathematical Foundations of Computing, IEEE Computer Society, 2005.

[2] Mark Braverman and Stephen Cook. Computing over the reals: Foundations for scientific computing. *Notices of the AMS*, 53(3):318–329, March 2006.

[3] Abbas Edalat. Domains for computation in mathematics,physics and exact real arithmetic. *The Bulletin of Symbolic Logic, Association for Symbolic Logic*, 3(4):401–452, December 1997.

[4] Pablo I. Fierens. *Towards a Chaotic Probability Model for Frequentist Probability*. PhD thesis, Cornell University, 2003.

[5] Pablo I. Fierens and Terrence L. Fine. Towards a frequentist interpretation of sets of measures. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*. Shaker Publishing, 2001.

[6] Pablo I. Fierens and Terrence L. Fine. Towards a chaotic probability model for frequentist probability. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*. Carleton Scientific, 2003.

[7] Peter Hertling. Computable analysis via representations. Notes for a presentation made at a Satellite Seminar of the Second International Conference on Computability and Complexity in Analysis, held in Kyoto, Japan, August 2005.

[8] Leandro Chaves Rêgo and Terrence L. Fine. Estimation of chaotic probabilities. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, 2005.

[9] Walter Rudin. *Functional Analysis*. McGraw-Hill, 1973.

[10] Glenn Shafer and Vladimir Vovk. *Probability and Finance. It's Only a Game!* Wiley Series in Probability and Statistics. John Wiley & Sons, 2001.

[11] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[12] Klaus Weihrauch. A simple introduction to computable analysis (2nd edition). Informatik Berichte 171, FernUniversitt Hagen, Hagen, July 1995.

Data-Based Decisions under Imprecise Probability and Least Favorable Models

Robert Hable
Department of Statistics, LMU Munich
Robert.Hable@stat.uni-muenchen.de

Abstract

Data-based decision theory under imprecise probability has to deal with optimisation problems where direct solutions are often computationally intractable. Using the Γ-minimax optimality criterion, the computational effort may significantly be reduced in the presence of a least favorable model. In 1984, A. Buja derived a necessary and sufficient condition for the existence of a least favorable model in a special case. The present article proves that essentially the same result is valid in case of general coherent upper previsions. This is done mainly by topological arguments in combination with some of L. Le Cam's decision theoretic concepts. It is shown how least favorable models could be used to deal with situations where the distribution of the data as well as the prior is assumed to be imprecise.

Keywords. Decision theory, robust statistics, imprecise probability, coherent upper previsions, Le Cam, equivalence of models, least favorable models.

1 Introduction

1.1 Motivation

Decision theory provides a formal framework for determining optimal actions under uncertainty on the states of nature. It has a wide range of potential areas of application which includes also statistical problems, for example. However, a serious problem in practical applications of decision theory is that the uncertainty often is too complex to be adequately described by a classical, i.e. precise, probability distribution. Ambiguity, i.e. the extent of deviation from ideal stochasticity, plays an important role in decision making that cannot be neglected. To take ambiguity into account properly, generalisations of the concept of probability have been developed, among others, by [24] (imprecise probability) and [25] (interval probability). Here, the probability of an event is no longer a number $p \in [0,1]$ but an interval $[\underline{p}, \overline{p}] \subset [0,1]$. These concepts are applied in a number of recent articles in decision theory, e.g. [3], [21] and [22].

Generalisations of probabilities as in [24] and [25] have a strong relationship with some concepts of robust statistics (cf. e.g. [20, §3.1.7]) - a fact which is frequently disregarded. Actually, [6] develops a concept of robust statistics (named "upper expectations") which lies between the concepts of [24] and [25]. [6] considers decision making which is explicitly data-based. This can be understood as a matter of its own as has been pointed out by [3]. In the spirit of the celebrated article [14], [6] characterises the existence of precise models which are simultaneously least favorable for a class of loss functions (or for a class of prior distributions):

[14] deals with hypothesis testing where a (rather special) upper prevision is tested against another one. This is equivalent to testing between certain sets of (precise) probabilities $\mathcal{M}_0$ and $\mathcal{M}_1$. [14] shows that there is a pair $(p_0, p_1) \in \mathcal{M}_0 \times \mathcal{M}_1$ which is least favorable: Testing between p_0 and p_1 is as hard as testing between $\mathcal{M}_0$ and $\mathcal{M}_1$ and, as a consequence, there is an optimal test between p_0 and p_1 which is also an optimal test between $\mathcal{M}_0$ and $\mathcal{M}_1$. That way, testing between $\mathcal{M}_0$ and $\mathcal{M}_1$ can be done by testing only between p_0 and p_1. This reduces the computational effort substantially. In fact, it is one of the most important drawbacks of data-based decision theory (including hypothesis testing) that the computational effort of direct solutions is frequently not manageable. Therefore, least favorablility has attracted enormous attention after the publication of [14]. For a review of [14] and the work following [14], confer [2]. In quite general data-based decision theory, where there are n states of nature (instead of two), an analogous question of that one solved by [14] is: Does there exist a model $(q_1, q_2, \ldots, q_n) \in \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_n$ which is simultaneously least favorable for a class of loss functions? This is not always the case but [6] proves a

necessary and sufficient condition for the existence of such simultaneously least favorable models.

Unfortunately, [6] contains an error which reduces its applicability significantly. The validity of the conclusions in [6] can only be guaranteed by adding a restrictive assumption on the involved upper previsions (cf. [10]).

The present article follows the lines of [6] - but within the concept of [24] which dispenses with σ-additivity. It is shown that the same result as in [6] is possible without any additional assumption on the involved (coherent) upper previsions. This demonstrates that, in [6], insistence on σ-additivity of probabilities happens to be an unnecessary burden (cf. also Remark 2.2).

By ignoring σ-additivity, we are in line with Le Cam's decision theoretic framework (cf. [15] and [16]), which provides us with some effective methods. Within this framework some terms (e.g. randomisations) are slightly generalised (cf. [16, §1] and [9, §4]).

Sections 2 and 3 develop the decision theoretic framework. Section 4 contains a generalisation of the LeCam-Blackwell-Sherman-Stein-Theorem which plays an important role in Section 5. In Section 5, the analogue to [6, Theorem 8.2] is proven which characterises the existence of least favorable models. This is the main theorem of the present article. Section 6 explains how least favorability could be used to deal with situations where the distribution of the data as well as the prior is assumed to be imprecise.

Since the content of this article might be obscured by the mathematical details, the following subsection presents a rather detailed outline.

1.2 Outline

In order to explain the decision theoretic setup we are concerned with, the classical decision theoretic setup is recalled at first:

There is a set Θ where each element $\theta \in \Theta$ represents a possible state of nature. We know that one state of nature will occur but we do not know which one it will be. Furthermore, there is a set $\mathbb{D}$ where each element $t \in \mathbb{D}$ is a decision we can choose. Depending on what state of nature θ occurs, every decision t leads to a loss $W_\theta(t)$. The goal is to choose a "good" decision so that the loss is as small as possible.

Sometimes, we might know a precise expectation π for the states of nature $\theta \in \Theta$. Then, we can choose the decision that minimises the expected loss

$$\int_\Theta W_\theta(t)\,\pi(dt)$$

Quite often, we can choose our decision on the base of an observation $y \in \mathcal{Y}$. For example, the observation y may be the outcome of an experiment. The distribution of the observation y might be a precise expectation q_θ which depends on the state of nature θ. That is $(q_\theta)_{\theta\in\Theta}$ is a model which describes the distribution of the observation y.

Such "data-based decision making" can be formalised by choosing a decision function $\delta : \mathcal{Y} \to \mathbb{D}, \quad x \mapsto \delta(y)$ which minimises

$$\int_\Theta \int_\mathcal{Y} W_\theta(\delta(y))\,q_\theta(dy)\,\pi(dt)$$

Decision theory commonly also deals with randomised decisions. Randomised decision procedures (randomisations) are defined in Subsection 2.1. Confer [4] for an introduction to these basic concepts of decision theory.

In the following, we are concerned with a more general decision theoretic setup because we also want to deal with imprecise probabilities:

Since the prior knowledge about the states of nature will frequently not be precise, we allow for a whole set $\mathcal{P}$ of possible precise expectations π. Also the knowledge about the distribution of the observation may only be imprecise so that there are sets $\mathcal{M}_\theta$ of possible precise expectations q_θ. While minimising the expected loss in case of precise expectations is widely accepted, there are several reasonable optimality criteria in case of imprecise expectations, confer [21] for a discussion of the most important ones. In the present article the so-called Γ-minimax criterion is used which represents a worst case consideration.[1] That is we choose a decision function δ (or rather a randomisation later on) which minimises the twofold upper expectation

$$\sup_{\pi\in\mathcal{P}} \int_\Theta \sup_{q_\theta\in\mathcal{M}_\theta} \int_\mathcal{Y} W_\theta(\delta(y))\,q_\theta(dy)\,\pi(dt)$$

Unfortunately, a direct solution of this problem is quite often computationally intractable. In Section 6, it is shown how the situation might become manageable: In the presence of a model $(\tilde{q}_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ which is simultaneously least favorable for $\mathcal{P}$ (or for a corresponding set of loss functions) the above minimisation problem may be solved by minimising

$$\sup_{\pi\in\mathcal{P}} \int_\Theta \int_\mathcal{Y} W_\theta(\delta(y))\,\tilde{q}_\theta(dy)\,\pi(dt)$$

However, such a least favorable model $(\tilde{q}_\theta)_{\theta\in\Theta}$ need not exist. In Section 5, a necessary and sufficient con-

[1] For the use of the Γ-minimax criterion in Bayesian analysis, cf. [23] and the literature cited therein.

dition for existence is proven (Theorem 5.4). This condition is formulated in terms of standard models.

Standard models are our main tool. They are introduced in Subsection 2.3. An important fact is that every model (consisting of precise expectations) is equivalent to a standard model. In Subsection 2.2, we define an equivalence relation on the set of all (precise) models $(q_\theta)_{\theta\in\Theta}$ according to which two (precise) models $(p_\theta)_{\theta\in\Theta}$ and $(q_\theta)_{\theta\in\Theta}$ are equivalent if the following is true: Observations of model $(p_\theta)_{\theta\in\Theta}$ can artificially be generated (by a randomisation) from observations of model $(q_\theta)_{\theta\in\Theta}$ and vice versa. Here and also as decision procedures, randomisations become important. By topological reasons, the term "randomisation" has to be slightly generalised in the present article (cf. Subsection 2.1). All these tools from decision theory (namely randomisations, equivalence of models, standard models) are presented in Section 2.

In Section 3, minimal Bayes risks are defined for precise models and for imprecise models as well. It is shown that minimal Bayes risks can be expressed in terms of standard models, which in fact is the reason why we use standard models.

Section 4 contains a generalisation of the LeCam-Blackwell-Sherman-Stein-Theorem, which is important in the proof of the main theorem, Theorem 5.4. Theorem 5.4 characterises the existence of simultaneously least favorable models.

1.3 Some Notation

This subsection collocates some notation which is used throughout the article.

Let $(\mathcal{Y},\mathcal{B})$ be a measurable space and $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ be the Banach space of all bounded Borel-measurable real functions $g : Y \to \mathbb{R}$ where $\|g\| = \sup_{y\in\mathcal{Y}} g(y)$. For a subset B of $\mathcal{Y}$, I_B denotes the characteristic function of B on $\mathcal{Y}$.

The set of all finitely additive signed measures $\mathrm{ba}(\mathcal{Y},\mathcal{B})$ can be identified with the dual space of $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$, i.e. the Banach space of all linear continuous real functionals on $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ where $\|\mu\| = \sup\big\{|\mu[g]| \,\big|\, g \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B}), \|g\| \le 1\big\}$ for all $\mu \in \mathrm{ba}(\mathcal{Y},\mathcal{B})$ (cf. [7, Theorem IV.5.1]). $\mu \in \mathrm{ba}(\mathcal{Y},\mathcal{B})$ is called *positive* if $\mu[g] \ge 0$ for every $g \ge 0$. This is denoted by $\mu \ge 0$.

Let Θ be an index set. Throughout the article, $(\overline{Q}_\theta)_{\theta\in\Theta}$ is a family of coherent upper previsions $\overline{Q}_\theta : \mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \to \mathbb{R}$ (cf. [24]). The corresponding sets of majorised linear previsions are denoted by $\mathcal{M}_\theta := \big\{q_\theta \in \mathrm{ba}(\mathcal{Y},\mathcal{B}) \,\big|\, q_\theta[g] \le \overline{Q}_\theta[g]\ \forall g \in \mathcal{L}_\infty(\mathcal{Y},\mathcal{B})\big\}$. Analogously to [25], $\mathcal{M}_\theta$ is called *structure*. $(\overline{Q}_\theta)_{\theta\in\Theta}$ is called *imprecise model* on $(\mathcal{Y},\mathcal{B})$. A family $(q_\theta)_{\theta\in\Theta}$ of linear previsions $q_\theta : \mathcal{L}_\infty(\mathcal{Y},\mathcal{B}) \to \mathbb{R}$ is called *precise model* on $(\mathcal{Y},\mathcal{B})$. These terms are adapted from the notion "statistical model". [6] and [15] use the term "experiment" instead of "model".

Let $(\mathcal{X},\mathcal{A})$ be another measurable space. $\mathcal{F} = (q_\theta)_{\theta\in\Theta}$ will always denote a precise model on $(\mathcal{Y},\mathcal{B})$, $\mathcal{E} = (p_\theta)_{\theta\in\Theta}$ will always denote a precise model on $(\mathcal{X},\mathcal{A})$. If $q_\theta \in \mathcal{M}_\theta$ for every $\theta \in \Theta$, we may also write $(q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ or $\mathcal{F} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$. Expressions of the form $(a_\theta)_{\theta\in\Theta}$ will often be abbreviated by $(a_\theta)_\theta$.

For some fixed $n \in \mathbb{N}$, put $\mathcal{U} := \big\{u \in \mathbb{R}^n \,\big|\, u = (u_{\theta_1},\dots,u_{\theta_n})',\ u_\theta \ge 0\ \forall\theta\in\Theta,\ u_{\theta_1}+\dots+u_{\theta_n} = 1\big\}$ and $\mathcal{C} := \mathbb{B}^{\otimes n} \cap \mathcal{U}$ where $\mathbb{B}^{\otimes n}$ is the Borel-σ-algebra of $\mathbb{R}^n$. For $\theta \in \Theta$, put $\iota_\theta : \mathcal{U} \to [0,1],\ u \mapsto u_\theta$ where u_θ is the θ-component of u.

2 Some Tools from Decision Theory

2.1 Randomisations

2.1.1 Introduction

Let $\mathcal{X}$ be a set of possible outcomes of an experiment and $\mathbb{D}$ be a set of possible decisions t. Then, a decision function may be a map $\delta : \mathcal{X} \to \mathbb{D}$ where $\delta(x) = t$ means: If x appears, choose action t. In addition, decision theory commonly deals with randomised decisions $\delta : \mathcal{X} \to \mathrm{ba}(\mathbb{D},\mathcal{D}),\ x \mapsto \tau_x$. Here, it is supposed that each τ_x is a linear prevision and that $\tau.[h] : x \mapsto \tau_x[h]$ lies in $\mathcal{L}_\infty(\mathcal{X},\mathcal{A})$ for every $h \in \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$. Then, $\delta(x) = \tau_x$ means: After observing x, start an auxiliary random experiment according to the distribution τ_x and choose that action d which is the outcome of the auxiliary random experiment.

For our purposes, we will need a slight generalisation. Note that every randomised decision function $x \mapsto \tau_x$ defines a map

$$\sigma : \mathrm{ba}(\mathcal{X},\mathcal{A}) \to \mathrm{ba}(\mathbb{D},\mathcal{D}),\ \ \mu \mapsto \sigma(\mu)$$

via

$$\sigma(\mu) :\ \ h \ \mapsto\ \ \sigma(\mu)[h] \ =\ \mu\big[\tau.[h]\big] \qquad (1)$$

It is easy to see that σ is

- linear
- positive: $\sigma(\mu) \ge 0$ for every $\mu \ge 0$
- normalised: $\|\sigma(\mu)\| = \|\mu\|$ for every $\mu \ge 0$

2.1.2 Definition

Let $(\mathcal{X},\mathcal{A})$ and $(\mathcal{Y},\mathcal{B})$ be measurable spaces. According to [15], a *randomisation* from $\mathcal{X}$ to $\mathcal{Y}$ is a linear,

positive and normalised map

$$T : \mathrm{ba}(\mathcal{X}, \mathcal{A}) \to \mathrm{ba}(\mathcal{Y}, \mathcal{B})$$

where "positive" means $T(\mu) \geq 0$ for every $\mu \geq 0$ and "normalised" means $\|T(\mu)\| = \|\mu\|$ for every $\mu \geq 0$. Let $\mathcal{T}(\mathcal{X}, \mathcal{Y})$ denote the set of all randomisations from $\mathcal{X}$ to $\mathcal{Y}$.

We also mark a class of randomisations of a very simple form: To this end, let κ be a map

$$\kappa : \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B}) \to \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}), \qquad g \mapsto \kappa(g)$$

so that there is some finite set $S \subset \mathcal{Y}$ and

$$\kappa(g) = \sum_{y \in S} g(y)\alpha_y \qquad \forall\, g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$$

where $\alpha_y \in \mathcal{L}_\infty(\mathcal{X}, \mathcal{A})$ $\forall\, y \in S$, $\alpha_y \geq 0$ $\forall\, y \in S$ and $\sum_{y \in S} \alpha_y \equiv 1$. Then,

$$\kappa^* : \mathrm{ba}(\mathcal{X}, \mathcal{A}) \to \mathrm{ba}(\mathcal{Y}, \mathcal{B}), \qquad \mu \mapsto \kappa^*(\mu)$$

where $\kappa^*(\mu)[g] = \mu[\kappa(g)]$ $\forall\, g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$, is called *restricted randomisation*. It is easy to see that every restricted randomisation is generated by a (very simple) randomised decision function via (1). Every restricted randomisation is in fact a randomisation , i.e. $\mathcal{T}_r(\mathcal{X}, \mathcal{Y}) \subset \mathcal{T}(\mathcal{X}, \mathcal{Y})$ where $\mathcal{T}_r(\mathcal{X}, \mathcal{Y})$ denotes the set of all restricted randomisations.

2.1.3 Topological Issues

Models which consist of imprecise probabilities are so extensive that sequential limit arguments are no longer adequate. So, we have to resort to topological arguments.

Let $\overline{Q} : \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B}) \to \mathbb{R}$ be a coherent upper prevision with structure $\mathcal{M} := \big\{q \in \mathrm{ba}(\mathcal{Y}, \mathcal{B}) \,\big|\, q[g] \leq \overline{Q}[g]\ \forall\, g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})\big\}$.

In addition to the norm-topology, $\mathrm{ba}(\mathcal{Y}, \mathcal{B})$ can also be provided with the $\sigma(\mathrm{ba}, \mathcal{L}_\infty)$-topology. This is the smallest topology so that

$$\mathrm{ba}(\mathcal{Y}, \mathcal{B}) \to \mathbb{R}, \qquad \mu \mapsto \mu[g]$$

is continuous for every $g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$.

Theorem 2.1 *$\mathcal{M}$ is $\sigma(ba, \mathcal{L}_\infty)$-compact.* (Cf. [24, §3.6.1].)

Remark 2.2 *According to Theorem 2.1, compactness of $\mathcal{M}$ comes for free. If we restricted $\mathcal{M}$ to σ-additive measures, we would have to impose additional assumptions to ensure compactness in reasonable topologies. So, insistence on σ-additivity appears to be a burden.*

$\mathcal{T}(\mathcal{X}, \mathcal{Y})$ can be provided with the topology of pointwise convergence on $\mathrm{ba}(\mathcal{X}, \mathcal{A}) \times \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$. This is the smallest topology so that

$$\mathcal{T}(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}, \qquad T \mapsto T(\mu)[g]$$

is continuous for every $\mu \in \mathrm{ba}(\mathcal{X}, \mathcal{A})$ and every $g \in \mathcal{L}_\infty(\mathcal{Y}, \mathcal{B})$. The following theorem is the reason why we use the generalisation of randomised procedures:

Theorem 2.3 *$\mathcal{T}(\mathcal{X}, \mathcal{Y})$ is a compact Hausdorff space.* (Cf. [16, Theorem 1.4.2].)

The following theorem indicates that the term "randomisation" has only been slightly generalised:

Theorem 2.4 *$\mathcal{T}_r(\mathcal{X}, \mathcal{Y})$ is dense in $\mathcal{T}(\mathcal{X}, \mathcal{Y})$.*

Proof: This is a consequence of [15, Theorem 1]. □

Especially, Theorem 2.4 implies that the randomised procedures defined via (1) are dense in $\mathcal{T}(\mathcal{X}, \mathcal{Y})$.

2.2 Sufficiency and Equivalence of Models

Let $\mathcal{E} = (p_\theta)_{\theta \in \Theta}$ be a precise model on $(\mathcal{X}, \mathcal{A})$ and $\mathcal{F} = (q_\theta)_{\theta \in \Theta}$ a precise model on $(\mathcal{Y}, \mathcal{B})$.

Analogously to [6], $(p_\theta)_{\theta \in \Theta}$ is called *sufficient* for $(q_\theta)_{\theta \in \Theta}$ if there is a randomisation $T \in \mathcal{T}(\mathcal{X}, \mathcal{Y})$ so that $T(p_\theta) = q_\theta$ $\forall\, \theta \in \Theta$.

This definition of "sufficiency" essentially goes back to [5]. It does not strictly coincide with the more common definition in terms of conditional expectations but, under suitable assumptions of regularity, the definitions do coincide (cf. [13]). At least, if the randomisation T is generated by a randomised function $x \mapsto \tau_x$ via (1), the above definition has a very descriptive interpretation:

Let x be an observation distributed according to p_θ. After observing x, start an auxiliary random experiment according to τ_x. Then, the outcome y of the auxiliary random experiment is distributed according to q_θ. That is, if we have observations of the model $(p_\theta)_\theta$, we can artificially generate observations of the model $(q_\theta)_\theta$ "by coin tossing".

$(p_\theta)_{\theta \in \Theta}$ and $(q_\theta)_{\theta \in \Theta}$ are called *equivalent* if they are mutually sufficient, i.e. there are some $T_1 \in \mathcal{T}(\mathcal{X}, \mathcal{Y})$, $T_2 \in \mathcal{T}(\mathcal{Y}, \mathcal{X})$ so that $T_1(p_\theta) = q_\theta$ $\forall\, \theta \in \Theta$ and $T_2(q_\theta) = p_\theta$ $\forall\, \theta \in \Theta$.

The descriptive interpretation of sufficiency already indicates that equivalent models essentially coincide from a decision theoretic point of view. Our definition of equivalence is in accordance with Le Cam's definition (cf. [9, §5.2]).

Let $(\overline{Q}_\theta)_{\theta \in \Theta}$ be an imprecise model with corresponding structures $\mathcal{M}_\theta$, $\theta \in \Theta$.

Analogously to [6], $(p_\theta)_{\theta\in\Theta}$ is called *worst-case-sufficient* for $(\overline{Q}_\theta)_{\theta\in\Theta}$ if $(p_\theta)_{\theta\in\Theta}$ is sufficient for some $(q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$. So, $(p_\theta)_{\theta\in\Theta}$ is worst-case-sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$ if and only if there is some $T \in \mathcal{T}(\mathcal{X},\mathcal{Y})$ so that $\forall\,\theta\in\Theta$

$$T(p_\theta)[g] \;\le\; \overline{Q}_\theta[g]\,, \qquad \forall\, g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$$

2.3 Standard Models

Let the index set Θ be finite with cardinality n.

In Subsection 2.2, we have defined an equivalence relation on the precise models with a fixed index set Θ. Each equivalence class contains a uniquely defined representative (called standard model later on) which has some nice properties.[2] This is the content of the following theorem.

Theorem 2.5 *Every precise model $\mathcal{F} = (q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$ admits a uniquely defined (σ-additive) probability measure $s^{\mathcal{F}}$ on $(\mathcal{U},\mathcal{C})$ so that $ds^{\mathcal{F}}_\theta = n\iota_\theta\, ds^{\mathcal{F}}$ defines a precise model $(s^{\mathcal{F}}_\theta)_{\theta\in\Theta}$ on $(\mathcal{U},\mathcal{C})$ which is equivalent to $\mathcal{F}$.* (Cf. [9, Theorem 6.5].)

Analogously to [6], $s^{\mathcal{F}}$ is called *standard measure* and $(s^{\mathcal{F}}_\theta)_{\theta\in\Theta}$ is called *standard (precise) model* of $\mathcal{F}$.

Standard models share two important properties:

- They are defined on the very nice measurable space $(\mathcal{U},\mathcal{C})$ (cf. Subsection 1.3).
- They consist of linear previsions s_θ which are σ-additive probability measures.

For the imprecise model $(\overline{Q}_\theta)_{\theta\in\Theta}$ with corresponding structures $\mathcal{M}_\theta$, we can uniquely define

$$\overline{S}[h] = \sup\left\{s^{\mathcal{F}}[h] \;\middle|\; \mathcal{F}\in(\mathcal{M}_\theta)_{\theta\in\Theta}\right\} \quad \forall\, h\in\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$$

$$\overline{S}_\theta[h] = \sup\left\{s^{\mathcal{F}}_\theta[h] \;\middle|\; \mathcal{F}\in(\mathcal{M}_\theta)_{\theta\in\Theta}\right\} \quad \forall\, h\in\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$$

$\overline{S}$ is called *standard upper prevision*, $(\overline{S}_\theta)_{\theta\in\Theta}$ is called *standard imprecise model* of $(\overline{Q}_\theta)_{\theta\in\Theta}$. Note that $\overline{S}$ is a coherent upper prevision on $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ and $(\overline{S}_\theta)_{\theta\in\Theta}$ is an imprecise model on $(\mathcal{U},\mathcal{C})$.

3 Minimal Bayes Risks

Let the index set $\Theta = \{\theta_1,\dots,\theta_n\}$ be finite with cardinality n and let π be a prior distribution on $(\Theta,2^\Theta)$, i.e. π is a linear prevision on $\mathcal{L}_\infty(\Theta,2^\Theta)$. Put $\pi_\theta := \pi[I_{\{\theta\}}]$.

[2] As stated in Subsection 2.2, equivalent models essentially coincide from a decision theoretic point of view. Therefore, every decision problem coincides with a "standard decision problem" where a standard model is involved. We will deduce properties of the original decision problem from the corresponding "standard decision problem" later on.

A *decision space* is a measurable space $(\mathbb{D},\mathcal{D})$ where $\mathbb{D}$ is the set of possible decisions. A loss function is a family $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$.

The measurable space $(\mathcal{Y},\mathcal{B})$ may represent the results of an experiment. According to [15], a *decision procedure* is a randomisation

$$\sigma : \mathrm{ba}(\mathcal{Y},\mathcal{B}) \;\to\; \mathrm{ba}(\mathbb{D},\mathcal{D})$$

i.e. $\sigma\in\mathcal{T}(\mathcal{Y},\mathcal{D})$.

Now, Bayes risks can be defined for precise models (Subsection 3.1) and for imprecise models (Subsection 3.2). The main goal of the present section is to express minimal Bayes risks in terms of standard measures (Theorem 3.2) and standard upper previsions (Theorem 3.4).

3.1 Precise Models

Let $(q_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{Y},\mathcal{B})$. For a decision procedure $\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})$ and a loss function $(W_\theta)_{\theta\in\Theta}\subset\mathcal{L}_\infty(\mathbb{D},\mathcal{D})$, the *risk function* of $(q_\theta)_{\theta\in\Theta}$ is

$$\sigma(q)[W] :\; \theta \;\mapsto\; \sigma(q_\theta)[W_\theta]$$

The *Bayes risk* is

$$\begin{aligned} R\big((q_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big) &= \pi\big[\sigma(q)[W]\big] = \\ &= \sum_{\theta\in\Theta}\pi_\theta\sigma(q_\theta)[W_\theta] \end{aligned}$$

Note that this definition coincides with the usual one if σ is defined by a randomised decision function via (1).

The minimal Bayes risk is the same if we let σ vary among the randomisations or the restricted randomisations:

Proposition 3.1

$$\begin{aligned} \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big) = \\ = \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big) \end{aligned}$$

Proof: The definition of the topology of pointwise convergence implies continuity of the map

$$\sigma \;\mapsto\; \big(\sigma(q_{\theta_1})[W_{\theta_1}]\,,\dots,\,\sigma(q_{\theta_n})[W_{\theta_n}]\big)$$

and, therefore, continuity of

$$\sigma \;\mapsto\; R\big((q_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big)$$

Since $\mathcal{T}_r(\mathcal{Y},\mathbb{D})$ is dense in $\mathcal{T}(\mathcal{Y},\mathbb{D})$ (Theorem 2.4), the statement follows. □

For $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$, put

$$K\big((W_\theta)_\theta\big) : u \mapsto \inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\iota_\theta(u) \qquad (2)$$

on $\mathbb{R}^n$ where $\iota_\theta(u) = u_\theta$ is the θ-component of $u \in \mathbb{R}^\Theta \cong \mathbb{R}^n$. Note that $K\big((W_\theta)_\theta\big)$ is concave and, therefore, continuous on $\mathbb{R}^n$. Hence, the restriction of $K\big((W_\theta)_\theta\big)$ on $\mathcal{U}$ is Borel-measurable and $s^{(q_\theta)_\theta}\big[K\big((W_\theta)_\theta\big)\big]$ is defined well where $s^{(q_\theta)_\theta}$ is the standard measure of $(q_\theta)_{\theta\in\Theta}$.

Theorem 3.2

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = s^{(q_\theta)_\theta}\big[K\big((W_\theta)_\theta\big)\big]$$

Proof: According to Theorem 2.5, the standard model $\big(s^{\mathcal{F}}_\theta\big)_{\theta\in\Theta}$ is equivalent to $\mathcal{F} := (q_\theta)_{\theta\in\Theta}$. That is $\big(s^{\mathcal{F}}_\theta\big)_{\theta\in\Theta}$ and $\mathcal{F}$ are mutual sufficient. So, a twofold application of Lemma 8.2 yields

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big(\mathcal{F},\sigma,(W_\theta)_\theta\big) = \\ = \inf_{\rho\in\mathcal{T}(\mathcal{U},\mathbb{D})} R\big((s^{\mathcal{F}}_\theta)_\theta,\rho,(W_\theta)_\theta\big)$$

and an application of Lemma 8.1 closes the proof. □

3.2 Imprecise Models

Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y},\mathcal{B})$ with corresponding structures $\mathcal{M}_\theta$, $\theta\in\Theta$, and standard upper prevision $\overline{S}$. For a decision procedure $\sigma \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ and a loss function $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$, the *risk function* of $(\overline{Q}_\theta)_{\theta\in\Theta}$ is

$$\theta \mapsto \sup_{q_\theta\in\mathcal{M}_\theta} \sigma(q_\theta)[W_\theta]$$

and the Bayes risk is

$$R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \sum_{\theta\in\Theta} \pi_\theta \sup_{q_\theta\in\mathcal{M}_\theta} \sigma(q_\theta)[W_\theta]$$

Hence,

$$R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \sup_{(q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta} R\big((q_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

These definitions includes that we have chosen the Γ-minimax optimality criterion which represents a worst case consideration (cf. Subsection 1.2) - as done in [14] and [6].

Now, we can derive the analogues of Proposition 3.1 and Theorem 3.2 in case of imprecise models:

Proposition 3.3

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big) = \\ = \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_{\theta\in\Theta},\sigma,(W_\theta)_{\theta\in\Theta}\big)$$

Proof: This is a direct consequence of Lemma 8.3 (a), Proposition 3.1 and Lemma 8.3 (b). □

Theorem 3.4

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \overline{S}\Big[K\big((W_\theta)_\theta\big)\Big]$$

Proof: This is a direct consequence of Lemma 8.3, Theorem 3.2 and the definition of the standard upper prevision. □

4 The General LeCam-Blackwell-Sherman-Stein-Theorem

This section contains a generalisation of the LeCam-Blackwell-Sherman-Stein-Theorem. We need this theorem in the proof of our main theorem, Theorem 5.4.

Let Θ be a finite index set. Let π be a prior distribution on $(\Theta,2^\Theta)$ so that $\pi_\theta := \pi[I_{\{\theta\}}] > 0 \ \ \forall\,\theta\in\Theta$. Let $(p_\theta)_{\theta\in\Theta}$ be a precise model on $(\mathcal{X},\mathcal{A})$ and $(\overline{Q}_\theta)_{\theta\in\Theta}$ an imprecise model on $(\mathcal{Y},\mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of structures. Let $s^{(p_\theta)_\theta}$ be the standard measure of $(p_\theta)_{\theta\in\Theta}$ and $\overline{S}$ the standard upper prevision of $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{U},\mathcal{C})$.

Let Ψ be the set of all functions $k \in \mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ such that there is some decision space $(\mathbb{D},\mathcal{D})$ and a loss function $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$ where $k(u) = \inf_{\tau\in\mathbb{D}} \sum_{\theta\in\Theta} n\pi_\theta W_\theta(\tau)\iota_\theta(u) \ \ \forall\, u\in\mathcal{U}$.

Theorem 4.1 *The following statements are equivalent:*

(a) $(p_\theta)_{\theta\in\Theta}$ *is worst-case-sufficient for* $(\overline{Q}_\theta)_{\theta\in\Theta}$.

(b) $s^{(p_\theta)_\theta}[k] \le \overline{S}[k] \qquad \forall\, k\in\Psi$

(c) *For every finite decision space* $(\mathbb{D},\mathcal{D})$ *and every loss function* $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$,

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((p_\theta)_\theta,\rho,(W_\theta)_\theta\big) \le \\ \le \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

(d) *For every decision space* $(\mathbb{D},\mathcal{D})$ *and every loss function* $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$,

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((p_\theta)_\theta,\rho,(W_\theta)_\theta\big) \le \\ \le \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

The proof of Theorem 4.1 is located in [11].

5 Least Favorable Models

Let the index set Θ be finite with cardinality n. Let π be a prior distribution on $(\Theta, 2^\Theta)$ so that $\pi_\theta := \pi[I_{\{\theta\}}] > 0 \ \ \forall\, \theta \in \Theta$. Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y}, \mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of structures. Let $(\mathbb{D}, \mathcal{D})$ be a fixed decision space and let $\mathcal{W}$ be a set of loss functions $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$.

Definition 5.1 $(q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *is called* least favorable (precise) model of $(\mathcal{M}_\theta)_{\theta\in\Theta}$ for $\mathcal{W}$ *if*

$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta, \sigma, (W_\theta)_\theta\big) = \\ = \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, (W_\theta)_\theta\big)$$

for every $(W_\theta)_{\theta\in\Theta} \in \mathcal{W}$. [3]

We are not primarily interested in a set of loss functions but in a set of prior distributions. However, a set of prior distributions can always be transformed into a set of loss functions (cf. Section 6).

For $\mathcal{F} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$, put

$$\Phi_\mathcal{F} := \big\{h \in \mathcal{L}_\infty(\mathcal{U}, \mathcal{C}) \ \big|\ s^\mathcal{F}[h] = \overline{S}[h]\big\}$$

where $s^\mathcal{F}$ is the standard measure of $\mathcal{F}$ and $\overline{S}$ is the standard upper prevision of $(\overline{Q}_\theta)_{\theta\in\Theta}$ on $(\mathcal{U}, \mathcal{C})$.

The following lemma is an easy consequence of the definitions. A written proof may be found in [11].

Lemma 5.2 $\Phi_\mathcal{F}$ *is a norm-closed convex cone in* $\mathcal{L}_\infty(\mathcal{U}, \mathcal{C})$.

For every $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D}, \mathcal{D})$, define $K\big((W_\theta)_\theta\big)$ as in (2).

$$\Psi_\mathcal{W} := \big\{K\big((W_\theta)_\theta\big) \ \big|\ (W_\theta)_\theta \in \mathcal{W}\big\} \ \subset\ \mathcal{L}_\infty(\mathcal{U}, \mathcal{C})$$

$\tilde{\Psi}_\mathcal{W}$ denotes the smallest norm-closed convex cone in $\mathcal{L}_\infty(\mathcal{U}, \mathcal{C})$ which contains $\Psi_\mathcal{W}$. The following lemma is a direct consequence of Theorem 3.2 and Theorem 3.4:

Lemma 5.3 $\mathcal{F} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *is least favorable for* $\mathcal{W}$ *if and only if*

$$s^\mathcal{F}[k] \ =\ \overline{S}[k] \qquad \forall\ k \in \Psi_\mathcal{W}$$

Theorem 5.4 is the analogue to [6, Theorem 8.2]. It characterises the existence of least favorable models in full generality.

[3] That is the minimal Bayes risk of the imprecise model is attained in the least favorable model which represents the worst-case. (This justifies the term "least favorable".) Remember that our definition of the Bayes risk corresponds to a worst-case consideration.

Theorem 5.4 *The following statements are equivalent:*

(a) *There is some* $\mathcal{F} := (q_\theta)_{\theta\in\Theta} \in (\mathcal{M}_\theta)_{\theta\in\Theta}$ *which is least favorable for* $\mathcal{W}$.

(b) $\overline{S}[k_1 + k_2] = \overline{S}[k_1] + \overline{S}[k_2] \qquad \forall\, k_1, k_2 \in \tilde{\Psi}_\mathcal{W}$

Proof:
(a)⇒(b): Statement (a) and Lemma 5.3 imply $\Psi_\mathcal{W} \subset \Phi_\mathcal{F}$. According to Lemma 5.2, $\tilde{\Psi}_\mathcal{W} \subset \Phi_\mathcal{F}$ and $k_1 + k_2 \in \Phi_\mathcal{F} \ \ \forall\, k_1, k_2 \in \tilde{\Psi}_\mathcal{W}$. Hence, for every $k_1, k_2 \in \tilde{\Psi}_\mathcal{W}$

$$\begin{aligned}\overline{S}[k_1 + k_2] \ &=\ s^\mathcal{F}[k_1 + k_2] = s^\mathcal{F}[k_1] + s^\mathcal{F}[k_2] = \\ &=\ \overline{S}[k_1] + \overline{S}[k_2]\end{aligned}$$

(a)⇐(b): Put $s[k] := \overline{S}[k] \ \ \forall\, k \in \tilde{\Psi}_\mathcal{W}$ and

$$s[k_1 - k_2] := s[k_1] - s[k_2] = \overline{S}[k_1] - \overline{S}[k_2]$$

for all $k_1, k_2 \in \tilde{\Psi}_\mathcal{W}$. Statement (b) implies that this is defined well. Hence, s is a linear functional on the vector space $\mathrm{lin}(\tilde{\Psi}_\mathcal{W}) = \tilde{\Psi}_\mathcal{W} - \tilde{\Psi}_\mathcal{W}$. For every $k = k_1 - k_2 \in \tilde{\Psi}_\mathcal{W} - \tilde{\Psi}_\mathcal{W} = \mathrm{lin}(\tilde{\Psi}_\mathcal{W})$,

$$\begin{aligned}s[k] \ &=\ \overline{S}[k_2 + k_1 - k_2] - \overline{S}[k_2] \leq \\ &\leq\ \overline{S}[k_2] + \overline{S}[k_1 - k_2] - \overline{S}[k_2] = \overline{S}[k]\end{aligned}$$

According to the Hahn-Banach-Theorem ([7, Theorem II.3.10]), s can be extended to a linear functional on $\mathcal{L}_\infty(\mathcal{U}, \mathcal{C})$ (again denoted by s) so that

$$s[h] \ \leq\ \overline{S}[h] \qquad \forall\, h \in \mathcal{L}_\infty(\mathcal{U}, \mathcal{C}) \tag{3}$$

(3) implies, that $s\big[I_\mathcal{U}\big] = 1$ and $s[\iota_\theta] = \frac{1}{n} \ \ \forall\, \theta \in \Theta$ (cf. Theorem 2.5). Then, $s_\theta : h \mapsto s[n\iota_\theta h]$ defines a precise model $(s_\theta)_{\theta\in\Theta}$ on $(\mathcal{U}, \mathcal{C})$. For every decision space $(\hat{\mathbb{D}}, \hat{\mathcal{D}})$ and every $(\hat{W}_\theta)_\theta \subset \mathcal{L}_\infty(\hat{\mathbb{D}}, \hat{\mathcal{D}})$,

$$\inf_{\rho\in\mathcal{T}(\mathcal{U},\hat{\mathbb{D}})} R\big((s_\theta)_\theta, \rho, (\hat{W}_\theta)_\theta\big) = s\big[K\big((\hat{W}_\theta)_\theta\big)\big] \tag{4}$$

according to Lemma 8.1 and

$$\inf_{\rho\in\mathcal{T}(\mathcal{U},\hat{\mathbb{D}})} R\big((s_\theta)_\theta, \rho, (\hat{W}_\theta)_\theta\big) \overset{(4)}{=} s\big[K\big((\hat{W}_\theta)_\theta\big)\big] \leq \\ \overset{(3)}{\leq}\ \overline{S}\big[K\big((\hat{W}_\theta)_\theta\big)\big] \ =\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\hat{\mathbb{D}})} R\big((\overline{Q}_\theta)_\theta, \sigma, (\hat{W}_\theta)_\theta\big)$$

according to Theorem 3.4. Hence, Theorem 4.1 implies that $(s_\theta)_{\theta\in\Theta}$ is worst-case-sufficient for $(\overline{Q}_\theta)_{\theta\in\Theta}$, i.e. there is some $T \in \mathcal{T}(\mathcal{U}, \mathcal{Y})$ so that $q_\theta := T(s_\theta) \in \mathcal{M}_\theta \ \ \forall\, \theta \in \Theta$. Finally for all $(W_\theta)_{\theta\in\Theta} \in \mathcal{W}$,

$$\begin{aligned}&\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta, \sigma, (W_\theta)_\theta\big) \ = \\ &\quad =\ \overline{S}\Big[K\big((W_\theta)_\theta\big)\Big] = s\Big[K\big((W_\theta)_\theta\big)\Big] \ = \\ &\quad \overset{(4)}{=}\ \inf_{\rho\in\mathcal{T}(\mathcal{U},\mathbb{D})} R\big((s_\theta)_\theta, \rho, (W_\theta)_\theta\big) \ \leq \\ &\quad \leq\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta, \sigma, (W_\theta)_\theta\big)\end{aligned}$$

where the last inequality follows from Lemma 8.2. □

6 Application of Least Favorable Models

Situations where we are faced with one precise prior distribution and a set of loss functions seem to be of secondary interest. More frequently, we are interested in situations where we are faced with an *imprecise* prior and one fixed loss function. However, the second issue can be treated as a special case of the first one:

Let Θ be a finite index set with cardinality n and $(W_\theta)_{\theta\in\Theta} \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$ be a loss function. Let $(\overline{Q}_\theta)_{\theta\in\Theta}$ be an imprecise model on $(\mathcal{Y},\mathcal{B})$ where $(\mathcal{M}_\theta)_{\theta\in\Theta}$ is the corresponding family of structures. Let $\overline{\Pi}$ be a coherent upper prevision on $\mathcal{L}_\infty(\Theta,2^\Theta)$ i.e. $\overline{\Pi}$ corresponds to a set of prior distributions $\mathcal{P} := \big\{\pi \in \mathrm{ba}(\Theta,2^\Theta) \,\big|\, \pi[a] \le \overline{\Pi}[a] \;\forall\, a \in \mathcal{L}_\infty(\Theta,2^\Theta)\big\}$.
For some $\pi \in \mathcal{P}$, put $\pi_\theta := \pi[I_{\{\theta\}}] \;\forall\, \theta \in \Theta$. Let σ be a randomisation. For the prior π, the Bayes risk is

$$R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \sum_{\theta\in\Theta} \pi_\theta \sigma(\overline{Q}_\theta)[W_\theta] =$$
$$= \frac{1}{n}\sum_{\theta\in\Theta} \sigma(\overline{Q}_\theta)[n\pi_\theta W_\theta] = R_0\big((\overline{Q}_\theta)_\theta,\sigma,(n\pi_\theta W_\theta)_\theta\big)$$

where $R_0\big((\overline{Q}_\theta)_\theta,\sigma,(n\pi_\theta W_\theta)_\theta\big)$ denotes the Bayes risk for the uniform prior π_0 defined by $\pi_0[I_\theta] = \frac{1}{n}$.

That is every prior can be absorbed in the loss function. So, we can transform the set $\mathcal{P}$ of priors π into a set $\mathcal{W}$ of loss functions $(n\pi_\theta W_\theta)_{\theta\in\Theta}$. Next, Theorem 5.4 yields a necessary and sufficient condition for the existence of a precise model which is simultaneously least favorable for the set of loss functions $\mathcal{W}$. We may also say that such a precise model is *simultaneously least favorable for the set of priors* $\mathcal{P}$.

The next theorem shows how least favorable models can be used to deal with situations where the distribution of the data as well as the prior is assumed to be imprecise. A decision procedure is optimal if it minimises the upper Bayes risk

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) \;=\; \sup_{\pi\in\mathcal{P}} R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

Theorem 6.1 *If $(\tilde{q}_\theta)_{\theta\in\Theta}$ is a simultaneously least favorable model of $(\mathcal{M}_\theta)_{\theta\in\Theta}$ for $\mathcal{P}$, there is a decision procedure $\tilde{\sigma} \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ which minimises*

$$R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

and also

$$R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

over $\mathcal{T}(\mathcal{Y},\mathbb{D})$.

Proof: For every $\sigma \in \mathcal{T}(\mathcal{Y},\mathbb{D})$ and $\pi \in \mathcal{P}$, put

$$\Gamma_1(\sigma,\pi) = R_\pi\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

and

$$\Gamma_2(\sigma,\pi) = R_\pi\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

It is easy to see that $\sigma \mapsto \Gamma_j(\sigma,\pi)$ is convex and lower semicontinuous for every $\pi \in \mathcal{P}$ and $j \in \{1,2\}$. Then, [8, Theorem 2] and simultaneous least favorability implies

$$\begin{aligned}
&\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \\
&= \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \sup_{\pi\in\mathcal{P}} \Gamma_1(\sigma,\pi) = \sup_{\pi\in\mathcal{P}} \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \Gamma_1(\sigma,\pi) \\
&= \sup_{\pi\in\mathcal{P}} \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \Gamma_2(\sigma,\pi) = \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} \sup_{\pi\in\mathcal{P}} \Gamma_2(\sigma,\pi) \\
&= \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) \qquad (5)
\end{aligned}$$

Lower semicontinuity of

$$\sigma \;\mapsto\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

and compactness of $\mathcal{T}(\mathcal{Y},\mathbb{D})$ ensure existence of some $\tilde{\sigma}$ which minimises $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$ (cf. [17, Theorem 3.7]). Additionally,

$$\begin{aligned}
R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\tilde{\sigma},(W_\theta)_\theta\big) \;&\le\; R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\tilde{\sigma},(W_\theta)_\theta\big) = \\
&= \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = \\
&\overset{(5)}{=} \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)
\end{aligned}$$

□

Remark 6.2 *It can easily be read off from the above proof that a decision procedure $\tilde{\sigma}$ which minimises $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$ also minimises $R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$. However, the reverse statement will not always be true.*[4] *So, it does not suffice to find a decision procedure $\hat{\sigma}$ which minimises $R_{\overline{\Pi}}\big((\tilde{q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$. It still has to be checked that $\hat{\sigma}$ really minimises $R_{\overline{\Pi}}\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$. Theorem 6.1 only states that there is a decision procedure which solves both minimisation problems.*

7 Concluding Remarks

In decision theory, straightforward updating may lead to suboptimal decisions if the data is distributed according to imprecise probabilities (cf. [3]). Therefore, data-based decision theory can be seen as a matter of its own. One of the major problems in data-based decision theory is that direct solutions of the

[4] In case of hypothesis testing, for example, this follows from [1, p. 162ff].

involved optimisation problems are quite often computationally intractable. Theorem 6.1 offers an opportunity to reduce the computational effort significantly if the imprecise model admits a least favorable (precise) model. Therefore, it is important to know for a given decision problem if such a least favorable model exists or not.

This question has been addressed by [6]. The concept of imprecise probability developed in [6] is very close to that one developed in [24]. From a mathematical point of view, the only difference is that [6] assumes that precise probabilities (i.e. linear previsions) have to be σ-additive. Surprisingly, this appears to be a burden which significantly reduces the applicability of [6].[5] The present article shows that the same result as in [6] is possible without any assumption on the involved (coherent) upper previsions if we dispense with σ-additivity.

This offers a general tool which makes it possible to reduce the computational effort in data-based decision theory under imprecision. However, further research has to be done for using it in concrete problems: As in [14], Theorem 5.4 is only concerned with the *existence* of a least favorable model but an algorithm for calculating least favorable models has not yet been developed. After [14], a lot of work was done to construct least favorable pairs in hypothesis testing for special cases (e.g. [19], [18], [12], [1]). In the much more general case of the present article, this is a matter of further research.

The present article might not only be interesting because of its results but also because of the applied tools: Getting around σ-additivity in the proofs of the present paper was possible by the use of notions and methods of [16]. This article is probably the first one which explicitly uses concepts of [16] in the theory of imprecise probability. Since these concepts were especially developed for large models, it is most likely that they can profitably be used in the theory of imprecise probability further on. Additionally, a theory of "sufficiency" is used which is not formulated in terms of conditional probabilities. In this way, a sufficiency theory for imprecise probabilities may be possible which is not affected by the problems which arise for conditional imprecise probabilities.

8 Appendix

Lemma 8.1 *Assume that s is a linear prevision on $\mathcal{L}_\infty(\mathcal{U},\mathcal{C})$ so that $s[\iota_\theta] = \frac{1}{n}\ \forall\,\theta\in\Theta$. Then, $s_\theta : h \mapsto s[n\iota_\theta h]$ defines a precise model $(s_\theta)_{\theta\in\Theta}$ on $(\mathcal{U},\mathcal{C})$ and*

$$\inf_{\rho\in\mathcal{T}(\mathcal{U},\mathbb{D})} R\big((s_\theta)_\theta,\rho,(W_\theta)_\theta\big) = s\big[K\big((W_\theta)_\theta\big)\big] \qquad (6)$$

for every decision space $(\mathbb{D},\mathcal{D})$ and every $(W_\theta)_\theta \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$. $K\big((W_\theta)_\theta\big)$ is defined as in (2).

For a proof of Lemma 8.1, confer [9, §6.3].

Lemma 8.2 *If a precise model $(p_\theta)_{\theta\in\Theta}$ on $(\mathcal{X},\mathcal{A})$ is sufficient for the precise model $(q_\theta)_{\theta\in\Theta}$ on $(\mathcal{Y},\mathcal{B})$, then*

$$\inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})} R\big((p_\theta)_\theta,\rho,(W_\theta)_\theta\big) \le \le \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

for every decision space $(\mathbb{D},\mathcal{D})$ and every $(W_\theta)_\theta \subset \mathcal{L}_\infty(\mathbb{D},\mathcal{D})$.

Proof: There is some $T\in\mathcal{T}(\mathcal{X},\mathcal{Y})$ so that $T(p_\theta) = q_\theta\ \forall\,\theta\in\Theta$. Therefore,

$$\begin{aligned}\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})}\sum_{\theta\in\Theta}\pi_\theta\sigma(q_\theta)[W_\theta] &= \\ = \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})}\sum_{\theta\in\Theta}\pi_\theta\sigma\big(T(p_\theta)\big)[W_\theta] &= \\ = \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})}\sum_{\theta\in\Theta}\pi_\theta\big(\sigma\circ T\big)(p_\theta)[W_\theta] &\ge \\ \ge \inf_{\rho\in\mathcal{T}(\mathcal{X},\mathbb{D})}\sum_{\theta\in\Theta}\pi_\theta\rho(p_\theta)[W_\theta]&\end{aligned}$$

because $\sigma\circ T\in\mathcal{T}(\mathcal{X},\mathbb{D})\ \forall\,\sigma\in\mathcal{T}(\mathcal{X},\mathbb{D})$. □

The following lemma is a consequence of the minimax theorem [8, Theorem 2]. Here, topological properties are crucial (Subsection 2.1.3). For a proof, confer [11].

Lemma 8.3

(a)
$$\inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = = \sup_{(q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta}\ \inf_{\sigma\in\mathcal{T}_r(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

(b)
$$\inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((\overline{Q}_\theta)_\theta,\sigma,(W_\theta)_\theta\big) = = \sup_{(q_\theta)_\theta\in(\mathcal{M}_\theta)_\theta}\ \inf_{\sigma\in\mathcal{T}(\mathcal{Y},\mathbb{D})} R\big((q_\theta)_\theta,\sigma,(W_\theta)_\theta\big)$$

Acknowledgements

I thank Thomas Augustin for valuable discussions. His suggestions have also greatly improved the readability of the article. Furthermore, I thank Helmut Rieder and Peter Ruckdeschel, who have drawn my attention to the work of Buja and Le Cam. I am also much indebted to the Cusanuswerk (Foundation of the Roman Catholic Church) for a Ph.D. scholarship. Finally, I thank the reviewers for their helpful comments.

[5] By topological reasons, insistence on σ-additivity enforces an additional, restrictive assumption on the involved (coherent) upper previsions (cf. Remark 2.2 and [10]).

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck & Ruprecht, Göttingen, 1998.

[2] T. Augustin. Neyman-Pearson Testing under Interval Probability by Globally Least Favorable Pairs - Reviewing Huber-Strassen Theory and Extending It to General Interval Probability. *Journal of Statistical Planning and Inference* 105:1–25, 2002.

[3] T. Augustin. On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View: A Cautionary Note on Updating Imprecise Priors. In: J.M. Bernard, T. Seidenfeld, M. Zaffalon (eds.) *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, Lugano*. Carleton Scientific, Waterloo, 31–45, 2003.
www.carleton-scientific.com/isipta/2003-toc.html

[4] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis. 2nd edition*. Springer-Verlag, New York, 1985.

[5] D. Blackwell. Comparison of Experiments. In: J. Neyman (ed.) *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 93–102, 1951.

[6] A. Buja. Simultaneously Least Favorable Experiments, Part I: Upper Standard Functionals and Sufficiency. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:367–384, 1984.

[7] N. Dunford, J.T. Schwartz. *Linear Operators, Part I: General Theory*. Wiley-Interscience, New York, 1957.

[8] K. Fan. Minimax Theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39:42–47, 1953.

[9] R. Hable. Decision Theory: Some of Le Cam's Concepts. *E-print*.
www.statistik.lmu.de/~hable/publications.html

[10] R. Hable. A Note on an Error in Buja (1984). *E-print*.
www.statistik.lmu.de/~hable/publications.html

[11] R. Hable. Supplements to the Article "Data-Based Decisions under Imprecise Probability and Least Favorable Models". *E-print*.
www.statistik.lmu.de/~hable/publications.html

[12] R. Hafner. Konstruktion robuster Teststatistiken. In: S. Schach, G. Trenkler (eds.) *Data Analysis and Statistical Inference*, Eul, Bergisch Gladbach, 145–160, 1992.

[13] H. Heyer. *Mathematische Theorie statistischer Experimente*. Springer-Verlag, Berlin, 1973.

[14] P.J. Huber, V. Strassen. Minimax Tests and the Neyman-Pearson Lemma for Capacities. *The Annals of Statistics*, 1:251–263, 1973.

[15] L. Le Cam. Sufficiency and Approximate Sufficiency. *The Annals of Mathematical Statistics*, 35:1419–1455, 1964.

[16] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.

[17] E.J. McShane, T.A. Botts. *Reals Analysis*. Van Nostrand, Princeton, (1959).

[18] F. Österreicher. On the Construction of Least Favorable Pairs of Distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43:49–55, 1978.

[19] H. Rieder. Least Favorable Pairs for Special Capacities. *The Annals of Statistics*, 5:909–921, 1977.

[20] B. Rüger. *Test- und Schätztheorie. Band II: Statistische Tests*. Oldenbourg-Verlag, München, 2002.

[21] M.C.M. Troffaes. Decision Making under Uncertainty Using Imprecise Probabilities *International Journal of Approximate Reasoning*, accepted, 2006.

[22] L.V. Utkin, T. Augustin. Powerful Algorithms for Decision Making under Partial Prior Information and General Ambiguity Attitudes. *4th International Symposium on Imprecise Probabilities and Their Applications*, Pittsburg, 2005.
www.sipta.org/isipta05/proceedings/046.html

[23] B. Vidakovic. Γ-minimax: a paradigm for conservative robust Bayesians. In: D.R. Insua, F. Ruggeri (eds.) *Robust Bayesian analysis*. Springer-Verlag, New York, 241–259, 2000.

[24] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London, 1991.

[25] K. Weichselberger. The Theory of Interval-Probability as a Unifying Concept for Uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000.

Climbing the Hills of Compiled Credal Networks

Rolf Haenni

Bern University of Applied Sciences, Switzerland
rolf.haenni@bfh.ch

University of Bern, Switzerland
haenni@iam.unibe.ch

Abstract

This paper introduces a new approximate inference algorithm for credal networks. The algorithm consists of two major steps. It starts by representing the credal network as a compiled logical theory. The resulting graphical structure is the basis on which the subsequent steepest-ascent hill-climbing algorithm operates. The output of the algorithm is an inner approximation of the exact lower and upper posterior probabilities.

Keywords. Credal Networks, Bayesian Networks, Credal Sets, Approximate Inference, Logical Compilation, Hill-Climbing, Local Search.

1 Introduction

Credal networks are like discrete Bayesian networks, except that they specify closed convex sets of probability mass functions, so-called *credal sets* [36], instead of single probability mass functions. They are usually *locally* (or *separately*) specified [1, 16], i.e. every network variable is associated with a collection of local conditional credal sets, which do not interfere with each other. It is possible to view a locally specified credal network as a set of Bayesian networks with the same directed acyclic graph [17].

In general, credal sets contain an infinite number of probability mass functions, but they are normally fully specified by a finite number of extreme points. These are the vertices of a convex polytope in the corresponding multi-dimensional space. In the case of binary variables, the polytopes coincide with intervals, which restricts the maximal number of necessary extreme points to two. In a Bayesian network, each polytope is restricted to a single (extreme) point.

Inference in a locally specified credal network usually means to derive lower and upper posterior probabilities from the *strong extension* [16], i.e. from the the largest joint credal set that satisfies *strong independence* [15]. Except for the particular case of binary variables in polytree-shaped networks [30], this is computationally extremely challenging, much more than classical inference in Bayesian network. The case of general categorical variables is NP–complete for polytree-shaped networks and NP^{PP}–complete for an unbounded induced treewidth, thus making inference in credal networks very inefficient [27].

In comparison with Bayesian networks, the additional computational complexity results from the potentially unbounded number of vertices needed to describe arbitrary credal sets. This can quickly outperform the benefits of applying local computation techniques to graphical models such as Bayesian networks. Local messages propagated through a credal network (resp. through the join tree obtained from a credal network) may thus possess the richness and complexity of the (global) joint credal set [10]. In fact, inference in credal networks is essentially a global multilinear optimization problem on top of the given graphical structure [18].

Facing the inherent computational complexity of credal networks, exact inference methods are only exceptionally suitable. One exception is the abovementioned case of binary variables in polytree-shaped networks, for which a polynomial-time algorithm exists [30]. All other exact methods (e.g. vertex enumeration, global optimization, and transformation algorithms) are only applicable to very small problem instances.

For large networks, approximate inference seems to be the most natural solution. There is a general distinction between *inner* and *outer* approximations, depending on whether the resulting interval is enclosed in the exact solution or vice versa. The quest for such approximate methods is currently one of the major research topics in the imprecise probability community, as the increasing number of corresponding publications in the last couple of years demonstrates, see e.g. [2, 4, 5, 7, 8, 9, 19, 20, 31, 32].

1.1 General Ideas

In this paper, we present a new approximate method for the inference problem in credal networks. The approach results from combining the following two basic techniques:

Logical Compilation. This is an emerging inference technique for Bayesian networks [12, 13, 14, 23, 45]. The general idea is to represent the graphical structure (topology) of the Bayesian network by a propositional theory. Possible local structures within the given CPTs can be exploited to simplify corresponding sentences of the theory [12, 14]. The resulting logical encoding is then *compiled* into an appropriate logical form called d-DNNF [25, 46], which supports all necessary operations to answer arbitrary queries (conditional probabilities) in polynomial time. The computational task is thus divided into an expensive (off-line) compilation phase and a fast (on-line) query-answering phase.

Hill-Climbing. This is a generic combinatorial optimization technique, which is widely used in many AI-related fields and applications [41]. The goal is to maximize (or minimize) a function $f : X \to \mathbb{R}$ through local search, where X is usually a discrete multi-dimensional state space. Local search means to jump from one configuration in the state space to a neighboring one, until a local maximum or possibly the global maximum is reached. An obvious heuristic for the selection of the neighboring configuration is to jump to the configuration with the steepest ascent of the respective value of f (*steepest-ascent hill-climbing*). The basic hill-climbing process is usually iterated with randomly generated starting points (*random-restart hill-climbing*), thus making it an interruptible anytime algorithm.

The idea of compiling a credal network in the same way as compiling a Bayesian network is quite obvious, but to our knowledge, this is still an unexplored approach. Pointing out this possibility is one of the goals of this paper.

Applying hill-climbing or other local search algorithms to approximate inference in credal networks is also quite obvious, as some of the existing approximation algorithms have demonstrated [4, 5, 6, 20]. Most of them are oriented towards the local propagation scheme in corresponding join trees [33, 43], in which each hill-climbing step requires the updating of the affected join tree messages. The hill-climbing procedure itself is guided by the current configuration of so-called *transparent* variables, whose role consists in selecting the actual vertices in the local credal sets.

1.2 Overview and Outline

In our method, we will also exploit the benefits of local computation in join trees, but only to compile the network structure into a d-DNNF during the inward phase [13]. The necessary information for the hill-climbing procedure is then available in a very simple and compact logical structure. For the current selection of vertices, this structure can then be used to efficiently compute or update the resulting posterior probability. Moreover, without much computational overhead, it is possible to determine the currently unselected vertex (i.e. the neighboring configuration) with the steepest ascent (resp. descent), which we can use as a heuristic to improve the performance of the local search.

After all, we get a simple but yet powerful steepest-ascent, random-restart hill-climbing algorithm to approximate inference in credal networks. By running the algorithm twice, once as a maximizing and once as a minimizing procedure, it produces good inner approximations of the exact probability bounds.

With respect to existing hill-climbing techniques for credal networks, our approach appears to be considerably simpler, as no complicated management of a bidirectional *double message system* is required, like e.g. in [6]. The logical representation is also inherently predestined to exploit existing local CPT regularities in the form of *context-specific independence* [3], *logical relationships* (pure or noisy), or *determinism* [12], for which existing methods typically use so-called *probability trees* [6, 9]. Finally, from the possibility of quickly finding the neighboring configuration with the steepest ascent (respectively descent), our method is likely to converge faster towards the exact results.

The rest of the paper is organized as follows. In Section 2, we give a short introduction to the main concepts of Bayesian and credal networks and the terminology used in this paper. Section 3 summarizes the compilation-based approach to inference in Bayesian (and credal) networks. Section 4 introduces hill-climbing and its application to compiled credal networks. This is the main part of the paper. The discussion and outlook in Section 5 concludes the paper.

2 Bayesian and Credal Networks

A *Bayesian network* (BN) is an efficient representation of a joint probability mass function over a set $\mathbf{X} = \{X_1, \ldots, X_n\}$ of variables [38]. We assume throughout this paper that all variables $X \in \mathbf{X}$ are categorical, i.e. their associated sets Ω_X of possible values are *finite*. The network itself consists of a directed acyclic graph (DAG), which represents the

direct influences among the variables, each of them attached to one node, and a set of conditional probability tables (CPT), which quantify the strengths of these influences. The whole BN represents a *joint probability mass function* $p : \Omega_{\mathbf{X}} \to [0, 1]$ over its variables in a compact manner by

$$p(\mathbf{X}) = \prod_{X \in \mathbf{X}} p(X|\mathbf{\Pi}(X)), \tag{1}$$

where $\mathbf{\Pi}(X)$ denotes the parents of node X in the DAG. Figure 1 depicts the BN for the "Dog-Problem" [11], which is often used in the literature for illustrative purposes. It consists of five binary variables F, B, L, D, and H, with corresponding CPTs $p(F)$, $p(B)$, $p(L|F)$, $p(D|F, B)$, and $p(H|D)$.

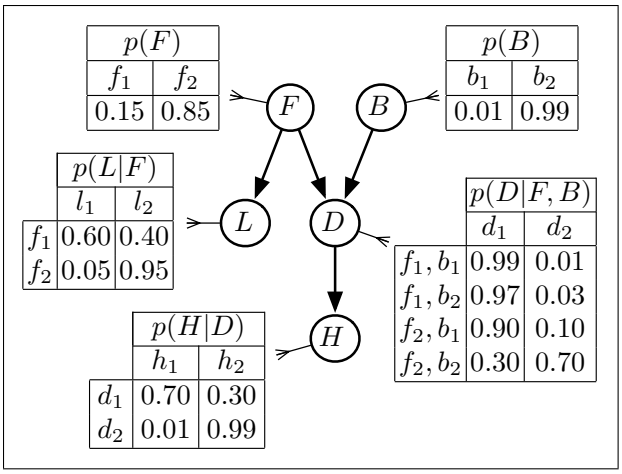


Figure 1: Example of a simple Bayesian network with five binary variables.

Inference in Bayesian networks means to compute the conditional probability $P(H{=}h \,|\, E_1{=}e_1, \ldots, E_r{=}e_r)$, or simply

$$P(h|\mathbf{e}) = \frac{P(h, \mathbf{e})}{P(\mathbf{e})}, \tag{2}$$

of a hypothesis $h \in \Omega_H$ for some observed evidence $\mathbf{e} = (e_1, \ldots, e_r) \in \Omega_{\mathbf{E}}$. We will call $H \in \mathbf{X}$ *query variable* and the elements of $\mathbf{E} = \{E_1, \ldots, E_r\} \subseteq \mathbf{X}$ *evidence variables*. To see how to solve the inference problem, let $\mathbf{Y} = \{Y_1, \ldots, Y_s\} \subseteq \mathbf{X}$ be an arbitrary subset of variables, $\mathbf{y} = (y_1, \ldots, y_s) \in \Omega_{\mathbf{Y}}$ a configuration of values $y_i \in Y_i$, and $\mathbf{Z} = \mathbf{X} \backslash \mathbf{Y}$. Then it is sufficient to compute

$$P(\mathbf{y}) = \sum_{\mathbf{z} \in \Omega_{\mathbf{Z}}} p(\mathbf{yz}) \tag{3}$$

twice, once with $\mathbf{Y} = \{H\} \cup \mathbf{E}$ and $\mathbf{y} = (h, \mathbf{e})$ to get the nominator and once with $\mathbf{Y} = \mathbf{E}$ and $\mathbf{y} = \mathbf{e}$ to get the denominator of the above formula. Note that the necessary sum-of-products involve exponentially many terms, but if the computations are performed *locally* in a join tree propagation or variable elimination process, it is almost always possible to replace it by a compact factorization [28, 33, 43]. Join trees are also useful to avoid redundant computations in the case of multiple queries or updates.

Credal networks (CN) are similar to Bayesian networks, but they relax the uniqueness assumption for the given probability values [16]. In a *locally* (or *separately*) specified CN, the CPT entries are replaced by corresponding conditional credal sets, on which no further restrictions are imposed [1]. A *credal set* for a variable $X \in \mathbf{X}$ is a closed convex set $K(X)$ of probability mass functions $p(X)$ [36]. Similarly, a *conditional credal set* $K(X|\boldsymbol{\pi})$ is a closed convex set of conditional probability mass functions $p(X|\boldsymbol{\pi})$, where $\boldsymbol{\pi} \in \Omega_{\mathbf{\Pi}(X)}$ is one particular assignment of values for the direct influences $\mathbf{\Pi}(X)$ of X. With

$$K(X|\mathbf{\Pi}(X)) = \{K(X|\boldsymbol{\pi}) : \boldsymbol{\pi} \in \Omega_{\mathbf{\Pi}(X)}\} \tag{4}$$

we denote the collection of all such conditional credal sets. This is what a CN needs to specify for all variables $X \in \mathbf{X}$.

Normally, a single conditional credal set $K(X|\boldsymbol{\pi})$ is specified and represented by a finite set

$$\mathrm{Ext}(K(X|\boldsymbol{\pi})) = \{p_1(X|\boldsymbol{\pi}), \ldots, p_m(X|\boldsymbol{\pi})\} \tag{5}$$

of extreme points $p_i(X|\boldsymbol{\pi})$. Geometrically, these extreme points are vertices of a polytope in the corresponding additive subspace of $[0, 1]^{|\Omega_X|}$. In the binary case, i.e. for $|\Omega_X| = 2$, the additive subspace of $[0, 1]^2$ is a simple straight line between $(0, 1)$ and $(1, 0)$, on which credal sets degenerate into intervals with at most two extreme points (the bounds of the intervals).

If we generalize the BN of Fig. 1 to a CN, we need to replace the rows in each CPT by corresponding (conditional) credal sets. Since all involved variables are binary, it is sufficient to specify *two* extreme points for each credal set. As an example, consider $K(H|D)$, which consists of the credal sets $K(H|d_1)$ and $K(H|d_2)$, and suppose that the precise values $p(H|d_i)$ from Fig. 1 are enlarged to sets of extreme points $\mathrm{Ext}(K(H|d_i)) = \{p_1(H|d_i), p_2(H|d_i)\}$ with the following values:

	Ext($K(H\|D)$)			
	$p_1(H\|D)$		$p_2(H\|D)$	
	h_1	h_2	h_1	h_2
d_1	0.70	0.30	0.80	0.20
d_2	0.01	0.99	0.03	0.97

Note that the particularity of binary variables allows us to specify the same information more compactly

by $p(h_1|d_1) \in [0.7, 0.8]$ and $p(h_1|d_2) \in [0.01, 0.03]$ (and therefore by $p(h_2|d_1) \in [0.2, 0.3]$ and $p(h_2|d_2) \in [0.97, 0.99]$), thus making the interval-shaped credal sets more visible. This is an appealing view, in which credal sets appear to be nothing but *probability intervals* or *interval-valued probabilities* [35, 37, 44, 48], but the simplicity of this view belies the fact that credal sets are more general than probability intervals, e.g. for variables with more than two values. Interval representations are also problematical when it comes to apply Bayes' rule or to propagate them through a network [6, 16].

For a given credal network, we use $K(\mathbf{X})$ to denote its *joint credal set*. Note that its actual definition depends on how the concept of independence is adopted for credal sets. In this paper, we follow the usual convention of *strong independence* [15], which allows us to define $K(\mathbf{X})$ to be the *strong extension* of the credal network, i.e. as the largest joint credal set such that every variable $X \in \mathbf{X}$ is strongly independent [16]. This set contains all possible joint probability mass functions, if we select corresponding elements $p(X|\boldsymbol{\pi})$ from each conditional credal set $K(X|\boldsymbol{\pi})$. Formally, we can write

$$K(\mathbf{X}) = \left\{ \prod_{X \in \mathbf{X}} p(X|\mathbf{\Pi}(X)) : p(X|\boldsymbol{\pi}) \in K(X|\boldsymbol{\pi}) \right\},$$

where $\boldsymbol{\pi}$ denotes respective configurations of $\mathbf{\Pi}(X)$. Note that each element $p(\mathbf{X}) \in K(\mathbf{X})$ can be seen as the joint probability mass function of a corresponding Bayesian network (on the same DAG).

The convexity of $K(\mathbf{X})$ guarantees its extreme points to result only from combinations of extreme points of each conditional credal set $K(X|\boldsymbol{\pi})$ [17], and this allows us to rewrite the above expression as

$$\mathrm{Ext}(K(\mathbf{X})) = \mathrm{CH} \left\{ \prod_{X \in \mathbf{X}} p(X|\mathbf{\Pi}(X)) : p(X|\boldsymbol{\pi}) \in \mathrm{Ext}(K(X|\boldsymbol{\pi})) \right\},$$

where CH stands for an algorithm to compute the convex hull of a set of points in a multi-dimensional space [26]. This property reflects the fact that inference in credal networks is reducible to computations of extreme points.

Inference for a given credal set $K(\mathbf{X})$, a query $h \in \Omega_H$, and some observations $\mathbf{e} \in \Omega_{\mathbf{E}}$ means to determine tight bounds over all possible probability values $P(h|\mathbf{e})$, i.e. to compute the *lower* posterior probability

$$\underline{P}(h|\mathbf{e}) = \min\{P(h|\mathbf{e}) : p(\mathbf{X}) \in K(\mathbf{X})\}, \qquad (6)$$

and the *upper* posterior probability

$$\overline{P}(h|\mathbf{e}) = \max\{P(h|\mathbf{e}) : p(\mathbf{X}) \in K(\mathbf{X})\}. \qquad (7)$$

To compute these values under the assumption of strong independence, we can again exploit the convexity of $K(\mathbf{X})$ to restrict the necessary search space to the finite set $\mathrm{Ext}(K(\mathbf{X}))$ of extreme points [17]. Note that if N denotes the total number of involved conditional credal sets, all of them described by k extreme points, then $\mathrm{Ext}(K(\mathbf{X}))$ may possess up to N^k elements, thus making the above minimization/maximization problems very difficult tasks. Except for polytree-shaped networks with binary variables, no algorithm can handle large credal networks exactly [27, 30].

3 Compiling Bayesian Networks

The goal of compiling a Bayesian or credal network is the construction of a logical representation φ, in which all the topological and context-specific information of the network is included in a compact and easily manageable form. This construction is a one-time preparatory step, which is intended to take place off-line. The resulting logical representation φ contains two types of propositional variables, the ones linked to the CPT entries and the ones linked to the individual values of the network variables. The corresponding sets of propositions are denoted by Θ and Δ, respectively.

To compute the probability $P(\mathbf{y})$ of a configuration $\mathbf{y} = (y_1, \ldots, y_s) \in \Omega_{\mathbf{Y}}$ w.r.t. $\mathbf{Y} = \{Y_1, \ldots, Y_s\} \subseteq \mathbf{X}$, which is the basic computational task to answer arbitrary probabilistic queries (see Equation 3 in Section 2), φ is transformed into $\varphi_{\mathbf{y}} = (\varphi|\mathbf{y})^{-\Delta}$ by first *conditioning* φ on $\mathbf{y}$ and then *eliminating* (or *forgetting*) from $\varphi|\mathbf{y}$ all Δ-variables. The remaining Θ-variables in $\varphi_{\mathbf{y}}$ are all of the form $\theta_{x|\boldsymbol{\pi}}$, i.e. each of them is linked to a CPT entry $p(x|\boldsymbol{\pi})$.

To ensure that the above-mentioned computational steps are always efficient, φ must be a so-called d-DNNF [25, 46].[1] A *negation normal form* (NNF) is a rooted, directed acyclic graph, whose leaves are labeled with the literals of a propositional language.[2] All other nodes denote either a logical AND or a logical OR. d-DNNFs are NNFs satisfying two important properties called *determinism* (d) and *decomposability* (D).[3] Fig. 2 depicts the d-DNNF φ_{h_1} for the Bayesian

[1] The suggestion of using d-DNNFs as a target compilation language for Bayesian networks goes back to [23]. The mathematical explanation in [45] backups this choice.

[2] Note that NNFs are *propositional directed acyclic graphs* (PDAG), for which the *simple-negation* property holds [46].

[3] NNFs, in which some propositional variables are implicitly known to be exclusive and exhaustive, should be regarded as corresponding *multi-state directed acyclic graphs* (MDAG), a generalization of PDAGs (and NNFs) to arbitrary categorical variables [47]. In the context of MDAGs, some properties (incl. determinism and decomposability) and some operations (incl. conditioning and variable elimination) are based on more gen-

network in Fig. 1 and the query $\mathbf{y} = h_1$. Note that the network node L has no impact on $P(h_1)$, which is why φ_{h_1} is not affected by variables of the form $\theta_{l_i|f_j}$ (they disappear while l_1 and l_2 are eliminated from $\varphi|h_1$). Similarly, φ_{h_1} does not contain variables of the form $\theta_{h_2|d_i}$ (they disappear while φ is conditioned on h_1).

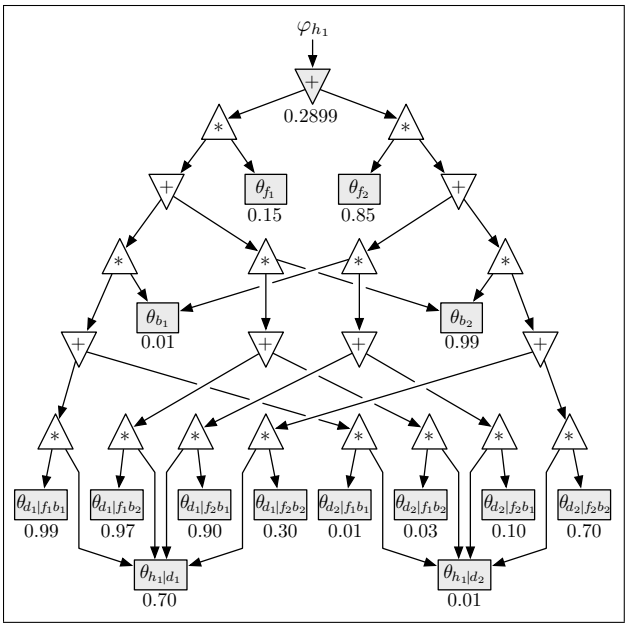


Figure 2: The d-DNNF obtained for the Bayesian network in Fig. 1 and the query $\mathbf{y} = h_1$. AND- and OR-nodes are denoted by $\triangle$ and $\triangledown$, respectively.

For a given d-DNNF $\varphi_{\mathbf{y}}$, it is easy to compute $P(\mathbf{y}) = P(\varphi_{\mathbf{y}})$ by simply propagating the conditional probabilities $p(x|\boldsymbol{\pi})$ from the leaves $\theta_{x|\boldsymbol{\pi}}$ upwards to the root of the DAG. At each OR-node, determinism allows the incoming values to be added, and at each AND-node, decomposability allows the incoming values to be multiplied, as indicated in Fig. 2 by the symbols + and $*$, respectively. The result we obtain at the root is $P(h_1) = P(\varphi_{h_1}) = 0.2899$.

Computing probabilities is thus another efficient operation supported by d-DNNFs. In other words, any given compiled Bayesian network φ allows us to efficiently compute all possible simple queries $P(\mathbf{y}) = P(\varphi_{\mathbf{y}})$. This in turn enables the efficient computation of all possible general queries $P(h|\mathbf{e})$, namely in terms of two simple queries $P(h, \mathbf{e}) = P(\varphi_{h,\mathbf{e}})$ and $P(\mathbf{e}) = P(\varphi_{\mathbf{e}})$. Note that $\varphi_{h,\mathbf{e}}$ is often simpler than $\varphi_{\mathbf{e}}$. Moreover, it is very likely that $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ (or any pair of related d-DNNFs) share a substantial number of common subgraphs.[4] In Fig. 2, for example, it turns out that φ_{f_1,h_1} corresponds to the left subgraph of the root node of φ_{h_1}, whereas only three additional nodes are required to construct φ_{b_1,h_1}, two of them pointing to respective subgraphs of φ_{h_1}. The sharing of common subgraphs is important, as it allows the bottom-up computation of several probabilities in one single pass. We will heavily exploit this when it comes to realize the selection of the steepest ascent in the random-restart hill-climbing algorithm of the following section.

For the compilation itself, there are two distinct classes of methods. The methods of the first class start from encoding the Bayesian network as a CNF ψ, which is then converted into a d-DNNF $\varphi = \text{CNF2dDNNF}(\psi)$, e.g. by using Darwiche's compiler [22, 24]. This is the classical compilation approach in the literature [12, 14, 42, 45].

The more recent methods of the second class, called *tabular compilation* methods [13], avoid the detour over a CNF. The idea is to run a simple variable elimination procedure over all network variables. More generally spoken, it is the application of the *fusion* (or *bucket elimination*) algorithm to a particular type of *semiring valuations* [34], in which the semiring consists of all Boolean functions w.r.t. the variables $\Theta \cup \Delta$ (respectively of all classes of equivalent logical representations). For appropriate input valuations, it is easy to show that the output of the algorithm is indeed a d-DNNF. The fact that any algebra of semiring valuations satisfies the general valuation algebra axioms (see Theorem 2 in [34]) allows this type of compilation to fully exploit the principle of *local computation*. The worst-case complexity (for both time and space) is thus identical to standard join tree algorithms for Bayesian networks, i.e. exponential in the network's induced treewidth (= size of the largest node in the join node). In fact, one can look at tabular compilation as a standard inward propagation in a join tree, where the evolving d-DNNF keeps trace of the effected computations [13, 21].

4 Hill-Climbing in Compiled Credal Networks

Let's assume now that a given credal network is compiled in the same way as a corresponding Bayesian network, i.e. as if the attached credal sets were precise values. We will now show how to use the resulting d-DNNF φ as a starting position for the inner approximation of lower and upper posterior probabilities $\underline{P}(h|\mathbf{e})$ and $\overline{P}(h|\mathbf{e})$, respectively. If the hypothesis h and the evidence $\mathbf{e}$ are given, the first step is clear, namely to transform φ into corresponding d-DNNFs $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ (see previous section). Note that the same $\varphi_{\mathbf{e}}$ can be used for several hypotheses as long as $\mathbf{e}$ remains unchanged.

eral definitions, but their basic functionalities and properties remain the same.

[4] This is a consequence of the linear running time of conditioning, which restrains the number of newly created nodes.

4.1 The Hill-Climbing Algorithm

To realize the approximation of $\underline{P}(h|\mathbf{e})$ and $\overline{P}(h|\mathbf{e})$ as a hill-climbing algorithms, the next thing to do is to define an appropriate search space. For this, we make use of the fact that both $\underline{P}(h|\mathbf{e})$ and $\overline{P}(h|\mathbf{e})$ result from corresponding extreme points of the joint credal set $K(\mathbf{X})$, i.e. from elements of the set $\mathrm{Ext}(K(\mathbf{X}))$. This set in turn is determined by the extreme points $\mathrm{Ext}(K(X|\boldsymbol{\pi}))$ of the local credal sets $K(X|\boldsymbol{\pi})$ at each node of the network (see Section 2).

To access individual elements of $\mathrm{Ext}(K(\mathbf{X}))$, we employ a strategy that is similar to the use of *transparent variables* in [4, 6], but here we will not integrate them as explicit nodes into the network structure. The idea is thus to consider discrete variables $T_{X|\boldsymbol{\pi}}$, one for each local credal set $K(X|\boldsymbol{\pi})$, where the role of each $T_{X|\boldsymbol{\pi}}$ is to select an extreme point of the credal set $K(X|\boldsymbol{\pi})$. If $k_{X|\boldsymbol{\pi}} = |\mathrm{Ext}(K(X|\boldsymbol{\pi}))|$ denotes the number of extreme points of the credal set $K(X|\boldsymbol{\pi})$, then $\Omega_{T_{X|\boldsymbol{\pi}}} = \{1, \ldots, k_{X|\boldsymbol{\pi}}\}$ is the set of possible values of $T_{X|\boldsymbol{\pi}}$. Furthermore, if $\mathbf{T}$ denotes the set of all such variables $T_{X|\boldsymbol{\pi}}$, then

$$\Omega_{\mathbf{T}} = \prod_{T_{X|\boldsymbol{\pi}} \in \mathbf{T}} \Omega_{T_{X|\boldsymbol{\pi}}} \tag{8}$$

denotes the set of all configurations with respect to $\mathbf{T}$. For a specific configuration $\mathbf{t} = \mathbf{s}t\mathbf{u} \in \Omega_{\mathbf{T}}$, in which t denotes the value of the transparent variable $T_{X|\boldsymbol{\pi}}$ in $\mathbf{t}$, we can write $p_t(X|\boldsymbol{\pi}) \in \mathrm{Ext}(K(X|\boldsymbol{\pi}))$ to select the corresponding extreme point of the credal set $K(X|\boldsymbol{\pi})$. Similarly, we write $p_{\mathbf{t}}(\mathbf{X})$ for the selected joint probability mass function, $P_{\mathbf{t}}(h|\mathbf{e})$ for induced posterior probabilities, and $P_{\mathbf{t}}(\varphi_{h,\mathbf{e}})$ and $P_{\mathbf{t}}(\varphi_{\mathbf{e}})$ for probabilities of a compiled network. This formal setting allows us to rephrase the definitions of lower and upper posterior probabilities in Equation 6 and 7 by

$$\underline{P}(h|\mathbf{e}) = \min_{\mathbf{t} \in \Omega_{\mathbf{T}}} P_{\mathbf{t}}(h|\mathbf{e}) = \min_{\mathbf{t} \in \Omega_{\mathbf{T}}} \frac{P_{\mathbf{t}}(\varphi_{h,\mathbf{e}})}{P_{\mathbf{t}}(\varphi_{\mathbf{e}})}, \tag{9}$$

$$\overline{P}(h|\mathbf{e}) = \max_{\mathbf{t} \in \Omega_{\mathbf{T}}} P_{\mathbf{t}}(h|\mathbf{e}) = \max_{\mathbf{t} \in \Omega_{\mathbf{T}}} \frac{P_{\mathbf{t}}(\varphi_{h,\mathbf{e}})}{P_{\mathbf{t}}(\varphi_{\mathbf{e}})}, \tag{10}$$

respectively, i.e. $\Omega_{\mathbf{T}}$ is the discrete search space, on which the following steepest-ascent, random-restart hill-climbing procedure operates. The details of the procedure are shown in Algorithm 1, which deserves some additional explanations:

- Lines 2–3 describe the preparation phase. Line 4 sets the current global maximum $P_{\max}$ to 0.
- The outer loop (lines 5–12) describes the "random-restart" part of the algorithm. It starts by selecting a random configuration $\mathbf{t} \in \Omega_{\mathbf{T}}$ in Line 8 and ends by updating the current value for the global maximum. We assume the existence of a global variable *maxRuns*, which determines the number of passes.
- The actual hill-climbing takes place in the inner loop (lines 7–10). The crucial step for this is the selection of $\mathbf{t}$'s best neighbor in the search space $\Theta_{\mathbf{T}}$ by calling the function `BestNeighbor` (Line 9). This is the "steepest-ascent" part of the algorithm, which will later be discussed in further details. If no neighbor improves the current local maximum $P_{\mathbf{t}} = P_{\mathbf{t}}(h|\mathbf{e})$, we expect `BestNeighbor` to return *nil*.[5]
- The current value of the local maximum is updated in Line 8. This involves the bottom-up computation of the probabilities $P_{\mathbf{t}}(\varphi_{h,\mathbf{e}})$ and $P_{\mathbf{t}}(\varphi_{\mathbf{e}})$ based on the current selection of extreme points $p_{\mathbf{t}}(X|\boldsymbol{\pi})$, from which the actual values $p_t(\theta_{x|\boldsymbol{\pi}})$ of all variables $\theta_{x|\boldsymbol{\pi}} \in \Theta$ are extracted. Note that only those parts of $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ need to be processed, which are affected by the transition from the old to the new configuration. Of course, common subgraphs of $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ are processed in one single pass.

Algorithm 1: `ApproxUpperProb`$(\varphi, h, \mathbf{e}, \mathbf{T})$

```
1  begin
2      φ_{h,e} ← (φ|h,e)^{-Δ};
3      φ_e ← (φ|e)^{-Δ};
4      P_max ← 0;
5      for i ← 1 to maxRuns do
6          t ← RandomConfiguration(T);
7          repeat
8              P_t ← P_t(φ_{h,e}) / P_t(φ_e);
9              t ← BestNeighbor(t, T, P_t, φ_{h,e}, φ_e);
10         until t = nil;
11         P_max ← max{P_max, P_t};
12     return P_max;
13 end
```

The corresponding minimization algorithm, i.e. the approximation of the lower posterior probability $\underline{P}(h|\mathbf{e})$, is almost identical, except for the initialization of the global maximum (Line 4), the selection of the best neighbor (Line 9), and the updating of the global maximum (Line 11). In the rest of this paper, we will therefore restrict our discussion to the maximization problem.

[5] To avoid getting stuck on a plateau (flat part of the search space), the algorithm should allow so-called *sideway moves* to states with equal values. This may cause infinite loops, but they can be avoided by keeping track of previously visited plateau states. For simplicity, we do not explicitly take care of these details in the proposed algorithm.

4.2 Selecting the Best Neighbor Efficiently

Let us now take a closer look at the problem of selecting the best neighbor of the actual configuration $\mathbf{t}$. For this, suppose that $t \in \Omega_{T_{X|\boldsymbol{\pi}}}$ is the current value of a transparent variable $T_{X|\boldsymbol{\pi}} \in \mathbf{T}$ in the actual configuration $\mathbf{t} = \mathbf{s}t\mathbf{u}$. Every configuration $\mathbf{t}' = \mathbf{s}t'\mathbf{u}$ with $t' \in \Omega_{T_{X|\boldsymbol{\pi}}}$ and $t' \neq t$ is then a possible neighbor of $\mathbf{t}$ in $\Omega_{\mathbf{T}}$. Selecting the best neighbor, i.e. the neighbor with the most significant improvement with respect to the actual local maximum $P_{\mathbf{t}} = P_{\mathbf{t}}(h|\mathbf{e})$, means thus to compute $P_{\mathbf{t}'} = P_{\mathbf{t}'}(h|\mathbf{e})$ for all such configurations $\mathbf{t}'$ and all transparent variables $T_{X|\boldsymbol{\pi}} \in \mathbf{T}$. The following algorithm shows a naïve solution for this simple idea.

Algorithm 2: `BestNeighbor`($\mathbf{t}$,$\mathbf{T}$,$P_{\mathbf{t}}$,$\varphi_{h,\mathbf{e}}$,$\varphi_{\mathbf{e}}$)

```
1  begin
2      t_max ← t;
3      foreach T_{X|π} ∈ T do
4          t ← value of T_{X|π} in t;
5          foreach t' ∈ Ω_{T_{X|π}} \ {t} do
6              t' ← replace t by t' in t;
7              P_{t'} ← P_{t'}(φ_{h,e}) / P_{t'}(φ_e);
8              if P_{t'} > P_t then
9                  t_max ← t';
10                 P_t ← P_{t'};
11     if t = t_max then return nil;
12     else return t_max;
13 end
```

The problem with this naïve solution is the repetitive probability calculation in the inner loop (Line 7). This can be avoided by pre-compiling $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ according to the following Shannon decomposition, in which $\varphi_{\mathbf{y}}$ denotes a general instantiation of φ to a vector $\mathbf{y}$ and $X \in \mathbf{X}$ the network variable affected by the current transparent variable $T_{X|\boldsymbol{\pi}}$:

$$\begin{aligned} P_{\mathbf{t}'}(\varphi_{\mathbf{y}}) &= \sum_{x \in \Omega_X} P_{\mathbf{t}'}(\theta_{x|\boldsymbol{\pi}}) P_{\mathbf{t}'}(\varphi_{\mathbf{y}}|\theta_{x|\boldsymbol{\pi}}) \\ &= \sum_{x \in \Omega_X} p_{t'}(x|\boldsymbol{\pi}) P_{\mathbf{t}}(\varphi_{\mathbf{y}}|\theta_{x|\boldsymbol{\pi}}). \end{aligned} \tag{11}$$

Note that in the second line of Equation 11, it is no longer necessary to explicitly generate the neighboring configurations $\mathbf{t}'$. In other words, if we first derive from $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ all possible instantiations $\varphi_{h,\mathbf{e}}|\theta_{x|\boldsymbol{\pi}}$ and $\varphi_{\mathbf{e}}|\theta_{x|\boldsymbol{\pi}}$, respectively, we can use Equation 11 to directly obtain the probabilities $P_{\mathbf{t}'}(h|\mathbf{e})$ of all neighboring configurations $\mathbf{t}'$, i.e. without actually generating them. In Algorithm 2, this can be realized by skipping Line 6 and by replacing the right hand side of Line 7 by corresponding versions of Equation 11.

4.3 Recapitulation and Complexity Analysis

To conclude this section, let's first recapitulate the individual steps of the proposed method and then discuss their respective running time complexities.

To make the above steepest-ascent scheme work for a given hypothesis h and the evidence $\mathbf{e}$, we first need to transform the compiled network φ into $\varphi_{h,\mathbf{e}}$ and $\varphi_{\mathbf{e}}$ and then into $\varphi_{h,\mathbf{e}}|\theta_{x|\boldsymbol{\pi}}$ and $\varphi_{\mathbf{e}}|\theta_{x|\boldsymbol{\pi}}$ for all $\theta_{x|\boldsymbol{\pi}} \in \Theta$. The result is a collection

$$\begin{aligned} \Phi_{h|\mathbf{e}} = &\{\varphi_{h,\mathbf{e}}, \varphi_{\mathbf{e}}\} \cup \{\varphi_{h,\mathbf{e}}|\theta_{x|\boldsymbol{\pi}} : \theta_{x|\boldsymbol{\pi}} \in \Theta\} \\ &\cup \{\varphi_{\mathbf{e}}|\theta_{x|\boldsymbol{\pi}} : \theta_{x|\boldsymbol{\pi}} \in \Theta\} \end{aligned} \tag{12}$$

of d-DNNFs, which are likely to overlap heavily. This is illustrated in Fig. 3 in the form of a d-DNNF with multiple roots.

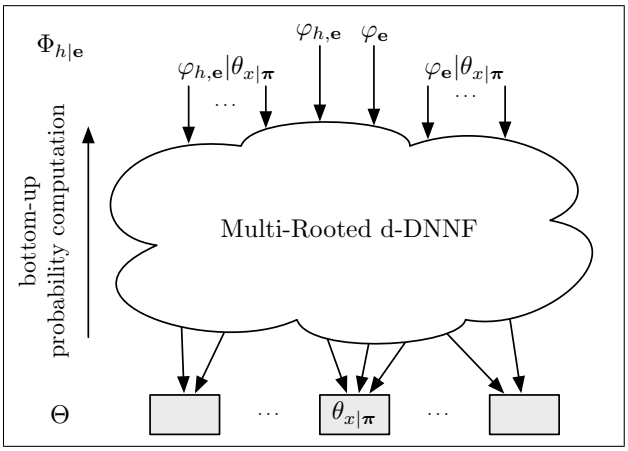


Figure 3: Probability computations in a multi-rooted d-DNNF with overlapping subgraphs.

To always keep the involved probabilities at each root up-to-date during the hill-climbing, we need to do the bottom-up probability computation only once at each hill-climbing step (i.e. at Line 8 of Algorithm 1), namely for the entire multi-rooted d-DNNF. The decision about the steepest ascent with respect to the current configuration $\mathbf{t}$ follows then from applying Equation 11 to all values t' that are incompatible with $\mathbf{t}$.

As discussed earlier, the worst-case running time and space complexity of the compilation phase is $O(2^d)$, where d denotes the network's induced treewidth for the given variable ordering. This is equivalent to the complexity of standard join tree algorithms for Bayesian networks. In other words, if $s = |\varphi|$ denotes the size (= number of edges) of the d-DNNF φ, then s reflects roughly the number of basic arithmetic operations (additions and multiplications) to be performed in the inward phase of a corresponding join tree propagation algorithm. Note that in the presence of strong local regularities in the form of context-specific independence, (pure or noisy) logical relationships, or

scarce CPTs, it is not untypical for the size s and therefore for the problem-specific complexity of the compilation phase to be much more favorable than $O(2^d)$.

The second preparatory step for the actual hill-climbing algorithm is the element-wise computation of the set $\Phi_{h|\mathbf{e}}$. For a given d-DNNF φ of size s, computing one such element requires $O(s)$ steps, which is a consequence of the fact that both conditioning and the particular type of variable elimination run in $O(s)$ time for d-DNNFs [25, 46]. Thus the total running time of the second step is $O(s{\cdot}|\Phi_{h|\mathbf{e}}|)$ and therewith $O(s{\cdot}|\Theta|)$, where $|\Theta|$ itself is proportional to both the number of network variables $n = |\mathbf{X}|$ and the corresponding maximal cardinality $c = \max\{|\Omega_X| : X \in \mathbf{X}\}$. This means that the worst-case running time of the entire preparatory phase is $O(c{\cdot}n{\cdot}2^d)$. This shows that the preparatory phase only depends on the network parameters c, d, and n, but not on the concrete local credal sets.

To analyze the running time of the actual hill-climbing algorithm, let S denote the total size of the multi-rooted d-DNNF on which the algorithm operates. Note that probability computations are supported by d-DNNFs in linear time, i.e. if K denotes the total number of extreme points over all locally specified credal sets (which correlates with the number of basic steps in the selection of the steepest ascent), then each individual hill-climbing step runs in $O(S{+}K)$ time. Since S is likely to be much larger than K, we can assume that the running time of the entire hill-climbing procedure is simply $O(\mathit{maxRuns}{\cdot}S)$. Due to the overlapping areas in the multi-rooted d-DNNF, S itself is often of the same order of magnitude as s.

5 Discussion and Conclusion

The method presented in this paper is a new technique to approximate inference in credal networks. The core of the approach is the idea of compiling the network into an appropriate logical form φ, which allows us to efficiently accomplish all necessary computational steps to answer probabilistic queries. Compilation techniques are increasingly applied to Bayesian networks, but the proposal to apply them to credal networks and to combine them with local search techniques is original.

With respect to existing approximation techniques for credal networks, let's point out some of the most important strengths of our approach.

- *Simplicity.* To make our approach work, only few simple procedures need to be implemented. The most important procedure is the compilation itself. For this, e.g. by using Nenok [39, 40], a generic framework for local computations in (semiring) valuation algebras, only few lines of code are necessary to handle the construction of the d-DNNF φ. Further procedures to implement are the operation of conditioning $\varphi|\mathbf{y}$ and the variable elimination $\varphi^{-\Delta}$. Both of them can be realized by simple recursions. The same holds for computing (and updating) the involved probabilities in the multi-rooted d-DNNF $\Phi_{h|\mathbf{e}}$, which turns out to be a classical postorder (bottom-up) traversal of a directed acyclic graph.

- *Flexibility.* The compiled logical form can be seen as a general recipe with precise instructions for the computation of all sorts of probabilities w.r.t. a given network. This is a very flexible and powerful starting position, which allows us to do all sorts of different things very easily, e.g. the efficient selection of the steepest ascent. The same structure could thus be used to solve other problems such as MAP or MPE.

- *Efficiency.* For a given multi-rooted d-DNNF, the updating of the probabilities during the hill-climbing process and the selection of the steepest ascent can be realized without any redundancy. The avoidance of redundancy can be enforced by exploiting local regularities already at the logical level. In fact, this is one of the key arguments for applying compilation techniques to Bayesian networks [12].

A couple of key questions have not yet been addressed in this paper. As corresponding implementations and testbeds are currently under development, we are not yet ready to say much about the empirical performance of the proposed method compared to existing methods. Other open questions concern the implementation of more sophisticated local search techniques such as *stochastic hill-climbing*, *simulated annealing*, or *genetic algorithms* [41]. These problems will be attacked in our subsequent work.

Acknowledgements

This research supported by the *Swiss National Science Foundation*, Project No. PP002-102652/1, and *The Leverhulme Trust*. Thanks to Michael Wachter for helpful discussions at the origin of this paper.

References

[1] A. Antonucci and M. Zaffalon. Locally specified credal networks. In *PGM'06, 3rd European Workshop on Probabilistic Graphical Models*, pages 25–34, Prague, Czech Republic, 2006.

[2] A. Antonucci, M. Zaffalon, J. S. Ide, and F. G. Cozman. Binarization algorithms for approximate updating in credal nets. In *STAIRS'06, 3rd European Starting AI Researcher Symposium*, pages 120–131, Riva del Garda, Italy, 2006.

[3] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *UAI'96, 12th Conference on Uncertainty in Artificial Inteligence*, pages 115–123, Portland, USA, 1996.

[4] A. Cano, J. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing on a tree of cliques. *IPMU'94, 5th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, LNCS 945, pages 978–983, Paris, France, 1994. Springer.

[5] A. Cano, J. M. Fernández-Luna, and S. Moral. Computing probability intervals with simulated annealing and probability trees. *Journal of Applied Non-Classical Logics*, 12(2):151–171, 2002.

[6] A. Cano, M. Gómez, S. Moral, and J. Abellán. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *International Journal of Approximate Reasoning*, 44(3):261–280, 2007.

[7] A. Cano and S. Moral. A genetic algorithm to approximate convex sets of probabilities. In *IPMU'96, 6th international Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 859–864, Granada, Spain, 1996.

[8] A. Cano and S. Moral. A review of propagation algorithms for imprecise probabilities. *ISIPTA'99, 1st Internation Symposium on Imprecise Probabilities and Their Applications*, pages 51–60, Ghent, Belgium, 1999.

[9] A. Cano and S. Moral. Using probability trees to compute marginals with imprecise probabilities. *International Journal of Approximate Reasoning*, 29(1):1–46, 2002.

[10] J. E. Cano, S. Moral, and J. F. Verdegay-López. Propagation of convex sets of probabilities in directed acyclic networks. *Uncertainty in Intelligent Systems*, pages 15–26. North-Holland, 1993.

[11] E. Charniak. Bayesian networks without tears. *AI Magazine*, 12(4):50–63, 1991.

[12] M. Chavira and A. Darwiche. Compiling Bayesian networks with local structure. In *IJCAI'05, 19th International Joint Conference on Artificial Intelligence*, Edinburgh, U.K., 2005.

[13] M. Chavira and A. Darwiche. Compiling Bayesian networks using variable elimination. In *IJCAI'07, 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007.

[14] M. Chavira, A. Darwiche, and M. Jaeger. Compiling relational Bayesian networks for exact inference. *International Journal of Approximate Reasoning*, 42(1–2):4–20, 2006.

[15] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk, Decision and Policy*, 5(2):165–181, 2000.

[16] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

[17] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2–3):167–184, 2005.

[18] F. G. Cozman and C. P. de Campos. Local computation in credal networks. In *ECAI'04, 16th European Conference on Artificial Intelligence, Workshop 22 on "Local Computation for Logics and Uncertainty"*, pages 5–11, Valencia, Spain, 2004.

[19] J. C. F. da Rocha and F. G. Cozman. Inference in credal networks with branch-and-bound algorithms. *ISIPTA'03, 3rd International Symposium on Imprecise Probabilities and Their Applications*, pages 480–493, Lugano, Switzerland, 2003.

[20] J. C. F. da Rocha, F. G. Cozman, and C. P. de Campos. Inference in polytrees with sets of probabilities. *UAI'03, 19th Conference on Uncertainty in Artificial Intelligence*, pages 217–224, Acapulco, Mexico, 2003.

[21] A. Darwiche. A differential approach to inference in Bayesian networks. *UAI'00, 16th Conference on Uncertainty in Artificial Intelligence*, pages 123–132, Stanford, USA, 2000.

[22] A. Darwiche. A compiler for deterministic, decomposable negation normal form. In *AAAI'02, 18th National Conference on Artificial Intelligence*, pages 627–634, Edmonton, Canada, 2002.

[23] A. Darwiche. A logical approach to factoring belief networks. *KR'02, 8th International Conference on Principles and Knowledge Representation and Reasoning*, pages 409–420, Toulouse, France, 2002.

[24] A. Darwiche. New advances in compiling CNF to decomposable negational normal form. In *ECAI'04, 16th European Conference on Artificial Intelligence*, Valencia, Spain, 2004.

[25] A. Darwiche and P. Marquis. A knowlege compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002.

[26] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, 2nd edition, 2000.

[27] C. P. de Campos and F. G. Cozman. The inferential complexity of Bayesian and credal networks. *IJCAI'05, 19th International Joint Conference on Artificial Intelligence*, pages 1313–1318, Edinburgh, U.K., 2005.

[28] R. Dechter. Bucket elimination: a unifying framework for reasoning. *Artificial Intelligence*, 113(1–2):41–85, 1999.

[29] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

[30] E. Fagiuoli and M. Zaffalon. 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[31] J. S. Ide and F. G. Cozman. IPE and L2U: Approximate algorithms for credal networks. In *STAIRS'04, 2nd European Starting AI Researcher Symposium*, pages 118–127, Valencia, Spain, 2004.

[32] J. S. Ide and F. G. Cozman. Approximate inference in credal networks by variational mean field methods. *ISIPTA'05, 4th International Symposium on Imprecise Probabilities and Their Applications*, pages 203–212, Pittsburgh, USA, 2005.

[33] J. Kohlas and P. P. Shenoy. Computation in valuation algebras. *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 5: Algorithms for Uncertainty and Defeasible Reasoning, pages 5–39. Kluwer Academic Publishers, Dordrecht, Netherlands, 2000.

[34] J. Kohlas and N. Wilson. Exact and approximate local computation in semiring induced valuation algebras. Technical Report 06–06, University of Fribourg, Switzerland, 2006.

[35] H. E. Kyburg. Interval-valued probabilities. *The Imprecise Probabilities Project*. IPP Home Page, available at http://ippserv.rug.ac.be, 1998.

[36] I. Levi. *The Enterprise of Knowledge*. The MIT Press, Cambridge, USA, 1980.

[37] J. Pearl. On probability intervals. *International Journal of Approximate Reasoning*, 2(3):211–216, 1988.

[38] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, USA, 1988.

[39] M. Pouly. Implementation of a generic architecture for local computation. *ECAI'04, 16th European Conference on Artificial Intelligence, Workshop 22 on "Local Computation for Logics and Uncertainty"*, pages 31–37, Valencia, Spain, 2004.

[40] M. Pouly. NENOK 1.1 user guide. Technical Report 06–02, Department of Informatics, University of Fribourg, Switzerland, 2006.

[41] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2003.

[42] T. Sang, P. Beame, and H. Kautz. Solving Bayesian networks by weighted model counting. In *AAAI'05, 20th National Conference on Artificial Intelligence*, volume 1, pages 475–482, Pittsburgh, USA, 2005.

[43] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. *UAI'88, 4th Conference on Uncertainty in Artificial Intelligence*, pages 169–198, Minneapolis, USA, 1988.

[44] B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7:95–120, 1992.

[45] M. Wachter and R. Haenni. Logical compilation of Bayesian networks. Technical Report iam-06-006, University of Bern, Switzerland, 2006.

[46] M. Wachter and R. Haenni. Propositional DAGs: a new graph-based language for representing Boolean functions. *KR'06, 10th International Conference on Principles of Knowledge Representation and Reasoning*, pages 277–285, Lake District, U.K., 2006. AAAI Press.

[47] M. Wachter and R. Haenni. Multi-state directed acyclic graphs. *CanAI'07, 20th Canadian Conference on Artificial Intelligence*, LNAI 4509, pages 464–475, Montréal, Canada, 2007.

[48] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2–3):149–170, 2000.

Quantile-Filtered Bayesian Learning for the Correlation Class

Hermann Held
Potsdam Institut for Climate Impact Research (PIK), 14412 Potsdam, Germany
hermann.held@pik-potsdam.de

Abstract

We introduce a new rule for Bayesian updating of classes of precise priors. The rule combines Walley's generalized Bayes rule with a filter based on prior quantiles of the observational evidence. We introduce this new "quantile-filtered Bayesian update rule" because in many situations, Walley's generalized Bayes rule reveals counter-intuitively non-informative, dilation-type results while an alternative rule, the maximum likelihood update rule after Gilboa and Schmeidler, is not robust against imprecise priors that are contaminated with spurious information. Our new quantile-based update rule addresses the former issue and fully resolves the latter. By the new rule we update an imprecise prior that was recently motivated by expert interviews with climate, ecosystem and economic modelers: a "correlation class" of precise priors with arbitrary correlation structure, however, prescribed precise marginals. For an insurance situation we demonstrate that under our new rule a set of clients would be insured that is disregarded under standard generalized Bayesian updating.

Keywords. Bayesian updating, Generalized Bayes rule, imprecise probability, robust Bayesian approach, modeling expert opinions, prescribed marginals, unknown correlation structure.

1 Introduction

Complex numerical models provide key working horses within climate, ecosystem and economic research and hence their output strongly influences the discussion on ecologically and economically sustainable climate policies. In turn, model output strongly depends on various tuning parameters which cannot fully be determined through objective data in general. For that reason Bayesian methods become increasingly popular in these fields as they would allow to incorporate subjective prior knowledge on model parameters, often aggregated from scattered sources of information in the brains of modelers, in a statistical analysis. A recent semi-formalized expert elicitation aimed at generic patterns of knowledge vs. ignorance in modelers' prior information on multivariate model parameters [8, 9]. As a key result modelers across disciplines stated to hold fundamentally more precise information on marginals than on the (higher order) correlation structure among parameters.

This key finding from above elicitation fueled our interest in Tchen's imprecise model [14] (further investigated in [2, 5, 6, 11, 12]) consisting of a class $\mathcal{P}$ of precise measures P the marginals of which would all equal certain prescribed marginals. We call this class "correlation class". When updating a correlation class along the lines of global Bayesian robustness [1], i.e. element-wise updating according to standard Bayes rule, then observing the extremes of ensuing answers as the prior varies over the class, we found non-informative imprecise posteriors over a wide range of potential observations y [8, 9].

This is in line with prior results on a similar $\mathcal{P}$ by [11] (see also Seidenfeld and Wasserman [13] for a discussion of such a dilation phenomenon where posterior bounds are dilated even for all possible measurements y). In case the set $\mathcal{P}$ is convex this updating procedure is equivalent to Walley's generalized Bayes rule [15]. We will call this element-wise updating and subsequent extremizing "GBR" throughout this article regardless of whether $\mathcal{P}$ is convex or not. (An alternative class displaying imprecise correlations is introduced in [10] characterized by a radially symmetric possibility measure. However as no results on Bayesian updating have been published for that class so far, we disregard it in the context of this article.)

Gilboa's and Schmeidler's maximum likelihood update rule [7] delivers much more informative results. Their rule is equivalent to applying GBR – not to $\mathcal{P}$ but – to the subset of those precise priors that would maximize the prior expectation of the evidence y. In [8, 9] that rule is generalized by not completely dis-

regarding those priors that would not maximize prior expectation of y but by giving any element of $\mathcal{P}$ an influence, weighted by its prior expectation of y. (y may either represent a single sample or a number of samples that can be combined to the multi-variate observation y.) However as against GBR, both likelihood update rules face the problem that spurious information may enter the final result: in case $\mathcal{P}$ contains an unjustified element that accidentally displays high prior expectation of y, this may result in a posterior that is more precise than for the uncontaminated version of $\mathcal{P}$.

For that reason here we present a new updating rule that combines important advantages of GBR and the latter two likelihood updating rules: (i) it is more informative than GBR and (ii) in case $\mathcal{P}$ is contaminated this contamination would not add spurious information to the posterior.

We are aware that there exists the further method of reducing the class of priors in view of evidence as described in [3, 4]. However the relation to our work appears intricate and its elucidation shall be outlined elsewhere.

This article is organized as follows. In Section 2 we introduce the new updating rule. In Section 3 we apply that rule to the briefly recapitulated imprecise prior in [8, 9] motivated by above expert elicitation. In Section 4 we regularize our prior by bounding the gradients of densities making up the imprecise prior. In Section 5 we offer an interpretation of our new updating rule that involves also concepts from classical statistics and therefore might be controversial. In Section 6 we compare the results of various updating methods from the point of view of an idealized insurance company. Finally, in Section 7 we summarize our findings and outline the most pressing issues from the point of view of a modeler.

2 The Quantile-Filtered Bayesian Update Rule

The crucial element of our new updating rule is the filter that acts on $\mathcal{P}$, before GBR is applied.

Definition 1 *Let $\mathcal{P}$ represent an imprecise prior made up by a non-empty set of precise priors. Let $Q \in]0,1[$. Let P_L denote the probability measure induced by a precise prior $P' \in \mathcal{P}$ and the precise likelihood L on the space of all potential observations Y. Then $\mathcal{Y}_{\mathcal{P}LYQ}$ is a* **generator of a Q-filtered Bayesian update rule (QFB)** *iff $\mathcal{Y}_{\mathcal{P}LYQ} : \mathcal{P} \to 2^Y$ with $\forall_{y\in Y}\forall_{P'\in\mathcal{P}} \;\; P_L(y \in \mathcal{Y}_{\mathcal{P}LYQ}(P')) \geq Q$.*

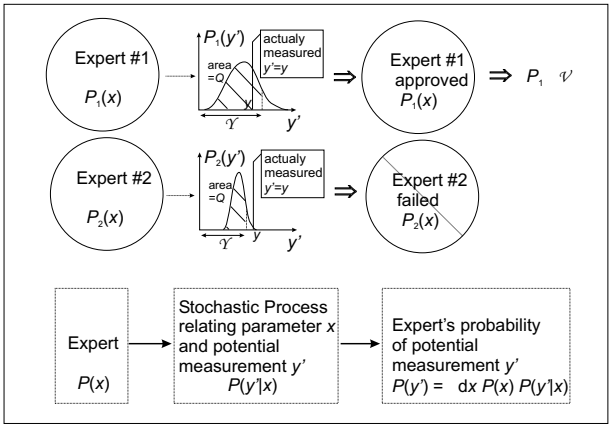


Figure 1: Scheme for the construction of the subset $\mathcal{V}$ in the class of priors. Any prior (here identified with a different "expert") induces – through a given likelihood – a probability measure on the space of potential measurements y' (bottom). Once the measurement has been realized, i.e. $y' := y$, one can disregard priors that display y outside of a quantile, characterized by a pre-set probability Q.

Hence $\mathcal{Y}$ maps $P' \in \mathcal{P}$ onto a prior ($\geq Q$)-quantile in observation space. As an illustrative example, in Figure 1, the two elements of $\mathcal{P}$, P_1, P_2, are mapped onto an interval $]-\infty, y'_{\max}]$ within the respective abscissa (the latter denoting the space of potential observations y').

Definition 2 *Let $\mathcal{P}, Q, L, P_L, Y$ as above and $\mathcal{Y}_{\mathcal{P}LYQ}$ the accompanying generator of a Q-filtered Bayesian update rule. Then $\mathcal{V}_{\mathcal{P}LYQ}$ is a* **Q-GBR-filter** *iff $\mathcal{V}_{\mathcal{P}LYQ} : Y \to 2^{\mathcal{P}}, \quad y \mapsto \{P' \in \mathcal{P} \mid y \in \mathcal{Y}_{\mathcal{P}LYQ}(P')\}$.*

Hence for given observation y, $\mathcal{V}(y)$ represents those priors for which y is not too "far-fetched" (see Figure 1).

Definition 3 *Let $\mathcal{V}$ be according to previous Defs. Then we call the operation GBR $\circ$ $\mathcal{V}$ a* **quantile filtered Bayesian learning rule** *(QFB).*

Before we discuss a desirable property of QFB w.r.t. contaminations, we would like to recall that GBR shares this property:

Theorem 1 *Let $\overline{\mathcal{U}}_{\mathrm{GBR}} : Y \otimes \mathcal{P} \to \mathbb{R}$ the "updating operator" maximizing the ensuing answer of Bayesian learning over the class of priors along GBR, and $\underline{\mathcal{U}}_{\mathrm{GBR}}$ the analogue minimization operator. Then $\forall_{y\in Y} \;\; \underline{\mathcal{U}}_{\mathrm{GBR}}(y, \mathcal{P} \cup \mathcal{P}_c) \leq \underline{\mathcal{U}}_{\mathrm{GBR}}(y, \mathcal{P}) \leq \overline{\mathcal{U}}_{\mathrm{GBR}}(y, \mathcal{P}) \leq \overline{\mathcal{U}}_{\mathrm{GBR}}(y, \mathcal{P} \cup \mathcal{P}_c)$.*

This relation simply follows from the fact that the sup(inf)-operator is monotonous w.r.t. set-extension.

It implies that a contamination $\mathcal{P}_c$ would not add spurious information to the posterior result. In general, such a relation is violated by the two likelihood updating rules mentioned before, but importantly it holds for QFB:

Theorem 2 *Let $\overline{\mathcal{U}}_{\mathrm{QFB}} : Y \otimes \mathcal{P} \to \mathbb{R}$ the "updating operator" maximizing the ensuing answer of Bayesian learning over the class of priors along QFB, and $\underline{\mathcal{U}}_{\mathrm{QFB}}$ the analogue minimization operator. Then $\forall_{y\in Y} \quad \underline{\mathcal{U}}_{\mathrm{QFB}}(y, \mathcal{P} \cup \mathcal{P}_c) \leq \underline{\mathcal{U}}_{\mathrm{QFB}}(y, \mathcal{P}) \leq \overline{\mathcal{U}}_{\mathrm{QFB}}(y, \mathcal{P}) \leq \overline{\mathcal{U}}_{\mathrm{QFB}}(y, \mathcal{P} \cup \mathcal{P}_c)$.*

This Theorem readily follows from the fact that the way the operator GBR $\circ\ \mathcal{V}$ acts on $P' \in \mathcal{P}$ does *not* depend on the other elements of $\mathcal{P}$. This is in contrast to the other two likelihood update rules for which the relative weight (the prior expectation of y) of P', compared to the other priors matters. We regard the fact that those Theorems hold as a key advantage of QFB and GBR. It now remains to show that QFB is significantly more informative than GBR in relevant cases.

3 Specification and updating of the correlation class

3.1 The imprecise prior and the likelihood

In order to keep the discussion as transparent as possible we decide on the simplest non-trivial $\mathcal{P}$ and likelihood possible. We consider the uncertain parameter $(x_1, x_2)^t \in \mathbb{R}^2$, the "observation" or "evidence" $y \in \mathbb{R}$. Furthermore for any element of $\mathcal{P}$, any of its two marginals should equal $N(\mu, \sigma^2)$, a Gaussian with mean μ and σ^2 variance[1]. A likelihood employed shall write $L(x_1, x_2) \equiv P(y|x_1, x_2) := N(\ x_1 + x_2, \sigma_\eta^2)(y)$, known to the modeler. From now on whenever results are not displayed in analytic form, we choose the specific parameter values $\mu = 1/2, \sigma = 1/4, \ := 1.05$ (as $\ = 1$ would lead to a degenerate and $|\ | \gg 1$ to a trivial case [8, 9]), $\sigma_\eta := \sigma/10$.

So far we have specified only marginals, hence we do not rule out multi-modal densities. However we find the subset of unimodal prior densities more convincing a model for generic prior expert knowledge. This is conveniently implemented by requiring that any prior shall be a 2D Gaussian, although admittedly hereby we potentially disregard too many priors. For pragmatic reasons, however, we stick to this computationally convenient case for the remainder of the article. It is shown in [8, 9] that then

[1] For a multivariate application, the first entry would represent a vector of means, the second the symmetric covariance matrix.

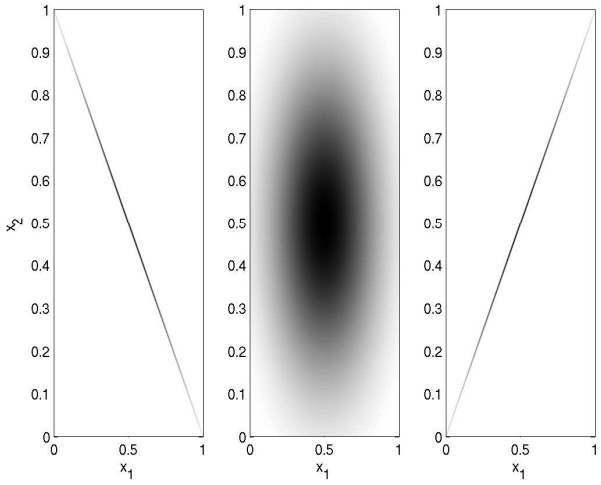


Figure 2: Three extreme representatives of the class of Gaussian priors with prescribed marginals. From left to right: maximally anticorrelated case ($f = -1$), uncorrelated case ($f = 0$), and maximally correlated case ($f = 1$) – for a definition of the parameter f see Eq. 1).

$$\mathcal{P} = \{P \mid \exists_{f\in[-1,1]}\ P\ \sim N((\mu,\mu)^t, \Sigma(f))\} \text{ with } \quad (1)$$
$$\forall_{f\in[-1,1]} \quad \Sigma(f) := \sigma^2 \begin{pmatrix} 1 & f \\ f & 1 \end{pmatrix}.$$

$f = 0$ represents standard Bayesian updating with an uncorrelated prior, $f = 1$ ($f = -1$) the fully (anti)correlated prior. Accompanying densities are displayed in Figure 2.

Finally we select the functional we are interested in – the *probability of ruin*:

Definition 4 *Let $P \in \mathcal{P}$. Let $x_1^* \in \mathbb{R}$. Then we define the probability of ruin as*

$$P^* := \int_{x_1^*}^{\infty} dx_1 \int_{-\infty}^{+\infty} dx_2\ P(x_1, x_2).$$

In the context of climate modeling, x_1^* could represent a well-known critical value of global mean temperature beyond which "catastrophic" global warming impacts may occur, and x_1, x_2 two uncertain climate model parameters.

3.2 Bayesian learning

In order to generate the posterior probability of ruin per precise prior, the posterior marginal for x_1 is key. In [8, 9] it is shown that

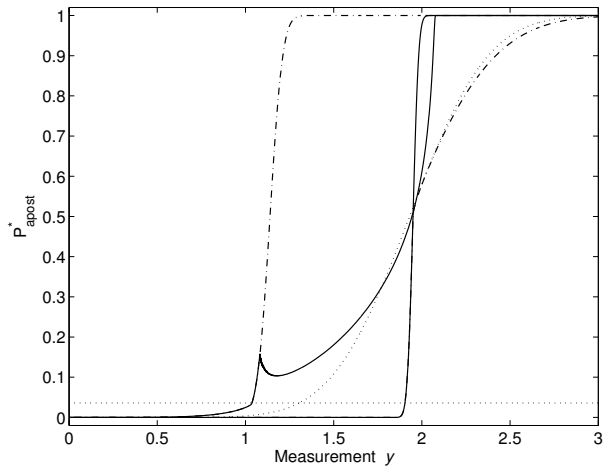


Figure 3: Probability of ruin (upper and lower) for the correlation class parameterized by the correlation coefficient $f \in [-1,1]$ after Eq. 1 for $\quad = 1.05, x_1^* = 0.95$, $\sigma_\eta = \sigma/10$. Horizontal dotted line: apriori value, curved dotted: (standard) uncorrelated case, dashed-dotted: GBR, solid: QFB for $Q = 98\%$, the lower probabilities of ruin for GBR and QFB coalescing. GBR reveals quasi non-informative posterior results for $y \in [1.3, 1.8]$. Quite the contrary the new QFB is informative for any $y \in \mathbb{R}$. For an expanded representation of the "avoided crossing" region around (2,1/2), see the following Figure.

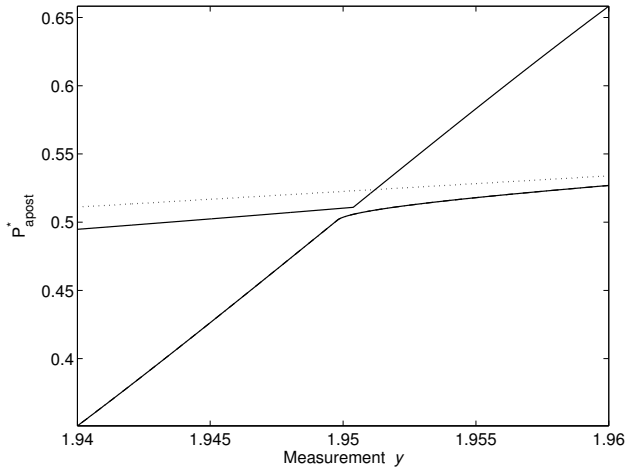


Figure 4: Expansion of the previous Figure's center that shows an "avoided crossing" of the solid lines. We explain this feature of almost precise posterior probability in Subsection 3.3 by an approximate symmetry in the transfer function $(x_1, x_2) \to y$ in combination with Gaussian symmetry.

$$\begin{aligned} P_{\text{post}}(x_1|y) &\sim N(\mu'(f,y), \sigma'^2(f,y)) \quad \text{with} \qquad (2)\\ \mu'(f,y) &= (\mu(1-(1-f)(\ \ -1)\ \sigma^2/\sigma_\eta^2)\\ &\quad +(f+\ \ )\ y\ \sigma^2/\sigma_\eta^2)\\ &\quad /(1+(1+2f\ \ +\ \ ^2)\ \sigma^2/\sigma_\eta^2),\\ \sigma'(f) &= \sigma\ \sqrt{\frac{1+(1-f^2)\ \sigma^2/\sigma_\eta^2}{1+(1+2f\ \ +\ \ ^2)\ \sigma^2/\sigma_\eta^2}}. \end{aligned}$$

We utilize this expression to calculate the posterior probability of ruin

$$P^*_{\text{apost}}(f,y) = \int_{x_1^*}^{\infty} N\left(\mu'(f,y), (\sigma'(f))^2\right)(x_1)\ dx_1. \quad (3)$$

From this we obtain the upper probability of ruin in the case of GBR by

$$\overline{P}^*_{\text{apost.GBR}}(y) = \sup_{f\in[-1,1]} \quad P^*_{\text{apost}}(f,y). \qquad (4)$$

For QFB we need to define generator of a Q-filtered Bayesian update rule $\mathcal{Y}$. As larger y will imply higher probabilities of ruin in general, we expect that the following prescription will lead to informative posteriors $(0 < Q < 1)$:

$$\begin{aligned} \forall_{f\in[-1,1[} \quad \mathcal{Y}(f) &:= \ ]-\infty, y_{\max}(f)] \quad \text{with} \qquad (5)\\ Q &=: \int_{-\infty}^{y_{\max}(f)} dy\ P_{y;f.\text{prior}}(y) \quad \text{and}\\ \mathcal{Y}(f=1) &:= \ ]-\infty, \infty[\qquad (6) \end{aligned}$$

Let $\mathcal{V}$ be the filter generated by $\mathcal{Y}$. Then latter equation ensures that for all y, $\mathcal{V}(y) \neq \emptyset$ (for a more extended discussion the reader is put off to the more "philosophical" Subsection 5.3 – here we just point to Theorem 1 which ensures that no spurious information is added when making a class of priors, subject to GBR, larger). In order to operationalize Eq. 5 we need $P_{y;f.\text{prior}}$. In [8, 9] we show

$$P_{y;f.\text{prior}} \sim N(\mu(1+\ \ ), \sigma^2(1+2\ \ f+\ \ ^2)+\sigma_\eta^2). \quad (7)$$

Then

$$\overline{P}^*_{\text{apost.QFB}}(y) = \sup_{f\in f(\mathcal{V}(y))} \quad P^*_{\text{apost}}(f,y) \qquad (8)$$

if $f(\mathcal{V}(y))$ denotes the set of f-values needed to parameterize $\mathcal{V}(y)$. (For $\underline{P}_{\text{apost}}$, "sup" is to be replaced by "inf" in above equations.)

We do not claim that our choice of $\mathcal{Y}$ generates the most informative QFB. Here we just would like to demonstrate that even a rather unsophisticated choice leads to much more informative results than GBR does.

The dependency of the probability of ruin on y is depicted in Figure 3 for GBR (dashed-dotted curves for upper and lower probability of ruin), the new QFB (solid curves) under a choice of $Q = 98\%$, for comparison also the assumption of independent parameters (uncorrelated case $f = 0$).

We observe that in general QFB is much more informative than GBR – i.e. the difference of upper and lower probability of ruin is smaller for QFB than for GBR. A bizarre feature can be observed for QFB however: the upper probability of ruin is not a monotonous function of y, a feature occurring in an even more pronounced way for the maximum likelihood update rule [8, 9]. There we attribute this to a certain degenerate feature within $\mathcal{P}$, related to $f = -1$ and becoming virulent at $y = (\ -1)x_1^* + 2\mu \approx 1.05$. We propose that such effects would vanish if a non-parametric class of priors were considered. A skeptic of our new method may now argue that also the superiority of QFB over GBR as displayed in Figure 3 may be a result of degenerate priors and would vanish under more regular imprecise priors. In the following Section we show that this is not the case, but QFB is robustly more informative even when we "regularize" $\mathcal{P}$. Before that, however, we would like to interpret the striking convergence of QFB-upper and lower probability of ruin to the value $1/2$ as displayed in Figure 3.

3.3 An "avoided crossing" for QFB

The fact that we use a Gaussian class of priors leads to a series of peculiar phenomena of which the "avoided crossing" of upper vs. lower solid curve at $\sim (2, 1/2)$ in Figure 3 may be of special interest. For readers that would like to focus more on the general statements of this article we suggest that they skip this Subsection and proceed directly with Section 4.

The key reason for the almost precise QFB posterior at $y \approx 2$ is easiest accessed in considering the following double limit of Eq. 3 on the posterior mean

$$\forall_{f\in]-1,1]}\ \forall_{y\in\mathbb{R}} \quad \lim_{\to 1}\lim_{\sigma_\eta\to 0}\mu'(f,y;\ \ ,\sigma_\eta) = \frac{y}{2}, \quad (9)$$

i.e. for the whole class, the posteriors will be centered at $y/2$ (with differing variances).

This implies that

$$\left\{\frac{y}{2} = x_1^*\right\} \Rightarrow \left\{\forall_{f\in]-1,1]}\ \lim_{\to 1}\lim_{\sigma_\eta\to 0} P^*_{\text{apost}}(f,y) = \frac{1}{2}\right\}. \quad (10)$$

As $f = -1$ is not element of the volume of confidence at $y/2 = x_1^*$, from this Eq. we conclude a precise posterior at that $y \approx 2$.

We now investigate how this exact prosterior dilutes into an avoided crossing for $\ = 1.05, \sigma_\eta = \sigma/10 = 1/40$. For this, it is important to note that Eq. 3 can be rewritten as

$$P^*_{\text{apost}}(f,y) = \int_{-\infty}^{y} dy'\ N\left(\frac{x_1^* - \mu_0(f)}{\mu_1(f)}, \frac{\sigma'(f)}{\mu_1(f)}\right)(y'), \quad (11)$$

whereby the two new functions $\mu_0(f) + y\mu_1(f) := \mu'(f,y)$ are determined by the (in y) linear relation Eq. 3. From Eq. 11 we learn that for any f, $P^*_{\text{apost}}(f,y)$ is an error function in y. Now we deduce the analytic form of the lower solid line before the crossing. After verifying $\partial\mu'/\partial f < 0$ (for $y > (1 + \ )\mu$) and $d\sigma'/df < 0$, we conclude that for given y, $P^*_{\text{apost}}(f,y)$ decreases with f. Hence the QFB lower bound is generated by the single posterior $P^*_{\text{apost}}(f = 1, y)$ for $y < y_c$. We define y_c as the "crossing value" $P^*_{\text{apost}}(f = 1, y_c) := 1/2 \Rightarrow y_c \approx 1.9497$, also compare to Figure 4.

While the lower QFB bound before the crossing is made up by a single f (i.e. a single prior) in terms of one single error function, the QFB upper bound is the envelope of error functions generated according to Eq. 11 from different f's. This is related to the fact that the upper bound per y is generated from the lower bound f_- of the interval of confidence $[f_-(y), 1]$ and $df_-/dy > 0$. However, locally in y, the upper bound can be related to one single f. We find numerically $f_-(y_c) \approx 1/2$ (in accordance with Figure 4, QFB excludes the uncorrelated case (dotted line $\Leftrightarrow f = 0 \notin [f_-(y_c), 1]$)). We can now address the following question: what parameters determine the width of the avoided crossing

$$P^*_{\text{apost.QFB.ac}} := P^*_{\text{apost}}(f_-(y_c), y_c) - \frac{1}{2}. \quad (12)$$

Let $\ f := 1 - f_-(y_c)$, i.e. the difference of the QFB upper bound f to the prior's f that generates the QFB lower bound. Utilizing Eq. 11 we then derive in first order perturbation theory

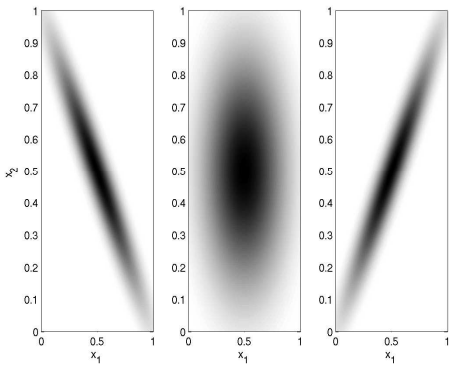


Figure 5: Extreme cases of priors after bounding the gradient. Left: $f = -f^*$, center: $f = 0$, right: $f = f^*$, $f^* \approx 0.95434$. The bound f^* was chosen such that the expert "can resolve not more than 5 items per typical marginal parameter scale" (in our case $[0, 1]$, spanning 4σ – for details see [8, 9]). For that reason, the densities displayed in the left and the right panel are smoother than their counterparts in Figure 2.

$$\lim_{\sigma_\eta \to 0} \quad P^*_{\text{apost.QFB.ac}} \approx \frac{1}{4\sqrt{\ }} (\ \ -1) \frac{x_1^* - \mu}{\sigma} \sqrt{1 - \quad f}. \tag{13}$$

Inserting the values of our example, we obtain $P^*_{\text{apost.QFB.ac}} \approx 0.01$, in accordance with the distance within the crossing displayed in Figure 4. The last equation also reveals that the avoided crossing becomes an exact crossing if x_1, x_2 influence y symmetrically, i.e. $\quad \to 1$, in accordance with Eq. 10.

4 Introducing a gradient filter

Following Walley [16] we regard it as meaningful to bound the gradient of densities within a class of priors. It is very questionable that in general an expert will hold such a sophisticated prior knowledge that bizarre density structures of arbitrary gradient could be distinguished in her or his brain. For our class this would imply to disregard priors with too large $|f|$.

Working with such a "regularized" class of priors comes with the additional advantage that effects like those at $y = (\ \ -1)x_1^* + 2\mu \approx 1.05$ may vanish as our class becomes more similar to a non-parametric, however, gradient-bounded class which the "imprecise community" may find more adequate for generic expert knowledge in the future.

The question now is how to restrict $|f|$. Following [8, 9] we argue that in general, an expert will not be able to distinguish more than 5 "major blocks" per parameter dimension. This idea is formalized in [8, 9]

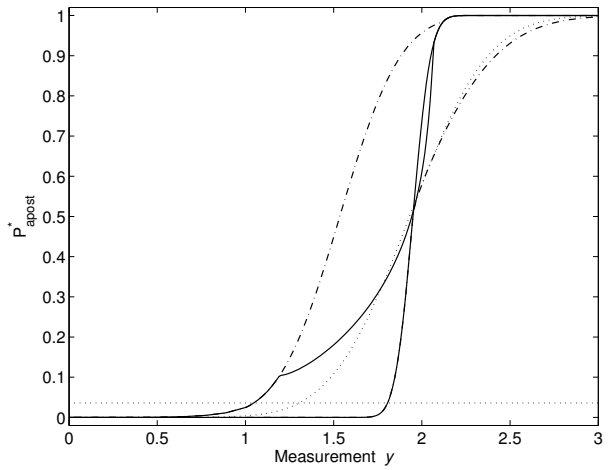


Figure 6: Upper and lower probabilties of ruin as in Figure 3, yet for bounded gradients of prior densities (dashed-dotted lines: GBR, solid lines: QFB, dotted curved line: updated uncorrelated precise prior, horizontal). Note that even for this class of priors "regularized" by gradient bounding (equivalent to $|f| \leq f^* \approx 0.95434$), QFB is more informative than GBR.

and leads to the prescription $|f| \leq 0.95434$. Figure 5 then represents the bounded-gradient counterpart of Figure 2. In the long run, this issue must ultimately be addressed by suitable expert elicitations and social experiments that would reveal the expert's "prior resolution."

In fact Figure 6 reveals that even after bounding the density gradient over $\mathcal{P}$ QFB stays qualitatively more informative than GBR. In addition, for this "regularized imprecise prior" now also QFB responds monotonously w.r.t. observation y what is more in line with intuition.

Hence QFB seems to combine both desirable features discussed in the Introduction: it is informative and it does not absorb spurious information (Theorem 2). For that reason we regard it as worthwhile to look for an interpretation of QFB. (The reader may start using QFB for pragmatic reasons even if she or he does not want to follow the assumptions in the interpretation given below.)

5 Interpretation and nesting of quantile-filtered Bayesian learning

5.1 Interpretation of QFB

We present one possible interpretation QFB that is based on the following two assumption:

(1) Any prior class of precise measures specified by an

expert contains "the adequate, yet un-identified" precise measure for that actual assessment;
(2) when considering the sequence of the expert's assessments over her or his life and transforming each "adequate precise prior" to a uniform prior by a suitable coordinate transformation, then the sequence of accordingly transformed "true states of the world" (the sequence of true parameter values) would behave as drawn from a uniform distribution.

Assumption 1 reminds of a situation in which a king needs to listen to a series of agents, knowing that only one agent has really been sent by the king's friend whereas the others are from "false friends".

Assumption 2 shall be illustrated by a special case first: suppose an expert performed a series of assessments $a_1, ..., a_n, ..., a_N$, whereby at each assessment a_n she or he would be asked for the probability of whether a certain "true" state of the world s_n belonged to a certain set S_n. We denote this probability as $P(s_n \in S_n)$ and we assume further that for any n, the expert would claim $P(s_n \in S_n) = p$. We now imagine that some time will have passed by and in the corse of history the true nature of $s_1, ..., s_N$ will have become public, i.e. the expert's customers will then be able to objectively determine the index function $\mathrm{ind}(s_n \in S_n)$ – that is 1 in case the statement is true and 0 otherwise. Then Assumption 2 requires that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathrm{ind}(s_n \in S_n) \to p \qquad (14)$$

in the sense of the law of large numbers. Hence we require that in a frequentistic sense the expert will have been neither over- nor under-confident, i.e. the life-averaged assessment prior p was "adequate."

Therefore in more general terms, Assumption 2 implies that in a world in which an expert would specify prior knowledge always as uniform distribution on [0,1] per assessment, later generations would find the histogram of true states of the world, the expert had assessed, converge to that uniform distribution over the life-span of the expert.

That way, we choose an interpretation of subjective probability that allows us to treat it not only as epistemic uncertainty, but also as aleatoric uncertainty, i.e., as a stochastic process that governs the relation of the expert to reality during her or his life. Those users that could accept such an interpretation of experts' knowledge have the chance to interpret the combination of "choose the parameter" and "predict, given that parameter, the measurement y" as a joint stochastic process. If the former is described by $P(x)$ and the latter by $P(y|x)$, then, given the expert's P:

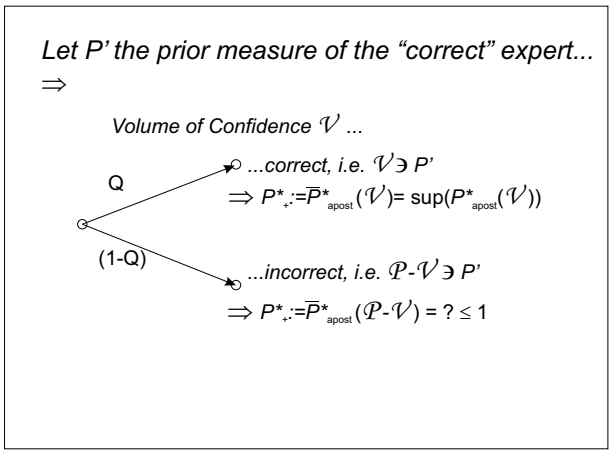


Figure 7: Nesting the classical volume of confidence $\mathcal{V}$ in a decision situation. In our frequentist's interpretation we can explicitly take care of the possibility that $\mathcal{V}$ may *not* contain the adequate prior. For that we utilize a probability tree, resulting in Eqs. 15 and 16.

$P(y) = \int dx\ P(x)\ P(y|x)$.

As in our interpretation, for any prior, $P(y)$ is generated by a stochastic process, it must be possible to evaluate the elements within the set of priors on the basis of the measurement utilizing frequentistic statistics. In particular we are interested in defining a classical volume of confidence within the set of priors as a filter, conditioned on y.

By construction $\mathcal{V}$ of Definition 2 is such a classical volume of confidence with the confidence value Q. Q can then be interpreted as follows: it represents a (lifetime averaged) lower bound for the relative frequency that an expert does include the "adequate" precise prior in $\mathcal{V} \subset \mathcal{P}$ in case for any inference situation such an "adequate prior" exists.

Given this interpretation we can make use of the fact that upper (lower) posterior probabilities of the event we are interested in are bounded functionals over $\mathcal{P}$ and ask whether we can somehow also account for those cases in which $\mathcal{V}$ fails, i.e. does not contain the "adequate prior".

5.2 Proposing a nesting formula

One may now ask how a decision-maker may deal with the fact that the volume of confidence does not hold with certainty but only with probability Q. If $Q \approx 1$, in many applications of classical tests, this aspect is simply ignored and the volume of confidence is dealt with as if it were certain.

However, here we would like to suggest an exact ap-

proach that explicitly takes care of those cases for which the volume of confidence fails, appearing with probability $(1 - Q)$. We "nest" the classical uncertainty $(1 - Q)$ into the Bayesian scheme by a probability-tree argument (see Figure 7).

Let P^*_+ and P^*_- the upper and lower probabilities of ruin derived, after the quantile filter has been applied before GBR. In case 1, the classical volume was correct, and $\overline{P}^*_{\text{apost}} = P^*_+$, being true with probability Q. In case 2, the classical volume was wrong, and we set $\overline{P}^*_{\text{apost}} = 1$ as a conservative estimate of that quantity, with probability $(1 - Q)$. (Analogously we can proceed with the *lower* probability of ruin.)

According to the thereby induced tree diagram,

$$\underline{P}^*_{\text{apost.QFB.nest}} = Q \cdot P^*_- + (1-Q) \cdot 0, \quad (15)$$

$$\overline{P}^*_{\text{apost.QFB.nest}} = Q \cdot P^*_+ + (1-Q) \cdot 1. \quad (16)$$

In the following, we will call the upper and lower probabilities of ruin "nested".

In case one subscribed to the two assumptions given in the beginning of this Section, one could interpret $\underline{P}^*_{\text{apost.QFB.nest}}$ and $\overline{P}^*_{\text{apost.QFB.nest}}$ as upper and lower bounds for relative frequencies of "ruins" over a sequence of equivalent assessments, in the limit of large numbers (of assessments). To the best of our knowledge, this is the first time the incompleteness of interval estimates is addressed.

5.3 Treatment of an empty $\mathcal{V}(y)$

How to proceed if y is such an "outlier" that $\mathcal{V}(y) = \emptyset$? One could proceed in saying that no expert were available, hence there were no information on P^*_{apost}. However, that lack of posterior information is counterintuitive. If the quantile filter is used together with GBR, we know that adding a prior to the class does not result in spurious information. Hence if $\mathcal{V}(y) = \emptyset$ we could add a prior P_a from the original class that is most informative, e.g. the maximum likelihood prior. No spurious information is added by re-introducing P_a due to Theorem 1. This is exactly the argument that was used when setting up Eq. 6.

We would like to illustrate what updating of the imprecise prior with our new rule QFB may mean in a decisions situation. Hence, before presenting the implementation of above combinations of learning rules and filters, we now introduce a stylized potential user of our ideas.

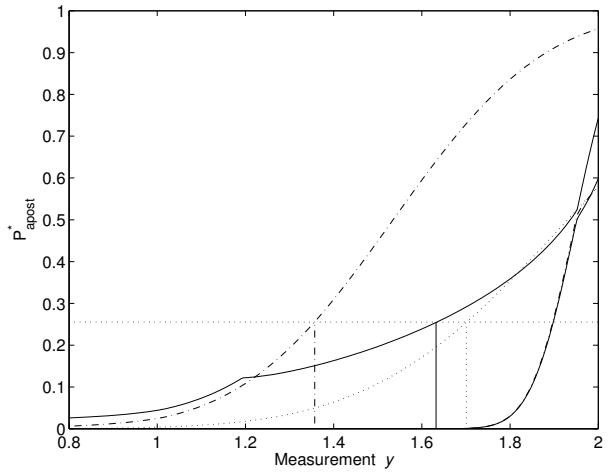


Figure 8: Inverting Figure 6: reading from a prescribed maximum probability of ruin (horizontal line) the accompanying maximum y^*, as needed for the stylized insurance problem. As against Figure 6, here we have involved the nesting correction for QFB, that amounts, however, only to an upwards shift of 0.02. We observe that according to QFB clients with characteristic $y \in [1.34, 1.66]$ could be insured in addition to GBR.

6 Various updating rules for a stylized insurance situation

Following [8, 9] we consider an admittedly rather stylized insurance company that plans to insure a fixed number of clients J each of which comes with a potential standard loss of 1, the behaviorally identical clients' willingness to pay (for a premium) of $2^{-1+1/\alpha} p^{1/\alpha}, \alpha := 3$, for the upper probability of ruin per client p. If the company asks for a residual upper probability for bankrupt, i.e. net loss, of 0.1%, then in a Gaussian approximation we obtain as upper probabilities of ruin allowed per client: 0.12927417 or 0.27004601 for $J = 30$ or $J = 100$ respectively.

With these numbers we enter the ordinate in Figure 6 and read the maximum characteristic y^* per client with which that client would still be insured. Within that Figure the concept of a maximum allowed y makes sense as all curves monotonously increase. In Figure 8 we further illustrate this inversion for the case of 30 clients, i.e. $\overline{P}^*_{\text{apost}} \approx 27\%$. The only difference is that for QFB we show the nesting-corrected results according to Eqs. 15 and 16 (for the upper probability of ruin, this amounts approximately to an addition of $1 - Q = 0.02$ that is almost negligible). Interestingly, clients with much higher y could be insured according to QFB than according to GBR.

We summarize threshold values y^* that denote the

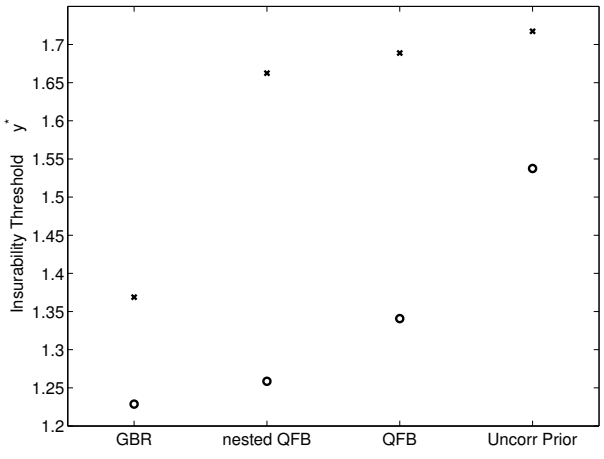


Figure 9: y^* as upper limit of y's with which clients would be insured: Circles: pooling with 30 clients; crosses: pooling with 100 clients. The abscissa indicates the four learning rules according to the tabular of this Section. (Any entry for $= 1.05, x_1^* = 0.95, \sigma_\eta = \sigma/10, Q = 98\%$.) According to QFB significantly more risky clients could be insured than for GBR.

maximum y with which a client would get insured in the following tabular:

	J updating rule	30	100
1	GBR	1.23	1.37
2	QFB after nesting	1.26	1.66
3	QFB before nesting	1.34	1.69
4	uncorrelated prior	1.54	1.72

As expected, the standard Bayesian updating (uncorrelated prior) is found on the optimistic (upper) end of y^*. Otherwise QFB significantly out-competes GBR in that it would allow the insurance company to tap a new class of clients. This tabular is visualized in Figure 9.

7 Summary and Conclusions

This article introduces a new rule for Bayesian updating of imprecise priors that can be represented by classes of precise priors. Our quantile-filted Bayesian learning rule (QFB) disregards those priors that would see the evidence y outside a certain Q-quantile before updating along (a modified version of) generalized Bayes' rule (GBR). The aspect of disregarding priors in view of the evidence before applying GBR is along the idea of Gilboa and Schmeidler to consider only those priors that would maximize prior probability of y. However, in contrast to their rule, QFB has the advantage that it does not add spurious information in case the imprecise prior is contaminated by a "wrong" precise prior. QFB and GBR share the latter advantage.

We demonstrate QFB for a (special version of a) class of precise priors with prescribed marginals and arbitrary correlations. Such a class has been motivated by a recent expert elicitation among modelers working along issues of climate policy advice in the broadest sense. We find that QFB is considerably more informative than GBR for this class.

This suggests an interpretation of QFB. (The reader may use QFB on pragmatic grounds even if she or he would not like to follow the interpretation given in this paragraph.) One possible interpretation assumes (1) that within the class of priors, one prior is the – un-identified – "adequate" and (2) that prior measure can be given a frequentistic interpretation: the life-averaged successes and failures of an expert. Then QFB would imply that with probability $\geq Q$, QFB would acknowledge this adequate prior within the GBR-step. A nesting correction would probabilistically capture the cases if which the adequate prior would be lost. This is possible as upper and lower posterior probabilities are bounded functionals over the set of priors. Hence a nesting-corrected QFB would reveal upper and lower bounds on frequencies of events when averaged over the life of an expert. Remarkably, even after nesting-correcting QFB, QFB remains much more informative than GBR. Hereby we would like to stress that our implementation of QFB is by no means optimized w.r.t. being as informative as possible. One could further optimize Q together with the quantile functional.

Finally we illustrate the effects of various updating rules for the example of a stylized insurance situation. Under QFB much more risky clients could be insured compared to GBR.

A skeptic may argue that any updating rule which disregards precise priors in view of the evidence before applying GBR would be logically inconsistent, as the evidence were used twice: firstly, the evidence is used to disregard priors from then applying GBR. Secondly, those priors that "have made it," are again treated in view of the evidence, namely by standard Bayes' rule.

This counter-argument would apply for Gilboa's and Schmeidler's rule as well as for QFB. Such type of discussion is beyond the scope of this paper, however,

we observe that society very often just behaves like that: it would listen more carefully to experts (i.e. precise priors) that have stated the evidence stronger in advance.

With this article we would like to fuel a discussion on the adequate update rule when updating classes of priors: is it allowed to disregard priors in view of the evidence before Bayesian updating? If yes, what is a meaningful filter? In addition, subsequent algorithms are needed to update imprecise priors that are much more precise on marginals than on (higher order) correlations. Those items seem to be crucial when modeling Bayesian updating of state-of-the-art models in politically influential modeling areas.

In any case it appears as stimulating and satisfying to see experts' relief when not being forced to specify precise measures but instead much less informative measures. We regard this observation as a key motivation for further investments in adequate imprecise models of prior knowledge and generalized Bayesian updating. This also implies the use of social data based choices of non-parametric priors and subsequent numerics.

Finally, we ultimately understand this constribution as an invitation to the "imprecise community" to develop a sound axiom system (as suggested by one of our reviewers) about updating and imprecision (in relation to information).

Acknowledgements

We would like to thank H. Küchenhoff for drawing our attention to the issue of "double-counting" the observational information in non-GBR updating rules. Furthermore we are grateful for the very valuable comments of three anonymous referees. The author has been supported by the Volkswagen Foundation under grant number II/78470.

References

[1] J. O. Berger, D. R. Insua, and F. Ruggeri. Bayesian robustness. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian analysis*, volume 125, pages 1–32. Lecture Notes in Statististics, Springer, New York, 2000.

[2] D. Berleant and Jianzhong Zhang. Using Pearson correlations to improve envelopes around the distribution of functions. *Reliable Computing*, 10(2):139–161, 2004.

[3] F. Coolen. Imprecise conjugate prior densities for the one-parameter exponential family of distributions. *Statistics & Probability Letters*, 16:337–342, 1993.

[4] F. Coolen. *PhD Thesis*. Eindhoven University of Technology, 1994.

[5] J. Dhaene, M. Denuit, M. J. Goovaerts, R. Kaas, and D. Vyncke. The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31:3–33, 2002.

[6] M. J. Frank, R. B. Nelsen, and B. Schweizer. Best-possible bounds for the distribution of a sum – a problem of Kolmogorov. *Probability Theory and Related Fields*, 74:199–211, 1987.

[7] I. Gilboa and D. Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59:33–49, 1993.

[8] H. Held, E. Kriegler, and T. Augustin. Bayesian learning for a class of priors with prescribed marginals. *preprint server http://www.stat.uni-muenchen.de/sfb386/*, 488, 2006.

[9] H. Held, E. Kriegler, and T. Augustin. Bayesian learning for a class of priors with prescribed marginals. *International Journal of Approximate Reasoning*, submitted.

[10] H. Held and T. Schneider von Deimling. Transformation of possibility functions in a climate model of intermediate complexity. *Advances in Soft Computing*, 6:337–345, 2006.

[11] M. Lavine, L. Wasserman, and R. L. Wolpert. Bayesian inference with specified prior marginals. *Journal of the American Statistical Association*, 86(416):964–971, 1991.

[12] M. Martel-Escobar, F. J. Vázquez-Polo, and A. Hernández-Bastida. Analysing the independence hypothesis in models for rare errors: an application to auditing. *Appl. Statist.*, 54(4):795–804, 2005.

[13] T. Seidenfeld and L. Wasserman. Dilation for convex sets of probabilities. *Annals of Statistics*, 21:1139–1154, 1993.

[14] A. H. Tchen. Inequalities for distributions with given marginals. *The Annals of Probability*, 8(4):814–827, 1980.

[15] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[16] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24:463–483, 1997.

Information Processing under Imprecise Risk with the Hurwicz criterion

Jean-Yves Jaffray
LIP6, UPMC-Paris 6
Jean-Yves.Jaffray@lip6.fr

Meglena Jeleva
GAINS, U. Maine and CES U. Paris 1
Jeleva@univ-paris1.fr

Abstract

An agent has Hurwicz criterion with pessimism-optimism index α under imprecise risk and adopts the root dictatorship version of McClennen's Resolute Choice in sequential decision situations, i.e. evaluates strategies at the root of the decision tree by the Hurwicz criterion and enforces the best strategy, thus behaving in a dynamically consistent manner. We address two questions raised by this type of behavior: (i) is information processed correctly? and (ii) to what extent do unrealized outcomes influence decisions (non-consequentialism)? Partial answers are provided by studying: (i) the random sampling of a binary variable, and finding the influence of the pessimism-optimism index to be decreasing with the sample size, and the optimal decision rule to asymptotically only depend on the relative frequencies observed; and (ii) an insurance problem in which the agent chooses his coverage at period two after observing the period one outcome (*accident* or *no accident*); when no accident happened, a seemingly irrelevant data - the first period deductible level- is found to be able to influence the second period insurance choice. We analyse this result in relation with the existence and value of the pessimism-optimism degree.

Keywords. Imprecise risk, Hurwicz criterion, resolute choice, non-consequentialism, learning

1 Introduction

This paper deals with the impact of information on the decisions of an agent whose beliefs concerning the events are imprecise and whose preferences are not in accordance with the Subjective Expected Utility (SEU) model. Precisely, we assume that preferences are representable by the Hurwicz criterion: the value of a decision is a weighted sum of its lowest possible expected value (pessimistic evaluation) and of its highest one (optimistic evaluation).

It is well known that a SEU maximizer has dynamically consistent preferences: future decisions which seem the best today will still be judged the best tomorrow; this justifies the determination of the optimal strategy by backward induction (sophisticated choice). Preferences as modelled by the Hurwicz criterion no longer verify this consistency property. Thus, sophisticated choice no longer guarantees a rational behavior: the selected strategy may well be dominated.

An alternative to sophisticated choice which ensures rationality is the version of McClennen's Resolute Choice (1990) where the best strategy at the root is continued at every node (root dictatorship). We adopt this model here: strategies are evaluated at the root of the decision tree by the Hurwicz criterion; the enforcement of the best strategy all along the tree automatically guarantees dynamic consistency.

The use of Resolute Choice in an imprecise probability environment raises a first important question: is information processed correctly in this model? The existence of phenomena such as dilation (ambiguity increase with new information, cf. Seidenfeld, Wasserman (1993)) makes the answer unclear. We provide a positive answer in a particular case by considering a situation where data are provided by the random sampling of a binary variable and decisions are bets on future values of that variable. This decision problem is closely related to simple hypothesis testing.

Optimal decision rules turn out to be based on observed frequencies (just as likelihood ratio tests) and the influence of the degree of pessimism fades progressively when samples become larger.

A distinctive, controversial feature of Resolute Choice is non-consequentialism: decisions may depend on seemingly irrelevant data such as unrealized outcomes. Since this is a theoretical result, the question arises whether this phenomenon is widespread in real world decision problems or not. As a first field of

investigation, we have chosen multi-period insurance contracting which constitutes an active research domain (Dionne, Doherty, Fombaron 2000). In this domain, up to now, the environment has invariably been described as a situation of risk (subjective or frequentist probabilities) and the model used is EU theory. However, for some risks, due to lacking or conflicting data, this assumption is highly irrealistic which is our justification for introducing imprecise risk in the case of a two-period insurance problem in which an individual has to chose his coverage for the second period after observing the first period outcome (*loss*, *no loss*). We apply Hurwicz's criterion together with a Resolute Choice behavior and determine to which extent unrealized outcomes influence optimal decisions. It turns out that such an influence indeed exists but only to a limited extent and for individuals who are neither extremely pessimistic, nor extremely optimistic.

2 Dynamic decision making in imprecise probabilities framework

2.1 Imprecise Risk

When facing common, general or personal, hazards, and in particular insurable hazards, most agents do not have a precise idea of their likelihoods. Statistics may be inexistent, unavailable or just neglected by the agent; also, important individual variations can exist. Thus, whatever the reasons, an agent may prove to be unable to ascribe specific probabilities to the relevant events in a significant manner.

On the other hand, he may feel more comfortable with associating with each event E a probability interval, $[P^-(E), P^+(E)]$; for instance, typical intervals would be: $[0.01, 0.10]$ for an event he considers as "very unlikely to happen but not impossible"; $[0.10, 0.30]$ for an event he judges "rather unlikely to happen"; and their union $[0.01, 0.30]$ for an event he just thinks "unlikely to happen".

If the agent moreover believes that there is a *true* probability P_0 on the events (which he is just not able to identify), these judgments are submitted to consistency rules, such as $P^+(E) \geq 1 - P^+(E^c)$ for complementary events E and E^c; this circumscribes P_0 to $\mathcal{P} = \{P : \text{for all } E,\ P(E) \in [P^-(E), P^+(E)]\}$, a subset of $\mathcal{L}$, set of all probabilities on the event set.

Such an agent uses an *imprecise probability* representation of uncertainty and, accordingly, makes decisions under *imprecise risk*.

2.2 The Hurwicz decision criterion

Various theories have been proposed for modelling decision making under imprecise risk. The most popular one (but not the only one, see § 2.3.4.) combines existing theories applying to the limiting cases of risk and complete ignorance.

(i) Under *risk*, the standard criterion is Expected Utility (EU). A decision maker (DM), believing the true probability to be P_0, ascribes to a decision δ value

$$U_{P_0}(\delta) = E_{P_0} u(\delta) = \textstyle\sum_x u(x) P_0(\delta^{-1}(x))$$

i.e., the expectation of the utilities of the outcomes x that δ may bring about depending on which event $\delta^{-1}(x)$ obtains;

(ii) Under *complete ignorance*, Hurwicz's criterion, proposed as early as 1951, ascribes to a decision δ a value which is a weighted sum of its worst and best possible outcomes, $\alpha m_\delta + (1-\alpha) M_\delta$; parameter α being interpreted as a degree of pessimism.

Suppose now that *complete ignorance* prevails in $\mathcal{P}$ and consider a DM for whom being only able to locate probability P_0 in a set $\mathcal{P}$ amounts to being uncertain about which of the values $U_P(\delta)$, P in $\mathcal{P}$, is the correct one. Then, this DM will look at the worse and best possible evaluations and, according to its degree of pessimism, will put more or less weight on the former or the later, which is expressed by the following formula:

$$V(\delta) = \alpha \inf_{P \in \mathcal{P}} E_P u(\delta) + (1-\alpha) \sup_{P \in \mathcal{P}} E_P u(\delta) \quad (1)$$

This criterion being the natural extension of the Hurwicz one to imprecise risk, we will preserve its denomination of "Hurwicz criterion". In a decision making context, the interest of a preference model depends crucially on its ability to induce *economically rational behavior*, which includes invulnerability to Dutch books and money-pumps (Schick 1986, Diecidue, Wakker 2002) in situations involving sequential choices. Obviously, economic rationality cannot be guaranteed by a criterion which does not increase with dominance - is not *monotone* - in some sense.

Under suitable topological assumptions ($\mathcal{P}$ a compact subset of a separable space), Hurwicz's criterion satisfies strict and weak monotonicity properties. If the expected utility of decision δ is strictly higher than that of decision d for every probability measure, i.e., $E_P u(\delta) > E_P u(d)$ for all $P \in \mathcal{P}$ (strict pointwise dominance on $\mathcal{P}$), then $\inf_{P \in \mathcal{P}} E_P u(\delta) > \inf_{P \in \mathcal{P}} E_P u(d)$, $\sup_{P \in \mathcal{P}} E_P u(\delta) > \sup_{P \in \mathcal{P}} E_P u(d)$,

and finally $V(\delta) > V(d)$; moreover, the weaker relation, $E_P u(\delta) \geq E_P u(d)$ for all $P \in \mathcal{P}$, implies $V(\delta) \geq V(d)$. In particular, if decision δ performs strictly better (resp. better) than decision d whatever happens, i.e., $u(\delta(e)) > (\geq)\ u(d(e))$ for every event e on which both δ and d are constant, then $E_P u(\delta) > (\geq)\ E_P u(d)$ for all $P \in \mathcal{P}$, hence $V(\delta) > (\geq)\ V(d)$.

On the other hand, if $E_P u(\delta) \geq E_P u(d)$ for all $P \in \mathcal{P}$, with $E_P u(\delta) > E_P u(d)$ for some $P \in \mathcal{P}$, it may none- theless happen that $\inf_{P\in\mathcal{P}} E_P u(\delta) = \inf_{P\in\mathcal{P}} E_P u(d)$ and $\sup_{P\in\mathcal{P}} E_P u(\delta) = \sup_{P\in\mathcal{P}} E_P u(d)$, hence that $V(\delta) = V(d)$; in particular, $u(\delta(e)) \geq u(d(e))$ for every e, plus $u(\delta(e)) > u(d(e))$ for some e, do not imply $V(\delta) > V(d)$. Note however that for every $\varepsilon > 0$, $V(\delta) > V(d-\varepsilon)$ and $V(\delta+\varepsilon) > V(d)$ will hold; thus, although not monotone, Hurwicz's criterion is, in a straightforward sense, $\varepsilon-$*monotone*.

These monotonicity properties are sufficient to make the model behave satisfactorily in one-shot decision problems. Multiple decision situations are a different matter, as illustrated in the following subsection.

2.3 Problems with dynamic decision making and the Resolute Choice solution

2.3.1 An illustrative example

Consider a DM who at time 1 (node A of the decision tree in Fig.1) has to choose between two decisions, Up_1 and $Down_1$; then, at time 2 (node B), provided he has chosen Up_1 and event E obtains, he has again a choice, Up_2 or $Down_2$, his gain further depending on the realization or not of some events, G or G^c and H or H^c; if at time 1 he has chosen Up_1 and event E^c obtains, or has chosen $Down_1$, there is no other choice to make. Gains are indicated next to the corresponding leaves of the tree.

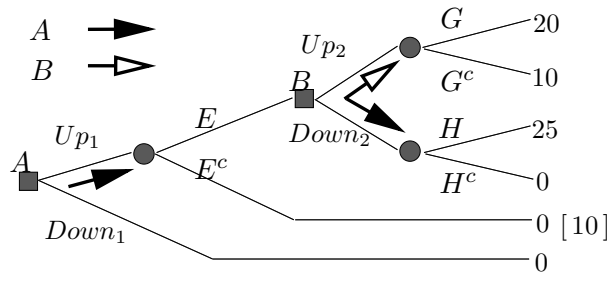


Figure 1: Dynamically inconsistent preferences

The DM's criterion is Hurwicz's, with the same parameters u and α, at both decision nodes, A and B. For the sake of simplicity we assume $\alpha = 1/2$, risk-neutrality ($u(x) = 2x$ for all x), and complete ignorance on the algebra of events generated by E, G and H; thus, $\mathcal{P} = \mathcal{L}$ and a strategy (at A), as well as a substrategy (at B), δ, giving outcomes $\delta(e)$ on events e has value $V(\delta) = \inf_e \delta(e) + \sup_e \delta(e)$.

At node A, the values of the three available strategies, $(Up_1, Up_2), (Up_1, Down_2)$, and $Down_1$ ((Up_1, Up_2) means Up_1 at node A; then Up_2 at node B if E happens; etc.) are, respectively, $V(Up_1, Up_2) = 20; V(Up_1, Down_2) = 25; V(Down_1) = 0$; thus the DM prefers $(Up_1, Down_2)$ to (Up_1, Up_2) (and to $Down_1$) in A.

However at node B he prefers substrategy (decision) Up_2 to substrategy $Down_2$ since $V(Up_2) = 30 > V(Down_2) = 25$; thus, if he takes decision Up_1 in A and event E happens, then, once arrived in B, he no longer considers $Down_2$ to be the best feasible action; his preferences are not *dynamically consistent*.

2.3.2 Resolute Choice

What are the decisions actually made by a DM with a logical mind, who is able to anticipate on his future actions (*sophistication*, as opposed to *myopia*), and is aware that his preferences are not dynamically consistent? Roughly, one can think of two different patterns of behavior.

(i) If his future choices are always dictated by his future preferences, then the DM should use *backward induction* in the decision tree: at each given decision node, knowing which substrategies would be triggered by each of his feasible actions, he can evaluate and compare them, according to his criterion, and choose the best available action. Coping locally in that way with his preferential inconsistencies unfortunately does not warrant him at the end (when arrived at the root of the tree) the selection of a strategy possessing a valuable global property. Indeed, going back to the example, the DM would be willing to pay up to 5 units to have the tree pruned and edge Up_2 suppressed in B. Consider then the augmented tree in which a new subtree offers this possibility to the DM; strategy $(Up_1, Down_2)$, which is still materially feasible, clearly strictly dominates the additional strategy, which is nonetheless chosen by the backward induction procedure. In general, the use of that behavioral procedure is always a potential source of unnecessary waste: it is not economically rational.

How can any waste be avoided? There is a straightforward way:

(ii) If the strategy which is judged best according to preferences at the root node is actually played, then, the criterion being used only once as in one-shot deci-

sion problems, the monotonicity of Hurwicz's criterion guarantees economic rationality. This dictatorship of the root node preferences means of course that future choices do not have to bear any relation with future preferences. More generally and less drastically, *Resolute Choice* (McClennen, 1990, p.260) only requires the achievement of a compromise strategy reflecting both present and future preferences; in McClennen's terms: "the theory of resolute choice is predicated on the notion that the single agent who is faced with making decisions over time can achieve a cooperative arrangement between his present self and his relevant future selves that satisfies the principle of intrapersonal optimality". Resolute Choice is not just a theoretical construct; it can be implemented in an operational way (see Jaffray-Nielsen 2006).

2.3.3 Non-consequentialism and unrealized outcomes

A feature of Resolute Choice is *non-consequentialism*: the choice at a given decision node, being induced by a strategy which depends on all the data in the decision tree, may in particular depend on those data which are outside the subtree rooted at that node; these elements are known as *unrealized outcomes*.

In Fig.1, if the best strategy in A, $(Up_1, Down_2)$, can be imposed, $Down_2$ is played in B. Modify now a single outcome, at the leaf following Up_1 and E^c, by changing 0 into 10; the best strategy in A is now (Up_1, Up_2) and Up_2 is played in B accordingly; thus the action taken in B depends on a unrealized outcome, the outcome at a leaf that is not part of the subtree rooted at B.

For an illuminating discussion of consequentialism see Machina (1989). Let us just note for the moment that, since, as seen above, economic rationality cannot provide arguments against non-consequentialism, any defense of consequentialism must rely on a different conception of rationality.

2.3.4 Alternative approaches

Resolute Choice should not be confused with consequentialist approaches to dynamic decision making, which have recourse to recursive models (see e.g. Epstein, Schneider 2003); such models are straightforwardly dynamically consistent and backward induction remains valid; on the other hand, economic rationality is not necessarily satisfied. Neither is it in the non-consequentialist approach, preserving a weak form od dynamic consistency of Hanany, Klibanoff (2006).

Another approach to dynamic decision making under uncertainty, called E-admissibility, has been suggested by Levi (1974) and discussed by Seidenfeld (2004). It works by first selecting all the last stage Bayes rules and then moving backwards repeating this selection stage by stage. In order to uniquely select a strategy in the remaining set, a secondary criterion, applied at the root node, is used. While more discriminating than Resolute Choice with root dictatorship, E-admissibility (with a suitable secondary criterion) still guarantees normative qualities such as nonnegative value of information.

Note that E-admissibility is a non-consequentialist solution in general. However, de Cooman and Troffaes (2005) prove the validity of dynamic programming (which amounts to consequentialism) in the particular case of sequential decision making in the absence of conditional decisions.

3 Learning with Resolute Choice

An urn contains red and black balls; the proportion of red balls is *either* p^- *or* p^+, where $0 < p^- < p^+ < 1$. The DM is told that: $n+1$ balls are going to be drawn one by one from the urn, with replacement; that he can make bets on the color of the $(n+1)^{th}$ being red; and that his decision of betting or not can be conditioned on the outcome of the n first draws. When betting, his stake is m and he will receive gain M if the $(n+1)^{th}$ is red. We assume $p^- < \frac{m}{M} < p^+$.

The DM conditions his bets on the outcomes of the n first draws by just specifying a *betting rule* $K_n \subseteq \{0, 1, .., n\}$, "$k \in K_n$" meaning: "if k balls among the n first drawn are red, bet (on red) at the $(n+1)^{th}$ draw".

One denotes $k_n = \min_{k \in K_n} k$.

The DM uses the Hurwicz criterion, is risk neutral ($u(x) = x$) and is resolute; he chooses his betting rule when learning the sample size n and before the observations begin.

We are interested in the evolution of the optimal betting rule when n tends to infinity.

A *betting behavior* is a sequence $(K_n)_{n \in \mathbb{N}}$. Betting behavior $(K_n)_{n \in \mathbb{N}}$ *weakly dominates* betting behavior $(K'_n)_{n \in \mathbb{N}}$ if for all $n \in \mathbb{N}$, $V(K_n) \geq V(K'_n)$; if, moreover, $V(K_n) > V(K'_n)$ for some value of $n \in \mathbb{N}$, then $(K_n)_{n \in \mathbb{N}}$ *dominates* $(K'_n)_{n \in \mathbb{N}}$. A betting behavior which is not dominated by any other is *admissible.* A betting behavior which weakly dominates all the others is *optimal.*

A betting behavior $(K_n)_{n \in \mathbb{N}}$ will be called *consistent* when its betting rules are all of the form $K_n = \{k_n, k_n + 1, .., n\}$ (i.e., betting if and only if at least

k_n red balls have been drawn).

Lemma 1 For a fixed n, let betting rules K_n and K'_n only differ in the case where k red balls are drawn: $k \in K_n$; $K'_n = K_n \backslash \{k\}$; then

$$V(K_n) > [=] V(K'_n) \Longleftrightarrow \frac{k}{n} > [=] L + \frac{1}{n} R$$

with $L = \dfrac{\ln \dfrac{1-p^-}{1-p^+}}{\ln \dfrac{p^+(1-p^-)}{(1-p^+)p^-}}$ and

$$R = \frac{\ln \left[\dfrac{\alpha}{1-\alpha} \times \dfrac{m - p^- M}{p^+ M - m} \right]}{\ln \dfrac{p^+(1-p^-)}{(1-p^+)p^-}}$$

N.B. The proofs of Lemma 1 and of the other results can be found in Jaffray, Jeleva (2007).

The following proposition is a direct application of Lemma 1.

Proposition 1 Consider betting behavior $(K_n)_{n\in\mathbb{N}}$, and let $k_n = \min_{k\in K_n} k$.

A necessary condition for the admissibility of $(K_n)_{n\in\mathbb{N}}$ is that
$\frac{k_n}{n} \to_{n\to\infty} L$
with L defined in lemma 1.

Proposition 2 The consistent betting behavior, $(K^*_n)_{n\in\mathbb{N}}$ where $K^*_n = \{k^*_n, k^*_n + 1, k^*_n + 2, ..., n\}$, and for each n, k^*_n is the smallest integer such that

$\frac{k^*_n}{n} \geq L + \frac{1}{n} R$ with L and R defined in emma 1.

is an *optimal betting behavior.*

Note that expression $\left[\dfrac{p^+}{1-p^+} \times \dfrac{1-p^-}{p^-} \right]^k \times \left[\dfrac{1-p^+}{1-p^-} \right]^n$ is a likelihood ratio; in fact the monotonicity properties of the Hurwicz criterion make likelihood ratio (possibly random) tests an admissible family as in the standard statistical decision theory (Neyman-Pearson lemma). For related results concerning hypothesis testing with imprecise probabilities on the parameter space, see Jaffray, Saïd (1994).

Note also that expression R , defined in emma 1, has a strong similarity with the term that would appear in a Bayesian model, which is $\dfrac{\ln \left[\dfrac{\pi}{1-\pi} \times \dfrac{m - p^- M}{p^+ M - m} \right]}{\ln \dfrac{p^+(1-p^-)}{(1-p^+)p^-}}$, with π the prior probability of p^- being the true proportion of red balls.

Let us finally emphasize the fact that, although all betting decisions are made only on the basis of a single *ex ante* evaluation, data are taken into account in a sensible way: for high values of n, the DM acts as if he used relative frequencies as estimators of probabilities; however, for smaller n, the degree of pessimism has some influence on the bets through the term R.

4 An application of Resolute Choice to Two-period Insurance Demand

In this section, we study a two-period insurance problem in which an individual has to choose his coverage at period 2 after observing the period 1 outcome ([*a*] *loss* [*occurred*] or *no loss* [*occurred*]).

An individual with initial wealth W faces a risk with a unique amount of potential loss $L < W$. This situation can be represented by a random variable X: if E is the event *loss (occurs)* and E^c the event *no loss*, $X(\omega) = L$ for $\omega \in E$ and $X(\omega) = 0$ for $\omega \in E^c$. The individual's information and/or beliefs allow him to assert that the probability of loss occurrence during a year is between p^- and p^+. The set of probability distributions which are consistent with the available information is:

$$\mathcal{P} = \left\{ P \in \mathcal{L} : P(E) \in \left[p^-, p^+ \right] \right\} \qquad (2)$$

where $\mathcal{L}$ denotes the set of all probability distributions on the relevant support.

Two periods of time are considered: in the first period, the individual has no insurance choice to make; for instance, he rents a car, and an insurance coverage with a deductible $K \leq L$ is automatically included in the contract. In the second period however, the individual has to decide if he will subscribe an insurance contract or not, for instance he will buy a car and has to decide whether or not he will take a theft insurance (which is not mandatory). We assume that only one insurance contract is available: it corresponds to full coverage and the premium is $\Pi < L$.

We assume that the individual needs to decide immediately, at the beginning of the first period, what his insurance policy will be; the reason may be, for instance, that he still has then other opportunities beside renting-then-buying a car and that their comparisons require accurate evaluations, or that he has to plan out his expenses in advance.

Individual preferences are represented by the Hurwicz criterion: a decision $\delta : \Omega \to \mathbb{R}$ is evaluated by functional V of formula (1) where u is a strictly increasing function.

In the simpler, one period situation, where there is no previous experience of loss, the set of strategies D contains only two elements, denoted: d, the individual subscribes an insurance contract, and $\bar{d}$, the individual does not buy any insurance. According to (1), these decisions have the following values:

$$V(d) = u(W - \Pi)$$
$$V(\bar{d}) = (\alpha p^+ + (1-\alpha)p^-)\,u(W-L) + (1 - \alpha p^+ - (1-\alpha)p^-)\,u(W)$$

and the decision to buy coverage depends on the pessimism-optimism index α and on the information precision in the following way:

$$V(d) \geq V(\bar{d}) \Leftrightarrow \alpha\left(p^+ - p^-\right) \geq \frac{u(W) - u(W-\Pi)}{u(W) - u(W-L)} - p^-.$$

Thus, a higher degree of pessimism and a greater imprecision both act in favor of the decision to buy insurance coverage.

4.1 Decisions evaluation

We now turn to the evaluation of the decisions of an individual who acquires additional information related to a period one potential loss. His decisions can then be conditioned on the realization of the loss in the first period. Our goal is to determine the influence of the first period loss realization on the second period decision as well as the impact of α on that decision. We further assume probabilistic independence of the successive events, i.e., that for any given probability $p \in [0,1]$, with E_i denoting the event "loss in period i", if $P(E_1) = p$ then $P(E_2/E_1) = p$ as well, hence $P(E_2) = p$ and $P(E_1 \cap E_2) = p^2$.

A strategy is now characterized by a pair of decisions: the first one conditional on the realization of E_1, and the second one on the realization of E_1^c. The set of possible strategies D consists then in four pairs of decisions: $D = \{dd,\ d\bar{d},\ \bar{d}d,\ \bar{d}\bar{d}\}$, where $dd = \{d$ if E_1, d if $E_1^c\}$, $d\bar{d} = \{d$ if E_1, $\bar{d}$ if $E_1^c\}$, ... The decision tree corresponding to this problem is given in Fig.2.

The evaluations of the strategies at the beginning of period one by the Hurwicz criterion are given in the following proposition. This evaluation requires the determination of the probabilities in $[p^-, p^+]$ at which the lowest and highest expected utility are achieved. It turns out that these probabilities may well differ from p^+ and p^- and depend on the strategy.

Proposition 3 If $\Pi,\ K,\ L,\ p^-,\ p^+$ are such that:

- $u(W-L-K) \leq \frac{1}{2p^-}\left[u(W-\Pi) + (2p^- - 1)u(W-K)\right]$

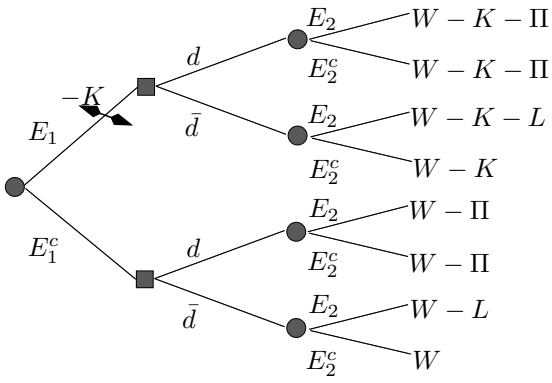


Figure 2: Insurance Demand Tree

- $p^* = \frac{1}{2} + \frac{u(W) - u(W-\Pi-K)}{2[u(W)-u(W-L)]}$
  verifies $p^* \in [p^-, p^+]$ and $p^* > \frac{1}{2}(p^- + p^+)$,

then the available decisions are evaluated as follows:

$V(dd) = A(p^+,p^-)u(W-\Pi-K) + (1 - A(p^+,p^-))\,u(W-\Pi)$;

$V(d\bar{d}) = A(p^*,p^-)u(W-\Pi-K) + B(p^*,p^-)u(W-L) + C(1-p^*, 1-p^-)u(W)$;

$V(\bar{d}d) = C(p^+,p^-)u(W-L-K) + B(p^+,p^-)u(W-K) + A(1-p^+, 1-p^-)u(W-\Pi)$

$V(\bar{d}\bar{d}) = C(p^+,p^-)u(W-L-K) + B(p^+,p^-) \times (u(W-K) + u(W-L)) + C(1-p^+, 1-p^-)u(W)$

where

$A(p,q) = \alpha p + (1-\alpha)q$,
$B(p,q) = \alpha p(1-p) + (1-\alpha)q(1-q)$,
$C(p,q) = \alpha p^2 + (1-\alpha)q^2$.

In very ambiguous situations, the requirements above are not too restrictive; for instance, in the limiting case of complete ignorance, that is, for $[p^-, p^+] = [0,1]$, these conditions reduce to $\Pi > K$.

From now on, we assume that these conditions are satisfied.

Note that the pessimistic evaluation of strategy $d\bar{d}$ is not achieved at the upper probability bound p^+: with p^* smaller than p^+ but close to it, the advantage of incurring period 1 loss K with the smaller probability p^* is not compensated by the disadvantage of incurring period 2 loss L with probability $(1-p^*)p^*$ greather than $(1-p^+)p^+$.

Let us now turn to a specific feature of the model: the relevance of unrealized outcomes.

Consider strategies dd and $d\bar{d}$. They differ by the

decision that follows the period 1 *no loss* event. The utilities involved in the direct comparison of these conditional decisions do not depend on K, and its value would be irrelevant in a consequentialist approach. However, with our criterion, $V(dd) - V\left(d\bar{d}\right) =$
$\alpha\left(p^{+}-p^{*}\right)u\left(W-\Pi-K\right)+$
$(1-\alpha p^{+}-(1-\alpha)p^{-})\,u\left(W-\Pi\right)$
$-\left(\alpha p^{*}(1-p^{*})+(1-\alpha)p^{-}(1-p^{-})\right)u\left(W-L\right)-$
$\left(\alpha(1-p^{*})^{2}+(1-\alpha)(1-p^{-})^{2}\right)u(W)$

The sign of the previous expression is indeterminate and depends on the value of K, which influences both the lowest utility $u\left(W-\Pi-K\right)$ and p^{*}. More precisely, the influence of K on the discrepancy between $V(dd)$ and $V\left(d\bar{d}\right)$ increases with the pessimism-optimism index α, since

$\frac{d\left[V(dd)-V\left(d\bar{d}\right)\right]}{dK} =$
$\alpha\{-\frac{dp^{*}}{dK}u\left(W-\Pi-K\right)-\left(p^{+}-p^{*}\right)u'\left(W-\Pi-K\right)$
$+\left(2p^{*}-1\right)\frac{dp^{*}}{dK}u(W-L)+2\left(1-p^{*}\right)\frac{dp^{*}}{dK}u(W)\}$

The reason why the comparison of $V(dd)$ and $V\left(d\bar{d}\right)$ depends on the irrelevant outcome K is that the Hurwicz criterion is a limiting form of a rank dependent utility (RDU) criterion and that in RDU theory (Quiggin 1982) the decision weight associated with a consequence depends on the rank of this consequence in the set of consequences of a given decision. Decisions dd and $d\bar{d}$ have $W-\Pi-K$ as a common consequence but while with dd, $W-\Pi-K$ is the worst consequence, this is no longer the case with $d\bar{d}$ for which it is $W-L$. Consequently, the decision weight of $u\left(W-\Pi-K\right)$ is not the same in the evaluation of dd and $d\bar{d}$, even if this consequence is obtained for the same event (E_1) with both decisions. Thus, the second period preference between insurance or not in the case where no loss occurred in the first period may depend on the deductible level which the individual would have paid had loss occurred.

4.2 A numerical example

The following example illustrates the impact of K on the optimal strategy[1].

We consider an individual with initial wealth $W = 1\,000\,000$ who faces the risk of a loss of amount $L = 40\,000$. Loss probability at each period, p, belongs to $[0.01,\ 0.7]$. The insurance premium for full coverage is $\Pi = 4\,000$. The utility function is assumed to be in the CRRA class (with constant relative risk aversion) that is $u(x) = \dfrac{x^{1-R}}{1-R}$; here, we take $R = 2$.

The sign of $V(dd) - V\left(d\bar{d}\right)$ depends on α and K as follows:

- for $\alpha \in [0, 0.22[$, $V(dd) - V\left(d\bar{d}\right) < 0$ for any $K \in [0, 40\,000]$;
- for $\alpha \in [0.22,\ 0.29[$, there exist $K^{*} < 40\,000$ such that $V(dd) - V\left(d\bar{d}\right) \leq 0$ for $K \leq K^{*}$ and $V(dd) - V\left(d\bar{d}\right) > 0$ for $K > K^{*}$;
- for $\alpha \in [0.29, 0.33[$, there exist K^{*} and K^{**} with $0 < K^{*} < K^{**} < 40\,000$ such that $V(dd) - V\left(d\bar{d}\right) < 0$ for $K^{*} < K < K^{**}$ and $V(dd) - V\left(d\bar{d}\right) \geq 0$ for $K \leq K^{*}$ and $K \geq K^{**}$;
- for $\alpha \in [0.33, 1]$, $V(dd) - V\left(d\bar{d}\right) > 0$ for any $K \in [0, 40\,000]$.

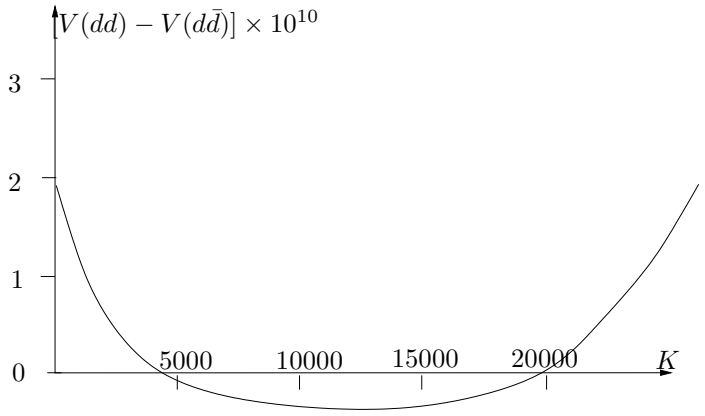


Figure 3: Choice Dependence on K for $\alpha = 0.31$

Let us now study the dependence of the optimal strategy on K and α.

- We start by comparing $d\bar{d}$ and $\bar{d}d$: $V\left(d\bar{d}\right) - V\left(\bar{d}d\right)$is a linear function of α; moreover,

  for $\alpha = 0$, as well as for $\alpha = 1$, $V\left(d\bar{d}\right) - V\left(\bar{d}d\right) > 0$ for any $K \in [0, 40000]$; thus, for any $\alpha \in [0, 1]$, $d\bar{d}$ is preferred to $\bar{d}d$.

- The same result is obtained for $d\bar{d}$ when compared with $\bar{d}\bar{d}$.

- The choice between dd and $\bar{d}\bar{d}$ depends on α in the following way:

  $V(dd) - V(\bar{d}\bar{d}) < 0$ for $\alpha \in [0, 0.003[$;

  $V(dd) - V(\bar{d}\bar{d}) > 0$ for $\alpha \in [0.003, 1]$.

Thus, for any $K \in [0, 40000]$, strategies $\bar{d}\bar{d}$ and $\bar{d}d$ are dominated so that, the best strategy is always either dd or $d\bar{d}$.

This dominance is due to the low insurance premium Π that corresponds here to a probability estimation of 0.1. In consequence, individuals prefer either to fully

[1] Numerical results are obtained with Mathematica 4.1.

insure in any case (if they are pessimistic enough) and thus benefit from the low premium, or to adapt their decision to the observed loss. Fig.4 shows the optimal strategy as a function of K and α. It appears that the optimal decision results from a trade-off between the attractivity of low price insurance and that of information depending decisions. For strong optimists, the information effect dominates, whereas for strong pessimists, the full coverage effect dominates. For intermediate values of α however, the deductible value K may influence choice: a high value of K can even influence all decisions by lowering the individual's expected wealth perspectives and acting in favor of full coverage.

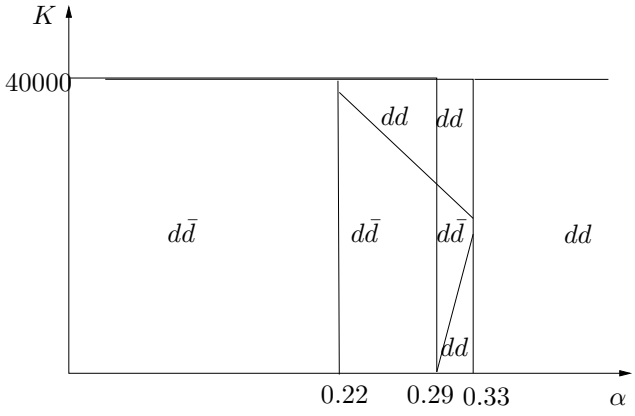


Figure 4: Choice Dependence on K and α

4.3 Optimal strategy for risk-neutral individuals

To emphasize the impact of the pessimism index α on the optimal insurance strategies, we now consider the case when $u(x) = x$. This allows us to isolate the influence of ambiguity attitude, characterized here by α, from that of the risk attitude, characterized by u.

Proposition 4 Consider a two period insurance problem, where the individual's imprecise information on the loss probability is given by an interval $[p^-, p^+]$ with $p^- < \frac{1}{2} < p^+$ and the insurance premium Π for full coverage is such that $\Pi \in [p^- L, p^+ L]$. The preferences of the individual are characterized by the Hurwicz criterion with $u(x) = x$. Then, he orders the different available strategies in the following way:

- $d\bar{d} \succsim \bar{d}d$ for any $\alpha \in [0,1]$;
- $dd \succsim \bar{d}\bar{d} \Leftrightarrow \alpha \geq \alpha^*$ with $\alpha^* = \frac{(\Pi - p^- L)}{(p^+ - p^-)L}$ where $\alpha^* < 1$;
- if $K = 0$, $dd \succsim d\bar{d} \Leftrightarrow \alpha \geq \alpha^{**}$ with $\alpha^{**} = \frac{(1-p^-)(\Pi - p^- L)}{(p^* - p^-)(\Pi - p^- L + L(1-p^*))}$ where $\alpha^{**} < 1$;

  if $K > 0$, both $dd \succsim d\bar{d}$ and $d\bar{d} \succsim dd$ are possible depending on the value of K.
- $d\bar{d} \succsim \bar{d}\bar{d} \Leftrightarrow \alpha \geq \alpha^{***}$ with $\alpha^{***} = \frac{p^-(\Pi - p^- L)}{(p^+ - p^*)(K+L) + (p^* - p^-)[L(p^* + p^-) - \Pi]}$ where $\alpha^{***} < 1$.

This proposition allows to determine, for $K = 0$ the impact of the pessimism index on the individual's optimal strategy. More precisely, in this case, $\alpha^{***} < \alpha^* < \alpha^{**}$ and $\bar{d}\bar{d}$ is the optimal strategy for $\alpha \in [0, \alpha^{***}[$, $d\bar{d}$ is the optimal strategy for $\alpha \in]\alpha^{***}, \alpha^{**}[$ and dd is the optimal strategy for $\alpha \in]\alpha^{**}, 1]$. For $\alpha = \alpha^{***}$, the individual is indifferent between $\bar{d}\bar{d}$ and $d\bar{d}$, and for $\alpha = \alpha^{**}$, he is indifferent between $d\bar{d}$ and dd.

To sum up, in this model, neither a very optimistic individual (α close to 0) nor a very pessimistic one (α close to 1) takes advantage of the information: his decisions do not depend on his period 1 observation. The reason is that, strong pessimists are trying above all to avoid the lowest possible consequences, which are here $W - L - K$ if E_1 and $W - L$ if E_1^c; choosing dd is the strategy that makes it possible. The opposite is true for strong optimists: they will prefer the decisions that allow the higher possible consequences, which are here $W - K$ if E_1 and W if E_1^c.

For moderate individuals, choice is less straightforward: for them, it is valuable both to avoid $W - L - K$ if E_1 (which however means renouncing to get $W - K$) and to preserve the possibility to obtain W if E_1^c (which however means risking to get $W - L$); this is only possible with $d\bar{d}$, and trade-offs, which depend on all the parameters (in particular on Π) may favor this strategy.

5 Conclusion

The preceding results demonstrate the operational tractability of the Resolute Choice dynamic adaptation of the Hurwicz criterion for decision making under imprecise risk. This model is able to process information correctly; in particular, for large samples, choices made show that the true probabilities are learned correctly although implicitly.

Also, the puzzling influence of unrealized outcomes appears as rather limited (only concerns individuals whose pessimism index belongs to a small range) and does not seem to lead to counter-intuitive decisions. It is moreover interesting to note that sensitivity to unrealized outcomes being excluded by Expected Utility theory, the Resolute Choice model has a flexibility that makes it attractive for descriptive purposes.

References

[1] de Cooman G., Troffaes M. C. M. (2005), "Dynamic programming for deterministic discrete-time systems with uncertain gain" *Int. J. Approximate Reasoning* 39(2-3): 257-278.

[2] Diecidue E., Wakker P. (2002), "Dutch books: Avoiding strategic and dynamic complications, and a comonotonic extension", *Mathematical Social Sciences*, 43, 135-149.

[3] Dionne G., Doherty N., Fombaron N. (2000), "Adverse Selection in Insurance Markets", Chapter 7 in *Handbook of Insurance*, ed. G. Dionne, Kluwer.

[4] Epstein L., Schneider M. (2003), "Recursive Multiple Priors", *J. Economic Theory* 113, 1-31.

[5] Hanany E., Klibanoff P. (2006), "Updating Preferences with Multiple Priors", working paper.

[6] Jaffray J.Y., Jeleva M. (2007), "Information Processing under Imprecise Risk with an Insurance Demand Illustration", working paper.

[7] Jaffray J.Y., Nielsen T. (2006), "Dynamic Decision Making without Expected Utility: an operational approach", *European J. Operational Research* 169-1, 226-246.

[8] Jaffray, J.Y., Said T. (1994), "Optimal Hypothesis Testing with a Vague Prior" in *Decision Theory and Decision Analysis: Trends and Challenges*, ed. Sixto Rios, Kluwer, 207-222.

[9] Levi, I. (1974), "On Indeterminate Probabilities", *J. Philosophy* 71 (13), 391-418.

[10] Mc Clennen E. (1990), *Rationality and Dynamic Choice*, Cambridge University Press.

[11] Machina M. (1989), "Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty", *J. Economic Literature* 27, 1622-1668.

[12] Quiggin J., (1982), "A Theory of Anticipated Utility", *J.Economic Behavior and Organization*, 3, 324-343.

[13] Schick F. (1986), "Dutch Bookies and Money Pumps", *J. Philosophy*, 83, 112-119.

[14] Seidenfeld, T. (2004), "A Contrast Between two Decision Rules for use with (Convex) Sets of Probabilities: G-Maximin Versus E-Admissibility", *Synthèse* 140, 69-88.

[15] Seidenfeld, T., Wasserman, L. (1993) "Dilation for Sets of Probabilities," *Annals of Statistics* 21 (3), 1139-1154.

Compositional Models of Belief Functions

Radim Jiroušek
Inst. of Inform. Th. and Autom.
Acad. Sci. of the Czech Republic
radim@utia.cas.cz

Jiřina Vejnarová
Inst. of Inform. Th. and Autom.
Acad. Sci. of the Czech Republic
vejnar@utia.cas.cz

Milan Daniel
Inst. of Computer Science
Acad. Sci. of the Czech Republic
milan.daniel@cs.cas.cz

Abstract

After it has been successfully done in probability and possibility theories, the paper is the first attempt to introduce the operator of composition also for belief functions. We prove that the proposed definition preserves all the necessary properties of the operator enabling us to define compositional models as an efficient tool for multidimensional models representation.

Keywords. Belief function, basic assignment, multidimensional frame of discernment, operator of composition, perfect sequence.

1 Introduction

Last years of the last century witnessed emergence of a new approach to efficient representation of multidimensional probability distributions. This approach, which is an alternative to Graphical Markov Modeling, is based on a simple idea: multidimensional distribution is *composed* from a system of low-dimensional (oligodimensional) distributions by repetitive application of a special operator of composition. This is also the reason why the models are called *compositional models*. In several papers, in which the properties of the operator and models were studied [3, 4, 5], it was shown (among others) that these models are, in a way, equivalent to Bayesian networks. Roughly speaking, *any multidimensional distribution representable by a Bayesian network can also be represented with approximately the same number of parameters (probabilities) in the form of a compositional models, and vice versa.*

Though Bayesian networks and compositional models represent the same class of distributions, they do not do it in the same way. Bayesian networks use *conditional distributions* whereas compositional models consist of *unconditional distributions*. Naturally, both types of models bear the same information but whilst some marginal distributions are explicitly expressed in compositional models, it may happen that their computation from a corresponding Bayesian network is rather computationally expensive. Therefore it appears that some of computational procedures designed for compositional models are (algorithmically) simpler than their Bayesian network counterparts.

The goal of this paper is to show that the operator of composition can also be introduced for belief functions. Moreover, we will show that it inherits the basic properties of its probabilistic pre-image and therefore it will enable us to introduce compositional models for multidimensional belief functions.

We will see that this approach enables us to represent, let us say, a 15-dimensional belief function as a sequence of 3 or 4-dimensional belief functions. Whilst representation of a 15-dimensional belief function is completely impossible (it would require in binary case $2^{2^{15}} = 2^{32k}$ numbers), representation of a 4-dimensional belief function requires only $2^{2^4} = 2^{16} = 64k$ numbers and therefore a model consisting of twelve 4-dimensional belief functions requires "only" $12 \times 2^{16} = 768k$ values.

Let us stress at the very beginning that this paper is the first one dealing with compositional models for belief functions. At this moment, we do not know what is the connection of the introduced operator of composition to different concepts of conditioning (and conditional independence) introduced for belief functions. The reader should realize that composition defined in this paper is different from that defined by Shenoy in [7]. His composition meets the requirements given by Shenoy's axioms (commutativity, associativity and distributivity) neither of which is met by the composition defined here. Therefore we do not know to what extent his principles of local computations are applicable to our model. This is one of many important open problems, some of which will be mentioned in Conclusions.

The reader familiar with the literature on belief functions is accustomed to the conjunctive rule of combination. Ben Yaghlane et al. [2] apply this rule to the set of marginal and conditional belief functions with the goal to compute a joint belief function in a way analogous to Bayesian networks (so-called Belief Chain Rule). This type of operation again substantially differs from the composition considered in this paper; the conjunctive rule of combination is commutative and associative. Moreover, in older papers, Xu and Smets consider only 2-dimensional belief functions, see e.g. [10].

Though the present paper is a contribution to belief function theory, we will not use the term of *belief function* any more in this paper. We are convinced that it will make the paper more legible for the reader when we will restrict our considerations to *basic belief assignments*, only. Therefore we will define a composition of basic assignments and show how to compose a sequence of simple basic assignments to get an assignment corresponding to a multidimensional belief function.

The contribution is organized as follows. In Section 2 we summarize basic notions, notation and introduce the operator of composition. Its basic properties can be found in Section 3, while Section 4 is devoted to more advanced properties. Finally, in Section 5 we introduce the notion of so-called *p*erfect sequences and demonstrate their importance.

2 Notation

Consider a finite index set $N = \{1, 2, \ldots, n\}$ and finite sets $\{\mathbf{X}_i\}_{i\in N}$. In this text we will consider *multidimensional frame of discernment*

$$\Omega = \mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n,$$

and its *subframes*. For $K \subset N$, $\mathbf{X}_K$ denotes a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \bigtimes_{i\in K} \mathbf{X}_i.$$

A *projection* of $x = (x_1, x_2, \ldots, x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1, i_2, \ldots, i_\ell\}$

$$x^{\downarrow K} = (x_{i_1}, x_{i_2}, \ldots, x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K | \exists x \in A : y = x^{\downarrow K}\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$. In this case $A^{\downarrow\emptyset} = \emptyset$.

In addition to the projection, in this text we will need also the opposite operation which will be called extension. By an *extension* of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ we will understand a set

$$A \otimes B = \{x \in \mathbf{X}_{K\cup L} : x^{\downarrow K} \in A \ \ \& \ \ x^{\downarrow L} \in B\}.$$

Consider a *basic* (*probability or belief*) *assignment* (or just assignment) m on $\mathbf{X}_N$, i.e.

$$m : \mathcal{P}(\mathbf{X}_N) \longrightarrow [0, 1]$$

for which $\sum_{A\subseteq\mathbf{X}_N} m(A) = 1$. For each $K \subset N$ its *marginal basic assignment* is defined (for each $B \subseteq \mathbf{X}_K$):

$$m^{\downarrow K}(B) = \sum_{A\subseteq\mathbf{X}_N : A^{\downarrow K} = B} m(A).$$

Having two basic assignments m_1 and m_2 on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively (we assume that $K, L \subseteq N$), we say that these assignments are *projective* if

$$m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L},$$

which occurs if and only if there exists a basic assignment m on $\mathbf{X}_{K\cup L}$ such that both m_1 and m_2 are marginal assignments of m.

Now, let us start considering how to define composition of two basic assignments. Consider two sets $K, L \subset N$. At this moment we do not pose any restrictions on K and L; they may be but need not be disjoint, one may be subset of the other. We even admit that one or both of them are empty[1]. Let m_1 and m_2 be basic assignments on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively.

Our goal is to define new basic assignment, denoted $m_1 \triangleright m_2$, which will be defined on $\mathbf{X}_{K\cup L}$ and will contain all of the information contained in m_1 and as much as possible of information of m_2 (for the exact meaning see properties (iii) and (iv) of Lemma 1). The required property is met by the following definition.

Definition 1 For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ a *composition* $m_1 \triangleright m_2$ is defined for all $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:

[**a**] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \otimes C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

[**b**] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[1] Notice that basic assignment m on $\mathbf{X}_\emptyset$ is defined $m(\emptyset) = 1$. Let us note that this is the only case where we accept $m(\emptyset) > 0$, otherwise $m(\emptyset) = 0$ according to the classical definitions of basic assignment and belief function, see [6].

[**c**] in all other cases

$$(m_1 \triangleright m_2)(C) = 0.$$

Remark Notice what this definition yields in the following degenerate situations:

- if $K \cap L = \emptyset$ then $m_1 \triangleright m_2 = m_1 \cdot m_2$ (recall that $m_2^{\downarrow\emptyset}(\emptyset) = 1$) — for details regarding this situation see Example 1;
- if $K \supseteq L$ then $m_1 \triangleright m_2 = m_1$.

3 Basic properties of composition

Lemma 1 *For arbitrary two basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ the following properties hold true:*

(i) *$m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K\cup L}$.*

(ii) $(m_1 \triangleright m_2)^{\downarrow K} = m_1$.

(iii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$.

(iv) *If $K \subseteq L$ then $m_2^{\downarrow K} \triangleright m_2 = m_2$.*

Proof. Let us first prove that for any $B \subseteq \mathbf{X}_K$

$$\sum_{A\subseteq\mathbf{X}_{K\cup L}:A^{\downarrow K}=B} (m_1 \triangleright m_2)(A) = m_1(B). \qquad (1)$$

Since, due to Definition 1, $(m_1 \triangleright m_2)(C) = 0$ for any $C \subseteq \mathbf{X}_{K\cup L} \setminus (\mathbf{X}_K \otimes \mathbf{X}_L)$ (in other words for $C \neq C^{\downarrow K} \otimes C^{\downarrow L}$) we see that

$$\begin{aligned}\sum_{A\subseteq\mathbf{X}_{K\cup L}:A^{\downarrow K}=B} &(m_1 \triangleright m_2)(A)\\ &= \sum_{A\subseteq\mathbf{X}_K\otimes\mathbf{X}_L:A^{\downarrow K}=B} (m_1 \triangleright m_2)(A)\\ &= \sum_{C\subseteq\mathbf{X}_L:C^{\downarrow K\cap L}=B^{\downarrow K\cap L}} (m_1 \triangleright m_2)(B \otimes C).\end{aligned}$$

To prove formula (1), we have to distinguish two situations depending on the value of $m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})$. If this value is positive then

$$\begin{aligned}\sum_{A\subseteq\mathbf{X}_{K\cup L}:A^{\downarrow K}=B} &(m_1 \triangleright m_2)(A)\\ &= \sum_{C\subseteq\mathbf{X}_L:C^{\downarrow K\cap L}=B^{\downarrow K\cap L}} \frac{m_1(B)\cdot m_2(C)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})}\\ &= \frac{m_1(B)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})} \sum_{C\subseteq\mathbf{X}_L:C^{\downarrow K\cap L}=B^{\downarrow K\cap L}} m_2(C)\\ &= \frac{m_1(B)}{m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})} m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L})\\ &= m_1(B).\end{aligned}$$

If $m_2^{\downarrow K\cap L}(B^{\downarrow K\cap L}) = 0$ then, according to Definition 1, there exists only one $A \subseteq \mathbf{X}_{K\cup L}$ for which $A^{\downarrow K} = B$ such that $(m_1 \triangleright m_2)(A)$ may be positive; namely $A = B \times \mathbf{X}_{L\setminus K}$. Therefore

$$\begin{aligned}\sum_{A\subseteq\mathbf{X}_{K\cup L}:A^{\downarrow K}=B} &(m_1 \triangleright m_2)(A)\\ &= (m_1 \triangleright m_2)(B \times \mathbf{X}_{L\setminus K})\\ &= m_1(B),\end{aligned}$$

Thus having proved that equality (1) holds true let us start proving assertions (i) – (iv).

ad (i) To prove that $m_1 \triangleright m_2$ is a basic assignment on $\mathbf{X}_{K\cup L}$ we have to show that for each $A \subseteq \mathbf{X}_{K\cup L}$ value $(m_1 \triangleright m_2)(A)$ is nonnegative (which is evident) and that the sum of all these values equals 1. The latter holds true, too, because (using equality (1))

$$\begin{aligned}\sum_{A\subseteq\mathbf{X}_{K\cup L}} &(m_1 \triangleright m_2)(A)\\ &= \sum_{B\subseteq\mathbf{X}_K} \sum_{A\subseteq\mathbf{X}_{K\cup L}:A^{\downarrow K}=B} (m_1 \triangleright m_2)(A)\\ &= \sum_{B\subseteq\mathbf{X}_K} m_1(B) = 1.\end{aligned}$$

ad (ii) The formula is another form of equality (1).

ad (iii) Let us first prove

$$m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L} \implies m_1 \triangleright m_2 = m_2 \triangleright m_1.$$

Consider any $A \subseteq \mathbf{X}_{K\cup L}$. If $A \not\subseteq \mathbf{X}_K \otimes \mathbf{X}_L$ then both $(m_1 \triangleright m_2)(A)$ and $(m_2 \triangleright m_1)(A)$ equal 0. Therefore we have to prove the implication only for $A \subseteq \mathbf{X}_K \otimes \mathbf{X}_L$. If $m_1^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = m_2^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$ then

$$\begin{aligned}(m_1 \triangleright m_2)(A) &= \frac{m_1(A^{\downarrow K})\cdot m_2(A^{\downarrow L})}{m_2^{\downarrow K\cap L}(A^{\downarrow K\cap L})}\\ &= \frac{m_1(A^{\downarrow K})\cdot m_2(A^{\downarrow L})}{m_1^{\downarrow K\cap L}(A^{\downarrow K\cap L})}\\ &= (m_2 \triangleright m_1)(A).\end{aligned}$$

In opposite when $m_1^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = m_2^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = 0$, both $m_1(A^{\downarrow K})$ and $m_2(A^{\downarrow L})$ must equal 0 and therefore (according to Definition 1) $(m_1 \triangleright m_2)(A) = (m_2 \triangleright m_1)(A) = 0$.

To prove the other side of the equivalence (i.e. $m_1 \triangleright m_2 = m_2 \triangleright m_1$ implies $m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$) it is enough to realize that if $m_1^{\downarrow K\cap L} \neq m_2^{\downarrow K\cap L}$ then also $m_1 \triangleright m_2 \neq m_2 \triangleright m_1$ because, due to already proved (item ii) of this assertion, $m_1^{\downarrow K\cap L} = (m_1 \triangleright m_2)^{\downarrow K\cap L}$ and $m_2^{\downarrow K\cap L} = (m_2 \triangleright m_1)^{\downarrow K\cap L}$.

Table 1: Basic assignments m_1 and m_2.

$A \subseteq \mathbf{X}_1$	$m_1(A)$
$\{a_1\}$	0.2
$\{b_1\}$	0.3
$\{a_1b_1\}$	0.5

$A \subseteq \mathbf{X}_2$	$m_2(A)$
$\{a_2\}$	0.6
$\{b_2\}$	0
$\{a_1b_2\}$	0.4

Table 2: Basic assignment $m_1 \triangleright m_2$.

$C \subseteq \mathbf{X}_{\{1,2\}}$	$C = C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}}$	$(m_1 \triangleright m_2)(C)$
$\{a_1a_2\}$	$\{a_1\} \otimes \{a_2\}$	0.12
$\{a_1b_2\}$	$\{a_1\} \otimes \{b_2\}$	0
$\{b_1a_2\}$	$\{b_1\} \otimes \{a_2\}$	0.18
$\{b_1b_2\}$	$\{b_1\} \otimes \{b_2\}$	0
$\{a_1a_2, a_1b_2\}$	$\{a_1\} \otimes \mathbf{X}_2$	0.08
$\{a_1a_2, b_1a_2\}$	$\mathbf{X}_1 \otimes \{a_2\}$	0.3
$\{a_1a_2, b_1b_2\}$		0
$\{a_1b_2, b_1a_2\}$		0
$\{a_1b_2, b_1b_2\}$	$\mathbf{X}_1 \otimes \{b_2\}$	0
$\{b_1a_2, b_1b_2\}$	$\{b_1\} \otimes \mathbf{X}_2$	0.12
$\{a_1a_2, a_1b_2, b_1a_2\}$		0
$\{a_1a_2, a_1b_2, b_1b_2\}$		0
$\{a_1a_2, b_1a_2, b_1b_2\}$		0
$\{a_1b_2, b_1a_2, b_1b_2\}$		0
$\left\{ \begin{matrix} a_1a_2, a_1b_2 \\ b_1a_2, b_1b_2 \end{matrix} \right\}$	$\mathbf{X}_1 \otimes \mathbf{X}_2$	0.2

ad (iv) This property follows directly from previously proved items (iii) and (ii). ∎

Let us now illustrate the operator of composition and its properties by two examples. The first shows what happens when $K \cap L = \emptyset$, the other demonstrates non-commutativity of the operator.

Example 1 Consider two basic assignments m_i (for $i = 1, 2$) on $\mathbf{X}_i = \{a_i, b_i\}$ specified in Table 1.[2] Since, in this case, $K \cap L$ is empty (recall that $m_2^{\downarrow\emptyset}(\emptyset) = 1$), composition simplifies to the expression

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow\{1\}}) \cdot m_2(C^{\downarrow\{2\}}).$$

Using Table 2, where the values of $m_1 \triangleright m_2$ are presented, the reader can easily check that $m_1 = (m_1 \triangleright m_2)^{\downarrow\{1\}}$, and since m_1 and m_2 are trivially projective also $m_2 = (m_1 \triangleright m_2)^{\downarrow\{2\}}$. ♦

[2]Let us note that, for the sake of simplicity, we use in examples $x_1 \dots x_n$ instead of $(x_1, \dots, x_n)$.

Example 2 Let for $i = 1, 2, 3$, $\mathbf{X}_i = \{a_i, b_i\}$ and let us consider the following basic assignments m_1 and m_2 on $\mathbf{X}_1 \times \mathbf{X}_2$ and $\mathbf{X}_2 \times \mathbf{X}_3$, respectively:

$$\begin{aligned} m_1(\mathbf{X}_1 \times \{a_2\}) &= 0.4, \\ m_1(\mathbf{X}_1 \times \mathbf{X}_2) &= 0.6, \\ m_2(\mathbf{X}_2 \times \{a_3\}) &= 0.5, \\ m_2(\mathbf{X}_2 \times \mathbf{X}_3) &= 0.5, \end{aligned}$$

the values of both basic assignments m_1 and m_2 on the remaining subsets being zero. From Definition 1 (case [a]) one can immediately see that both $(m_1 \triangleright m_2)(A)$ and $(m_2 \triangleright m_1)(A)$ can be positive only for those $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ for which

$$A^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \{a_2\} \quad \text{or} \quad A^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \mathbf{X}_2,$$

and

$$A^{\downarrow\{2.3\}} = \mathbf{X}_2 \times \{a_3\} \quad \text{or} \quad A^{\downarrow\{2,3\}} = \mathbf{X}_2 \times \mathbf{X}_3.$$

There are only two such sets

$$A_1 = \mathbf{X}_1 \times \mathbf{X}_2 \times \{a_3\} \quad \text{and} \quad A_2 = \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3.$$

For these sets we get

$$\begin{aligned} &(m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{a_3\}) \\ &\qquad = \tfrac{m_1(\mathbf{X}_1 \times \mathbf{X}_2) \cdot m_2(\mathbf{X}_2 \times \{a_3\})}{m_2^{\downarrow\{2\}}(\mathbf{X}_2)} \\ &\qquad = \tfrac{0.6 \cdot 0.5}{1} = 0.3, \end{aligned}$$

$$\begin{aligned} &(m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3) \\ &\qquad = \tfrac{m_1(\mathbf{X}_1 \times \mathbf{X}_2) \cdot m_2(\mathbf{X}_2 \times \mathbf{X}_3)}{m_2^{\downarrow\{2\}}(\mathbf{X}_2)} \\ &\qquad = \tfrac{0.6 \cdot 0.5}{1} = 0.3, \end{aligned}$$

and similarly

$$\begin{aligned} &(m_2 \triangleright m_1)(\mathbf{X}_1 \times \mathbf{X}_2 \times \{a_3\}) \\ &\qquad = \tfrac{m_2(\mathbf{X}_2 \times \{a_3\}) \cdot m_1(\mathbf{X}_1 \times \mathbf{X}_2)}{m_1^{\downarrow\{2\}}(\mathbf{X}_2)} \\ &\qquad = \tfrac{0.5 \cdot 0.6}{0.6} = 0.5, \end{aligned}$$

$$\begin{aligned} &(m_2 \triangleright m_1)(\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3) \\ &\qquad = \tfrac{m_2(\mathbf{X}_2 \times \mathbf{X}_3) \cdot m_1(\mathbf{X}_1 \times \mathbf{X}_2)}{m_1^{\downarrow\{2\}}(\mathbf{X}_2)} \\ &\qquad = \tfrac{0.5 \cdot 0.6}{0.6} = 0.5. \end{aligned}$$

From case [b] of Definition 1 we will get yet another focal element for $m_1 \triangleright m_2$, namely

$$A_3 = \mathbf{X}_1 \times \{a_2\} \times \mathbf{X}_3,$$

Table 3: Composed basic assignments.

	$(m_1 \triangleright m_2)(A)$	$(m_2 \triangleright m_1)(A)$
A_1	0.3	0.5
A_2	0.3	0.5
A_3	0.4	0

for which

$$A_3^{\downarrow\{1,2\}} = \mathbf{X}_1 \times \{a_2\} \quad \text{and} \quad A_3^{\downarrow\{3\}} = \mathbf{X}_3.$$

Since $m_2^{\downarrow\{2\}}(A_3^{\downarrow\{2\}}) = 0$ and $A_3^{\downarrow\{3\}} = \mathbf{X}_3$ we get

$$\begin{aligned}(m_1 \triangleright m_2)(\mathbf{X}_1 \times \{a_2\} \times \mathbf{X}_3) &= m_1(\mathbf{X}_1 \times \{a_2\}) \\ &= 0.4.\end{aligned}$$

Notice that there does not exist such a focal element for $m_2 \triangleright m_1$, as $m_1^{\downarrow\{2\}}(A_3^{\downarrow\{2\}}) > 0$.

Both the composed basic assignments $m_1 \triangleright m_2$ and $m_2 \triangleright m_1$ are outlined in Table 3 (recall once more that for all other $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$ different from those included in Table 3 both assignments equal to 0). ♦

As said in the Introduction, operator of composition was originally introduced in probability theory. A basic assignment m degenerates into a probability distribution if all its focal elements are singletons (in other words: $m(A) > 0 \implies |A| = 1$).

In agreement with [6] we will call such assignments *Bayesian basic assignments*. It would be strange if the operator of composition we have introduced in this paper would not coincide with the probabilistic one if applied to Bayesian basic assignments. Fortunately, it is not the case.

Lemma 2 *Let m_1 and m_2 be Bayesian basic assignments on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively, for which*

$$m_2{}^{\downarrow K\cap L}(A) = 0 \implies m_1{}^{\downarrow K\cap L}(A) = 0 \qquad (2)$$

for any $A \subseteq \mathbf{X}_{K\cup L}$. Then $m_1 \triangleright m_2$ is a Bayesian basic assignment.

Proof. To prove that a basic assignment $m_1 \triangleright m_2$ is Bayesian, it is enough to show that if $A \subseteq \mathbf{X}_{K\cup L}$ is not a singleton then $(m_1 \triangleright m_2)(A) = 0$.

Consider any $A \subseteq \mathbf{X}_{K\cup L}$, and two different elements $x, y \in A$. Since $x \neq y$ then either $x^{\downarrow K} \neq y^{\downarrow K}$ or $x^{\downarrow L} \neq y^{\downarrow L}$ (or both). Therefore either $A^{\downarrow K}$ or $A^{\downarrow L}$ is not a singleton and therefore $m_1(A^{\downarrow K}) \cdot m_2(A^{\downarrow L}) = 0$. This means that if $m_2{}^{\downarrow K\cap L}(A^{\downarrow K\cap L}) > 0$ then, due to Definition 1, $(m_1 \triangleright m_2)(A) = 0$.

If $m_2{}^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = 0$ then, because we assume the validity of implication (2), $m_1{}^{\downarrow K\cap L}(A^{\downarrow K\cap L}) = 0$ and therefore also $m_1(A^{\downarrow K}) = 0$. Therefore, according to Definition 1, $(m_1 \triangleright m_2)(A) = 0$, too. ∎

Remark The reader should however notice that the definition of the operator of composition for Bayesian basic assignments is not fully equivalent to the definition of composition for probabilistic distributions. They equal to each other only in case that the probabilistic version is defined. This is anchored in Lemma 2 by assuming the implication (2). In case it does not hold, the probabilistic operator is not defined whilst its belief version introduced in this paper is always defined. Nevertheless, in this case, the result is not a Bayesian assignment. We shall illustrate it by a simple example.

Example 3 Let $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ be as in the previous example and consider the following Bayesian basic assignments m_1 and m_2 on $\mathbf{X}_1 \times \mathbf{X}_2$ and $\mathbf{X}_2 \times \mathbf{X}_3$, respectively:

$$\begin{aligned}&m_1(\{a_1a_2\}) = m_1(\{a_1b_2\}) \\ &\qquad = m_1(\{b_1a_2\}) = m_1(\{b_1b_2\}) = 0.25, \\ &m_2(\{a_2a_3\}) = m_2(\{a_2b_3\}) = 0.5, \\ &m_2(\{b_2a_3\}) = m_2(\{b_2b_3\}) = 0.\end{aligned}$$

Let us compute $m_1 \triangleright m_2$ for singletons $\{x_1x_2x_3\} \in \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$. If $x_2 = a_2$ then

$$\begin{aligned}&(m_1 \triangleright m_2)(\{x_1a_2x_3\}) \\ &\qquad = \tfrac{m_1(\{x_1a_2\}) \cdot m_2(\{a_2x_3\})}{m_2^{\downarrow 2}(\{a_2\})} \\ &\qquad = \tfrac{0.25 \cdot 0.5}{1} = 0.125.\end{aligned}$$

For a singleton $\{x_1b_2x_3\}$ we get

$$(m_1 \triangleright m_2)(\{x_1b_2x_3\}) = 0,$$

because $m_2^{\downarrow 2}(\{b_2\}) = 0$. In this case, however, we get

$$\begin{aligned}(m_1 \triangleright m_2)(\{x_1b_2\} \times \mathbf{X}_3) &= m_1(\{x_1b_2\}) \\ &= 0.25.\end{aligned}$$

This means that in this case there are 6 focal elements of $m_1 \triangleright m_2$, namely 4 singletons:

$$\{x_1a_2x_3\}, \text{ for } x_1 \in \mathbf{X}_1, x_3 \in \mathbf{X}_3,$$

and 2 two-element sets

$$\{x_1b_2\} \times \mathbf{X}_3, \text{ for } x_1 \in \mathbf{X}_1.$$

Let us remark that in contrast to $m_1 \triangleright m_2$, $m_2 \triangleright m_1$ is a Bayesian basic assignment, because whenever

Table 4: Basic assignments m_1 and m_2.

$A \subseteq \mathbf{X}_1$	$m_1(A)$
$\{a_1\}$	0.5
$\{a_1, b_1\}$	0.5

$A \subseteq \mathbf{X}_2$	$m_2(A)$
$\{a_2\}$	0.4
$\{a_2, b_2\}$	0.6

$m_1^{\downarrow\{2\}}(x_2) = 0$ then also $m_2^{\downarrow\{2\}}(x_2) = 0$. Basic assignment $m_1 \triangleright m_2$ has 4 focal elements:

$$\begin{aligned}(m_2 \triangleright m_1)(\{a_1a_2a_3\}) &\\ &= (m_2 \triangleright m_1)(\{a_1a_2b_3\}) \\ &= (m_2 \triangleright m_1)(\{b_1a_2a_3\}) \\ &= (m_2 \triangleright m_1)(\{b_1a_2b_3\}) = 0.25. \quad \blacklozenge\end{aligned}$$

Remark In Examples 2 and 3 we showed that the operator of composition is not commutative. From the following example we shall see that this operator is neither associative.

Example 4 Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be as in previous examples and let us consider the following three basic assignments m_1, m_2 defined on $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, as suggested in Table 4 and m_3 have only one focal element, namely

$$m_3(\mathbf{X}_1 \times \mathbf{X}_2) \quad = \quad 1.$$

Then

$$\begin{aligned}(m_1 \triangleright m_2)(\{a_1a_2\}) &= 0.2, \\ (m_1 \triangleright m_2)(\{a_1\} \times \mathbf{X}_2) &= 0.3, \\ (m_1 \triangleright m_2)(\mathbf{X}_1 \times \{a_2\}) &= 0.2, \\ (m_1 \triangleright m_2)(\mathbf{X}_1 \times \mathbf{X}_2) &= 0.3,\end{aligned}$$

due to Definition 1 (the values on remaining sets being again zero) and $(m_1 \triangleright m_2) \triangleright m_3 = m_1 \triangleright m_2$ according to Lemma 1 property (iv). On the other hand

$$\begin{aligned}(m_2 \triangleright m_3)(\mathbf{X}_1 \times \{a_2\}) &= 0.4, \\ (m_2 \triangleright m_3)(\mathbf{X}_1 \times \mathbf{X}_2) &= 0.6.\end{aligned}$$

Now, computing $m_1 \triangleright (m_2 \triangleright m_3)$ we obtain

$$\begin{aligned}(m_1 \triangleright (m_2 \triangleright m_3))(\{a_1\} \times \mathbf{X}_2) &= 0.5, \\ (m_1 \triangleright (m_2 \triangleright m_3))(\mathbf{X}_1 \times \{a_2\}) &= 0.2, \\ (m_1 \triangleright (m_2 \triangleright m_3))(\mathbf{X}_1 \times \mathbf{X}_2) &= 0.3,\end{aligned}$$

which evidently differs from $(m_1 \triangleright m_2) \triangleright m_3$ (see Table 5). $\blacklozenge$

Table 5: Composed basic assignments.

	$(m_1 \triangleright m_2) \triangleright m_3$	$m_1 \triangleright (m_2 \triangleright m_3)$
$\{a_1a_2\}$	0.2	0
$\{a_1\} \times \mathbf{X}_2$	0.3	0.5
$\mathbf{X}_1 \times \{a_2\}$	0.2	0.2
$\mathbf{X}_1 \times \mathbf{X}_2$	0.3	0.3

4 Advanced properties of composition

In this section we are going to study properties which were proved for probabilistic version of the operator of composition and which are applied when proving important theorems regarding compositional models. Unless expressed explicitly otherwise in this section we will assume m_1, m_2, m_3 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \mathbf{X}_{K_3}$, respectively.

Lemma 3 *Let m_1, m_2, m_3 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \mathbf{X}_{K_3}$, respectively. If $K_1 \supseteq (K_2 \cap K_3)$ then*

$$(m_1 \triangleright m_2) \triangleright m_3 = (m_1 \triangleright m_3) \triangleright m_2. \tag{3}$$

Proof. The goal is to prove that for any $C \subseteq \mathbf{X}_{K_1 \cup K_2 \cup K_3}$

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = ((m_1 \triangleright m_3) \triangleright m_2)(C). \tag{4}$$

We will have to distinguish five special cases.

A. $C \neq C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$.
This is the simplest situation because in this case both sides of formula (4) equal 0 due to Definition 1 (case [c]).

B. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$ & $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}), m_3^{\downarrow K_1 \cap K_3}(C^{\downarrow K_1 \cap K_3}) > 0$.
In this case it is enough to realize that (under the given assumptions) $K_3 \cap (K_1 \cup K_2) = K_3 \cap K_1$ and, analogously, $K_2 \cap (K_1 \cup K_3) = K_2 \cap K_1$. Then we see that both sides of formula (4) again coincide:

$$\begin{aligned}((m_1 \triangleright m_2) \triangleright m_3)(C) &\\ &= \frac{m_1(C^{\downarrow K_1}) \cdot m_2(C^{\downarrow K_2})}{m_2^{\downarrow K_2 \cap K_1}(C^{\downarrow K_2 \cap K_1})} \\ &\quad \cdot \frac{m_3(C^{\downarrow K_3})}{m_3^{\downarrow K_3 \cap (K_1 \cup K_2)}(C^{\downarrow K_3 \cap (K_1 \cup K_2)})},\end{aligned}$$

$$\begin{aligned}((m_1 \triangleright m_3) \triangleright m_2)(C) &\\ &= \frac{m_1(C^{\downarrow K_1}) \cdot m_3(C^{\downarrow K_3})}{m_3^{\downarrow K_3 \cap K_1}(C^{\downarrow K_3 \cap K_1})} \\ &\quad \cdot \frac{m_2(C^{\downarrow K_2})}{m_2^{\downarrow K_2 \cap (K_1 \cup K_3)}(C^{\downarrow K_2 \cap (K_1 \cup K_3)})}.\end{aligned}$$

C. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$, $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) > 0 = m_3^{\downarrow K_1 \cap K_3}(C^{\downarrow K_1 \cap K_3})$. In this case, if $C^{\downarrow K_3 \setminus K_1} \neq \mathbf{X}_{K_3 \setminus K_1}$ then both sides of formula (4) equal 0, because, due to Definition 1, both assignments $m_1 \triangleright m_2$ and $(m_1 \triangleright m_3) \triangleright m_2$ equal 0. Therefore consider $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes \mathbf{X}_{K_3 \setminus K_1}$. For this we get from Definition 1

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = (m_1 \triangleright m_2)(C^{\downarrow K_1 \cup K_2}).$$

For the right-hand side of formula (4) we get

$$(m_1 \triangleright m_3)(C^{\downarrow K_1 \cup K_3}) = m_1(C^{\downarrow K_1})$$

and therefore

$$((m_1 \triangleright m_3) \triangleright m_2)(C) = (m_1 \triangleright m_2)(C^{\downarrow K_1 \cup K_2}).$$

D. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$, $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) = 0 < m_3^{\downarrow K_1 \cap K_3}(C^{\downarrow K_1 \cap K_3})$. The proof is analogous to that under item C.

E. $C = C^{\downarrow K_1} \otimes C^{\downarrow K_2} \otimes C^{\downarrow K_3}$, $m_2^{\downarrow K_1 \cap K_2}(C^{\downarrow K_1 \cap K_2}) = 0 = m_3^{\downarrow K_1 \cap K_3}(C^{\downarrow K_1 \cap K_3})$. It is obvious from Definition 1 that both sides of formula (4) equal 0 for all C but for $C = C^{\downarrow K_1} \otimes \mathbf{X}_{K_2 \setminus K_1} \otimes \mathbf{X}_{K_3 \setminus K_1}$. For this special case, however,

$$((m_1 \triangleright m_2) \triangleright m_3)(C) = m_1(C^{\downarrow K_1}),$$
$$((m_1 \triangleright m_3) \triangleright m_2)(C) = m_1(C^{\downarrow K_1}). \quad \blacksquare$$

Lemma 4 *Let m_1, m_2 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$, respectively. If $K_1 \cup K_2 \supseteq L \supseteq K_1$ then*

$$(m_1 \triangleright m_2)^{\downarrow L} = m_1 \triangleright m_2^{\downarrow K_2 \cap L}.$$

Proof. Consider first $B \subseteq \mathbf{X}_L$ such that $m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2}) > 0$. For this B we get

$$\begin{aligned}
&(m_1 \triangleright m_2)^{\downarrow L}(B) \\
&\quad = \sum_{A \subseteq \mathbf{X}_{K_1 \cup K_2} : A^{\downarrow L} = B} (m_1 \triangleright m_2)(A) \\
&\quad = \sum_{A \subseteq \mathbf{X}_{K_1} \otimes \mathbf{X}_{K_2} : A^{\downarrow L} = B} (m_1 \triangleright m_2)(A) \\
&\quad = \sum_{A \subseteq \mathbf{X}_{K_1} \otimes \mathbf{X}_{K_2} : A^{\downarrow L} = B} \frac{m_1(A^{\downarrow K_1}) \cdot m_2(A^{\downarrow K_2})}{m_2^{\downarrow K_1 \cap K_2}(A^{\downarrow K_1 \cap K_2})} \\
&\quad = \sum_{C \subseteq \mathbf{X}_{K_2} : C^{\downarrow L \cap K_2} = B^{\downarrow L \cap K_2}} \frac{m_1(B^{\downarrow K_1}) \cdot m_2(C)}{m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2})} \\
&\quad = \frac{m_1(B^{\downarrow K_1})}{m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2})} \sum_{C \subseteq \mathbf{X}_{K_2} : C^{\downarrow L \cap K_2} = B^{\downarrow L \cap K_2}} m_2(C) \\
&\quad = \frac{m_1(B^{\downarrow K_1}) m_2^{\downarrow L \cap K_2}(B^{\downarrow L \cap K_2})}{m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2})} \\
&\quad = (m_1 \triangleright m_2^{\downarrow L \cap K_2})(B).
\end{aligned}$$

If $m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2}) = 0$ for some $B \subseteq \mathbf{X}_L$, then there is only one $A \subseteq \mathbf{X}_{K_1 \cup K_2}$ such that $A^{\downarrow K_1} = B^{\downarrow K_1}$ for which $(m_1 \triangleright m_2)(A)$ may be positive, namely $A^* = B^{\downarrow K_1} \otimes \mathbf{X}_{K_2 \setminus K_1}$ with $(m_1 \triangleright m_2)(A^*) = m_1(B^{\downarrow K_1})$. Thus if $B = B^{\downarrow K_1} \otimes \mathbf{X}_{L \setminus K_1}$,

$$\begin{aligned}
(m_1 \triangleright m_2)^{\downarrow L}(B) &= \sum_{A \subseteq \mathbf{X}_{K_1 \cup K_2} : A^{\downarrow L} = B} (m_1 \triangleright m_2)(A) \\
&= (m_1 \triangleright m_2)(A^*) = m_1(B^{\downarrow K_1}) \\
&= (m_1 \triangleright m_2^{\downarrow K_2 \cap L})(A^{*\downarrow L}) \\
&= (m_1 \triangleright m_2^{\downarrow K_2 \cap L})(B).
\end{aligned}$$

If $B \neq B^{\downarrow K_1} \otimes \mathbf{X}_{L \setminus K_1}$ and $m_2^{\downarrow K_1 \cap K_2}(B^{\downarrow K_1 \cap K_2}) = 0$ then

$$(m_1 \triangleright m_2)^{\downarrow L}(B) = 0 = (m_1 \triangleright m_2^{\downarrow K_2 \cap L})(B). \quad \blacksquare$$

Lemma 5 *Let m_1, m_2 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$, respectively. If $K_1 \cup K_2 \supseteq L \supseteq K_1 \cap K_2$ then*

$$(m_1 \triangleright m_2)^{\downarrow L} = m_1^{\downarrow K_1 \cap L} \triangleright m_2^{\downarrow K_2 \cap L}.$$

Proof. We will compute the required marginal assignment in two steps. In the first step we will employ Lemma 4, then (iv) of Lemma 1 and finally Lemma 3:

$$\begin{aligned}
(m_1 \triangleright m_2)^{\downarrow K_1 \cup L} &= m_1 \triangleright m_2^{\downarrow K_2 \cap L} \\
&= (m_1^{\downarrow K_1 \cap K_2} \triangleright m_1) \triangleright m_2^{\downarrow K_2 \cap L} \\
&= (m_1^{\downarrow K_1 \cap K_2} \triangleright m_2^{\downarrow K_2 \cap L}) \triangleright m_1.
\end{aligned}$$

The last expression will be further marginalized with the help of Lemma 4 and afterwards the final form will be received with application of Lemma 3 and (iv) of Lemma 1.

$$\begin{aligned}
(m_1 \triangleright m_2)^{\downarrow L} &= \left((m_1^{\downarrow K_1 \cap K_2} \triangleright m_2^{\downarrow K_2 \cap L}) \triangleright m_1\right)^{\downarrow L} \\
&= (m_1^{\downarrow K_1 \cap K_2} \triangleright m_2^{\downarrow K_2 \cap L}) \triangleright m_1^{\downarrow K_1 \cap L} \\
&= (m_1^{\downarrow K_1 \cap K_2} \triangleright m_1^{\downarrow K_1 \cap L}) \triangleright m_2^{\downarrow K_2 \cap L} \\
&= m_1^{\downarrow K_1 \cap L} \triangleright m_2^{\downarrow K_2 \cap L}. \quad \blacksquare
\end{aligned}$$

Lemma 6 *Let m_1, m_2 be basic assignments on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}$, respectively. Then*

$$m_1 \triangleright m_2 = m_1 \triangleright (m_1 \triangleright m_2)^{\downarrow K_2}.$$

Proof. Due to (ii) of Lemma 1 assignments m_1 and $(m_1 \triangleright m_2)^{\downarrow K_2}$ are projective and therefore (due to property (iii) of the same lemma) these arguments may be commuted

$$\begin{aligned}
m_1 \triangleright (m_1 \triangleright m_2)^{\downarrow K_2} &= (m_1 \triangleright m_2)^{\downarrow K_2} \triangleright m_1 \\
&= (m_1^{\downarrow K_1 \cap K_2} \triangleright m_2) \triangleright m_1,
\end{aligned}$$

where the last modification is made on the basis of Lemma 5. The last expression meets the assumptions of Lemma 3 and therefore we can exchange second and third arguments, from which the required expression is got by application of (iv) of Lemma 1:

$$(m_1^{\downarrow K_1\cap K_2} \triangleright m_2) \triangleright m_1 = (m_1^{\downarrow K_1\cap K_2} \triangleright m_1) \triangleright m_2 = m_1 \triangleright m_2. \quad \blacksquare$$

5 Compositional models

Now we are starting to consider repetitive application of the operator of composition with the goal to create a multidimensional model. Since the operator is neither commutative nor associative we have always to specify in which order the oligodimensional assignments are composed together. To make the formulas more lucid we will omit brackets in case that the operator is to be applied from left to right, i.e., in what follows

$$m_1 \triangleright m_2 \triangleright m_3 \triangleright \ldots \triangleright m_{n-1} \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots \triangleright m_{n-1}) \triangleright m_n.$$

Moreover, we will always assume m_i be basic assignment on $\mathbf{X}_{K_i}$.

The reader familiar with some papers on probabilistic or possibilistic compositional models knows that one of the most important notions of this theory is that of a so-called *perfect sequence*, which will be now introduced also for a sequence of basic assignments.

Definition 2 A generating sequence of basic assignments m_1, $m_2, \ldots, m_n$ is called *perfect* if

$$\begin{aligned} m_1 \triangleright m_2 &= m_2 \triangleright m_1, \\ m_1 \triangleright m_2 \triangleright m_3 &= m_3 \triangleright (m_1 \triangleright m_2), \\ &\vdots \\ m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n &= m_n \triangleright (m_1 \triangleright \ldots \triangleright m_{n-1}). \end{aligned}$$

From the practical point of view it is also important to have a tool enabling us to recognize whether a generating sequence is perfect or not. For this one can take advantage of the following assertion.

Lemma 7 *A generating sequence $m_1, m_2, \ldots, m_n$ is perfect* iff *the pairs of basic assignments m_j and $(m_1 \triangleright \ldots \triangleright m_{j-1})$ are projective, i.e. if*

$$m_j^{\downarrow K_j\cap(K_1\cup\ldots\cup K_{j-1})} = (m_1 \triangleright \ldots \triangleright m_{j-1})^{\downarrow K_j\cap(K_1\cup\ldots\cup K_{j-1})},$$

for all $j = 2, 3, \ldots, n$.

Proof. This assertion is proved just by a multiple application of assertion (iii) of Lemma 1:

$$\begin{aligned} &m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K_2\cap K_1} = m_2^{\downarrow K_2\cap K_1}, \\ &m_1 \triangleright m_2 \triangleright m_3 = m_3 \triangleright (m_1 \triangleright m_2) \\ &\quad \iff (m_1 \triangleright m_2)^{\downarrow K_3\cap(K_1\cup K_2)} = m_3^{\downarrow K_3\cap(K_1\cup K_2)}, \\ &\quad \vdots \\ &m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n = m_n \triangleright (m_1 \triangleright \ldots \triangleright m_{n-1}) \\ &\quad \iff (m_1 \triangleright \ldots \triangleright m_{n-1})^{\downarrow K_n\cap(K_1\cup\ldots\cup K_{n-1})} \\ &\qquad = m_n^{\downarrow K_n\cap(K_1\cup\ldots\cup K_{n-1})}. \quad \blacksquare \end{aligned}$$

From Definition 2 one can hardly see what are the properties of the perfect sequences; the main one is expressed by the following characterization theorem.

Theorem 1 *A generating sequence of basic assignments $m_1, m_2, \ldots, m_n$ is perfect* iff *all the assignments from this sequence are marginal to the composed basic assignment $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$:*

$$(m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n)^{\downarrow K_j} = m_j,$$

for all $j = 1, \ldots, m$.

Proof. The fact that all assignments m_j from a perfect sequence are marginals of $(m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n)$ follows from the fact that $(m_1 \triangleright \ldots \triangleright m_j)$ is marginal to $(m_1 \triangleright \ldots \triangleright m_n)$ (due to (ii) of Lemma 1) and m_j is marginal to $m_j \triangleright (m_1 \triangleright \ldots \triangleright m_{j-1}) = m_1 \triangleright \ldots \triangleright m_j$.

Suppose now that for all $j = 1, \ldots, n$, m_j are marginal assignments to $m_1 \triangleright \ldots \triangleright m_n$. It means that all the assignments from the sequence are pairwise projective, and that each m_j is projective with any marginal assignment of $m_1 \triangleright \ldots \triangleright m_n$, and consequently also with $m_1 \triangleright \ldots \triangleright m_{j-1}$. So we get that

$$m_j^{\downarrow K_j\cap(K_1\cup\ldots\cup K_{j-1})} = (m_1 \triangleright \ldots \triangleright m_{j-1})^{\downarrow K_j\cap(K_1\cup\ldots\cup K_{j-1})}$$

for all $j = 2, \ldots, n$, which is equivalent, due to Lemma 7, to the fact that $m_1, \ldots, m_n$ is perfect. $\blacksquare$

Graphical Markov models (or rather decomposable models) are recalled by the following (almost trivial) assertion, which resembles assertions concerning decomposable models.

Theorem 2 *Let a generating sequence of pairwise projective assignments $m_1, m_2, \ldots, m_n$ be such that $K_1, K_2, \ldots, K_n$ meets the well-known running intersection property:*

$$\forall j = 2, 3, \ldots, n \quad \exists \ell (1 \leq \ell < j) \\ \text{such that} \quad K_j \cap (K_1 \cup \ldots \cup K_{j-1}) \subseteq K_\ell.$$

Then $m_1, m_2, \ldots, m_n$ is perfect.

Proof. Due to Lemma 7 it is enough to show that for each $j = 2, \ldots, n$ basic assignment m_j and the composed assignment $m_1 \triangleright \ldots \triangleright m_{j-1}$ are projective. Let us prove it by induction.

For $j = 2$ the required projectivity is guaranteed by the fact that we assume pairwise projectivity of all $m_1, \ldots, m_n$. So we have to prove it for general $j > 2$ under the assumption that the assertion holds for $j - 1$, which means (due to Theorem 1) that all $m_1, m_2, \ldots, m_{j-1}$ are marginal to $m_1 \triangleright \ldots \triangleright m_{j-1}$. Since we assume that $K_1, \ldots, K_n$ meets the running intersection property, there exists $\ell \in \{1, 2, \ldots j-1\}$ such that $K_j \cap (K_1 \cup \ldots \cup K_{j-1}) \subseteq K_\ell$. Therefore $(m_1 \triangleright \ldots \triangleright m_{j-1})^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})}$ and $m_\ell^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})}$ are the same marginals of $m_1 \triangleright \ldots \triangleright m_{j-1}$ and therefore they have to equal to each other:

$$(m_1 \triangleright \ldots \triangleright m_{j-1})^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})} \\ = m_\ell^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})}.$$

However we assume that m_j and m_ℓ are projective and therefore also

$$(m_1 \triangleright \ldots \triangleright m_{j-1})^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})} \\ = m_j^{\downarrow K_j \cap (K_1 \cup \ldots \cup K_{j-1})}. \quad \blacksquare$$

It should be stressed at this moment that running intersection property of $K_1, K_2, \ldots, K_n$ is a sufficient condition guaranteeing a perfectness of a generating sequence of pairwise projective assignments. By no means this condition is necessary as it will be shown in the following example.

Example 5 Simple example is given by two basic assignments m_1 and m_2 from Example 1 (recall that they are defined on $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and their values can be found in Table 1) and the third assignment $m_3 = m_1 \triangleright m_2$ (see Table 2). Considering sequence m_1, m_2, m_3, it is evident that $K_1 = \{1\}, K_2 = \{2\}, K_3 = \{1, 2\}$ do not meet the running intersection property. And yet the sequence m_1, m_2, m_3 is perfect because all the assignments are marginal (or equal) to $m_1 \triangleright m_2 \triangleright m_3$. Notice that if we chose any other basic assignment $\hat{m}_3$ on $\mathbf{X}_{\{1,2\}}$ different from $m_3 = m_1 \triangleright m_2$, the generating sequence $m_1, m_2, \hat{m}_3$ would not be perfect any more. So we see that perfectness of a sequence is not only a structural property connected with the properties of $K_1, K_2, \ldots, K_n$ but depends also on specific values of the respective basic assignments. ♦

The last assertion shows that each generating sequence defining a compositional model $m_1 \triangleright \ldots \triangleright m_n$ can be transformed into a perfect sequence. It means, any basic assignment representable by a generating sequence $m_1, m_2, \ldots, m_n$ can be represented also by a perfect sequence $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_n$

Theorem 3 *For any generating sequence $m_1, m_2, \ldots, m_n$ the sequence $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_n$ computed by the following process*

$$\begin{aligned} \hat{m}_1 &= m_1, \\ \hat{m}_2 &= \hat{m}_1^{\downarrow K_2 \cap K_1} \triangleright m_2, \\ \hat{m}_3 &= (\hat{m}_1 \triangleright \hat{m}_2)^{\downarrow K_3 \cap (K_1 \cup K_2)} \triangleright m_3, \\ &\vdots \\ \hat{m}_n &= (\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_{n-1})^{\downarrow K_n \cap (K_1 \cup \ldots K_{n-1})} \triangleright m_n \end{aligned}$$

is perfect and

$$m_1 \triangleright \ldots \triangleright m_n = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_n.$$

Proof. The perfectness of the sequence $\hat{m}_1, \ldots, \hat{m}_n$ follows immediately from Lemma 7 and from the definition of this sequence as

$$\hat{m}_i^{\downarrow K_i \cap (K_1 \cup \ldots \cup K_{i-1})} \\ = (\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_{i-1})^{\downarrow K_i \cap (K_1 \cup \ldots \cup K_{i-1})}$$

yields projectivity of $(\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_{i-1})$ and $\hat{m}_i$.

Let us prove

$$m_1 \triangleright \ldots \triangleright m_n = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_n$$

by mathematical induction. Since $m_1 = \hat{m}_1$ by definition, it is enough to show that

$$m_1 \triangleright \ldots \triangleright m_i = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i$$

implies also

$$m_1 \triangleright \ldots \triangleright m_{i+1} = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_{i+1}.$$

In the following computations we will use the fact that due to Lemma 5

$$(\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i)^{\downarrow K_{i+1} \cap (K_1 \cup \ldots K_i)} \triangleright m_{i+1} \\ = ((\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i) \triangleright m_{i+1})^{\downarrow K_{i+1}}$$

and afterwards we will employ Lemma 6.

$$\begin{aligned}
&\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_{i+1} \\
&\quad = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i \triangleright \\
&\qquad \left((\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i)^{\downarrow K_{i+1} \cap (K_1 \cup \ldots K_i)} \triangleright m_{i+1} \right) \\
&\quad = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i \triangleright ((\hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i) \triangleright m_{i+1})^{\downarrow K_{i+1}} \\
&\quad = \hat{m}_1 \triangleright \ldots \triangleright \hat{m}_i \triangleright m_{i+1} = m_1 \triangleright \ldots \triangleright m_i \triangleright m_{i+1},
\end{aligned}$$

where the last modification is an application of the inductive assumption. ∎

6 Conclusions

Graphical Markov Models were designed to enable description of real-life problems by probabilistic models. Since we are getting into problems when coping with computational complexity of probabilistic models, all the more so problems naturally appear when applying belief function models, for which there do not exist distribution functions; we have to represent them by set functions defined on the whole power set of the frame of discernment $\Omega = \mathbf{X}_N$. So, inspired by the original probabilistic approach the paper is the first attempt to build up compositional models of multidimensional belief functions. We have defined belief function operator of composition manifesting all the main characteristics of its probabilistic pre-image. Even more, there is one point in which the belief function operator of composition is superior to the probabilistic one: thanks to the ability of belief functions to model total ignorance, the operator of composition is for basic assignments always defined, which is not the case in the probabilistic framework.

In the paper we have proved the basic properties of the operator necessary to introduce compositional models and their most important special case, perfect sequence models. Naturally, there are still many open problems to be solved. The most important one is a design of efficient computational procedures for this type of models. It is also necessary to clarify interrelations between the operator of composition and conditional independence. This problem is not easy because in the framework of belief functions there exist several notions corresponding to stochastic conditional independence.

At this moment we know very little about similarities and differences between the described compositional models and other multidimensional models such as [1, 2, 7], as well as about the relation between the compositional models developed for belief functions and those introduced in possibility theory [8, 9].

Acknowledgements

The research was partially supported by GA AV ČR under grants A2075302 (Jiroušek), A100750603 (Vejnarová) and 1ET100300419 (Daniel) and MŠMT under grants 1M0572 and 2C06019 (Jiroušek, Vejnarová).

References

[1] R. G. Almond. *Graphical Belief Modelling.* Chapman & Hall, London, 1995.

[2] E. Ben Yaghlane, Ph. Smets, and K. Mellouli. Directed evidential networks with conditional belief functions. *Proc. of ECSQARU 2003*, LNAI 2711, Springer, pp. 291–305, 2003.

[3] R. Jiroušek. Composition of probability measures on finite spaces. *Proc. of UAI'97*, (D. Geiger and P. P. Shenoy, eds.). Morgan Kaufmann Publ., San Francisco, California, pp. 274–281, 1997.

[4] R. Jiroušek. Graph modelling without graphs. *Proc. of IPMU'98*, (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, pp. 809–816, 1988.

[5] R. Jiroušek. Marginalization in composed probabilistic models. *Proc. of UAI'00* (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California, pp. 301–308, 2000.

[6] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[7] P. P. Shenoy. Binary join trees for computing marginals in the Shenoy-Shafer architecture. *Int. J. of Approximate Reasoning*, **17**, (2-3), pp. 239–263, 1997.

[8] J. Vejnarová. Composition of possibility measures on finite spaces: preliminary results. In: *Proc. of IPMU'98,* (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, 1998, pp. 25–30.

[9] J. Vejnarová. Possibilistic independence and operators of composition of possibility measures. In: M. Hušková , J. Á. Víšek, P. Lachout (eds.) *Prague Stochastics'98,* JČMF, 1998, pp. 575–580.

[10] H. Xu, and Ph. Smets. Reasoning in evidential networks with conditional belief functions. *Int. J. of Approximate Reasoning*, **14**, (2-3), pp. 155–185, 1996.

Enhancement of Natural Extension

Igor Kozine
Risø National Laboratory
Denmark
igor.kozine@risoe.dk

Victor Krymsky
Ufa State Aviation Technical University
Russia
kvg@mail.rb.ru

Abstract

The theory of imprecise previsions admits the use of a wide variety of statistical evidence. Nevertheless, some existing evidence, for example, in reliability applications, cannot be utilized by models developed within its framework. In the pursuit of reducing imprecision, any available evidence should become an input to modeling. It is suggested to take a different look at the natural extension, the basic constructive step in the theory. It is shown that natural extension can be viewed as a problem belonging to the realm of variational calculus, which opens up new perspectives for obtaining tighter intervals.

Keywords. Imprecise probability, statistical reasoning, natural extension, variational calculus, reliability analysis

1 Introduction

In spite of the existence of a number of risk/reliability and other applied models built on imprecise statistical reasoning, only a few of them have ever been used in practice – and then only hesitantly –, the rest remaining firmly in the academic realm. Do they lack adequate promotion by their practitioners, or are there other primary obstacles that prevent them from being widely applied? We believe that the main obstacle to the practical application of imprecise statistical models is thoroughly familiar to the group of experts who practise interval computations: it is namely the rapid growth in imprecision that occurs when intervals are propagated through mathematical models.

Should this state of affairs be regarded as unalterable, or can this weakness in the model be remedied? If the growth in imprecision is due to a deficiency in the model, what is its basic cause in mathematical terms, and how can we attempt to develop a more adequate model?

A cause of the large imprecision in computed previsions should be sought in the mechanism producing the previsions. It is called natural extension and it may be seen as the basic constructive step in statistical reasoning; it enables us to construct new coherent previsions from old ones [1].

Natural extension can appear in different forms. Four forms of it were described in [2]. Each of them has pros and cons in the context of a specific application. The use of a proper form can substantially facilitate inference and computation of the probability measures of interest.

We suggest taking a different look at the natural extension, an approach which opens up new perspectives for obtaining tighter intervals.

It is shown that natural extension can be viewed as the problem of finding an extremal of a functional, a problem which belongs to the realm of variational calculus. If this path is followed, the modeller can utilise more versatile information than is possible with the natural extension suggested by Walley [1] and Kuznetsov [3]. For example, as demonstrated in this paper, bounds on probability density functions and their derivatives can be utilised by the new form of natural extension, which is an effective way of obtaining tighter bounds of statistical measures.

2 Different Forms of Natural Extension

Suppose there is a continuous random variable, for example, a lifetime X of a component or system defined on the sample space $[0,T]$ and information about this variable is represented as a set of n interval-valued expectations of functions $f_1(X),...,f_n(X)$. Denote these expectations $\overline{a}_i = \overline{M}(f_i)$ and $\underline{a}_i = \underline{M}(f_i)$, $i = 1,...,n$, where $\overline{a}_i$ and $\underline{a}_i$ upper and lower bounds for the expectations, correspondingly. For computing new expected values $\overline{M}(g)$ and $\underline{M}(g)$ of a function $g(X)$ from the available information, natural extension can be used in the following primal form:

$$\left.\begin{aligned} \underline{M}(g) &= \inf_{P} \int_0^T g(x)\rho(x)dx \\ \overline{M}(g) &= \sup_{P} \int_0^T g(x)\rho(x)dx \end{aligned}\right\} \quad (1)$$

subject to

$$\left.\begin{aligned} \underline{a}_i \le \int_0^T f_i(x)\rho(x)dx \le \overline{a}_i, i \le n \\ \int_0^T \rho(x)dx = 1, \rho(x) \ge 0 \end{aligned}\right\} \quad (2)$$

Here the infimum and supremum are taken over the set P of all admissible (matching the constraints) probability density functions $\rho(x)$ satisfying conditions (2). Solutions (1) exist if all the constraints (2) form a non-empty subset $P_0 \subseteq P$. If the subset P_0 is empty, this means that the set of evidence is conflicting. If all the evidence is interval-valued (this is a particular case of imprecise evidence), then two interval-valued judgements on the same prevision are called conflicting if they do not intersect.

It should be noted that problems (1)-(2) are linear and the dual optimization problems can be written for them. For $\underline{M}(g)$, for example, the dual problem is the following [2], [3]:

$$\underline{M}(g) = \sup_{c_0, c_i, d_i} \left(c_0 + \sum_{i=1}^{n} \left(c_i \underline{a}_i - d_i \overline{a}_i \right) \right)$$

subject to $c_0 \in \mathbf{R}$, $c_i, d_i \in \mathbf{R}_+$, $i = 1,...,n$, and for any $0 \le x \le T$, $c_0 + \sum (c_i - d_i) f_i(x) \ge g(x)$

Values $\overline{M}(g)$ and $\underline{M}(g)$ are often called upper and lower previsions and functions $f_i(X)$ and $g(X)$ are called gambles. Note that the lower and upper previsions $\overline{M}(g)$ and $\underline{M}(g)$ can be regarded as the bounds for an unknown precise prevision $M(g)$ which is called a linear prevision.

Natural extension is a general mathematical procedure for calculating new previsions from initial judgements. It produces a coherent overall model from a certain collection of imprecise probability judgements and may be seen as the basic constructive step in interval-valued statistical reasoning.

The crux of optimisation problems (1)-(2) is that their solutions obtained as a result of solving linear programs are defined on the family of degenerate probability distributions[1], which are included on equal footing in the set of all admissible probability distributions over which the solution is sought. As proven in [2], solving these optimisation problems on the set of all admissible probability distributions gives the same solution as that obtained on only the set of degenerate distributions:

$$\rho^*(x) = \sum_{k=1}^{n+1} c_k \delta(x, x_k), \quad (3)$$

where $c_k \in \mathbf{R}_+$, $\sum_{k=1}^{n+1} c_k = 1$, and $\delta(x, x_k)$ is the Dirac function which has unit area concentrated in the immediate vicinity of point x_k.

By substituting the degenerate class of densities (3) into objective function (1) for $\underline{M}(g)$ and constraints (2) we obtain

$$\underline{M}(g) = \inf_{c_k, x_k} \sum_{k=1}^{n+1} c_k g(x_k) \quad (4)$$

subject to

$$\left.\begin{aligned} \sum_{k=1}^{n+1} c_k = 1, \quad c_k \ge 0, k = 1,..., n+1, \\ \underline{a}_i \le \sum_{k=1}^{n+1} c_k f_i(x_k) \le \overline{a}_i, i \le n \end{aligned}\right\} \quad (5)$$

We refer to the natural extension (4)-(5) as the degenerate form.

All this would simply be mathematical subtlety – that is, of little interest to practitioners – if it did not give us a clue to deriving more precise previsions of interest for continuous random variables. For some variables it is often not realistic to assume that the probability masses are concentrated in a few points as opposed to being continuously distributed over the set of possible outcomes. In reliability applications, probability masses of time to failure cannot (except in very special cases) concentrate in a very few points of the positive real line. Ignoring this evidence is one of the causes (we hold it to be the root cause) of high imprecision in reliability applications as well as in other applications.

Example 1.

The sample set of a continuous random variable X is an interval $[0, T]$. The only available information about X is point-valued probability b of finding its value within an

[1] The probability distribution of a continuous random variable is referred to as degenerate if the probability masses are concentrated in a finite number of points belonging to the continuous set of possible states

interval $[\underline{q},\overline{q}] \subseteq [0,\ T]$. That is, $\Pr(x \in [\underline{q},\overline{q}]) = b$. What are the lower and upper bounds for the expected value of X?

Natural extension in its primal form appears as follows:

$$\underline{M}(X) = \min_{\mathrm{P}} \int_0^T x\rho(x)\mathrm{d}x \quad \text{subject to}$$

$$\int_0^T \rho(x)\mathrm{d}x = 1,\ \rho(x) \geq 0, \text{ and } \int_0^T I_{[\underline{q},\overline{q}]}(x)\rho(x)dx = b,$$

where $I_{[\underline{q},\overline{q}]}(x) = 1$ if $x \in [\overline{q},\underline{q}]$ and $I_{[\underline{q},\overline{q}]}(x) = 0$ otherwise.

Its counterpart in the degenerate form, as follows from (4)-(5), is the optimization problem

$$\underline{M}(X) = \inf_{c_i, x_i}(c_1 x_1 + c_2 x_2) \quad \text{subject to}$$

$$c_1 + c_2 = 1, c_i \geq 0 \text{ and } c_1 I_{[\underline{q},\overline{q}]}(x_1) + c_2 I_{[\underline{q},\overline{q}]}(x_2) = b\ .$$

From the constraints it can be concluded that $c_1 I_{[\underline{q},\overline{q}]}(x_1) + c_2 I_{[\underline{q},\overline{q}]}(x_2) = b$ holds only if $x_1 \notin [\underline{q},\overline{q}]$, $x_2 \in [\underline{q},\overline{q}]$ and $c_2 = b$, which entails $c_1 = 1 - b$. Plugging c_1 and c_2 into the objective function brings us to the simple optimisation problem

$$\underline{M}(X) = \inf_{x_i}\left((1-b)x_1 + bx_2\right).$$

The infimum is attained with $x_1 = 0$ and $x_2 = \underline{q}$, that is, $\underline{M}(X) = b\underline{q}$.

Thus, the probability distribution function delivering the infimum to the objective function degenerates into the one with probability masses concentrated in two points $x_1 = 0$ and $x_2 = \underline{q}$ with masses (1-*b*) and *b*, correspondingly. This case is presented in Fig. 1.

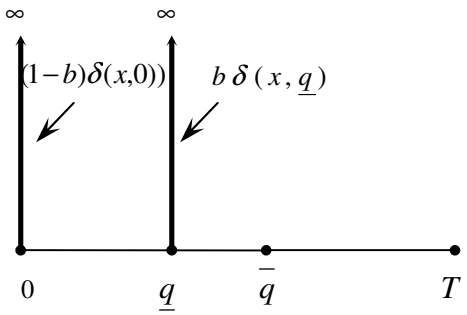


Figure 1: Degenerate probability distribution function providing the infimum to the objective function

3 An example Where Natural Extension Fails to Utilise Evidence

An attempt to mitigate the influence of degenerate probability distributions on the solutions and to obtain tighter bounds was undertaken in [4]. No significant effect was obtained through the introduction of judgements on the skewness and unimodality of the distributions as, in this case, the peaks of degenerate distributions simply become repositioned and probability masses become redistributed among the peaks. The nature of the distributions defining the solutions remained unchanged.

In the attempt to achieve tighter bounds, it seems natural to try to constrain the underlying probability distributions such that they rule out the degenerate distributions. This can be done through imposing a restriction on the upper bound of the probability density function. (This device is not new and was used, for example, in [5] and [6].) In some practical situations, such bounds can be elicited from experts. For example, in reliability applications, the expert could be asked: "What is the largest possible percentage of failures per year for a given component with a definite age?" In other cases, such bounds can be obtained from the statistical data or from a physical model of the corresponding phenomenon.

Once an upper bound to the probability density function is known, it can be used to restrict the set of feasible probability distributions and rule out the degenerate ones. Let us introduce such an upper bound $K \in \mathbf{R}_+$ on the values of the probability density function, i.e.,

$$0 \leq \rho(x) \leq K = const \quad \text{for } \forall x. \qquad (6)$$

Since the overall probability over the interval $[0,T]$ is equal to 1, the upper bound K satisfies the inequality $KT \geq 1$.

By bounding the density function, the set of constraints to optimisation problem (1)-(2) is complemented by inequality (6) which, as it turns out, complicates the optimisation problem drastically.

It is chiefly through duality theory that a linear program can be viewed in its proper perspective and solved. For primal problem (1)-(2) complemented by constraint (6), the dual optimisation problem has the infinite number of dual variables. This is because there are as many dual variables as primal constraints, and in our case the inequality $\rho(x) \leq K$ is to be regarded as denoting an infinite set of constraints $\rho(x_i) \leq K$ $i=1, \ldots, n$, $n \to \infty$. Thus, not being able to employ the dual form of natural extension nor its degenerate form, we become devoid of the key

mechanism for the construction of coherent imprecise models, natural extension.

One would anyway arrive at this stopping point in case of trying to use non-linear constraints, as real-life statistical evidence in many cases cannot be confined to linear constraints.

In the section below we suggest taking a different look at the primal form of natural extension (1)-(2), an approach which opens up new perspectives for obtaining tighter intervals.

4 Natural Extension as a Problem of the Calculus of Variations

The mathematical program (1)-(2) can be modified slightly to make it amenable to the calculus of variations. The calculus is based on the statement that we can always apply a small change $\pm\delta\rho(x)$ to a function $\rho(x)$. (Here $\delta\rho(x)$ denotes a variation of $\rho(x)$, and the symbol δ should not be confused with the Dirac function). Applying variation $\pm\delta\rho(x)$ to a function $\rho(x)$ has the consequence that $\rho(x)$ can become negative, which is in contradiction with the inequality $\rho(x)\geq 0$.

The requirement $\rho(x)\geq 0$ can be satisfied differently by introducing another function $z(x)$ for which

$$\rho(x) = z^2(x). \tag{7}$$

We then have to replace $\rho(x)$ by $z^2(x)$ in the expressions for the objective functions and constraints.

The other inequalities in constraints (2) are turned into equalities by introducing yet other unknown functions $u_{(i,1)}(x)$ and $u_{(i,2)}(x)$, $i=1,\ldots,n$, such that

$$\int_0^T f_i(x)\cdot\left(z^2(x)-u_{(i,1)}^2(x)\right)dx = \underline{a}_i, \tag{8}$$

$$\int_0^T f_i(x)\cdot\left(z^2(x)+u_{(i,2)}^2(x)\right)dx = \overline{a}_i, \tag{9}$$

More information on this technique can be found, for example, in [7].

After having made the above changes, the problem of finding the lower and upper bounds for $M(g)$ now has $z(x)$, $u_{(i,1)}(x)$ and $u_{(i,2)}(x)$, $i=1,2,\ldots,n$, as decision variables. Thus the original problem (1)-(2) turns into the following:

$$\inf_{z(x)}\int_0^T g(x)z^2(x)dx \text{ and } \sup_{z(x)}\int_0^T g(x)z^2(x)dx \tag{10}$$

subject to (8), (9) and

$$\int_0^T z^2(x)dx = 1. \tag{11}$$

Optimization problem (10) subject to (8), (9) and (11) is another form of natural extension amenable to variational calculus. Constraints like (8), (9) and (11), which are integrals of some unknown functions, are called *isoperimetric constraints* [8].

The conventional way of solving problem (10) subject to (8), (9) and (11) is to replace it with an unconstrained optimization problem. In this case the integrand of the objective function

$$F(z,x) = g(x)z^2(x) \tag{12}$$

is replaced by

$$F^*(z,u_{(i,1)},u_{(i,2)},x) = g(x)z^2(x)+\lambda_0 z^2(x)+ \sum_{i=1}^n \lambda_i f_i(x)\left(z^2(x)-u_{(i,1)}^2(x)\right)+\sum_{i=1}^n \lambda_{i+n} f_i(x)\left(z^2(x)+u_{(i,2)}^2(x)\right) \tag{13}$$

where $\lambda_i \in \mathbf{R}, i=0,\ldots 2n$, are (unknown) Lagrange multipliers that could be derived from a system of the Euler-Lagrange equations (see below) complemented by equations-constraints (8), (9) and (11).

The unconstrained optimization problem, which is to be solved now, appears as follows:

$$\inf_{z,u_{(i,1)},u_{(i,2)}}\int_0^T F^*(z,u_{(i,1)},u_{(i,2)},x)dx. \tag{14}$$

For an unconstrained optimization problem the solutions satisfying the necessary condition of optimality can be derived from the Euler-Lagrange equations [8]. For problem (14) these equations take the following form:

$$\left.\begin{aligned}\frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial z} &= \frac{d}{dx}\left(\frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial \dot{z}}\right)\\ \frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial u_{(i,1)}} &= \frac{d}{dx}\left(\frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial \dot{u}_{(i,1)}}\right)\\ \frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial u_{(i,2)}} &= \frac{d}{dx}\left(\frac{\partial F^*(z,u_{(i,1)},u_{(i,2)})}{\partial \dot{u}_{(i,2)}}\right)\end{aligned}\right\} \tag{15}$$

where $\dot{z} = dz/dx;\ \dot{u}_{(i,1)} = du_{(i,1)}/dx,\quad \dot{u}_{(i,2)} = du_{(i,2)}/dx.$

By plugging (13) into (15) we obtain

$$z(x)\cdot\left(g(x)+\lambda_0+\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x)\right)=0 \qquad (16)$$

$$\lambda_i\cdot u_{(i,1)}(x)=0\,,\ i=1,\ldots,n, \qquad (17)$$

$$\lambda_{i+n}\cdot u_{(i,2)}(x)=0\,,\ i=1,2,\ldots,n. \qquad (18)$$

Let us examine equation (16). It holds if $z(x)=0$ for all $x\in[0,T]$. But this would be in contradiction with constraints (8), (9) and (11). Thus $z(x)\neq 0$, at least in some points or possibly inside some subintervals of $[0,T]$. From (16) for those points where $z(x)\neq 0$ it holds that

$$g(x)+\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x)=-\lambda_0\,. \qquad (19)$$

Let us now denote $\xi(x)=g(x)+\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x)$, and consider as an example $g(x)=I_{[0,x_1]}(x)$ and all the other gambles $f_i(x),\ i=1,2,\ldots,n$ as linear functions. This case is depicted in Fig. 2.

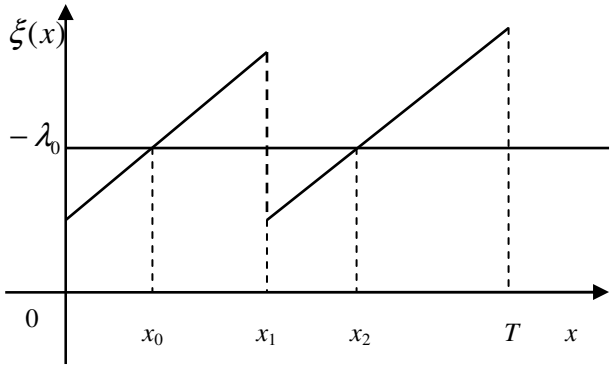


Figure 2: An example of $\xi(x)$ satisfying the necessary condition of optimality

In order to satisfy constraint (11) and to hold equation (16) true the probability density function delivering an extremum to the objective function $\int_0^T F^*(z,u_{(i,1)},u_{(i,2)},x)dx$ can only be degenerate, i.e., concentrated in the three points x_0, x_1 and x_2. This is formalised as $\rho(x)=z^2(x)=\sum_{k=0}^{2}c_k\cdot\delta(x,x_k)$, where $\sum_{k=0}^{2}c_k=1$. Thus, we have arrived at the case where optimal solutions belong to the family of degenerate distributions.

5 Utilising Boundary Constraints with the Variational Form of Natural Extension

Let us now turn back to the case where a boundary to the probability density function is known and we would like to utilise this knowledge to reduce, as we expect, imprecision in the probabilistic measures of interest. That is, we will seek inf and sup of the objective function (1) subject to constraints (2), (6). To solve this new problem, an approach based on the following theorem is proposed.

Theorem 1. *If for any interval* $\alpha\le x\le\beta,\ 0\le\alpha<\beta\le T$ *and for any* $h_0,h_1,\ldots,h_n\in\mathbf{R}$ *it holds that*

$$g(x)\neq h_0+\sum_{i=1}^{n}h_i f_i(x),$$

then probability density function $\rho(x)$*, on which inf and sup are attained in problems (1) subject to (2) and (6), is a step-wise function whose values are either 0 or K.*

Proof. In this problem we have two direct constraints on the density function: $\rho(x)\ge 0$ and $\rho(x)\le K$. To adjust the constraints to the calculus of variations, we introduce some new functions $z(x)$ and $v(x)$ such that $\rho(x)=z^2(x)$ and

$$z^2(x)+v^2(x)=K \qquad (20)$$

Thus, we have a new optimisation problem with objective function (10) subject to (8), (9), (11) and (20).

With respect to noted above, newly introduced equality (20) should be referred to as holonomic constraint.

As we did it earlier, the primal problem with holonomic and isoperimetric constraints is replaced by a new unconstrained optimization problem

$$\inf_{z,v,u_{(i,1)},u_{(i,2)}}\int_0^T F^{**}(z,v,u_{(i,1)},u_{(i,2)},x)dx\,, \qquad (21)$$

where

$$F^{**}(z,v,u_{(i,1)},u_{(i,2)})=g(x)z^2(x)+\lambda^*(x)\cdot\left(z^2(x)+v^2(x)-K\right)$$
$$+\lambda_0 z^2(x)+\sum_{i=1}^{n}\lambda_i f_i(x)\left(z^2(x)-u_{(i,1)}^2(x)\right)$$
$$+\sum_{i=1}^{n}\lambda_{i+n}f_i(x)\left(z^2(x)-u_{(i,2)}^2(x)\right), \qquad (22)$$

and $\lambda^*(x),\lambda_0,\lambda_1,\ldots,\lambda_{2n}$ are (unknown) Lagrange multipliers. Note that $\lambda^*(x)$ is to be a function of x because it is multiplied by a holonomic constraint, while $\lambda_0,\lambda_1,\ldots,\lambda_{2n}$ are constants because they correspond to isoperimetric constraints [7].

For an unconstrained optimization problem the solutions satisfying the necessary condition of optimality can be derived from the Euler-Lagrange equations [8]. By applying the Euler-Lagrange equations, as we did for (14), we arrive at the following set of equalities:

$$z(x)\cdot\left(g(x)+\lambda^*(x)+\lambda_0+\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x)\right)=0, \quad (23)$$

$$\lambda^*(x)v(x)=0, \quad (24)$$

$$\lambda_i\cdot u_{(i,1)}(x)=0, \quad i=1,2,\ldots,n, \quad (25)$$

$$\lambda_{i+n}\cdot u_{(i,2)}(x)=0, \quad i=1,2,\ldots,n. \quad (26)$$

Let us an interval $[\alpha,\beta]\subseteq[0,T]$ is that on which $z(x)\neq 0$. How would $z(x)$ behave on this interval and what values would it take?

According to (23), in those points *x* where $z(x)\neq 0$ it holds that

$$\lambda^*(x)=-g(x)-\lambda_0-\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x). \quad (27)$$

And according to (23), in those points *x* where $\lambda^*(x)\neq 0$ it holds that $v(x)=0$, which in turn, according to (20), results in $\rho(x)=z^2(x)=K$.

From (27) it follows that if

$$g(x)\neq -\lambda_0-\sum_{i=1}^{n}(\lambda_i+\lambda_{i+n})f_i(x),$$

then $\lambda^*(x)\neq 0$ and $\rho(x)=K$.
By denoting $h_0=-\lambda_0$ and $h_i=-(\lambda_i+\lambda_{i+n})$ we can rewrite

$$g(x)\neq h_0+\sum_{i=1}^{n}h_i f_i(x),$$

which was to be proven.

The theorem enables us to reduce the original variational optimization problem to an easier problem of optimizing a multivariate function under algebraic constraints.

Indeed, let $[x_0,x_1),[x_2,x_3),[x_4,x_5),\ldots,[x_{2m},x_{2m+1})$ be the intervals on which $\rho(x)=K\neq 0$, and let $[x_1,x_2),[x_3,x_4),[x_5,x_6),\ldots,[x_{2m+1},T)$ be the intervals on which $\rho(x)=0$. Let us denote

$$G(x_j,x_{j+1})=\int_{x_j}^{x_{j+1}}g(x)dx \quad (28)$$

$$\Phi_i(x_j,x_{j+1})=\int_{x_j}^{x_{j+1}}f_i(x)dx, \quad i=1,2,\ldots,n. \quad (29)$$

Then, problem (1) subject to constraints (2), (6) takes the following form:

$$\left.\begin{aligned}\underline{M}(g)&=\min_{x_0,x_1,\ldots}\left\{K\cdot\sum_{j=0}^{m}G(x_{2j},x_{2j+1})\right\}\\ \overline{M}(g)&=\max_{x_0,x_1,\ldots}\left\{K\cdot\sum_{j=0}^{m}G(x_{2j},x_{2j+1})\right\}\end{aligned}\right\} \quad (30)$$

subject to

$$K\cdot\sum_{j=0}^{m}(x_{2j+1}-x_{2j})=1 \quad (31)$$

$$\underline{a_i}\leq K\cdot\sum_{j=0}^{m}\Phi_i(x_{2j},x_{2j+1})\leq\overline{a_i}, \quad i=1,2,\ldots,n. \quad (32)$$

If the number of intervals *m* is known, this optimization problem can be solved by using standard numerical techniques such as gradient methods, simplex-based search methods, genetic algorithms, etc. In simple cases, the solution can be found in an analytical form.

How can we find *m*? One idea is to start with the smallest value *m*, corresponding to having one interval with nonzero density, and to solve the optimization problem with this *m*. Then, increase *m* by 1 and solve the problem again, etc. Repeat the process until when for a new *m* you get exactly the same optimising function $\rho(x)$ as for the previous *m* – this will mean that a further subdivision of intervals will probably not change the value of objective function (1).

Example 2. *Utilising knowledge on the boundary of the density function*

In this example, the statistical evidence about a random value *X* we have at hand is a boundary *K* on the probability density function and, as in Example 1, $\Pr(x\in[\underline{q},\overline{q}])=b$. What are the lower, $\underline{M}(X)$, and upper bounds, $\overline{M}(X)$, for the expected value of *X*?

It is found that increasing *m* step by step by 1 starting from 0 does not change the optimising density function $\rho(x)$ after *m* exceeds 1. That is, the solution of problem (30)-(32) must be sought for *m*=1. (Note that *m*=1 corresponds to having two intervals on which the probability density function is different from 0.)

Depending on the disposition of $\underline{q}$ within the interval [0,*T*], the probability density function delivering the minimum to the objective function is calculated differently.

If $\underline{q} \geq \frac{1-b}{K}$,

$$\rho(x) = \begin{cases} K \text{ for } 0 \leq x \leq \frac{1-b}{K} \\ 0 \text{ for } \frac{1-b}{K} < x < \underline{q} \\ K \text{ for } \underline{q} \leq x \leq \underline{q} + \frac{b}{K} \\ 0 \text{ for } \underline{q} + \frac{b}{K} < x \leq T \end{cases},$$

and for $\underline{q} < \frac{1-b}{K}$:

$$\rho(x) = \begin{cases} K \text{ for } 0 \leq x \leq \underline{q} + \frac{b}{K} \\ 0 \text{ for } \underline{q} + \frac{b}{K} < x < \bar{q} \\ K \text{ for } \bar{q} \leq x \leq \bar{q} + \frac{1-\underline{q}K-b}{K} \\ 0 \text{ for } \bar{q} + \frac{1-\underline{q}K-b}{K} < x \leq T \end{cases}$$

Let us assume that $\underline{q} \geq \frac{1-b}{K}$. Then it can be concluded that optimization problem (30)-(32) for the lower bound becomes as follows:

$$\underline{M}(X) = \min_{x_i} K \frac{x_1^2 - x_2^2 + x_3^2 - x_2^2}{2} \text{ subject to}$$

$$K(x_1 - x_0) + K(x_3 - x_2) = 1$$
$$K(x_3 - x_2) = b .$$

The next step is to plug the constraints into the objective function and observe that minimum is attained if $x_2 = \underline{q}$. After doing this, we obtain

$$\underline{M}(X) = \min_{x_0} \frac{1}{2K} + x_0(1-b) + b\left(\underline{q} - \frac{1-b}{K}\right).$$

It is not difficult to see that the minimum is attained if $x_0 = 0$. Thus

$$\underline{M}(X) = \frac{1}{2K} + \frac{b\left(\underline{q}K - (1-b)\right)}{K}.$$

The probability density function delivering the minimum is shown in Fig. 3.

For the case when $\underline{q} < \frac{1-b}{K}$ the solution is

$$\underline{M}(X) = \frac{1}{2K} + \frac{\bar{q}K - (\underline{q}K + b) \cdot \left(1 + (\bar{q} - \underline{q})K - b\right)}{K}.$$

The solution to $\overline{M}(X)$ can be obtained in a similar way to that for the lower bound.

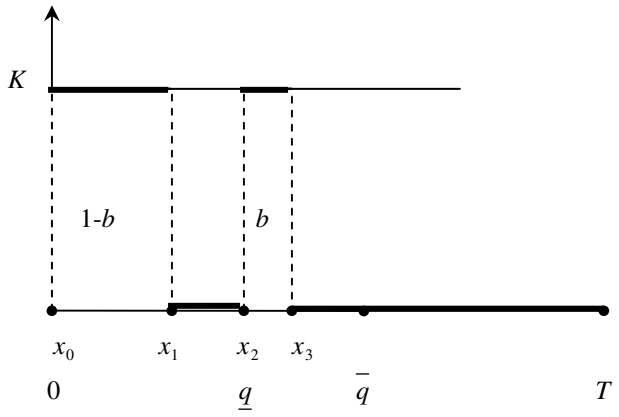


Figure 3: Bounded probability density function providing the infimum to the objective function

6 Bounded densities and their derivatives

In attempting to achieve tighter bounds, one can impose constraints on the derivatives (or their absolute values) of probability density functions. So, now we suppose that one has at hand an upper bound on the value of the probability density function and an upper bound on its derivative absolute value. Any other assumptions concerning the actual shape of the distribution are not introduced.

Once the additional upper bound is known, it can be used to restrict the set of admissible probability distributions and rule out the functions which derivatives take excessively high values.

Let us denote $M \in \mathbf{R}_+$ an upper bound on the value of the probability density absolute value, i.e., for $\forall x$

$$\left| d\rho(x) / dx \right| \leq M = const. \quad (33)$$

In the variational calculus set-up, now we seek inf and sup of the objective function (1) subject to constraints (2), (6) and (33). To solve this new problem, an approach based on the following theorem is proposed.

Theorem 2. *If for any interval* $\alpha \le x \le \beta,\ 0 \le \alpha < \beta \le T$ *and for any* $h_0, h_1, ..., h_n \in \mathbf{R}$ *it holds that*

$$g(x) \ne h_0 + \sum_{i=1}^{n} h_i f_i(x),$$

then the probability density function $\rho(x)$*, on which inf and sup are attained in problems (1) subject to (2), (6) and (33), is a stepwise linear function* $\rho(x) = \pm Mx + C$ *whose values are bounded by K from above*.

Proof. The logic of the proof is similar to that used to prove Theorem 1. The proof can be found in [9] which has been submitted for publication.

An example of the density function satisfying Theorem 2 is depicted in Fig. 4.

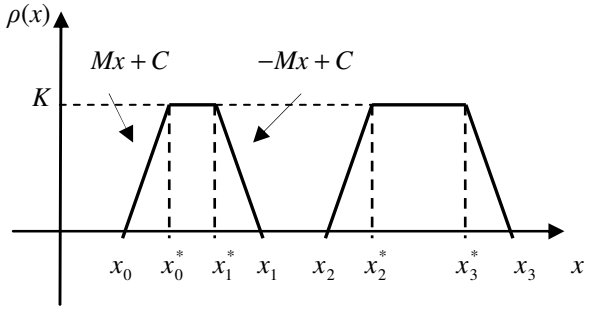


Figure 4: An example of the density function satisfying Theorem 2

The points in Fig. 5 marked with asterisks have the following values: $x_0^* = x_0 + K/M$, $x_1^* = x_1 - K/M$, $x_2^* = x_2 + K/M$, $x_3^* = x_3 - K/M$

Theorem 2 enables to reduce the original variational optimization problem to an easier one. This can be done because the shape of the density function is now known. Unknown are the points $x_0, x_1, ..., x_0^*, x_1^*, ...$, which become the parameters of the density function $\rho(x) = \rho(x_0, x_1, ..., x_0^*, x_1^*, ..., x)$ and the decision variables in the new optimisation problem.

Let $[x_0, x_1), [x_2, x_3), [x_4, x_5), ..., [x_{2m}, x_{2m+1})$ be the intervals on which $\rho(x) \ne 0$. They can be interpreted as lower bases of the trapezoids (see Fig. 5). The upper bases of the trapezoids are the intervals $[x_0 + K/M, x_1 - K/M), [x_2 + K/M, x_3 - K/M), ...,$ $[x_{2m} + K/M, x_{2m+1} - K/M)$. And $[x_1, x_2), [x_3, x_4), ...,$ $[x_{2m+1}, x_{2m+2})$ are the intervals on which $\rho(x) = 0$.

Now the optimisation problem appears as follows:

$$\left.\begin{aligned} \underline{M}(g) &= \min_{x_0, x_1, ...} \left\{ \sum_{j=0}^{m} G^*(x_{2j}, x_{2j+1}) \right\} \\ \overline{M}(g) &= \max_{x_0, x_1, ...} \left\{ \sum_{j=0}^{m} G^*(x_{2j}, x_{2j+1}) \right\} \end{aligned}\right\} \quad (34)$$

subject to

$$K \cdot \sum_{j=0}^{m} (x_{2j+1} - x_{2j} - K/M) = 1, \quad (35)$$

$$\underline{a_i} \le \sum_{j=0}^{m} \Phi_i^*(x_{2j}, x_{2j+1}) \le \overline{a_i}, \quad i = 1,2,..., . \quad (36)$$

where

$$G^*(x_j, x_{j+1}) = M \int_{x_j}^{x_j + K/M} g(x)(x - x_j)dx -$$

$$M \int_{x_{j+1} - K/M}^{x_{j+1}} g(x)(x - x_{j+1})dx + K \int_{x_j + K/M}^{x_{j+1} - K/M} g(x)dx$$

$$\Phi_i^*(x_j, x_{j+1}) = M \int_{x_j}^{x_j + K/M} f_i(x)(x - x_j)dx -$$

$$M \int_{x_{j+1} - K/M}^{x_{j+1}} f_i(x)(x - x_{j+1})dx + K \int_{x_j + K/M}^{x_{j+1} - K/M} f_i(x)dx .$$

Example 3. *Unbounded probability density function and bounded absolute values of its derivative*

Let us consider an example in which $|d\rho(x)/dx| \le M$ is the only restriction on $\rho(x)$. What are the bounds on the expected value $M(X)$ of the corresponding random variable?

In this example $g(x)=x$, which implies that everywhere $g(x) \ne h_0$ meaning that theorem 2 can be applied.

Note first that as the condition $\rho(x) \le K$ is not imposed on the density function, the trapezoidal shape of the density is changed to the triangular one.

Let us start with $m=0$ corresponding to having one interval $[x_0, x_1)$ on which the probability density function is different from 0 and denote $y = x_1 - x_2$.

Here we have only one isoperimetric constraint: $\int_0^T \rho(x)dx = M\frac{y^2}{4} = 1$, where $M\frac{y^2}{4}$ is the area the triangle. Hence $y = \frac{2}{\sqrt{M}}$, or $x_1 = x_0 + \frac{2}{\sqrt{M}}$.

The formula for the expected value $M(X)$ takes the form:

$$M(X)=\int_0^T x\rho(x)dx=\int_{x_0}^{x_0+1/\sqrt{M}} x(Mx-Mx_0)dx+$$
$$\int_{x_0+1/\sqrt{M}}^{x_0+2/\sqrt{M}} x\left(-Mx+M(x_0+2/\sqrt{M})\right)dx.$$

And further

$$M(X)=\frac{M}{3}\left((x_0+1/\sqrt{M})^3-x_0^3\right)-\frac{Mx_0}{2}\left((x_0+1/\sqrt{M})^2-x_0^2\right)$$
$$-\frac{M}{3}\left((x_0+2/\sqrt{M})^3-(x_0+1/\sqrt{M})^3\right).$$
$$+\frac{M(x_0+2/\sqrt{M})}{2}\left((x_0+2/\sqrt{M})^2-(x_0+1/\sqrt{M})^2\right).$$

In the following we will keep in mind that $x_0\ge 0$, $x_1\le T$, and hence $x_0\le T-\frac{2}{\sqrt{M}}$.

It is easy to see that the smallest value of $M(X)$ is attained when $x_0=0$, so $\underline{M}(X)=\frac{1}{\sqrt{M}}$.

Similarly, to obtain $\overline{M}(X)$, we take the largest possible value of x_0, i.e. $x_0=T-\frac{2}{\sqrt{M}}$, which brings us to $\overline{M}(X)=T-\frac{1}{\sqrt{M}}$.

If we take m=1 and do manipulations similar to the above, we find that the solutions do not change.

Example 4. *Bounded probability density function and bounded absolute values of its derivative*

Now we have two constraints (6) and (33), i.e., $0\le\rho(x)\le K$ and $|d\rho(x)/dx|\le M$. The question to answer is still the same: What are the bounds on the expected value $M(X)$?

As we keep the function $g(x)=x$ introduced in Example 3, theorem 2 can be also applied for this case.

Start with m=0. Here we have only one isoperimetric constraint (the area of the trapezoid equalised to 1): $\int_0^T \rho(x)dx=\frac{x_1-x_0+x_1-x_0-2K/M}{2}K=1,$ hence $x_1=1/K+K/M+x_0$.

The formula for the expected value $M(X)$ takes the form:

$$M(X)=\int_0^T x\rho(x)dx=\int_{x_0}^{x_0+K/M} x(Mx-Mx_0)dx+K\int_{x_0+K/M}^{x_0+1/K} xdx+$$
$$\int_{x_0+1/K}^{x_0+1/K+K/M} x(-Mx+M/K+K+Mx_0)dx.$$

And finally,

$$M(X)=\frac{M}{3}\left((x_0+K/M)^3-x_0^3\right)-\frac{Mx_0}{2}\left((x_0+K/M)^2-x_0^2\right)+$$
$$\frac{K}{2}\left((x_0+1/K)^2-(x_0+K/M)^2\right)$$
$$-\frac{M}{3}\left((x_0+1/K+K/M)^3-(x_0+1/K)^3\right).$$
$$+\frac{M(x_0+1/K)+K}{2}\left((x_0+1/K+K/M)^2-(x_0+1/K)^2\right)$$

As $x_0\ge 0$ and $x_1\le T$, hence $x_0\le T-\frac{1}{K}-\frac{K}{M}$.

The smallest value of $M(X)$ is attained when $x_0=0$, hence $\underline{M}(X)=\frac{1}{2K}+\frac{K}{2M}$.

Similarly, to obtain $\overline{M}(X)$, we take the largest possible value of x_0, i.e., $x_0=T-\frac{1}{K}-\frac{K}{M}$, hence $\overline{M}(X)=(T-1/K)+\frac{1}{2K}=T-\frac{1}{2K}-\frac{K}{2M}$.

If we take m=1 and do manipulations similar to the above, we find that the solutions do not change.

7 What is Still Dissatisfying?

There are at least two remaining problems with applying imprecise statistical reasoning to reliability analysis.

In reliability analysis, the pivotal characteristic is time to failure (or time between failures if a system is repairable), and a failure in a system can occur at any point of the lifetime. In contrast, the model presupposes that failures can take place only within some specific intervals but not at any point. This is because probability masses are not continuously distributed during the lifetime. In spite of bringing in more statistical evidence about time to failure, the situation does not seem to be remedied.

The other principle obstacle to reliability applications is the bounding condition on gambles, which in practice means dealing with bounded random values. That is, applying the reasoning to reliability implies that time to

failure is a bounded random value. Let us say, one must know the maximum time a system can survive in order to apply the theory. This is that what can hardly be known for certainty. Furthermore, as technical systems undergo preventive maintenance and are put out of operation based on volitional decisions rather than after observing their full inoperability, knowing the point behind which they become irrecoverable, and even defining what it means, make the bound on time to failure meaningless.

8 Summary and Conclusions

The usefulness of interval-valued statistical characteristics depends both on how tight the bounds are and on how easy they are to compute. The tightness of the bounds depends in turn on the amount of information available and that which can be utilised by the method, and on the method itself. The more relevant information the modeller has at hand and the greater the amounts of it that can be utilised by the model, the tighter the bounds are. We have been aiming at enhancing natural extension so that it could utilise a wider variety of statistical evidence, some of which is easy to acquire but not easy to utilise.

As has been demonstrated, natural extension can be viewed as the problem of finding an extremal of a functional, a problem which belongs to the realm of variational calculus. If this path is followed, the modeller can utilise more versatile information than is possible with the natural extension suggested by Walley [1] and Kuznetsov [3]. The present paper has demonstrated that imposing a restriction on the upper bound of the probability density function of a random value is an effective way of obtaining tighter bounds of statistical measures.

In some cases, common sense and intuition may suggest that the underlying distribution is for instance differentiable in any point or symmetrical without specifying a particular shape. Utilising this kind of evidence may drastically reduce imprecision in the resultant interval-valued statistical characteristics, and, it is clear, this evidence is acquired at a low cost; in some cases it can be gained at no effort.

We have been attempting to demonstrate in relation to the approach based on variational calculus that there is room for improvement without having to use unreliable data and introduce debatable assumptions as a means of obtaining reasonably precise results.

In the pursuit of robust reliability assessments, the next facing challenge is to update the existing reliability models so that they can take account of additional evidence, evidence that until now has not been requested owing to the models' incapacity to utilise it. The fact that there is currently a substantial amount of alternative evidence at our disposal presents other challenges. For example, what kind of evidence is worth using in order to facilitate computations and make substantial headway in terms of tighter bounds? What constraints are most beneficial for what models? These are directions in which, we suggest, further work with the calculus of variations ought to proceed.

Acknowledgement

Participation of V. Krymsky in the work described was partially supported by NATO grant CBP.NR.NREV 982410.

References

[1] Walley P. (1991) *Statistical reasoning with imprecise probabilities*. Chapman and Hall. New York.

[2] Utkin L. and Kozine I. (2001) *Different faces of the natural extension.* In: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA '01, pp. 316-323.

[3] Kuznetsov V. (1991) *Interval statistical models*. Radio and Sviaz. Moscow. (In Russian).

[4] Utkin L. (2002) *Imprecise calculation with the qualitative information about probability distributions*. In: Proceedings of the conference on Soft Methods in Probability and Statistics. Eds. P. Grzegorzewski, O. Hryniewicz and M.A. Gil, Physica-Verlag, Heidelberg, New York, pp. 164-169.

[5] Smith E.J. (1995) *Generalized Chebychev inequalities: theory and applications in decision analysis*. Operations Research, Vol. 43(5), pp. 807-825.

[6] Kozine I.O., Krymsky V.G. (2003) *Reducing uncertainty by imprecise judgements on probability distributions: Application to system reliability*. In: Proceedings. 3rd International symposium on imprecise probabilities and their applications (ISIPTA '03), Lugano (CH), 14-17 July 2003. Bernard, J.-M.; Seidenfeld, T.; Zaffalon, M. (eds.), (Carleton Scientific, Waterloo (CA), 2003) (Proceedings in Informatics, 18), pp. 335-344.

[7] Ivanov V.A., Faldin N.V. (1981) *Theory of optimal control systems*. Nauka Publ. Moscow. (In Russian).

[8] Gelfand I.M., Fomin S.V. (2000) *Calculus of variations.* Dover Publ. New York, 240 p.

[9] Kozine I., and Krymsky V. (2007) Computing interval-valued statistical characteristics: What is the stumbling block for reliability applications? Submitted to the Journal of Reliable Computing

On $\sigma-$additive robust representation of convex risk measures for unbounded financial positions in the presence of uncertainty about the market model

Volker Krätschmer *
Institute of Mathematics, Berlin University of Technology, 10623 Berlin, Germany
E-mail: KRAETSCH@math.tu-berlin.de

Abstract

Recently, Frittelli and Scandolo ([7]) extend the notion of risk measures, originally introduced by Artzner, Delbaen, Eber and Heath ([1]), to the risk assessment of abstract financial positions, including pay offs spread over different dates, where liquid derivatives are admitted as financial instruments, and unbounded fincial positions are also allowed. Convex risk measures may be viewed as convex upper previsions for unbounded gambles, a notion originally introduced by Pelessoni and Vicig [16]. The paper deals with $\sigma-$additive robust representations of convex risk measure, that means envelope theorems in terms of $\sigma-$additive probability measures. We shall focus on the aspect that the investor is faced with uncertainty about the market model. It turns out that the results may be applied for the case that a market model is available, and that they encompass as well as improve criteria obtained for robust representations of convex risk measures in the genuine sense ([2], [5], [13]).

Keywords. Convex risk measures, convex upper previsions, model uncertainty, $\sigma-$additive robust representation, Greco's representation theorem, Fatou property, inner Daniell stone theorem, general Dini theorem, strong $\sigma-$additive robust representation, Simons' lemma, nonsequential Fatou property, Krein-Smulian theorem.

1 Introduction

The notion of risk measures has been introduced by Artzner, Delbaen, Eber and Heath (cf. [1]) as the key concept to found an axiomatic approach for risk assessment of fincancial positions. Technically, risk measures are functionals defined on sets of financial positions, satisfying some basic properties to qualify riskiness consistently. An outcome of such a functional, that means the risk of a position, is usually interpreted as the capital requirement of the position to become an acceptable one. Genuinely, risk measures has been defined for one-period positions. Recently Fritelli and Scandolo ([7]) provide a general framework which extends considerations to abstract financial positions including pay off streams with liquid derivatives as hedging positions. Applied to the risk assessment of pay off streams such general risk measures are used for an a priori qualification, which means to take the static perspective. In contrary the dynamic risk assessment take into account adjustments time after time. Readers who interested in this topic are referred to e.g. [6], [17], [21].

The main goal of this paper is to investigate risk measures ρ which admit a robust representation of the form

$$\rho(X) = \sup_{\Lambda}(-\Lambda(X) - \beta(\Lambda)),$$

where X denotes a financial position, Λ a linear form on the set of financial positions, and β stands for a penalty function on the set of linear forms. Special attention will be paid to the problem when these representing linear forms may in turn be represented by ($\sigma-$additive) probability measures. We shall speak of a robust representation of ρ by probability measures or a $\sigma-$additive robust representation. Necessarily, only so-called convex risk measures, that means risk measures which are convex mappings, may have such a robust representation. The basic assumption of this paper is that the investors are uncertain about the market model underlying the outcomes of the financial positions. Within this setting a robust representation by probability measures offered an additional economic interpretation of the risk measures. As suggested by Föllmer and Schied (cf. [5]) such a representation means that an investor has a set of possible market models in mind, and evaluates the worst expected losses together with some penalty costs for misspecification w.r.t. these models. In particular an investor with such a risk measure may be viewed as

*This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

risk- and ambiguity-averse (cf. [19]).

The problem of $\sigma-$additive robust representation of convex risk measures in the genuine sense has been completely solved in the case that the investors have market models at hand. Ruszczynski and Shapiro showed that convex risk measures always admit robust representations by probability measures if for any real p every integrable mapping of order p is available (cf. [18]). However the used methods can not be applied to essentially bounded positions. Drawing on methods from functional analysis, Delbaen as well as Föllmer and Schied succeeded in giving a full characterization (cf. [2], [5]) by the so-called Fatou property. As pointed out by Delbaen, the Fatou property fails to be sufficient in general when the investor is faced with model uncertainty. Moreover, the problem of $\sigma-$additive robust representation is still open when a market model is not available. Restricting considerations to bounded one-period positions, Föllmer and Schied suggested a strict sufficient criterion, Krätschmer showed that it is in some sense also necessary, and he adds some more general conditions ([13]).

This paper may be viewed as a continuation of the studies by in [5] as well as in [13]. The generalizations will be proceeded into several directions. First of all multiperiod positions and liquid hedging instruments will be allowed. Secondly we shall drop the assumptions that only bounded positions are traded. This is in accordance with empirical evidences that the distributions of risky assets often show heavy tails. Thirdly we want to investigate the issue of strong robust representations by probability measures in the sense that the optimization involved in the $\sigma-$additive robust representation has a solution. Finally, the criteria should encompass the results already derived within a fixed market model.

The paper is organized as follows. Section 2 introduces the concept of Frittelli and Scandolo to define risk measures in general, and some representation results of risk measures will be presented as starting points for the following investigations. The general criterion is offered in section 3, extending a former result in [13] to unbounded positions, within a nontopological framework. It will be used for strong robust representations of risk measures by probability measures in section 4. We shall succeed in giving a complete solution. In particular the aboved mentioned strict criterion by Föllmer and Schied will turn out to be necessary and sufficient. Moreover, within a given market model the solution by Jouini, Schachermayer and Touzi (in [8]) may be recognized. Afterwards, section 5 deals with the question when the Fatou property might be used as a sufficient condition. In presence of a market model the results may be used to retain the above mentioned equivalent characterization by Delbaen as well as Föllmer and Schied. In general, as a rule a nonsequential counterpart is more suitable unless in some special cases.

The proofs of the results presented within this paper use several arguments from functional analyis, particularly from convex and superconvex analysis, as well as from abstract measure and integration theory. They are very technical in nature and must be omitted due to limitations of scope. The interested reader is kindly referred to the working paper version [15].

2 Some basic representations of convex risk measures

Let us fix a set Ω. Financial positions will be expressed by mappings $X \in \mathbb{R}^{\Omega}$. As a special case $\Omega = \widetilde{\Omega} \times \mathbb{T}$ with $\widetilde{\Omega}$ denoting a set of scenarios, equipped with a family $(\mathcal{F}_t)_{t\in\mathbb{T}}$ of $\sigma-$algebras, and $\mathbb{T}$ being a time set, we may consider financial positions $X \in \mathbb{R}^{\Omega\times\mathbb{T}}$ with $X(\cdot,t)$ being $\mathcal{F}_t-$measurable for every $t \in \mathbb{T}$. They may be viewed as discounted pay off streams, liquidated at the dates from the time set. In the case of $\mathbb{T} = \{1\}$ we shall speak of **one-period positions**. The available financial positions are gathered by a nonvoid vector subspace $\mathfrak{X} \subseteq \mathbb{R}^{\Omega}$ containing the constants. Sometimes we shall in addition assume that $X \wedge Y := \min\{X, Y\}, X \vee Y := \max\{X, Y\} \in \mathfrak{X}$ for $X, Y \in \mathfrak{X}$. In this case $\mathfrak{X}$ is a so-called Stonean vector lattice. For the space of bounded positions from $\mathfrak{X}$ the symbol $\mathfrak{X}_b$ will be used. Furthermore let us fix a vector subspace $\mathfrak{C} \subseteq \mathfrak{X}$ of financial positions for hedging, including the constants. This means that we may take into account liquid derivatives like put and call options as financial instruments. In particular, in the case of pay off streams we may also allow investments and disinvestments varying over the time. The financial positions are associated with a positive linear function $\pi : \mathfrak{C} \to \mathbb{R},\ \pi(1) = 1$, where $\pi(Y)$ stands for the initial costs to obtain Y. In the seminal paper by Artzner et al. in [1] considerations are restricted to one-period positions and π being the identity on $\mathbb{R}$. Let us now introduce the concept of risk measures suggested by Frittelli and Scandolo in [7]. As for one-period positions we may choose the axiomatic viewpoint, defining a **risk measure w.r.t.** π to be a functional $\rho : \mathfrak{X} \to \mathbb{R}$ which satisfies the properties

- **monotonicity:**
  $\rho(X) \leq \rho(Y)$ for $X \geq Y$
- **translation invariance w.r.t.** π**:**
  $\rho(X + Y) = \rho(X) - \pi(Y)$ for $X \in \mathfrak{X}, Y \in \mathfrak{C}$

The meaning of these conditions may be transferred from the genuine concept of risk measures. Moreover, it can be shown that a risk measure ρ w.r.t. π satisfies $\rho(X) = \inf\{\pi(Y) \mid Y \in \mathfrak{C}, \rho(X+Y) \leq 0\}$ for any $X \in \mathfrak{X}$ ([7], Proposition 3.6). Regarding $\rho^{-1}(]-\infty, 0])$ as the acceptable positions, an outcome $\rho(X)$ expresses the infimal costs to hedge it. This retains the original meaning of risk measures as capital requirements.

In the following we shall focus on so-called **convex risk measures**, defined to mean risk measures which are convex mappings. Convexity is a reasonable condition for a risk measure due to its interpretation that diversification should not increase risk. From the technical point of view convexity is a necessary property for the desired dual representations of risk measures. Convex risk measures may be viewed as convex upper previsions as introduced in [16]. More precisely, if $\bar{\mathrm{P}}$ denotes a convex upper prevision on the gambles from $\mathfrak{X}$, then ρ defined by $\rho(X) := \bar{\mathrm{P}}(-X)$ is a convex risk measure w.r.t. the identity on $\mathbb{R}$.

Let us now fix a convex risk measure $\rho : \mathfrak{X} \to \mathbb{R}$ w.r.t. π. It is associated with $\beta_\rho : \mathfrak{X}^* \to]-\infty, \infty]$, defined by

$$\beta_\rho(\Lambda) = \sup_{X \in \mathfrak{X}} (-\Lambda(X) - \rho(X)) = \rho^*(-\Lambda),$$

where $\mathfrak{X}^*$ gathers all real linear forms on $\mathfrak{X}$, and ρ^* denotes the so-called Fenchel-Legendre transform of ρ. It is easy to verify that every Λ from the domain $\beta_\rho^{-1}(\mathbb{R})$ of β_ρ has to be a positive linear form extending π. The standard tools from convex analysis provide basic representation results for ρ with β_ρ as a penalty function.

Proposition 1 *Let $\mathfrak{X}_+^{*\pi}$ denote the space of all positive linear forms on $\mathfrak{X}$ which extend π, and let τ be any topology on $\mathfrak{X}$ such that $(\mathfrak{X}, \tau)$ is a locally convex topological vector space with topological dual $\mathfrak{X}'$. Then $\rho(X) = \max_{\Lambda \in \mathfrak{X}_+^{*\pi}} (-\Lambda(X) - \beta_\rho(\Lambda))$ for every $X \in \mathfrak{X}$. Moreover, $\rho(X) = \sup_{\Lambda \in \mathfrak{X}_+^{*\pi} \cap \mathfrak{X}'} (-\Lambda(X) - \beta_\rho(\Lambda))$ holds for every $X \in \mathfrak{X}$ if and only if ρ is lower semicontinuous w.r.t. τ.*

The proof may be found in [15] (AppendixB).

The aim of the paper is to improve the representation results by allowing only representing linear forms which are in turn representable by $\sigma-$additive probability measures. For notational purposes let us introduce the counterpart of β_ρ w.r.t. the probability measures on the $\sigma-$algebra $\sigma(\mathfrak{X})$ on Ω generated by $\mathfrak{X}$

$$\alpha_\rho : \mathcal{M}_1 \to]-\infty, \infty], \ \mathrm{P} \mapsto \sup_{X \in \mathfrak{X}} (-E_{\mathrm{P}}[X] - \rho(X)).$$

Here $\mathcal{M}_1$ is defined to consist of all $\sigma-$additive probability measures P on $\sigma(\mathfrak{X})$ such that all positions from $\mathfrak{X}$ are P$-$integrable, and $E_{\mathrm{P}}[X]$ denotes the expected value of X w.r.t. P. We shall speak of a **robust representation by probability measures from $\mathcal{M}$** or a **$\sigma-$additive robust representation of ρ w.r.t. $\mathcal{M}$** if $\mathcal{M} \subseteq \mathcal{M}_1$ nonvoid, and the representation $\rho(X) = \sup_{\mathrm{P} \in \mathcal{M}} (-E_{\mathrm{P}}[X] - \alpha_\rho(\mathrm{P}))$ holds for every $X \in \mathfrak{X}$. As an immediate consequence of Proposition 1 we obtain a first characterization of such representations.

Proposition 2 *Let F be a vector space of bounded countably additive set functions on $\sigma(\mathfrak{X})$ which separates points in $\mathfrak{X}$ such that each $X \in \mathfrak{X}$ is integrable w.r.t. any $\mu \in F$. Then in the case that the set $\mathcal{M}_1(F)$ of all $\mathrm{P} \in \mathcal{M}_1 \cap F$ with $E_{\mathrm{P}}|\mathfrak{C} = \pi$ is nonvoid*

$$\rho(X) = \sup_{\mathrm{P} \in \mathcal{M}_1(F)} (-E_{\mathrm{P}}[X] - \alpha_\rho(\mathrm{P})) \text{ for all } X \in \mathfrak{X}$$

if and only if ρ is lower semicontinuous w.r.t. weak topology $\sigma(\mathfrak{X}, F)$ on $\mathfrak{X}$ induced by F.

Remark 1 *Retaking assumptions and notations from Proposition 2, ρ admits a robust representation in terms of $\mathcal{M}_1(F)$ if F contains the Dirac measures, and if $\liminf_i \rho(X_i) \geq \rho(X)$ holds for every net $(X_i)_{i \in I}$ in $\mathfrak{X}$ which converges pointwise to some X from $\mathfrak{X}$.*

In general the lower semicontinuity of ρ w.r.t. the topology from Proposition 2 is not easy to verify. Therefore we are looking for more accessible conditions. The considerations will be based on the crucial step to reduce the investigations to bounded financial positions. That means ρ should admit a $\sigma-$additive robust representation if and only if the restriction to the bounded positions does so. In the case that $\mathfrak{X}$ is in addition a Stonean vector lattice this may be achieved via Greco's representation theorem (cf. [11], Theorem 2.10 with Remark 2.3) if the linear forms from the domain of β_ρ are representable as asymmetric Choquet integrals w.r.t. a finitely additive probability measure (cf. [15], Lemma 6.5). The reader may consult the monograph [4] for the concept of asymmetric Choquet integrals w.r.t. isotone set functions. Fortunately, drawing on Greco's representation theorem again, we might express this condition equivalently by the property that the **cutting condition** $\lim_{n \to \infty} \rho(-\lambda(X-n)^+) = \rho(0)$ $((X-n)^+ := (X-n) \vee 0)$ is satisfied for every $\lambda > 0$ and any nonnegative $X \in \mathfrak{X}$ (cf. [15], Proposition 6.6). To summarize

Proposition 3 *Let $\mathfrak{X}$ be a Stonean vector lattice, and let $\lim_{n \to \infty} \rho(-\lambda(X-n)^+) = \rho(0)$ be fulfilled for every*

$\lambda > 0$ *and every nonnegative* $X \in \mathfrak{X}$. *Then for any nonvoid* $\mathcal{M} \subseteq \mathcal{M}_1$ *the following statements are equivalent*

.1 $\rho(X) = \sup_{\mathrm{Q} \in \mathcal{M}} (-E_{\mathrm{Q}}[X] - \alpha_\rho(\mathrm{Q}))$ *for all bounded* $X \in \mathfrak{X}$

.2 $\rho(X) = \sup_{\mathrm{Q} \in \mathcal{M}} (-E_{\mathrm{Q}}[X] - \alpha_\rho(\mathrm{Q}))$ *for all* $X \in \mathfrak{X}$.

The cutting condition will be the basic assumption for the general representation result of the paper. Essentially, it says that for a seller of a derived call option the risk of a loss tends to the risk of inactivity with increasing strike price. Notice that the cutting condition is redundant if all positions in $\mathfrak{X}$ are bounded.

Before going into the development of criteria for $\sigma-$additive representations let us collect some necessary conditions. In the case that the positions from $\mathfrak{X}$ are essentially bounded mappings w.r.t. a reference probability measure of a given market model the so-called Fatou property plays a prominent role. Adapting this concept, we shall say that a risk measure ρ fulfills the **Fatou property** if the inequality $\liminf_{n\to\infty} \rho(X_n) \geq \rho(X)$ holds whenever $(X_n)_n$ is a uniformly bounded sequence in $\mathfrak{X}$ which converges pointwise to some bounded $X \in \mathfrak{X}$. The Fatou property implies obviously that $\rho|\mathfrak{X}_b$ is **continuous from above**, defined to mean $\rho(X_n) \nearrow \rho(X)$ for $X_n \searrow X$. Both conditions coincide if $\sup_n X_n \in \mathfrak{X}$ for any uniformly bounded sequence $(X_n)_n$ in $\mathfrak{X}$.

Proposition 4 *Let* ρ *admit a* $\sigma-$*additive robust representation w.r.t. some nonvoid* $\mathcal{M} \subseteq \mathcal{M}_1$, *then* ρ *satisfies the Fatou property, and* $\rho|\mathfrak{X}_b$ *is continuous from above.*

The proof may be found in [15] (section 7).

3 Robust representation of convex risk measures by inner regular probability measures

Throughout this section let $\mathfrak{X}$ be a Stonean vector lattice, and let $\mathfrak{L} \subseteq \mathfrak{X}$ denote any Stonean vector lattice which contains $\mathfrak{C}$ as well as generates $\sigma(\mathfrak{X})$, and which induces the set system $\mathcal{S}$ consisting of all $\bigcap_{n=1}^{\infty} X_n^{-1}([x_n, \infty[)$, where $X_n \in \mathfrak{L}$ nonnegative, bounded, $x_n > 0$. Additionally, let E consist of all bounded $\sup_n Y_n$, where $(Y_n)_n$ is a sequence of nonnegative bounded positions from $\mathfrak{L}$.

One might think of an investor who is not aware of his or her preferences on the entire space $\mathfrak{X}$ but only on the subspace $\mathfrak{L}$. Let us also assume that he or she has a class of possible market models in mind yielding a $\sigma-$additive robust representation of $\rho|\mathfrak{L}$. Then for the modelling of the preferences on the whole set $\mathfrak{X}$ of available positions it might be useful for the investor to have conditions to hand which lead to a risk assessment consistent with her or his risk- and ambiguity-aversity expressed by the $\sigma-$additive robust representation of $\rho|\mathfrak{L}$.

First of all, in view of the inner Daniell-Stone theorem (cf. [11], Theorem 5.8, final remark after Addendum 5.9) every probability measure $\mathrm{P} \in \mathcal{M}_1$ has to be inner regular w.r.t. $\mathcal{S}$, i.e. $\mathrm{P}(A) = \sup_{A \supseteq B \in \mathcal{S}} \mathrm{P}(B)$ for every $A \in \sigma(\mathfrak{X})$. So within this setting we are dealing with robust representations of ρ by probability measures from $\mathcal{M}_1(\mathcal{S})$ defined to consist of all probability measures belonging to $\mathcal{M}_1$ which are inner regular w.r.t. $\mathcal{S}$ and which represent π on $\mathfrak{C}$. As a consequence we obtain the following necessary condition for a $\sigma-$additive robust representation of ρ (cf. [15], section 7).

Proposition 5 *If* ρ *has a robust representation w.r.t. some* $\mathcal{M} \subseteq \mathcal{M}_1$, *then* $\rho(X) = \sup_{X \leq Y \in E} \inf_{Y \geq Z \in \mathfrak{X}} \rho(Z)$ *for every bounded nonegative* $X \in \mathfrak{X}$.

Imposing the cutting condition, it remains to focus on the nonnegative bounded positions for representation purposes due to Proposition 3 and the translation invariance of ρ. Then by the necessary regularity from Proposition 5 the restriction of ρ to the bounded positions has to be already determined by the values of ρ at the bounded positions from $\mathcal{L}$. Moreover, a $\sigma-$additive robust representation might be guaranteed if the following properties are satisfied

(*) $\Lambda|\mathcal{L}$ is representable by a probability measure from $\mathcal{M}_1(\mathcal{S})$ for $\beta_\rho(\Lambda) < \infty$,

(**) $\alpha_\rho(\mathrm{P}) = \sup_{Y \in \mathcal{L}} (-E_{\mathrm{P}}[Y] - \rho(Y))$ for $\alpha_\rho(\mathrm{P}) < \infty$.

Property $(*)$ means that the investor's risk assessment of the positions from $\mathfrak{L}$ relies on a class of possible market models. Consequently the penalty of misspecification should only take into account the values of ρ at the positions from $\mathfrak{L}$, as stated in property $(**)$.

The general representation result w.r.t. inner regular probability measures encloses conditions which imply the properties $(*), (**)$.

Theorem 1 *Let* Δ_c $(c \in]-\rho(0), \infty[)$ *gather all* P *from* $\mathcal{M}_1(\mathcal{S})$ *with* $\alpha_\rho(\mathrm{P}) \leq c$, *and let* ρ *satisfy the following properties.*

(1) $\lim_{n\to\infty}\rho(-\lambda(X-n)^+)=\rho(0)$ *for every nonnegative* $X\in\mathfrak{X}$ *and* $\lambda>0$,

(2) $\rho(X)=\sup_{X\le Y\in E}\inf_{Y\ge Z\in\mathfrak{X}}\rho(Z)$ *for all nonnegative bounded* $X\in\mathfrak{X}$,

(3) $\rho(X_n)\searrow\rho(X)$ *for any isotone sequence* $(X_n)_n$ *of bounded positions* $X_n\in\mathfrak{L}$ *with* $X_n\nearrow X\in\mathfrak{L}$, X *bounded,*

(4) $\inf_{Y\ge Z\in\mathfrak{X}}\rho(Z)=\inf_{Y\ge Z\in\mathfrak{L}}\rho(Z)$ *for* $Y\in E$.

Then we may state:

.1 The initial topology $\tau_{\mathfrak{L}}$ *on* $\mathcal{M}_1(\mathcal{S})$ *induced by the mappings* $\psi_X:\mathcal{M}_1(\mathcal{S})\to\mathbb{R},\ \mathrm{P}\mapsto E_{\mathrm{P}}[X]$, *(*$X\in\mathfrak{L}$*) is completely regular and Hausdorff.*

.2 Each Δ_c $(c\in]-\rho(0),\infty[)$ *is compact w.r.t.* $\tau_{\mathfrak{L}}$, *and furthermore for every* Λ *from the domain of* β_ρ *there is some* $\mathrm{P}\in\mathcal{M}_1(\mathcal{S})$ *with* $\Lambda|\mathfrak{L}=E_{\mathrm{P}}|\mathfrak{L}$ *and* $\alpha_\rho(\mathrm{P})\le\beta_\rho(\Lambda)$.

.3 $\rho(X)=\sup_{\mathrm{P}\in\mathcal{M}_1(\mathcal{S})}(E_{\mathrm{P}}[-X]-\alpha_\rho(\mathrm{P}))$ *for all* $X\in\mathfrak{X}$.

Statement .1 is borrowed from [14] (p.12 there), the proof of the remaining parts of Theorem 1 may be found in [15] (section 7).

Remarks 1 *Assumption (1) is just the cutting condition as discussed in the previous section, whereas assumption (2) is the necessary regularity condition from Proposition 5. The continuity property (3) combined with the cutting condition yield property (*). In view of Theorem 2 property (*) is even equivalent with (1), (3). Finally the assumptions (1), (4) imply property (**). Moreover, the conditions* $(*),(**)$ *together are equivalent with the assumptions* $(1),(3),(4)$.

Remarks 2 *Let us point out some special situations where the assumptions on* ρ, *imposed in Theorem 1, may be simplified:*

.1 If $\mathfrak{X}$ *is restricted to bounded positions, then assumption (1) is redundant. Also (2), (4) hold in general in the case* $\mathfrak{X}=\mathfrak{L}$.

.2 Assumption (3) is fulfilled in general whenever $\mathfrak{L}_{+b}$, *consisting of all nonnegative bounded* $X\in\mathfrak{L}$, *is a so-called* **Dini cone**, *i.e.* $\inf_n\sup_{\omega\in\Omega}X_n(\omega)=\sup_{\omega\in\Omega}\inf_n X_n(\omega)$ *for any antitone sequence* $(X_n)_n$ *in* $\mathfrak{L}_{+b}$ *with pointwise limit in* $\mathfrak{L}_{+b}$. *The most prominent Dini cones are the cones of nonnegative upper semicontinuous and nonnegative continuous real-valued mappings on compact Hausdorff spaces due to the general Dini lemma (cf. [9], Theorem 3.7).*

.3 If $E\subseteq\mathfrak{X}$, *then assumptions (1), (2) read as follows:*

(1) $\rho(X)=\sup_{X\le Y\in E}\rho(Y)$ *for all nonnegative bounded* $X\in\mathfrak{X}$,

(2) $\rho(Y)=\inf_{Y\ge Z\in\mathfrak{L}}\rho(Z)$ *for* $Y\in E$.

Let us now consider some special situations where Theorem 1 might be used.

Remark 2 *Let* $\Omega=\widetilde{\Omega}\times\mathbb{T}$ *with* $\widetilde{\Omega}$ *denoting a set of scenarios, equipped with a metrizable topology* $\tau_{\widetilde{\Omega}}$ *as well as the induced* $\sigma-$*algebra* $\mathcal{B}(\widetilde{\Omega})$, *and* $\mathbb{T}$ *being a time set, endowed with a separably metrizable topology* $\tau_{\mathbb{T}}$ *as well as the generated* $\sigma-$*algebra* $\mathcal{B}(\mathbb{T})$. *Furthermore let* $\mathfrak{X}$ *consist of all bounded real-valued mappings on* $\widetilde{\Omega}\times\mathbb{T}$ *which are measurable w.r.t. the product* $\sigma-$*algebra* $\mathcal{B}(\widetilde{\Omega})\otimes\mathcal{B}(\mathbb{T})$, *and let* $\mathfrak{L}$ *be the set of all bounded real-valued mappings on* $\widetilde{\Omega}\times\mathbb{T}$ *which are continuous w.r.t. the product topology* $\tau_{\widetilde{\Omega}}\times\tau_{\mathbb{T}}$. *Finally* $\mathcal{S}$ *is defined to gather the closed subsets of* $\widetilde{\Omega}\times\mathbb{T}$ *w.r.t. the metrizable topology* $\tau_{\widetilde{\Omega}}\times\tau_{\mathbb{T}}$. *Using the introduced notations,* $\sigma(\mathfrak{X})=\mathcal{B}(\Omega)\otimes\mathcal{B}(\mathbb{T})$, *the product* $\sigma-$*algebra of* $\mathcal{B}(\Omega)$ *and* $\mathcal{B}(\mathbb{T})$, *is generated by* $\mathcal{S}$, $\mathfrak{L}\subseteq\mathfrak{X}$, *and we may restate Theorem 1 with* E *being the space of all bounded nonnegative lower semicontinuous mappings on* $\Omega\times\mathbb{T}$. *This version generalizes an analogous result for the one-period positions (cf. [13], Theorem 2)*

We may also utilize Theorem 1 for cadlag positions.

Remark 3 *Let* $\mathbb{T}=[0,T],\mathfrak{C}=\mathbb{R}$, *let* $(\mathcal{F}_t)_{t\in\mathbb{T}}$ *be a filtration of* $\sigma-$*algebras on some nonvoid set* $\widetilde{\Omega}$, *and let* $\mathfrak{X}$ *be the set of cadlag positions, i.e. mappings* $X\in\mathbb{R}^{\widetilde{\Omega}\times\mathbb{T}}$ *such that* $X(\cdot,t)$ *is* $\mathcal{F}_t-$*measurable for every* $t\in\mathbb{T}$ *and* $X(\omega,\cdot)$ *is a cadlag function for any* $\omega\in\widetilde{\Omega}$. *Then* $\sigma(\mathfrak{X})$ *is the so-called optional* $\sigma-$*algebra. We may associate for stopping times* $S_1,S_2,\ S_1\le S_2$, *the stochastic interval* $[S_1,S_2[$, *defined by* $[S_1,S_2[(\omega,t):=1$ *if* $S_1(\omega)\le t<S_2(\omega)$, *and* $[S_1,S_2[(\omega,t):=0$ *otherwise.* $\mathfrak{I}$ *stands for the set of all such stochastic intervals. It can be shown that* $\sigma(\mathfrak{X})$ *is generated by the stochastic intervals* $[S,\infty[$ *(cf. [3], IV, 64).*

For $\mathfrak{L}$ *let us choose the vector space spanned by the stochastic intervals* $[S,\infty[$. *Using the introduced notations, we may restate Theorem 1.*

Remark 4 *Recently, convex risk measures has been used as objectives of optimization problems like e.g. the investment for asset allocations or the choice of consumption-investment plans, when the investor is risk- and ambiguity-averse (cf. e.g. [19], [22]). Then Theorem 1 provides not only a criterion which recognizes an investor with such an risk attitude, but it*

might be also the starting point to get on to tracks of robust expected utility maximization. In particular the compactness statement .2 of Theorem 1 may allow to employ duality methods for the optimization problems.

4 Strong $\sigma-$additive robust representation of convex risk measures

We want to look for conditions which induce a strong robust representation of ρ by probability measures in the sense that

$$\rho(X) = \max_{\mathrm{P}\in\mathcal{M}_1} (-E_{\mathrm{P}}[X] - \alpha_\rho(\mathrm{P}))$$

holds for any $X \in \mathfrak{X}$. The considerations are reduced to a Stonean vector lattice $\mathfrak{X}$ being stable w.r.t. countable convex combinations of antitone sequences of financial positions. In this case the following result gives a complete answer to the problem of strong robust representations.

Theorem 2 *Let $\mathfrak{X}$ be a Stonean vector lattice and let us assume that for every antitone sequence $(X_n)_n$ in $\mathfrak{X}$ with $X_n \searrow 0$ and each sequence $(\lambda_n)_n$ in $[0,1]$ with $\sum_{n=1}^{\infty} \lambda_n = 1$ there is some pointwise limit $\sum_{n=1}^{\infty} \lambda_n X_n$ of $(\sum_{n=1}^{m} \lambda_n X_n)_m$ belonging to $\mathfrak{X}$. Then the following statements are equivalent:*

.1 $\rho(X) = \max_{\mathrm{P}\in\mathcal{M}_1} (-E_{\mathrm{P}}[X] - \alpha_\rho(\mathrm{P}))$ holds for every $X \in \mathfrak{X}$.

.2 $\rho(X_n) \searrow \rho(X)$ for $X_n \nearrow X$.

.3 Λ is representable by a probability measure from $\mathcal{M}_1$ for $\beta_\rho(\Lambda) < \infty$.

The implication .2 $\Rightarrow$ .3 may be concluded from Theorem 1, whereas .3 $\Rightarrow$ .1 is trivial due to Proposition 1. The proof of the implication .1 $\Rightarrow$ .2 may be found in [15] (section 9), its crucial tool is Simons' lemma (cf. [20], Lemma 2). For application of this result we need the assumed stability w.r.t. countable convex combinations of positions.

Remark 5 *The continuity property .2 in Theorem 2 is implied by a technically simplier one, which is even equivalent in many cases (cf. [15], Theorem 4.1).*

For bounded one-period positions, Theorem 2 enables us to give an equivalent characterization of convex risk measures that admit strong robust representations by probability measures.

Corollary 1 *Let $\mathcal{F}$ denote some $\sigma-$algebra on Ω, and let $\mathfrak{X}$ consist of all bounded $\mathcal{F}-$measurable real-valued mappings. Then the following statements are equivalent:*

.1 $\rho(X) = \max_{\mathrm{P}\in\mathcal{M}_1} (-E_{\mathrm{P}}[X] - \alpha_\rho(\mathrm{P}))$ holds for every $X \in \mathfrak{X}$

.2 $\rho(X_n) \searrow \rho(X)$ for $X_n \nearrow X$.

Originally, the implication .2 $\Rightarrow$ .1 of Corollary 1 may be found in [5], whereas the full equivalence has been shown the first time in [13].

In the case of $\mathfrak{X} = \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$, the set of all essentially bounded mappings w.r.t. a reference probability measure P on a $\sigma-$algebra $\mathcal{F}$, we may retain immediately the equivalent characterization of strong robust representations for ρ shown in [8], where the identity on $\mathbb{R}$ has been chosen for the price functional π. Note that the condition $\rho(X_n) \searrow \rho(X)$ for $X_n \nearrow X$ P $-$a.s. is equivalent with the property $\rho(X_n) \searrow \rho(X)$ for $X_n \nearrow X$.

Corollary 2 *Let $\mathfrak{X} = \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$, and let ρ satisfy $\rho(X) = \rho(Y)$ for $X = Y$ P a.s.. Then the representation $\rho(X) = \max_{\mathrm{Q}\in\mathcal{M}_1} (-E_{\mathrm{Q}}[X] - \alpha_\rho(\mathrm{Q}))$ holds for all $X \in \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ if and only if $\rho(X_n) \searrow \rho(X)$ for $X_n \nearrow X$ P $-$a.s..*

Remark 6 *Besides the potential for robust expected utility maximization as emphasized in Remark 4, Theorem 2 has significance from the practical point of view. In many cases the calculation of outcomes of risk measures has to be employed by numerical optimization algorithms, and the most customary ones assume the existence of solutions. Therefore Theorem 2 can be used to check whether the desired algorithms may be applied.*

5 Representation of convex risk measures by probability measures and the Fatou properties

In Proposition 4 we have indicated the Fatou property and continuity from above as necessary conditions for a $\sigma-$additive robust representation of the convex risk measure ρ. They are even sufficient if a market model is available for the investor, choosing the identity on $\mathbb{R}$ for the price functional (cf. [5]). As pointed by Delbaen (in [2]), they are not sufficient in general for a robust representation of ρ by ($\sigma-$additive) probability measures, even if $\mathfrak{X}$ contains bounded positions only. It will turn out by the investigations within this

section that in the case of uncertainty about the market model the nonsequential counterpart of the Fatou property takes over partly the role that the Fatou property plays when a reference probability measure is given. We shall say that ρ satisfies the **nonsequential Fatou property** if $\liminf_i \rho(X_i) \geq \rho(X)$ holds whenever $(X_i)_{i\in I}$ is a uniformly bounded net in $\mathfrak{X}$ which converges pointwise to some bounded $X \in \mathfrak{X}$.

At least for the sufficiency of the Fatou property we need further assumptions on the space $\mathfrak{X}$ of available positions. Since both Fatou properties are related to the pointwise topology on the space $B(\Omega)$, gathering the bounded real-valued mappings on Ω, we shall impose additional assumptions on this topology. The idea is to modify in view of Proposition 1 the classical proofs for the case of a reference probability measure, using again the Krein-Smulian theorem. Justified by success we shall use the following conditions.

(5.1) For any $r > 0$, every $Z \in \mathfrak{X}_b$ from the closure of $A_r := \{X \in \mathfrak{X}_b \mid \rho(X) \leq 0, \ \sup_{\omega\in\Omega} |X(\omega)| \leq r\}$ w.r.t. the topology of pointwise convergence on $\mathfrak{X}_b$ is the pointwise limit of a sequence in A_r.

(5.2) The sets $B_r := \{X \in \mathfrak{X}_b \mid \sup_{\omega\in\Omega} |X(\omega)| \leq r\}$ $(r > 0)$ are closed w.r.t. the topology of pointwise convergence on $B(\Omega)$.

Assumption (5.1) provides an important special situation when the Fatou property and its nonsequential counterpart are equivalent.

Lemma 1 *Under (5.1) ρ satisfies the nonsequential Fatou property if and only if it fulfills the Fatou property.*

The proof is enclosed in section 9 of [15].

Remark 7 *The sequential condition (5.1) is closely related with the concepts of double limit relations. For a comprehensive exposition the reader is referred to [12]. In general one may try to apply double limit relations to $\mathfrak{X}_b$ and suitable sets of bounded countably additive set functions on $\sigma(\mathfrak{X})$.*

We are now ready for the main result of this section.

Theorem 3 *Let either $\mathfrak{X} = \mathfrak{X}_b$ or $\mathfrak{X}$ be a Stonean vector lattice such that $\lim_{n\to\infty} \rho(\lambda(X - n)^+) = \rho(0)$ holds for any nonnegative $X \in \mathfrak{X}$, $\lambda > 0$. Furthermore let $\alpha^{-1}(\mathbb{R}) \neq \emptyset$. Consider the following statements:*

.1 ρ satisfies the nonsequential Fatou property.

.2 ρ has a $\sigma-$additive robust representation w.r.t. $\mathcal{M}_1$.

.3 ρ fulfills the Fatou property.

If (5.2) is valid, then .1 $\Rightarrow$ .2 $\Rightarrow$ .3, and all statements are equivalent provided that condition (5.1) holds in addition.

The proof may be found in section 9 of [15].

Remark 8 *The nonsequential Fatou property is not necessary for a $\sigma-$additive representation of risk measures. Take for example $\mathfrak{X}$ the space of all bounded Borel-measurable mappings on $\mathbb{R}$, and define ρ by $\rho(X) = -E_{\mathrm{P}}[X]$, where P denotes any probability measure which is absolutely convex w.r.t. the Lebesgue-Borel measure on $\mathbb{R}$. Obviously, on one hand ρ is a convex risk measure w.r.t. the identity on $\mathbb{R}$, having a trivial $\sigma-$additive robust representation. On the other hand, consider the net $(X_i)_{i\in I}$ of all indicator mappings of the cofinite subsets of $\mathbb{R}$, directed by set inclusion. It converges pointwise to 0, but unfortunately $\liminf_i \rho(X_i) = -1 < 0 = \rho(0)$.*

In the case of an at most countable Ω, we have a simplified situation which admits an application of the full Theorem 3. The reason is that then the topology of pointwise convergence on the space $B(\Omega)$ is metrizable.

Corollary 3 *Let Ω be at most countable, and let $\mathfrak{X} \subseteq B(\Omega)$ be sequentially closed w.r.t. the pointwise topology on $B(\Omega)$. Then ρ has a robust representation by probability measures from $\mathcal{M}_1$ if and only it satisfies the Fatou property, or equivalently, if and only if ρ is continuous from above.*

Remark 9 *Let a market model with reference probability measure P be given, and let $\mathfrak{X} := \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ be the space of all $\mathrm{P}-$essentially bounded mappings on Ω. Furthermore ρ is supposed to be a convex risk measure w.r.t. the identity on $\mathbb{R}$, satisfying $\rho(X) = \rho(Y)$ for $X = Y$ $\mathrm{P}-a.s.$. We may apply the full Theorem 3 (cf. section 9 in [15]) to retain an equivalent characterization of the robust representations for ρ which may be found in [5] (Theorem 4.31). More precisely, if $\mathcal{M}_1(\mathrm{P})$ denotes the set of probability measures on $\mathcal{F}$ which are absolutely continuous w.r.t. P, then the following statements are equivalent.*

.1 $\rho(X) = \sup_{\mathrm{Q}\in\mathcal{M}_1(\mathrm{P})} (-E_{\mathrm{Q}}[X] - \alpha_\rho(\mathrm{Q}))$ for all X from $\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$.

.2 $\rho(X_n) \nearrow \rho(X)$ for $X_n \searrow X$ $\mathrm{P}-a.s.$.

.3 $\liminf_{n\to\infty} \rho(X_n) \geq \rho(X)$ whenever $(X_n)_n$ is a uniformly $\mathrm{P}-$essentially bounded sequence in $\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathrm{P})$ with $X_n \to X$ $\mathrm{P}-a.s.$.

It is unclear whether we may avoid in Theorem 3 condition (2.2) in order to guarantee a $\sigma-$additive robust representation of convex risk measures by the nonsequential Fatou property. Moreover, the nonsequential Fatou property is unsatisfactory in the way that it does not work for trivial representations like those indicated in Remark 8. However, we may only guarantee a sufficient substitution by the Fatou property under the quite restrictive condition (2.1). So it seems that in presence of model uncertainty the Fatou property and its nonsequential counterpart are appropriate conditions for $\sigma-$additive representations of convex risk measures in quite exceptional situations only, like an at most countable Ω.

Acknowlegdements

The author would like to thank Freddy Delbaen and Alexander Schied for helpful hints. He is also indepted to Heinz König for good advice and continuous exchange of ideas from the fields of measure theory and superconvex analysis. Finally, he is grateful for some remarks by two anonymous referees which have helped to improve an earlier draft.

References

[1] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath. Coherent measures of risk. Math. Finance 9, 203-228, 1999.

[2] F. Delbaen. Coherent Measures of Risk on General Probability Spaces. In: K. Sandmann, P.J. Schönbucher (Eds.). "Advances in Finance and Stochastics". Springer, Berlin, 2002, pp. 1-37.

[3] C. Dellacherie, P.-A. Meyer. "Probabilities and Potential". North-Holland, Amsterdam, 1978.

[4] D. Denneberg. "Non-Additive Measure and Integral". Kluwer, Dordrecht, 1994.

[5] H. Föllmer, A. Schied. "Stochastic Finance". de Gruyter, Berlin, New York, 2004 (2nd ed.).

[6] M. Fritelli, E. Rosazza Gianin. Dynamic Convex Risk Measures. In: G. Szego (Ed). "Risk Measures for the 21st Century". Wiley, New York, 2004, pp. 227-248.

[7] M. Frittelli, G. Scandolo. Risk measures and capital requirements for processes. Math. Finance 16, 589-612, 2006.

[8] E. Jouini, W. Schachermayer and N. Touzi. Law invariant risk measures have the Fatou property. Advances in Mathematical Economics 9, 49-71, 2006.

[9] H. König. On some basic theorems in convex analysis. In: B. Korte (Ed.). "Modern Applied Mathematics - Optimization and Operations Research". North-Holland, Amsterdam,1982, pp. 107-144.

[10] H. König. "Measure and Integration". Springer, Berlin et al., 1997.

[11] H. König. Measure and Integration: Integral representations of isotone functionals. Annales Universitatis Saraviensis 9, 123-153, 1998.

[12] H. König, N. Kuhn. Angelic spaces and the double limit relation. J. London Math. Soc. 35, 454-470, 1987.

[13] V. Krätschmer. Robust representation of convex risk measures by probability measure. Finance and Stochastics 9, 597-608, 2005.

[14] V. Krätschmer. Compactness in spaces of inner regular measures and a general Portmanteau lemma. SFB 649 discussion paper 2006-081, downloadable at http://sfb649.wiwi.hu-berlin.de.

[15] V. Krätschmer. On $\sigma-$additive representations of convex risk measures for unbounded positions in the presence of uncertainty about the market model. SFB 649 discussion paper 2007-010, downloadable at http://sfb649.wiwi.hu-berlin.de.

[16] R. Pelessoni and P. Vicig. Convex imprecise previsions. Reliab. Comput. 9, 465-485, 2003.

[17] F. Riedel. Dynamic Coherent Risk Measures. Stochastic Processes and Applications 112, 185-200, 2004.

[18] A. Ruszczynski, A. Shapiro. Optimization of convex risk functions. Math. Oper. Res. 31, 433-451, 2006.

[19] A. Schied, Optimal investments for risk- and ambiguity-averse preferences: a duality approach. Finance Stoch 11, 107 - 129, 2007.

[20] S. Simons. A convergence theorem with boundary. Pacific J. Math. 40, 703-708, 1972.

[21] S. Weber. Distribution-invariant dynamic risk measures, information, and dynamic consistency. Math. Finance 16, 419-442, 2006.

[22] W. Wittmüß. Robust Optimization of Consumption with Random Endowment, SFB 649 discussion paper 2006-063, downloadable at http://sfb649.wiwi.hu-berlin.de.

Updating and Testing Beliefs: An Open Version of Bayes' Rule

Elmar Kriegler
Department of Engineering and Public Policy, Carnegie Mellon University
Potsdam Institute for Climate Impact Research
elmar@cmu.edu

Abstract

Developing models to describe real systems is a challenge because it is difficult to assess and control the residual between the two entities. Bayesian updating of a belief about model accuracy across an ensemble of available models can lead to spurious results, since the application of Bayes' rule presupposes that an accurate model is contained in the ensemble with certainty. We present a framework in which this assumption can be dropped. The basic idea is to extend Bayes' rule to the exhaustive, but unknown space of all models, and then contract it again to the known set of models by making best/worst case assumptions for the remaining space. We show that this approach leads to an ε-contamination model for the posterior belief, where the ε-contamination is updated along with the distribution of belief across available models. In essence, the ε-contamination provides an additional test on the accuracy of the overall model ensemble compared to the data, and will grow rapidly if the ensemble fails such a test. We demonstrate our concept with an example of autoregressive processes.

Keywords. Bayesian updating, prediction, model accuracy, ε-contamination model, AR process

1 Introduction

A vital part of the scientific endeavor consists in developing models for real systems. Obviously, a model can never be an identical copy of a real system, but rather a proxy to understand a limited set of system features, on the basis of which future observations of these features may be predicted. In order to construct a useful model, it is important to control the residual between model and real system in a way that allows the model to have some predictive accuracy. Therefore, it is extremely helpful if the real system can be studied in laboratory experiments where the experimenter can force her ways on it to test the model. However, controlling the residual becomes an enormous challenge if the real system is not accessible to laboratory studies. The situation is further exacerbated if available observations cover only a small part of the phase space. The climate system and computer models of it are a perfect example of this situation, and we will have this example in mind in what follows.

In such cases, model quality is usually assessed with a mixture of scientific knowledge about the system and statistical inference from system measurements. Here, we focus on model accuracy to predict certain system features. While several definitions of accuracy can be found in the literature (e.g. in terms of bias), we use our own definition tailored to an application to dynamic systems characterized by noisy time series. We say that a model S is *accurate* (to make predictions) if it can describe the observables Y of a (few) system feature(s) of interest – not of the entire system – up to an additive Gaussian iid process ϵ, i.e., $Y - S = \epsilon \sim N(0, \sigma)$. The choice of a Gaussian iid process for the residual between model and data is a common, but subjective assumption, and should be regarded as part of the model formulation. In principle, our approach could be applied with another choice of stationary stochastic process for the residual.

Assume we have an ensemble of model hypotheses $M(\theta)$, with $\theta \in \Theta$ indexing the available models, and some system data $\hat{y}$, from which we want to learn about the relationship between $M(\theta)$ and an accurate model S. If a probability $P(M(\theta) = S|\hat{y})$ is sought, we have to turn to Bayesian statistics. In our case, this requires to

1. estimate a likelihood function $\mathcal{L}(\theta; \hat{y}) \sim \rho(\hat{y}|\theta)$, i.e., the probability of observing $\hat{y}$ for a given model $M(\theta)$ under the assumption that $M(\theta)$ constitutes an accurate model (which defines the likelihood function given our knowledge about the residual $Y - S = \epsilon$), and
2. updating it with a prior probability density $\rho(\theta)$:

$\Omega \to \mathbb{R}_0^{+1}$ to derive the posterior probability density $\rho(\theta|\hat{y})$ that model $M(\theta)$ constitutes an accurate model S.

However, this approach requires us to make the assumption that S is contained in our model ensemble with certainty as evident from Bayes' rule

$$\rho(\theta|\hat{y}) = \frac{\mathcal{L}(\theta;\hat{y})\rho(\theta)}{\rho(\hat{y})} \quad (1)$$
$$\text{with} \qquad \rho(\hat{y}) := \int_\theta \mathcal{L}(\theta;\hat{y})\rho(\hat{y})d\theta\,,$$

where the denominator assures that the posterior probability is normalized. In the following, we may call the fact that $\int_\theta \rho(\theta)d\theta = \int_\theta \rho(\theta|\hat{y})d\theta = 1$ *closed world assumption*.

We believe that this assumption is at odds with the open nature of the scientific endeavor, where a set of possible models $\{M(\theta)|\theta \in \Theta\}$ imagined at some initial time is usually expanded as more data is obtained. More precisely, the model development process consists in (I) expanding the set Θ of known models, and (II) updating our belief about model accuracy across Θ. Obviously, only the later type (II) learning can be described in terms of Bayesian learning. The former type (I) may be informed by Bayesian inference, but seems to be complementary to it, since it relates to the emergence of positive belief in an area of the model space that was not supported by the prior belief.

Acknowledging this fundamental difference, we will not attempt to force type (I) learning in terms of expanding Θ into the Bayesian updating framework. Instead, we aim at the more modest goal to include an indicator for the necessity of type (I) learning into the updating process. This is important because naive application of Bayesian learning without contemplating the possibility that the entire model ensemble $\{M(\theta)|\theta \in \Theta\}$ might not contain an accurate model can lead to spurious results. As the amount of data $\hat{y}$ increases, the likelihood function tends to sharpen, and updating by means of Equation (1) will decrease the spread of the posterior belief that a given model $M(\theta)$ coincides with the accurate model S. Hence, an analyst ignoring anything else will converge in his belief on some model $M(\theta) = S$. As a consequence, his predictions of real system features, based on his converging belief, will grow more and more (over)confident – although off the mark – as the data accumulates. This paradoxical behavior is a direct consequence of the closed world assumption. It is therefore desirable to drop this assumption, and directly include an indicator for $S \notin \{M(\theta)|\theta \in \Theta\}$ in the updating process. In this paper, we present such a framework.

A similar concern about Bayesian learning on model quality and the subsequent use of posterior beliefs for prediction of future observations has been raised by Draper [2] and, more recently, Goldstein and Rougier [3, 4]. Draper criticizes the practice of neglecting structural uncertainty, and proposes to extend prior and likelihood to the space of possible model structures. His approach [2] leads to an increased spread of the posterior on the model ensemble. Goldstein and Rougier highlight the importance to assess the discrepancy between the ensemble of available models and the 'ideal' model which captures the system up to an additive noise term. They coined the term 'reified' for the 'ideal' reference model. Obviously, the idea of a 'reified model' is closely related to what we call accurate model here. In [3, 4], Goldstein and Rougier propose to address model discrepancy by including a meta-model of it in the updating process, and offer guidelines how such a meta-model might be constructed. This is a very challenging task. As indicated above, we take a different approach. We do not try to find a positive expression for model discrepancy, or the extension of prior and likelihood to the space of possible model structures, but rather seek to include an indicator for the negative result that model discrepancy impinges on the predictive accuracy of the model ensemble.

The paper is organized as follows. Section 2 presents a simple example of autoregressive (AR) processes in which the application of standard Bayesian updating is shown to fail if the model hypotheses have limited accuracy to predict the real system. Section 3 contains the core of the paper, detailing our derivation of an open version of Bayes' rule that allows to drop the closed world assumption. This rule is put into operation for our example of AR processes in Section 4. We conclude by highlighting the challenges for an application of the open Bayes' rule to real world problems in Section 5.

2 Limitations of Bayes' rule: Example of autoregressive (AR) processes

Let us assume the following dynamic 'real system' evolving over n time steps.

$$Y(n) = (\xi_1, \alpha_1^*\xi_1 + \xi_2, X_3, ..., X_n) \quad (2)$$
$$\text{with} \qquad X_t = \alpha_1^* X_{t-1} + \alpha_2^* X_{t-2} + \xi_t\,,\ t \geq 3 \quad (3)$$
$$\xi_t \sim N(0, \sigma_\xi^*) \quad \text{iid process (white noise)}\,,$$

[1] For the sake of simplicity, we assume throughout the paper that $\Omega \subseteq \mathbb{R}^n$ is a continuous space, and that a prior probability measure $P : \sigma(\Theta) \to [0,1]$ over a σ-field of Θ is continuous, i.e., can be described by a probability density on Θ.

where we require the AR(2) process X_t to be stationary. An AR(2) process described by Equation (3) is stationary iff $\alpha_1^* + \alpha_2^* < 1$, $\alpha_2^* - \alpha_1^* < 1$, and $|\alpha_2^*| < 1$. For the sake of simplicity, we have neglected any measurement error in observing the real system, and therefore can identify it directly with the observable $Y(n)$. Let us further assume that our ensemble of model hypotheses for $Y(n)$ is restricted to a closed set of stationary AR(1)-process with noise term $\xi_t \sim N(0, \sigma_\xi^*)$:

$$\{M(\alpha_1) := (\xi_1, X_2', ..., X_n') \mid X_t' = \alpha_1 X_{t-1}' + \xi_t\,, \quad t \geq 2\,,\ -\bar{\alpha} \leq \alpha_1 \leq \bar{\alpha}\,,\ \bar{\alpha} := 0.995\}\,. \quad (4)$$

Obviously, the model ensemble contains an accurate model S if $\alpha_1^* \in [-\bar{\alpha}, \bar{\alpha}]$ and $\alpha_2^* = 0$. In this case, we find $S := M(\alpha_1^*) = Y(n)$. We will discuss below whether there can be an accurate model in the ensemble if $\alpha_2^* \neq 0$.

After having received a realization $\hat{y}(n) = (\hat{y}_1, ..., \hat{y}_n)$ of $Y(n)$, we can apply Bayesian updating to our prior belief about the accuracy of the model hypotheses $M(\alpha_1)$ as defined by a probability density $\rho(\alpha_1)$. Without loss of generality, let the prior $\rho(\alpha_1)$ be uniformly distributed on $[-\bar{\alpha}, \bar{\alpha}]$. As shown in Appendix A, the likelihood of having obtained the realization $\hat{y}(n)$ from an AR(1)-process with propagator α_1 is given by

$$\mathcal{L}(\alpha_1; \hat{y}(n)) \sim N\left(\frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}, \frac{\hat{\sigma}(n)}{\sqrt{1-\hat{\beta}(n)}}\right)\,, \quad (5)$$

where $\hat{\alpha}(n)$, $\hat{\sigma}(n)$, and $\hat{\beta}(n)$ are estimated from the observed time series $\hat{y}(n)$ as defined in Equation (26), (27), and (28), respectively. Hence, application of Bayes rule (Equation 1) with a uniform prior for α_1 yields the following posterior probability density on $[-\bar{\alpha}, \bar{\alpha}]$:

$$\rho(\alpha_1|\hat{y}(n)) = \frac{\exp\left(-\frac{1-\hat{\beta}(n)}{2\hat{\sigma}(n)^2}\left(\alpha_1 - \frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}\right)^2\right)}{\int_{-\bar{\alpha}}^{\bar{\alpha}} \exp\left(-\frac{1-\hat{\beta}(n)}{2\hat{\sigma}(n)^2}\left(\alpha_1 - \frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}\right)^2\right) d\alpha_1}\,. \quad (6)$$

Equation (6) can be used to test the effect of the closed world assumption on the Bayesian updating process. For the experiment, we generated 200 realizations of time series $\hat{y}(n)$ with length $n = 5000$ for four different AR(2)-processes with $\sigma_\xi^* = 1$, $\alpha_1^* = 0.866$ and $\alpha_2^* = \{-0.9, -0.3, 0, 0.06\}$. Note that in the asymptotic limit $n \to \infty$ any AR(k)-process is normally distributed $\sim N\left(0, \sigma/\sqrt{1 - \sum_{i=1}^k \alpha_i \rho_i}\right)$, with ρ_i the autocorrelation of lag i [7]. Therefore, removing the time index from the observations renders AR-processes of different order indistinguishable from each other. It is in this sense, that we can calculate an AR(1)-equivalent of an AR(2)-process with propagators α_1 and α_2. The AR(1)-equivalent yielding a normal distribution with identical standard deviation in the asymptotic limit has the propagator

$$\alpha_{\text{equiv}} = \sqrt{\alpha_1\rho_1 + \alpha_2\rho_2} = \sqrt{\alpha_1^2 \frac{1+\alpha_2}{1-\alpha_2} + \alpha_2^2}\,. \quad (7)$$

For the four different AR(2)-processes chosen above we find AR(1)-equivalents with propagators $\alpha_{\text{equiv}} = \{0.922, 0.703, 0.866, 0.922\}$. It can be seen that the asymptotic distribution of the two AR(2)-processes with $\alpha_2^* = -0.9$ and $\alpha_2^* = 0.06$ are indistinguishable.

We have considered AR(2)-processes with very pronounced tails compared to ξ, because we are interested in the ability of the model ensemble $M(\alpha_1)$ to predict in particular the tails of the distributions. In practice, a good prediction of the tails is often what matters most. Note that it follows from Equation (7) that there will exist an accurate model S in the ensemble of model hypotheses $M(\alpha_1)$ even if the real system is described by an AR(2)-process with $\alpha_2 \neq 0$ - if we are only interested in predicting the asymptotic distribution of future observations. It will be interesting to see whether Bayesian updating is capable to converge to the propagator of the AR(1)-equivalent model.

Figure 1 shows the result of Bayesian updating for the four different AR(2)-processes. We have updated the posterior belief about α_1 (see Equation 6) after each 20 new observations. Shown is the development of the 90% confidence limits for the mean value of the posterior distribution. The confidence limits were derived from the sample of 200 time series used in the updating process. It can be seen that the posterior mean converges to the correct value of $\alpha_1^* = 0.866$ (horizontal solid line) in the case where the real system is described by an AR(1) process ($\alpha_2^* = 0$). Convergence is still good if only a small deviation from the AR(1) assumption is considered ($\alpha_2^* = 0.06$). In this case, the posterior mean converges to the propagator $\alpha_{\text{equiv}} = 0.922$ of the AR(1)-equivalent process. However, if the deviation from the AR(1) assumption is negative and increases in magnitude ($\alpha_2^* = -0.3$), the posterior belief converges to a biased value below α_{equiv}. In the extreme case $\alpha_2^* = -0.9$, Bayesian learning leads to a spurious result. Since the posterior distribution has contracted strongly after several thousand observations (see black dots on the right axis), the updating procedure has settled on the wrong region of α_1-space with very high confidence. This is a direct consequence of the closed world assumption.

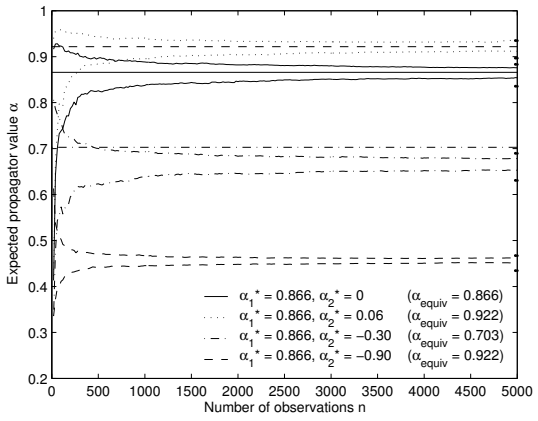


Figure 1: Updated belief about the propagator α of an hypothetical AR(1)-process after n observations. The lower and upper bound of the 90% confidence interval for the mean value of the posterior belief (derived from the sample of 200 time series) are plotted. Horizontal lines indicate the propagator value α_{equiv} of the equivalent AR(1)-process in the asymptotic limit. α_{equiv} for the AR(2)-process with $\alpha_2^* = -0.9$ is identical to the case $\alpha_2^* = 0.06$. Black dots on the right axis indicate the range between the 5% and 95% quantiles of the posterior belief after 5000 observations.

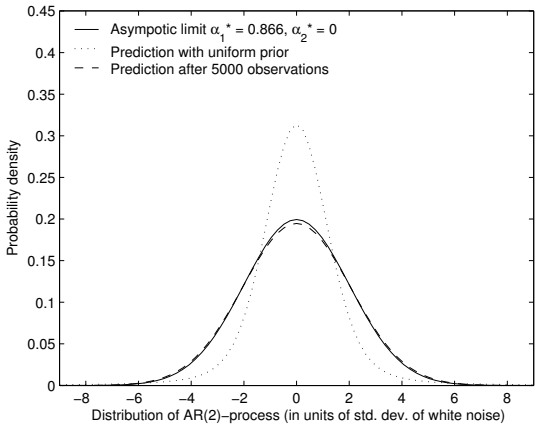


Figure 2: Predictions based on the belief about α_1 for the case $\alpha_1^* = 0.866$ and $\alpha_2^* = 0$. The solid line shows the asymptotic distribution of the 'real' AR(1) process. The dotted line shows the prediction before any learning occurred (based on a uniform prior for $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$). The updated prediction after 5000 observations (dashed line) lies almost exactly on the asymptotic distribution of the 'real' system.

We briefly assess the consequences for predicting the distribution of system observations y in the asymptotic limit. As mentioned above, we know that the asymptotic distribution of an AR(1) process for given values of α_1 and σ_ξ is defined by $\rho(y|\alpha_1) \sim N(0, \sigma_\xi/\sqrt{1-\alpha_1^2})$. Hence, if our belief about α_1 is described by the posterior $\rho(\alpha_1; \hat{y}(n))$, our prediction for the distribution of system observations based on past data $\hat{y}(n)$ is given by

$$\rho(y|\hat{y}(n)) = \int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\ \rho(\alpha_1; \hat{y}(n))\ d\alpha_1\,. \quad (8)$$

Figure 2 shows predictions for the case of learning from a realization of the AR(1)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = 0$. The dotted line depicts the prediction on the basis of the uniform prior, before any learning occurred. Interestingly, the assumption of the uniform prior strongly underestimates the probability mass in the flanks of the distribution. The example shows that in general it is not warranted to associate the uniform prior with a conservative (or non-informative) choice of belief. After the uniform prior is updated with observations $\hat{y}(n)$ the predictions converge very quickly to the asymptotic distribution of the 'real' system. Figure 2 shows that the prediction after 5000 observations is nearly identical with the 'real' distribution.

While Bayesian learning was very successful for the

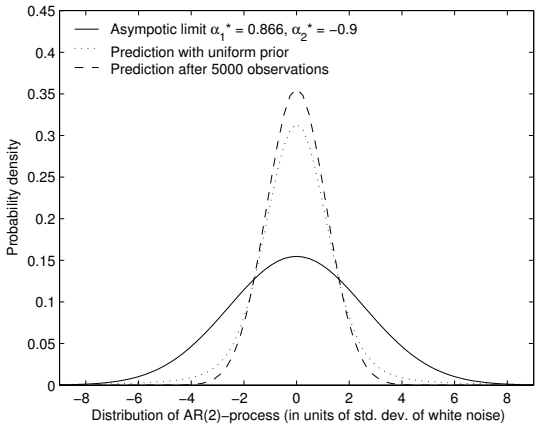


Figure 3: Predictions based on the belief about α_1 for the case $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. Solid, dotted and dashed lines as described in Figure 2.

case where the 'real' system is part of the ensemble of model hypotheses, the situation is markedly different for an AR(2)-process which strongly deviates from the AR(1)-assumption ($\alpha_2^* = -0.9$). As depicted in Figure 3, updating with observations $\hat{y}(n)$ leads to a further decrease in variance of the prediction as compared to the initial prediction based on the uniform prior. This is exactly the opposite of what should happen, because the asymptotic distribution of the 'real' system exhibits a much larger spread than both the initial and informed prediction. As is apparent from Figure 3, these spurious predictions strongly underes-

timate the tails of the distribution, and may therefore provide a false sense of security. What makes matters worse is that no amount of additional data will be able to rectify the situation. In contrast, the posterior belief will continue to sharpen, and the spread of the prediction will further decrease. This example shows that the closed world assumption underlying Bayes' rule can lead to spurious beliefs and predictions.

3 Extension of Bayes rule: dropping the closed world assumption

Given the spurious results that can emerge from a naive application of Bayes' rule, we are looking for an extension of Bayesian updating that includes an indicator for the overall accuracy of the model ensemble to reproduce 'real' system observations. This would allow us to drop the assumption that an accurate model S has to be included in the set of available models $M(\theta),\ \theta \in \Theta$ with certainty. A natural first step in this direction is to extend Bayes rule to a larger space $\Omega \supset \Theta$ for which the assumption $S = M(\omega)$ will be true for at least one $\omega^* \in \Omega$. Similar extensions are also the starting points for the proposals by Draper [2], and Goldstein and Rougier [4]. We assert that such a hypothetical space Ω is constituted by the space of all models, known and unknown. We think of Ω as a continuous vector space with large, but finite dimension that contains the parameter vectors ω for a large, but finite list of time-discrete[2] equations and relations. A model ensemble, i.e., a reduced list of parameterized equations, is characterized in this space by fixing the parameter values in some dimensions (collected in ψ), and allowing to vary – within bounds – the remaining parameters θ. Hence, a choice of model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$, defines a Cartesian product $\Omega = \Theta \times \Psi$, where the parameters $\psi \in \Psi$ are fixed at ψ_0, and only $\theta \in \Theta$ can be varied.

So far, we have gained little because the nature of models in the residual space $\Theta \times (\Psi - \{\psi_0\})$ is completely unknown to us. Thus, our prior belief about the accuracy of unknown models in that space is vacuous. Fortunately, imprecise probability theory allows to capture a vacuous belief without having to assess the cardinality of its underlying space [9]. This is simply done by the vacuous probability model $\mathcal{V}(\Theta \times (\Psi - \{\psi_0\}))$ comprising the set of all probability distributions with support on $\Theta \times (\Psi - \{\psi_0\})$ [8, Chapter 2.9.1]. Since the complement space $\Theta \times \{\psi_0\}$ has zero measure, $\mathcal{V}(\Theta \times (\Psi - \{\psi_0\}))$ is identical to $\mathcal{V}(\Theta \times \Psi)$ almost everywhere. Therefore, we will continue to use the latter vacuous probability model in what follows.

We assume that our prior belief about the 'known' model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$, i.e. more precisely, the set of models considered for our particular assessment, is described by $\nu(\theta, \psi_0)$. How should we combine this prior belief with the vacuous belief on the complementary unknown space? It seems to be a precondition of human agency that we assign non-zero probability to our conception of the 'real world' even though it exists on a space with zero measure. Thus, when it comes to considering the unknown, our prior belief on the space of all models will be degenerate,

$$\nu(\theta, \psi) \in p_0\ \nu(\theta, \psi_0)\ \delta(\psi - \psi_0) + (1 - p_0)\ \mathcal{V}(\Theta \times \Psi)\ , \quad (9)$$

where $\delta(\psi - \psi_0)$ denotes the Dirac measure which concentrates all probability mass on $\psi = \psi_0$, i.e., the set of models available to us. The probability $0 \leq p_0 \leq 1$ weighs our prior belief across the two different domains of knowledge, and may be associated with the prior level of confidence that the model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$ can accurately describe the 'real' system features of interest. For $p_0 = 1$, we completely ignore the possibility that the accurate model may still be unknown. This choice reflects the closed world assumption underlying the standard application of Bayesian learning. In the other extreme, $p_0 = 0$, we are completely lost in the unknown, and cannot expect to learn anything from whatever data we receive. Here, we suggest to choose p_0 as to reflect a typical confidence level used in statistics, e.g., $p_0 = 0.95$. However, the choice of p_0 will not influence the posterior belief significantly (see Equation 15) as long as it is not set to the extreme values of 0 or 1.

Since we cannot talk in positive terms about what we do not know, we are not searching for the posterior belief $\nu(\theta, \psi|\hat{y}(n))$ on the space of all models, but rather for its marginal distribution $\rho(\theta|\hat{y}(n))$ on the subspace of known models. After receiving an observed time series $\hat{y}(n)$, Bayes' rule gives us the following expression for the marginal posterior belief:

$$\rho(\theta|\hat{y}(n)) = \frac{\int_\Psi \mathcal{L}(\theta, \psi; \hat{y}(n))\ \nu(\theta, \psi)\ d\psi}{\int_{\Psi \times \Theta} \mathcal{L}(\theta, \psi; \hat{y}(n))\ \nu(\theta, \psi)\ d\psi d\theta}\ . \quad (10)$$

We follow the usual practice to normalize the likelihood on the space of *known* models to one. Hence, we divide both the nominator and denominator by the maximum likelihood $\mathcal{L}(\theta', \psi_0; \hat{y}(n))$ that we find on the model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$. Inserting the prior belief described by Equation (9) into above

[2] The assumption of time-discrete equations accounts for the numerical implementation of time-continuous differential equations. It shall also extend to other, e.g. spatial, dimensions if partial differential equations are concerned. Thus, we are thinking of computer models here.

expression, we then find

$$\rho(\theta|\hat{y}(n)) \quad \in \quad \frac{p_0\,\mu_{\mathcal{L}}(\theta) + (1-p_0)\,\mathcal{V}_{\mathcal{L}}(\Theta)}{\int_\Theta\,(p_0\,\mu_{\mathcal{L}}(\theta) + (1-p_0)\,\mathcal{V}_{\mathcal{L}}(\Theta)\,)\,d\theta}\,, \quad (11)$$

$$\mu_{\mathcal{L}}(\theta) \quad := \quad \frac{\mathcal{L}(\theta,\psi_0;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\;\nu(\theta,\psi_0)\,,$$

$$\mathcal{V}_{\mathcal{L}}(\Theta) \quad := \quad \int\limits_\Psi \frac{\mathcal{L}(\theta,\psi;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\;\mathcal{V}(\Theta\times\Psi)d\psi\,,$$

where $\mathcal{V}_{\mathcal{L}}(\Theta)$ is the unknown set of marginals on Θ that emerge from multiplying all prior probability distributions on $\Theta\times\Psi$ with an unknown likelihood function. Note that it is not a vacuous probability model itself, since its elements are not normalized. However, this set of marginals is contained in the set of all probability distributions on Θ multiplied by the range of values covered by the likelihood ratio, i.e.

$$\mathcal{V}_{\mathcal{L}}(\Theta) \quad \subset \quad [0,\mathcal{L}^*(n)]\cdot\mathcal{V}(\Theta)$$

$$\text{with} \quad \mathcal{L}^*(n) \quad := \quad \max_{(\theta,\psi)\in\Theta\times\Psi}\frac{\mathcal{L}(\theta,\psi;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\,, \quad (12)$$

Here, the zero lower bound of the interval accounts for the fact that there will certainly be a model with zero likelihood in the space of all models. Note that the nominator of $\mathcal{L}^*(n)$ describes the likelihood function on the entire model space prior to normalization, and therefore can take any value in $\mathbb{R}_0^+$. In the following, we will replace $\mathcal{V}_{\mathcal{L}}(\Theta)$ in the extended Bayes' rule (11) by its superset $[0,\mathcal{L}^*(n)]\cdot\mathcal{V}(\Theta)$ due to greater methodological convenience. This substitution will give us outer bounds on the set of posterior probabilities, but we assert that the associated information loss is minimal. As an example, consider the asymptotic case $n\to\infty$ for which the likelihood function will concentrate around the accurate model at the point (θ^*,ψ^*) $(\mathcal{L}(\theta,\psi;\hat{y}(n))\to\delta(\theta-\theta^*)\;\delta(\psi-\psi^*))$. Then, $\mathcal{V}_{\mathcal{L}}(\Theta)$ will contain only functions proportional to $\delta(\theta-\theta^*)$, which constitutes a considerably smaller set than the functions proportional to $\mathcal{V}(\Theta)$. However, since we are completely ignorant about the location of θ^*, we need to consider $\delta(\theta-\theta^*)$ for all possible values $\theta^*\in\Theta$, which coincides with the set of extreme points of $\mathcal{V}(\Theta)$.

The set of Dirac measures $\delta(\theta-\tilde{\theta})\;\delta(\psi-\tilde{\psi})$, $(\tilde{\theta},\tilde{\psi})\in\Theta\times\Psi$ comprises the extreme points of the vacuous probability model $\mathcal{V}(\Theta\times\Psi)$ on the entire model space. They are all we need to calculate the extreme points of the imprecise posterior probability given by Equation (11) [8, Theorem 8.4.8]. They also tell us that

$$0\le\int_\Theta\;V_{\mathcal{L}}(\Theta)\;d\theta\le\mathcal{L}^*(n)\,, \quad (13)$$

where the upper bound is achieved for the Dirac prior $\delta(\theta-\theta^*)\;\delta(\psi-\psi^*)$.

We are now in the position to separate the extended Bayes rule (11) into a term concerned with updating our prior belief on the model ensemble $M(\theta,\psi_0)$, $\theta\in\Theta$ (the original Bayes' rule), and a term that summarizes the contribution from the residual space of unknown models.

$$\rho(\theta|\hat{y}(n))\in$$

$$(1-\varepsilon(\lambda,p_0))\,\frac{\mu_{\mathcal{L}}(\theta)}{\int_\Theta\,\mu_{\mathcal{L}}(\theta)\,d\theta}+\varepsilon(\lambda,p_0)\,\mathcal{V}(\Theta)\,, \quad (14)$$

$$\varepsilon(\lambda,p_0) \quad := \quad \frac{(1-p_0)\,\lambda}{p_0\;\int_\Theta\;\mu_{\mathcal{L}}(\theta)\,d\theta+(1-p_0)\,\lambda}\,, \quad (15)$$

$$\text{with} \quad \lambda\in[0,\mathcal{L}^*(n)]\,.$$

Since the contamination $\varepsilon(\lambda,p_0)$ increases with λ, the most conservative posterior belief – encompassing the set of posterior probabilities for all possible choices of λ – is obtained in the limit $\lambda\to\mathcal{L}^*(n)$. Therefore, we focus in the following on the most conservative case, for which the vacuous probability model is mixed into the posterior belief with contamination $\varepsilon(\mathcal{L}^*,p_0)$. The ε-contamination model in Equation (14) has been investigated extensively in the context of robust Bayesian and imprecise probability approaches (see, e.g., [1, 5]). It is a very tractable model, since it can be easily characterized by its set of extreme points or its coherent lower probability which constitutes a belief function. Note that we can recover the standard case of Bayesian learning under the closed world assumption from Equations (14) and (15) by choosing $p_0=1$, implying $\varepsilon(\mathcal{L}^*,p_0)=0$. For $p_0\in(0,1)$, the 'contamination' $\varepsilon(\mathcal{L}^*,p_0)$ of our posterior belief will grow with increasing $\mathcal{L}^*(n)$ (see Equation 15). What can we say about $\mathcal{L}^*(n)$, and how will it behave as a function of our observations $\hat{y}(n)$?

In general, we expect the likelihood $\mathcal{L}(\theta,\psi;\hat{y}(n))$ to be largest at an unknown point (θ^*,ψ^*) where an accurate model S is located. The probability that it will be otherwise becomes infinitesimal as the number of observations $n\to\infty$. Hence, we assert that $\mathcal{L}^*(n)$ is obtained at the point (θ^*,ψ^*). Given the definition of an accurate model in the introduction, we know that $Y_t-M(\theta^*,\psi^*)=\epsilon_t\sim N(0,\sigma)$ at this point. Thus, for a given observation $\hat{y}(n)=(\hat{y}_1,...,\hat{y}_n)$, we construct a random variable

$$L^*(n) \quad := \quad \frac{\frac{1}{\sqrt{2\pi^n}\sigma^{2n}}\;\exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^n\;\epsilon_t^2\right)}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\,,$$

$$= \quad \exp\left(-\frac{1}{2}\;(\sum_{t=1}^n\;\frac{\epsilon_t^2}{\sigma^2}-\hat{s}(\theta'))\right) \quad (16)$$

$$\text{with} \quad \hat{s}(\theta'):=\sum_{t=1}^n\;\frac{(\hat{y}_t-M(\theta',\psi_0)_t)^2}{\sigma^2}\,,$$

where the denominator (respectively the second term in the exponent) includes the likelihood of the 'best' model $M(\theta', \psi_0)$ (respectively the least square sum of its residual) in our ensemble of available models (compare Equation 12). Hence, our quantity of interest, i.e., the realization $\mathcal{L}^*(n)$ of $L^*(n)$, depends on the actual realization $(\hat{\epsilon}_1, ..., \hat{\epsilon}_n)$ as well as the residual $\hat{s}(\theta')$ of the 'best' model $M(\theta', \psi_0)$. While we can calculate $\hat{s}(\theta')$ after having received the observation $\hat{y}(n)$, we cannot access the realization $\hat{\epsilon}$ of the residual between $\hat{y}(n)$ and the unknown accurate model $M(\theta^*, \psi^*)$. We only know that $\epsilon \sim N(0, \sigma)$ is an iid Gaussian process, and its variance is distributed as χ^2:

$$s(n) := \sum_{t=1}^{n} \frac{\epsilon_t^2}{\sigma^2} \sim \chi^2_{n-1} \,. \tag{17}$$

Therefore, we only can try to derive a useful estimator $E(L^*(n))$ of $\mathcal{L}^*(n)$ from the asymptotic χ^2_{n-1} distribution of $s(n)$:

$$E(L^*(n)) = \exp\left(-\frac{1}{2}\,\left(E(s(n)) - \hat{s}(\theta')\right)\right) \tag{18}$$

Such an estimator $E(L^*(n))$ will be useful for our purpose, if it discriminates between the two cases where an accurate model S is contained in the ensemble, i.e., it exists $\tilde{\theta} \in \Theta$ with $S = M(\tilde{\theta})$, and where it is not. In the former case, we can assume for large numbers of observations that S will coincide with the 'best' model $M(\theta', \psi_0)$ which exhibits the maximum likelihood on the space of available models Θ. In the latter case ($S \neq M(\theta', \psi_0)$), we assert for large numbers n of observations that the residual $\hat{s}(\theta')$ will grow faster than any estimator $E(s(n))$ constructed from a χ^2_{n-1} distribution, i.e., $E(s(n)) - \hat{s}(\theta') \to -\infty$ for $n \to \infty$, and thus $E(L^*(n)) \to \infty$ and $\varepsilon(E(L^*), p_0) \to 1$.

It remains to investigate the asymptotic behavior for the case $S = M(\theta', \psi_0)$, for which the residual between the 'best' model and the data will also be a realization of an iid Gaussian process $N(0, \sigma)$. Hence $\hat{s}(\theta')$ will constitute a draw from the same χ^2_{n-1} distribution on which $E(s(n))$ is based. Since χ^2_{n-1} becomes approx. normal for $n \to \infty$, it can be seen that $s(n) - s(\theta')$ will also be approx. normally distributed with zero mean and variance $\to \infty$. This shows that the estimator $E(s(n))$ needs to be carefully chosen in order to avoid a situation where $\varepsilon(E(L^*), p_0)$ can take any value between 0 and 1, *if* $S = M(\theta')$. Therefore, we select a q-quantile of the χ^2_{n-1}-distribution

$$\int_0^{qs} \chi^2_{n-1}(s)\; ds := q \,,$$

as estimator $E(s(n))$. The quantile qs will be larger than $\hat{s}(\theta)$ with probability q, if the accurate model S is contained in the model ensemble. We use qs to define our estimator $E(L^*(n))$ of $\mathcal{L}^*(n)$ in Equation 15), i.e.,

$$E(L^*(n)) := e^{-\frac{n-1}{2}\left(\frac{qs}{n-1} - \frac{\hat{s}(\theta')}{n-1}\right)} \,. \tag{19}$$

Equation (19) constitutes the final building block for our extension of Bayes' rule that allows us to drop the closed world assumption. This open version of Bayes' rule is summarised by Equations (14), (15) (with $\mathcal{L}^*(n)$ replaced by the estimator $E(L^*(n))$), and (19). It should be noted that the extended Bayes' rule depends on the choice of confidence level q for the upper limit of the variance of the residual $\hat{\epsilon}$. This makes it clear that in our attempt to account for the space of unspecified models, we allowed classical statistics to enter our otherwise Bayesian approach through the backdoor. For large n, the introduction of a contamination term in the posterior belief amounts to a hypothesis test on our best model $M(\theta')$. In this case, $E(L^*(n))$ will jump rapidly from zero to a very large number, when the residual of our best model $M(\theta')$ crosses the upper limit qs at the q-confidence level (see Equation 19). This will cause the contamination term $\varepsilon(E(L^*), p_0)$ to jump from 0 to 1 (see Equation 15). Therefore, our choice of $E(L^*(n))$ can lead to strong fluctuations in the contamination term if the residual of the best model $M(\theta')$ is hovering around the upper limit qs. The responsiveness of the contamination term can be reduced by replacing the linear scaling of the exponent of $E(L^*(n))$ with increasing number of observations by a sublinear function. We suggest that this is most effectively done by using the scaling of the χ^2_{n-1} distribution for increasing degrees of freedom, and offer the following heuristic expression as an alternative choice:

$$E(L^*(n)) := e^{-\frac{1}{2}(qs-n+1)\,\left(\frac{qs}{n-1} - \frac{\hat{s}(\theta')}{n-1}\right)} \,. \tag{20}$$

4 Prediction with ε-contamination: Example of AR processes continued

We now put the conceptual framework developed in the previous section into operation for our example of AR processes. The setup is identical to what was described in Section 2. For applying our open version of Bayes' rule to this updating problem, we need to calculate the development of the contamination $\varepsilon(E(L^*), p_0)$ for time series of observations $\hat{y}(n)$ with increasing length. We do this for the random sample of 200 time series from Section 2, and for both choices of $E(L^*(n))$ proposed in Equations (19) and (20). We use a prior weight $p_0 = 0.95$ on our model ensemble $M(\alpha_1)$, $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$, and choose a confidence level of $q = 0.99$ to determine the upper limit qs on the residual variance of the accurate model.

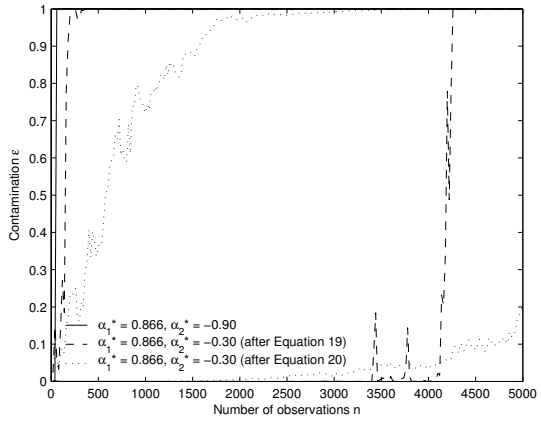


Figure 4: Behavior of lower and upper bounds of the 90% confidence interval for the ε-contamination. The ε-contamination for the models $\alpha_1^* = 0.866$ and $\alpha^2 = \{0, 0.06\}$ falls immediately to zero and stays there throughout the 5000 observations. In contrast, the ε-contamination for the model with $\alpha_2^* = -0.9$ jumps quickly to one (after 100 observations).

The lower 5% and upper 95% quantile limits (deduced from the sample of 200 time series) for the value of the contamination $\varepsilon(E(L^*), p_0)$ are shown in Figure 4. The contamination is zero for the cases in which standard Bayesian updating did well. Hence, in these cases our posterior belief about model accuracy and the associated prediction of the asymptotic distribution of system observations is identical to what we have found in Section 2. In the remaining two cases where standard Bayesian updating failed, the situation is markedly different. For $\alpha_2 = -0.9$, the contamination rapidly approaches $\varepsilon(E(L^*), p_0) = 1$, rendering our posterior belief vacuous after 100 observations at the latest. For the less extreme case $\alpha_2 = -0.3$, the increase in contamination is much slower, reflecting the results shown in Figure 1 that the posterior belief about the true propagator value remains in the vicinity of the AR(1)-equivalent propagator for several thousand observations. In that boundary case, the contamination term based on Equation (19) can fluctuate indeed strongly up to $n = 4000$ observations depending on the actual time series. The alternative contamination term based on Equation (20) offers a smoother response (see Figure 4), but on the downside responds slower to pick up the lack of accuracy in the model ensemble. We suggest that the proper choice of contamination term will depend on the application.

We now investigate the consequences of the growing contamination for the posterior belief in the case $\alpha_2^* = -0.9$. Our main question is whether the associated predictions of the asymptotic distribution of system observations can anticipate quickly the possibility of strong tails that was missed by standard Bayesian updating (see Figure 3). The analysis will also illustrate how the ε-contamination model can be used in statistical inference.

Due to the mixture with the vacuous probability model $\mathcal{V}(A_1)$, $A_1 = [-\bar{\alpha}, \bar{\alpha}]$, the posterior belief as expressed in Equation (14) is imprecise. Since it includes Dirac measures, the set of posterior probabilities can be depicted as a band of cumulative distributions (CDFs), but not as density band. The upper and lower CDFs set up by the ε-contaminated posterior belief model are given by (using the shorthand $\varepsilon^* := \varepsilon(E(L^*), p_0)$):

$$\underline{F}(\alpha_1; \hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\alpha_1} \rho(\alpha_1'|\hat{y}(n))\, d\alpha_1' + \varepsilon^*\, H(\alpha_1 - \bar{\alpha}) \quad (21)$$

$$\overline{F}(\alpha_1; \hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\alpha_1} \rho(\alpha_1'|\hat{y}(n))\, d\alpha_1' + \varepsilon^*\,, \quad (22)$$

where H denotes the Heavyside function which adds the missing probability mass at the upper bound of the support for α_1. It is important to note that the distribution band defined by $\underline{F}(\alpha_1; \hat{y}(n))$ and $\overline{F}(\alpha_1; \hat{y}(n))$ is not equivalent to the ε-contamination model, but a true superset of it. Every distribution contained in the ε-contamination model will be contained in the distribution band, but not vice versa

Figure 5 shows the change of posterior distribution band with increasing number of observations of an AR(2) process with $\alpha_2^* = -0.9$. It can be seen that the imprecision in the posterior belief increases quickly with observations. After $n = 80$ observations the posterior belief becomes vacuous, and the associated distribution band would cover the entire graph. At this point, any predictive power has been lost.

We take a closer look on the prediction of the asymptotic distribution of system observations $\rho(y|\hat{y}(n))$ in Figure 6. The prediction is again imprecise, and its lower and upper bound can be calculated on the basis of Equation (8) by recalling that these bounds are set up by the Dirac measures contained in the vacuous probability model $\mathcal{V}(A_1)$. Those Dirac measures allocate the probability mass carried by the contamination $\varepsilon^* := \varepsilon(E(L^*), p_0)$ at a value of α that minimizes respectively maximizes the contribution to $\rho(y|\hat{y}(n))$.

$$\underline{\rho}(y|\hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\, \rho(\alpha_1|\hat{y}(n))\, d\alpha_1 + \varepsilon^* \min_{\alpha_1 \in [-\bar{\alpha}, \bar{\alpha}]} \rho(y|\alpha_1)\,. \quad (23)$$

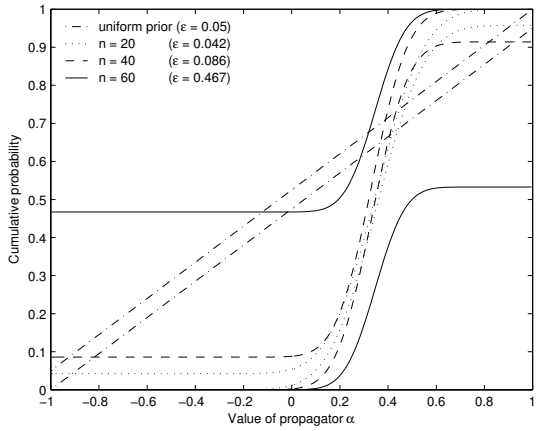


Figure 5: Cumulative posterior distribution bands for the propagator α_1 learned from a realization of the AR(2)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. The distribution band for $n = 80$ observations is vacuous and covers the entire graph.

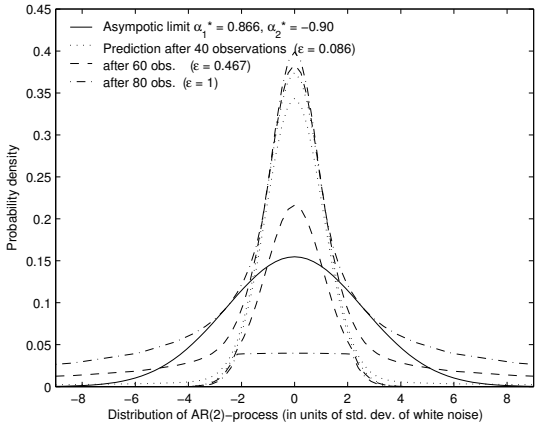


Figure 6: Predictions for the asymptotic probability distribution of system observations for the AR(2)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. The quickly growing ε-contamination destroys the predictive accuracy of the model ensemble after 80 observations.

$$\overline{\rho}(y|\hat{y}(n)) \;=\; (1-\varepsilon^*)\int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\;\rho(\alpha_1|\hat{y}(n))\;d\alpha_1 \;+\; \varepsilon^* \max_{\alpha_1\in[-\bar{\alpha},\bar{\alpha}]} \rho(y|\alpha_1)\,. \qquad (24)$$

Figure 6 shows the predicted bounds on the asymptotic distribution of system observations. It can be seen that the imprecision in the prediction grows quickly, and its range covers the tails after $n = 60$ observations. The full asymptotic distribution is contained in the predicted range after $n = 80$ observations when the posterior belief has become vacuous. At this point, the analyst employing our open version of Bayes' rule will have noticed that it is time to engage in type (I) learning as defined in the introduction, and to try to extend the set of models that she considers (e.g., to the set of all AR(2)-processes).

5 Conclusions

We have presented a framework for updating belief about prediction accuracy across an ensemble of available models using observations of the system that those models are supposed to predict. While following the Bayesian approach to learning, we have dropped the assumption that an accurate model – predicting the system observations up to an iid Gaussian process – is contained in the model ensemble with certainty as would be required by Bayes' rule in its conventional form. This is an achievement because the closed world assumption can lead to spurious beliefs about model accuracy and false predictions, as was demonstrated with an example of AR processes. By drawing on elements of imprecise probability theory and the knowledge of asymptotic distributions for large samples, we established an open version of Bayes' rule that extends its consideration to the unknown space of unspecified models, and thus includes the possibility that an accurate model might not be contained in the set of available models. Under the open Bayes' rule, the posterior belief takes on the form of an ε-contamination model, where the contamination ε is updated along with the prior belief on the set of available models. A growing contamination will indicate limited accuracy of the entire model ensemble, and will eventually lead to a vacuous posterior belief. In this way, false predictions due to limitations of the models under consideration can be avoided as was demonstrated again with an example of AR processes.

Also the method presented here has proven successful – in a stylized example – to discriminate between cases where standard Bayesian updating works well, and where it fails, this paper can offer only a proof of concept. It will require further research to investigate how the open Bayes' rule works in practice. In a next step we intend to apply it to the comparison of the 20th century temperature record with a simple climate model parameterized in terms of key quantities influencing the temperature response [6]. As a matter of concern, we will have to analyze whether the open Bayes' rule in its current form is too discriminative as it may discount every model that cannot explain the data up to an additive Gaussian process. In practice, such a strong requirement is hard to fulfill, not the least because the observations might be overlaid by a systematic non-Gaussian error due to changing

measurement practices over time. This, however, is a general problem for model validation and model-based prediction, and by no means limited to the application of the open Bayes' rule. In these cases it may be unavoidable to attempt adding and updating a positive model for the discrepancies between actual measurements and 'ideal' measurements (and, if necessary, between actual model and 'ideal' model) to the analysis as has been proposed by [4]. In any case, the open Bayes' rule can be a valuable tool to assess whether such additions are bearing fruit.

A Calculation of the likelihood for the AR(1) propagator α

Let an AR(1) process be defined by $X_1 = \xi_1$, $X_t = \alpha X_{t-1} + \xi_t$, $t \geq 2$, and $\xi_t \sim N(0, \sigma_\xi)$. Estimators $\alpha(n)$ for the propagator α and $s(n)$ for the variance of the AR(1) process are defined in terms of the observation $Y(n) = (X_1, ..., X_n)$ after n time steps

$$s(n) = \frac{1}{n-1} \sum_{t=1}^{n} X_t^2 \,, \tag{25}$$

$$\alpha(n) = \frac{1}{n-1} \frac{\sum_{t=2}^{n} X_t X_{t-1}}{s(n)} \,. \tag{26}$$

Here, we deviate from the standard choice of these estimators [7] by omitting the subtraction of the sample mean ($1/n \sum_{t=1}^{n} X_t \to 0$ for $n \to \infty$) in the estimator for the variance, and by inflating the estimator for the propagator by $n/(n-1)$. The reason for this is that the distribution of those estimators for a given choice of α, σ_ξ can be calculated easily:

$$\begin{aligned} &\rho(\alpha(n), s(n) | \alpha, \sigma_\xi) \\ &\quad \sim e^{-\frac{1}{2\sigma_\xi^2}\left(\sum_{t=2}^{n} (X_t - \alpha X_{t-1})^2 + X_1^2\right)} \\ &\quad = e^{-\frac{(n-1)\ s(n)}{2\sigma_\xi^2}\left(1 + \alpha^2 - 2\alpha\alpha(n) - \frac{\alpha^2\ X_n^2}{(n-1)\ s(n)}\right)} \,. \end{aligned}$$

Once we have observed an actual realization $\hat{y}(n) = (\hat{y}_1, ..., \hat{y}_n)$, fixing the values of the estimators at $\hat{\alpha}(n)$ and $\hat{V}(n)$, we can calculate a likelihood function $\mathcal{L}(\alpha; \hat{y}(n)) \sim \rho(\hat{\alpha}(n), \hat{V}(n) | \alpha, \sigma_\xi)$ for the propagator α of the underlying AR(1) process (assuming that σ_ξ is known). With

$$\hat{\sigma}(n) := \frac{\sigma_\xi}{\sqrt{(n-1)\ \hat{s}(n)}} \,, \tag{27}$$

$$\hat{\beta}(n) := \frac{\hat{y}_n^2}{(n-1)\ \hat{s}(n)} \,, \tag{28}$$

we find

$$\begin{aligned} \mathcal{L}(\alpha; \hat{y}(n)) &\sim e^{-\frac{1}{2\hat{\sigma}(n)^2}\left((\alpha - \alpha(n))^2 - \hat{b}(n)\alpha^2\right)} \\ &\sim N\left(\frac{\hat{\alpha}(n)}{1 - \hat{\beta}(n)}, \frac{\hat{\sigma}(n)}{\sqrt{1 - \hat{\beta}(n)}}\right) \,. \end{aligned}$$

Acknowledgment: Elmar Kriegler is supported by a Marie Curie Outgoing International Fellowship funded by the European Commission under the Sixth Framework Programme (Contract #MOIF-CT-2005-008758). Infrastructure for his research was partly provided by the Climate Decision Making Center (CDMC) located in the Department of Engineering and Public Policy. This Center has been created through a cooperative agreement between the National Science Foundation (SES-0345798) and Carnegie Mellon University.

References

[1] J. O. Berger. An overview of robust Bayesian analysis. Technical Report #93-53C, Purdue University, 1993.

[2] D. Draper. Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, 57:45–97, 1995.

[3] M. Goldstein and J. C. Rougier. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM J. Sci. Comput.*, 26:467–487, 2004.

[4] M. Goldstein and J. C. Rougier. Reified Bayesian modelling and inference for physical systems. *J. Stat. Plan. and Inf.*, forthcoming as discussion paper, 2007.

[5] T. Herron, T. Seidenfeld, and L. Wasserman. Divisive conditioning: Further results on dilation. *Philosophy of Science*, 64:411–444, 1997.

[6] E. Kriegler. *Imprecise probability analysis for integrated assessment of climate change*. PhD thesis, University of Potsdam, 256pp, 2005.

[7] H. von Storch and F. W. Zwiers. *Statistical analysis in climate research*. Cambridge University Press, Cambridge, 1999.

[8] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[9] P. Walley. Inference from multinomial data: Learning about a bag of marbles. *J. Roy. Stat. Soc. B*, 58:3–57, 1996.

Estimating Probability Distributions by Observing Betting Practices

Dr. C. Lynch
National University of Ireland, Galway
caroline.lynch@nuigalway.ie

Prof. D. Barry
University of Limerick
vpa@ul.ie

Abstract

A bookmaker takes bets on a two-horse race, attempting to minimise expected loss over all possible outcomes of the race. Profits are controlled by manipulation of customers' betting behaviour; in order to do this, we need some information about the probability distribution which describes how the customers will bet. We examine what information initial customers' betting behaviour provides about this probability distribution, and consider how to use this to estimate the probability distribution for remaining customers.

Keywords. EM Algorithm, bookmaker, horse race, Markov decision process.

1 Introduction

A bookie takes bets on a contest for which there are only two possible outcomes, which we will label as A and B. The bookie wishes to maximise his minimum expected profit purely by manipulation of customers' betting practice. A gambler enters the bookie's shop seeking to place a wager on this contest. Let p denote the gambler's probability that outcome A will occur. The bookie quotes odds of O_1 against outcome A and of O_2 against outcome B. This means that a winning wager of one unit on outcome A produces a return of $O_1 + 1$ while a winning wager of one unit on outcome B produces a return of $O_2 + 1$. Hence a wager on outcome A will be attractive to the gambler if

$$p(O_1 + 1) \geq 1$$

or equivalently

$$p \geq \frac{1}{O_1 + 1} = \theta_1.$$

Similarly a wager on outcome B will be attractive to the gambler if

$$1 - p \geq \frac{1}{O_2 + 1} = \theta_2$$

or equivalently

$$p \leq 1 - \theta_2.$$

The quantities θ_1 and θ_2 are called the bookie's quoted probabilities for outcome A and B respectively. Hence the strategy for an individual gambler is simple - he places a wager on any outcome for which his probability exceeds that quoted by the bookie.

It should be noted that the "quoted probabilities", θ_1 and θ_2, described above, are not probabilities, in the sense that their sum will generally be greater than 1. In fact, they may more properly be described as upper probabilities, as defined in [13].It can be shown that it is never to the advantage of the bookie to have these upper probabilities sum to less than 1, as in Lemma 2 of [14]. Thus, for the remainder of this paper, we shall assume that $\theta_1 + \theta_2 \geq 1$.

We also assume that the quoted odds, O_1 and O_2, are positive. This follows naturally from the requirement that a wager of 1 unit leads to a return of $O_1 + 1$ on outcome A or $O_2 + 1$ on outcome B, and the customer is unlikely to wager more than the expected return. By the definitions of θ_1 and θ_2, this means that θ_1 and θ_2, in turn, are positive. These conditions on θ_1 and θ_2 ensure the coherence of the upper probabilities in this case.

We idealise the bookie's shop by assuming that the bookie sells two types of tickets - one which guarantees a return of one unit should outcome A occur and costs θ_1, and one which guarantees a return of one unit should outcome B occur and costs θ_2. This avoids sure loss, as the customer's only options are to bet on A or B, individually; to bet on both would ensure a loss for the customer, as he would be required to bet an amount $\theta_1 + \theta_2$, greater than his (guaranteed) return of 1. We also assume that the bookie knows, before opening the book, that N customers will consider a wager on the contest and that their probabilities $p_1, p_2, ..., p_N$ of outcome A occurring behave like a random sample from a probability

distribution. Finally, we assume that customers can buy at most one of each type of ticket and that the bookie is free to alter the quoted probabilities after each customer leaves.

The bookie seeks to manipulate customers' betting behaviour as best suits himself; this depends, however, on knowing the probability distribution from which the customers' probabilities $p_1, p_2, ..., p_N$ derive. After first considering the optimal procedure when the distribution is known, we will consider how the bookie may estimate this probability distribution using information derived from customers' betting practices.

2 Distribution Known

Assuming the distribution of customers' probabilities to be known, the optimal algorithm for the bookie to follow is the "Dynamic Programming" Algorithm, as described in Barry & Hartigan[2]. This iterative algorithm depends on knowledge of the customers' probability distribution, F, the number of customers left to come, n, and the current "*state of the book*", i.e. the amount of the bookie's profit on outcome A, denoted a, and on outcome B, denoted b, if the book was closed at that instant, i.e. no more bets were taken.

Assuming knowledge of these quantities, the algorithm is then given as follows;

$$R_n(a,b) = \frac{a+b}{2} + P_n(d)$$

where

$$P_n(d) = P_{n-1}(d)$$

$$+\max_{\theta_1,\theta_2} \left\{ \begin{array}{l} [1\text{-}F(\theta_1)][\theta_1 - \frac{1}{2} + P_{n-1}(d-1) - P_{n-1}(d)] \\ +F(1\text{-}\theta_2)[\theta_2 - \frac{1}{2} + P_{n-1}(d+1) - P_{n-1}(d)] \end{array} \right\}$$

with $P_0(d) = -\frac{|d|}{2}$ and $d = a - b$. Here, $R_n(a,b)$ denotes the expected value of the bookie's final minimum profit between both outcomes. The algorithm gives the bookie a method for deriving optimal quoted probabilities for the next customer, given the current state of the book, n customers left to go and F known.

The above equation for $P_n(d)$ describes how it depends on the previous value, $P_{n-1}(d)$, then adds a term, maximised over θ_1 and θ_2, which describes the profit accruing if the customer bets on A, with probability $1 - F(\theta_1)$, and if the bet is on B - with probability $F(1-\theta_2)$; the only two possible bets.

3 Strategy for Distribution Unknown

Having determined a strategy for F known, we must consider how to estimate F when it is unknown. We subdivide the interval [0,1] into r subintervals of equal width - the choice of the value of r will be discussed in Section 4. We then estimate F by means of a histogram, with r intervals. For each of these r intervals, the height of the histogram will be determined by the probability assigned to that interval, π_j. This probability will be determined by the betting behaviour of the customers, as described hereafter. $F(\theta)$ may then be determined by the formula

$$F(\theta) = \left\{ \begin{array}{cc} r\theta\pi_1 & 0 \le \theta \le \frac{1}{r} \\ \pi_1 + (r\theta - 1)\pi_2 & \frac{1}{r} \le \theta \le \frac{2}{r} \\ \pi_1 + \pi_2 + (r\theta - 2)\pi_3 & \frac{2}{r} \le \theta \le \frac{3}{r} \\ \vdots & \vdots \\ \sum_{i=1}^{j} \pi_i + (r\theta - j)\pi_{j+1} & \frac{j}{r} \le \theta \le \frac{j+1}{r} \\ \vdots & \vdots \\ \sum_{i=1}^{r-1} \pi_i + (r\theta - r + 1)\pi_r & \frac{r-1}{r} \le \theta \le 1 \end{array} \right.$$

3.1 Estimation of F

This involves the EM Algorithm; we have an estimation, and a maximisation, step.

3.1.1 Estimation

Each customer's betting pattern gives us information about their value of p, as follows;

Bet on Horse A	$\theta_1 \le p \le 1$
Bet on Horse B	$0 \le p \le 1 - \theta_2$
No Bet	$1 - \theta_2 < p < \theta_1$

We denote the lower limit of the range in which p falls by a_1^k for customer k and the upper limit by a_2^k such that $a_1^k \le a_2^k$. We also denote the lower and upper limit of each of the subintervals of [0,1], I_j, by $[L_j, R_j]$, with $R_j = L_{j+1}$. We have an indicator function, X_{jk}, defined as follows;

$$X_{jk} = \left\{ \begin{array}{cc} 1 & p \in [L_j, R_j] \\ 0 & \text{otherwise} \end{array} \right.$$

In this case, the log likelihood function is given by

$$\ell = \sum_{k=1}^{N} \sum_{j=1}^{r} X_{jk} \log \pi_j.$$

Given the customer's behaviour, we have a range for the customer's probability - i.e. $a_1^k \le p \le a_2^k$. Let us call this information Y_k.

We will seek to maximize the Expected value of the log likelihood, given this information, i.e.

$$E(\ell|Y) = \sum_{k=1}^{N}\sum_{j=1}^{r} E(X_{jk}|Y_k)\log\pi_j.$$

$$\begin{aligned} E(X_{jk}|Y_k) &= P(X_{jk}=1|Y_k) \\ &= P(p\in[L_j,R_j]|p\in[a_1^k,a_2^k]) \\ &= \frac{P(p\in[L_j,R_j]\cap[a_1^k,a_2^k])}{P(p\in[a_1^k,a_2^k])} \\ &= \frac{P(p\in[L_j,R_j]\cap[a_1^k,a_2^k])}{\sum_{i=1}^{r}P(p\in[L_i,R_i]\cap[a_1^k,a_2^k])} \end{aligned}$$

We have

$$P(p\in[L_j,R_j]\cap[a_1^k,a_2^k]) = \pi_j\times\frac{l_{jk}}{R_j-L_j}$$

where l_{jk} is the length of $[L_j,R_j]\cap[a_1^k,a_2^k]$ and is given by

$$l_{jk} = \begin{cases} 0 & R_j\le a_1^k \\ R_j-a_1^k & L_j\le a_1^k\le R_j\le a_2^k \\ a_2^k-a_1^k & L_j\le a_1^k\le a_2^k\le R_j \\ R_j-L_j & a_1^k\le L_j\le R_j\le a_2^k \\ a_2^k-L_j & a_1^k\le L_j\le a_2^k\le R_j \\ 0 & L_j\ge a_2^k \end{cases}$$

3.1.2 Maximisation

Next, we seek to maximise the expected value of the log likelihood function. The Maximum Likelihood Estimate for π_j is given by

$$\hat{\pi}_j = \frac{\sum_{k=1}^{N}E(X_{jk}|Y_k)}{N}.$$

Each of the subintervals of [0,1] was assigned an initial probability, π_j^1. For simplicity, this initial probability was the same for each subinterval, assuming the Uniform distribution, so that, with r subintervals, the initial values of π_j^1 are given by

$$\sum_{j=1}^{r}\pi_j^1 = 1 \Rightarrow \pi_j^1 = \frac{1}{r}, \forall j.$$

This initial probability was then updated by observing each customer's behaviour.

As will be seen in this, and the next, subsection, we will now divide our customers into three groups; the very first will be used to initialise the information matrix, the second group will be used for the purpose of maximising the information we may obtain, leaving us with the third and final group for maximising profit, once F has been satisfactorily estimated. One of the questions with which we will be concerned is how many customers should be allocated to each group.

3.2 Early Customers

For the first few customers, the odds are chosen so as to maximise the information obtained.

We derive the information matrix, I, using the formula

$$I_{ij} = E\Big[-\frac{\partial^2\ell}{\partial\hat{\pi}_i\partial\hat{\pi}_j}\Big]$$

where ℓ is the log likelihood, defined as before.

As described in the previous section, we decided to divide the interval [0,1] into a number of subintervals, each of which was assigned a probability, π_j, which was updated by observation of customers' behaviour.

As before, we may express the log likelihood function as

$$\begin{aligned} \ell &= \sum_{k=1}^{N}\sum_{j=1}^{r} X_{jk}\log\pi_j \\ &= \sum_{k=1}^{N}\Big[\sum_{j=1}^{r-1}X_{jk}\log\pi_j + X_{rk}\log\Big(1-\sum_{j=1}^{r-1}\pi_j\Big)\Big] \\ &= \sum_{j=1}^{r-1}\Big(\sum_{k=1}^{N}X_{jk}\Big)\log\pi_j \\ &\quad +\Big(\sum_{k=1}^{N}X_{rk}\Big)\log\Big(1-\sum_{j=1}^{r-1}\pi_j\Big) \end{aligned}$$

as $\pi_r = 1-\sum_{j=1}^{r-1}\pi_j$.

Hence, we find that

$$\frac{\partial\ell}{\partial\pi_j} = \sum_{k=1}^{N}\Big[\frac{X_{jk}}{\pi_j} - \frac{X_{rk}}{1-\sum_{j=1}^{r-1}\pi_j}\Big]$$

and

$$\frac{\partial^2\ell}{\partial\pi_i\partial\pi_j} = -\frac{\sum_{k=1}^{N}X_{jk}}{\pi_j^2}\delta_{ij} - \frac{\sum_{k=1}^{N}X_{rk}}{(1-\sum_{j=1}^{r-1}\pi_j)^2},$$

where $\delta_{ij} = \{1 \text{ if } i=j, 0 \text{ otherwise}\}$. Thus, we have

$$E\Big[-\frac{\partial^2\ell}{\partial\pi_i\partial\pi_j}\Big] = \frac{N\delta_{ij}}{\pi_j} + \frac{N}{1-\sum_{j=1}^{r-1}\pi_j}.$$

The entries are added for each successive customer.

Having calculated the information matrix, we use it to choose the odds for each of the customers before F

is determined. Firstly, both θ_1 and θ_2 are set at $\frac{1}{r}$, for convenience of programming. The information matrix is recalculated for each combination of θ_1 and θ_2, each being incremented in steps of $\frac{1}{r}$. Finally, that combination of odds which maximises the determinant of the information matrix is used for the next customer, so long as it satisfies the condition $\theta_1 + \theta_2 \geq 1$. In practice, this condition was satisfied by every optimal combination of odds. This procedure is repeated for each of the customers in turn.

The optimal number of customers used to estimate F is found by inspection. This procedure is described later.

After each of these customers bets, our estimate of F is updated using the EM Algorithm, as described previously. Finally, we must initialise the information matrix.

3.3 Initialisation

3.3.1 Odds for the Initial Customers

In order to initialise the information matrix, the theta-values for the first few customers are chosen according to the following plan;

If we divide the interval [0,1] into r equally-spaced subintervals, placing the theta-values on the divisions of these subintervals will give us precise information about the distribution of probability within these subintervals. The optimal value of r is found by inspection, and is described subsequently.

We do not need to set either a theta-value equal to 1, which guarantees no bets, or equal to 0, which guarantees a bet from any customer.

Bearing these points in mind, we set the theta-values for the first customer as

$$\theta_1 = \theta_2 = \frac{r-1}{r}$$

We then take each theta-value down by a value $\frac{1}{r}$ in turn for each of the next few customers.

3.3.2 No. of Customers in this Group

As customers bet on Horse A with probability $1 - F(\theta_1)$, the value of θ_1 will provide us with information about the probabilities of the subintervals above θ_1. Thus, this value provides us with information about the subintervals at the upper end of the interval [0,1].

Similarly, customers bet on Horse B with probability $F(1-\theta_2)$. Thus, the value of θ_2 provides us with information about the probabilities of the subintervals below $1 - \theta_2$, and thus provides us with information about the subintervals at the lower end of the interval [0,1].

So the theta-values for the very first customer tell us something about the probability in the first, and last, subintervals. Each successive customer's set of theta-values tells us about an additional subinterval. Finally, we only need information about $(r-1)$ subintervals, as we know that the probabilities sum to 1 in total. Altogether, this tells us that we need $r-2$ customers in the first group, to initialise the information matrix.

4

Choice of No. of Subintervals and No. of Customers to Use in Estimation

Firstly, as discussed previously, we divide the interval [0,1] into r subintervals, to each of which is assigned a probability, so as to estimate F.

We now need to determine

1. the optimal number of subintervals, and also
2. the optimal number of customers, as a percentage of the total (assumed known), whose odds we should use in order to maximise the information matrix, as described in the previous section. This number is in addition to the $r-2$ customers used in the beginning to initialise the information matrix.

These were estimated simultaneously, by calculating profits for the same Dynamic Programming profit function for a variety of combinations of **(1)** numbers of subintervals and **(2)** numbers of customers used in estimation of F, and choosing the combination which proved best overall. The *state of the book* for a particular outcome denotes the bookie's profit if that outcome occurs. Let A_n denote the state of the book for outcome A, and B_n the state of the book for outcome B, when n of the N customers remain. In the strategy which our bookie uses, we have

$$\tilde{\theta}_1\Big(\frac{A_n - B_n}{n}\Big) \text{ and } \tilde{\theta}_2\Big(\frac{A_n - B_n}{n}\Big),$$

which are chosen to maximise

$$\min\{E[A_0|A_n = a, B_n = b], E[B_0|A_n = a, B_n = b]\}.$$

This gives the function to be maximised as

$$\min\Big\{\frac{d}{2} - n(1-\theta_1)[1-F(\theta_1)] + n\theta_2 F(1-\theta_2),$$

$$-\frac{d}{2} + n\theta_1[1-F(\theta_1)] - n(1-\theta_2)F(1-\theta_2)\Big\},$$

where $d = a - b$. This is the objective function which was used in the simulation study, a summary of whose results follows. It assumes that the quoted probabilities remain constant for all n remaining customers, and calculates the final expected profit if A occurs as the income from those customers who bet an amount θ_2 on B, with probability $F(1-\theta_2)$, less the outgoing return of 1 unit to those who bet an amount θ_1 on A, with probability $1 - F(\theta_1)$. A similar calculation determines the final expected profit if B occurs. The algorithm involves the calculation of the minimum of these two expected final profits, given the current state of the book.

It will be noted that this is a different algorithm to the optimal one discussed in Section 2; as discussed in Barry & Hartigan[2], this algorithm provides an easier method for calculation of the quoted probabilities, without excessive penalty in terms of the bookie's profit.

The measure of which combination of **(1)** and **(2)** proved best was provided by obtaining the mean profit, over fifty replications in each case, for each individual combination. The difference between each of these values and the maximum value over all combinations was then obtained for each distribution. This was repeated for each of $N = 100, 500$ and 1500 customers, and for each of five distributions; namely,

1. Uniform i.e. $F(\theta) = \theta$
2. $$F2(\theta) = \begin{cases} 0 & 0 \le \theta \le \frac{1}{8} \\ 2\left(\theta - \frac{1}{8}\right) & \frac{1}{8} \le \theta \le \frac{5}{8} \\ 1 & \frac{5}{8} \le \theta \le 1 \end{cases}$$
3. $$F3(\theta) = \begin{cases} 0 & 0 \le \theta \le \frac{1}{4} \\ 2\theta - \frac{1}{2} & \frac{1}{4} \le \theta \le \frac{3}{4} \\ 1 & \frac{3}{4} \le \theta \le 1 \end{cases}$$
4. $$F4(\theta) = \begin{cases} 0 & 0 \le \theta \le \frac{3}{4} \\ 4\theta - 3 & \frac{3}{4} \le \theta \le 1 \end{cases}$$
5. $$F5(\theta) = \begin{cases} 2\theta & 0 \le \theta \le \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} \le \theta \le \frac{3}{4} \\ \frac{1}{2} + 2\left(\theta - \frac{3}{4}\right) & \frac{3}{4} \le \theta \le 1 \end{cases}$$

We found the maximum difference, for each combination of number of subintervals and percentage of customers, over the five distributions. This represents the maximum loss per customer. Thus, we use the combination which provides the smallest value of maximum loss. The maximum loss per customer over all distributions is shown in the following tables.

Maximum Difference over Five Distributions

	% of Customers to Estimate F		
Intervals	0	1	2
1	0.0711	0.0715	0.0714
2	0.0358	0.0357	0.0356
3	0.0229	0.0216	0.0259
4	0.0259	**0.0181**	0.0195
5	0.03296	0.03624	0.03621
6	0.03305	0.03005	0.0295
7	0.03139	0.03453	0.03369
8	0.03531	0.03939	0.03728
10	0.03357	0.03355	0.0361

N=100

	% of Customers to Estimate F		
Intervals	0	1	2
1	0.015	0.015	0.015
2	0.009	0.009	0.009
3	0.0049	0.0054	0.0054
4	0.0035	**0.0034**	0.0036
5	0.0044	0.0057	0.0066
6	0.0049	0.0047	0.0051
7	0.005	0.005	0.0048
8	0.004	0.004	0.004
10	0.004	0.004	0.004

N=500

	% of Customers to Estimate F		
Intervals	0	1	2
1	0.005	0.0049	0.0049
2	0.0031	0.0031	0.0031
3	0.0016	0.0018	0.0018
4	0.0015	0.0015	0.0015
5	0.0015	0.0019	0.0022
6	0.0019	0.0019	0.0019
7	0.0019	**0.0012**	0.0017
8	0.0015	0.0015	0.0015
10	0.0015	0.0014	0.0014

N=1500

From these tables, we may see that the optimal % of customers for maximisation of the information matrix in all cases is 1%; higher percentages are not shown here, as they led to greater loss. We also see that the optimal number of subintervals into which to divide the interval [0,1] is 4 for $N = 100$ and 500, and 7 intervals for $N = 1500$. The optimal combination is that which minimises the difference shown in the above tables.

We may further see from these tables, however, that the maximum difference over all distributions decreases, for each combination, as the total number of customers increases- demonstrating that, for larger numbers of customers, there is reduced loss in using a non-optimal combination.

5 Summary and Conclusions

In summary, the method described in this paper provides us with a means of estimating the overall distribution of customers' probabilities, based solely on the betting practices of relatively few initial customers, which provide us with interval estimates of these probabilities. This proves a highly useful tool when distributions are unknown.

Further work on this topic might include the examination of whether it is possible to incorporate an element of profit maximisation into the stage where F is being determined. Another obvious extension of the work is to the case where there are more than two possible outcomes; however, each extra outcome leads to multiple extra possibilities for the customer, who may bet on any individual outcome, or possibly on a combination of them. As well as leading to a much more complicated model, this gives rise to the possibility of incoherence, and to the incurrence of sure loss; care needs to be taken in this scenario.

Acknowledgements

This paper is based on part of the first author's PhD thesis, completed at the University of Limerick under the supervision of Prof. D. Barry, whose help and support have proven invaluable.

References

[1] Aoki, M. (1967). *Optimization of Stochastic Systems.* Academic Press, New York

[2] Barry, D. & Hartigan, J. A. (1996). "The Minimax Bookie." *J. Appl. Prob.* **33** 1093-1107

[3] Bellman, R. (1957) a. *Dynamic Programming.* Princeton University Press, Princeton, NJ

[4] Bellman, R. (1967) b. *Introduction to the Mathematical Theory of Control Processes.* Academic Press, New York

[5] Blackwell, D. (1976). "The Stochastic Processes of Borel Gambling and Dynamic Programming." *Annals of Statistics* **4** 370-374

[6] Dubins, L. E. & Savage, L. J. (1965). *How to Gamble if you Must.* McGraw-Hill, New York

[7] Henery, R. J. (1984). "An Extreme-value Model for Predicting the Results of Horse Races." *Appl. Statist.* **33** 125-133

[8] Henery, R. J. (1985). "On the Average Probability of losing Bets on Horses with given Starting Price Odds." *J. R. Statist. Soc.* A **148** 342-349

[9] Hoerl, A. E. & Fallin, H. K. ((1974). "Reliability of subjective evaluations in a high incentive situation." *J. R. Statist. Soc.* A **137** 227-230

[10] Plackett, R. L. (1975). "The Analysis of Permutations." *Appl. Statist.* **24** 193-202

[11] Rieder, U. (1976). "On optimal policies and martingales in dynamic programming." *J. Appl. Prob.* **13** 507-518

[12] Whittle, P. (1982). *Optimization Over Time: Dynamic Programming and Stochastic Control.* Academic Press, New York

[13] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall

[14] Barry, D. & Lynch, C. (2006). "The Minimax Bookie: The Two-Horse Case." *Adv. in Appl. Prob.* **Vol. 38** No. 4

An independence concept under plausibility function

Marcello Mastroleo
Dip. Matematica e Informatica
Università di Perugia
mastroleo@dipmat.unipg.it

Barbara Vantaggi
Dip. Metodi e Modelli Matematici
Università "La Sapienza" Roma
vantaggi@dmmm.uniroma1.it

Abstract

Starting from considering different definitions of conditioning for decomposable measures, in particular for totaly monotone measures (belief functions) and totaly alternating measures (plausibility functions), we provide a concept of independence which covers some natural properties. In particular, we characterize the proposed independence for plausibility functions and we check some relevant properties. Relationships with other notions studied in literature are shown.

Keywords. Totaly monotone measures, Plausibility, Conditioning, Independence.

1 Introduction

The subtle notion of conditioning is controversial in several contexts, for example for non-additive measures and more specifically for plausibility and belief functions (which are totally alternating and monotone, respectively). We consider a general axiomatic definition of conditional measures proposed in [7]: the conditional measure is *directly* defined as a function on a set of conditional events which satisfies a suitable set of axioms. In this framework conditional measures are seen as a primitive notion, analogously to conditional probability according to de Finetti approach [16]. Among the conditional measures we deal with conditional plausibility and belief functions.

The theory of belief functions, also known as Dempster-Shafer theory [18] and theory of evidence, aims to model degree of belief. It can be regarded as a generalization of the probability approach and many interpretation have been proposed: a belief function can be seen as a particular lower probability or it can be derived from probability where a probability space is mapped by a one-to-many mapping on another space.

In [10] it is shown a sort of converse property of the fact that a belief function is a specific lower probability: a lower probability obtained as extension of a suitable coherent conditional probability is a belief function.

Starting from this general framework some other well-known definitions arise naturally (see also [6]). A comparison of these different conditioning operators has been carried out in [12] from another point of view by looking to a comparative setting and, more precisely, by studying local representability of ordinal relations defined on a finite algebra.

In particular, we refer to a well-known definition (see [19, 26]) of conditional belief, which can be obtained from the above one as particular case and, for any given conditioning event H, it can be seen as the dual function of a conditional plausibility.

Actually, we refer to a generalization of the definition provided in [26] that allows to deal also with events of zero plausibility. Then, the problem of dealing with "partial assessments" on (not necessarily structured) domains, containing only elements of interest, is faced. In any real situation the events of interest, and those in which the field expert or the decision maker has information, give rise usually to an arbitrary set. For this reason we need a notion of consistency, which allows to check whether a partial assessment is the restriction of a conditional belief function (or a conditional plausibility) [6]. A characterization of both consistent conditional plausibility and conditional belief, in terms of a suitable class of plausibility functions, is carried out: a conditional belief/plausibility is not always singled out by a unique unconditional measure.

In such framework, we study an important concept for uncertainty reasoning, which is independence. In probabilistic theory this condition has been deeply studied (see e.g. [15, 9, 32]); moreover such notion has been studied also in other non-probabilistic frameworks [1, 4, 11, 30, 31, 35] and in particular in upper and lower probabilities theory (see, for example,

[5, 8, 13, 34]).

However, the concept of independence has not been widely treated in belief theory (see [2, 3, 27, 29]). In addition to the theoretical reasons for the study of independence, there are also practical interest: many computational tasks can be simplified by using independence notion.

In this paper we propose a definition of independence for conditional plausibility, which can be reformulated by means of duality property also for belief functions. This notion covers some natural properties also in the case of events with degree equal to 0 or 1. In particular, we show that such independence notion implies logical independence. This is an intuitive implication: in fact *if an event is "logically" related to another one, the two events must be dependent* under *any* uncertainty measure. Handling logical constraints is interesting also from a practical point of view, since in many real applications (e.g. in finance, economics, medicine) variables are suitably linked (see e.g. [20, 21]). Then, we get this natural implication, which does not need to be required explicitly as in [2]. Actually, our definition of independence is inspirited to that one given for coherent conditional probability in [9, 32] and that for conditional possibilities in [11, 24].

We give a numerical characterization of the proposed definition of independence that helps to compare our definition with other ones given in literature [2, 3, 27, 29].

In Section 2, we introduce conditional plausibility and in Section 2.1 (through duality) conditional belief; we briefly deal with a consistency notion for partial assessments.

In Section 3 we provide an independence notion firstly for plausibility and then for belief function. We study its main properties by comparing it also with other notions introduced in literature.

2 Conditional plausibility and belief functions

Usually in literature conditional measures are presented as a derived notion of unconditional ones, but this is a restrictive view of conditioning. It is instead essential to adopt a general definition of generalized $(\oplus, \odot)$-decomposable conditional (uncertainty) measure (introduced in [10]). The peculiarity of this approach consists in the fact that conditional measures are directly defined on a suitable set of conditional events.

Moreover, by specifying the two operations $\oplus, \odot$ we obtain some particular conditional uncertainty measures. In particular, by taking the usual sum and product, respectively, we get a conditional probability (in the sense of de Finetti [16]), while for $\oplus = \max$ and $\odot = \min$ we obtain conditional possibility [4, 23]. In [10] it is shown that for specific operations conditional belief functions can be seen as particular generalized decomposable measures.

In order to revisit the belief functions and their connections with the so-called "imprecise" probabilities and with extensions of coherent conditional probabilities, we recall firstly some basic notions and then some results given in [10]. An assessment p on a set of conditional events $\mathcal{C}$ is a coherent conditional probability iff there exists a conditional probability P on the product of an algebra $\mathcal{E}$ and an additive set (closed under finite unions) $\mathcal{H} \subseteq \mathcal{E} \setminus \{\emptyset\}$ such that $\mathcal{C} \subseteq \mathcal{E} \times \mathcal{H}$ and the restriction of P on $\mathcal{C}$ coincides with p (i.e. $P(E|H) = p(E|H)$ for any $E|H \in \mathcal{C}$).

Given an *arbitrary* set $\mathcal{C}$ of conditional events, a *coherent lower conditional probability* on $\mathcal{C}$ is a nonnegative function $\underline{P}$ such that there exists a non-empty *dominating family* $\mathcal{P} = \{P(\cdot|\cdot)\}$ of *coherent* conditional probabilities on $\mathcal{C}$ whose lower envelope is $\underline{P}$, that is, for every $E|H \in \mathcal{C}$,

$$\underline{P}(E|H) = \inf_{\mathcal{P}} P(E|H)\,.$$

In particular, by taking $\mathcal{C}$ as a set of unconditional events, we get a coherent lower probability.

It is well known that a belief function (totaly monotone measure) is a lower probability; the following result proved in [10] shows the converse property: a lower probability obtained as extension of a suitable coherent conditional probability is a belief function.

Theorem 1 *Let $\mathcal{D} = \{H_1, \ldots, H_n\}$ be a finite set of pairwise incompatible events. Denoting by $\mathcal{K}$ the additive set spanned by them, and given an algebra $\mathcal{A} \supset \mathcal{K}$, put $\mathcal{C} = \mathcal{A} \times \mathcal{K}$. If $P(\cdot)$ is a coherent probability on $\mathcal{D}$, let $\mathcal{P}$ be the class of coherent conditional probabilities $P(\cdot|\cdot)$ extending $P(\cdot)$ on $\mathcal{C}$. Consider, for $E|K \in \mathcal{C}$, the lower probability*

$$\underline{P}(E|K) = \inf_{\mathcal{P}} P(E|K)\,; \qquad (1)$$

then for any $K \in \mathcal{K}$ the function $\underline{P}(\cdot\,|K)$ is a belief function on $\mathcal{A}$.

The involved set $\mathcal{D}$ is not a consequence of some particular circumstances, but it is always possible to find it, as shown by the following theorem [10] (Section 3.1.):

Theorem 2 *Let $\mathcal{A}$ be a finite algebra and Bel be a belief function on $\mathcal{A}$. Then, there exists a partition*

$\mathcal{D} = \{H_1, \ldots, H_n\}$ *of* Ω *and a (coherent) probability on* $\mathcal{D}$ *such that the lower envelope of the class of coherent conditional probabilities* $P(\cdot|\cdot)$ *extending* $P(\cdot)$ *on* $\mathcal{C} = \mathcal{A} \times \mathcal{K}$ *(*$\mathcal{K}$ *is the additive set generated by* $\mathcal{D}$*) coincides with Bel on* $\mathcal{A}$*.*

The following axioms are naturally derived (see [6]):

Definition 1 *Let* $\mathcal{E}$ *be an algebra and* $\mathcal{H} \subseteq \mathcal{E} \setminus \{\emptyset\}$ *an additive set. A function Pl defined on* $\mathcal{C} = \mathcal{E} \times \mathcal{H}$ *is a* conditional plausibility *if it satisfies the following conditions*

i) $Pl(E|H) = Pl(E \wedge H|H)$*;*

ii) $Pl(\cdot|H)$ *is a plausibility function* $\forall H \in \mathcal{H}$*;*

iii) For every $E \in \mathcal{E}$ *and* $H, K \in \mathcal{H}$

$$Pl(E \wedge H|K) = Pl(E|H \wedge K) \cdot Pl(H|K).$$

Moreover, given a conditional plausibility, a conditional belief *function* $Bel(\cdot|\cdot)$ *is defined by duality as follows: for every event* $E|H \in \mathcal{C}$

$$Bel(E|H) = 1 - Pl(E^c|H).$$

It is possible to see that the above axiomatization extends the Dempster's rule, i.e.

$$Bel(F|H) = 1 - \frac{Pl(F^c \wedge H)}{Pl(H)},$$

for all H such that $Pl(H) > 0$ (from condition *iii*)). When all the conditioning events have positive plausibility, i.e. $\Omega \in \mathcal{H}$ and $Pl(H|\Omega) > 0$ for any $H \in \mathcal{H}$, the above notions of conditional plausibility and conditional belief coincide with that given in [19, 26]. In fact, if $Pl(H) > 0$ it follows $Bel(F|H) = \frac{Pl(H) - Pl(F^c \wedge H)}{Pl(H)} = \frac{Bel(F \vee H^c) - Bel(H^c)}{Pl(H)}$.

2.1 Coherent conditional belief

By regarding a conditional plausibility function as a $(\oplus, \odot)$-decomposable measure, it is possible to study the structure underlying the conditional measure and to build an algorithm to check the consistency (with the model of reference) of a partial assessment.

In the following we denote by $\mathcal{F} = \{E_1|F_1, E_2|F_2, \ldots, E_m|F_m\}$ an arbitrary finite set of conditional events, by $\mathcal{E}$ the algebra generated by $\{E_1, F_1, \ldots, E_m, F_m\}$ and by $\mathcal{K}$ the additive set generated by the set of the conditioning events $\{F_1, \ldots, F_m\}$.

Definition 2 *A function* $f(\cdot|\cdot)$ *on an arbitrary finite set* $\mathcal{F}$ *is a* coherent *conditional belief (plausibility) if there exists* $\mathcal{C} \supset \mathcal{F}$*, with* $\mathcal{C} = \mathcal{E} \times \mathcal{K}$ *such that* $f(\cdot|\cdot)$ *can be extended from* $\mathcal{F}$ *to* $\mathcal{C}$ *as a* conditional belief (conditional plausibility)*.*

The following theorem [6] characterizes (coherent) conditional belief functions in terms of a class of plausibilities $\{Pl_1, ..., Pl_m\}$.

Theorem 3 *Let* $\mathcal{F} = \{E_1|F_1, E_2|F_2, \ldots, E_m|F_m\}$ *be an arbitrary finite set of conditional events and denote by* $\mathcal{E} = \{H_1, H_2, ..., H_n\}$ *the algebra generated by* $\{E_1, \ldots, E_m, F_1, \ldots, F_m\}$ *and* $H_0^0 = \vee_{j=1}^m F_j$*. For a real function Bel on* $\mathcal{F}$ *the following statements are equivalent:*

(a) $Bel : \mathcal{F} \to [0, 1]$ *is a coherent conditional belief assessment;*

(b) there exists (at least) a class $\mathcal{L} = \{Pl_\alpha\}$ *of plausibility functions such that* $Pl_\alpha(H_0^\alpha) = 1$ *and* $H_0^\alpha \subset H_0^\beta$ *for all* $\beta < \alpha$*, where* H_0^α *is the greatest element of* $\mathcal{K}$ *for which* $Pl_{(\alpha-1)}(H_0^\alpha) = 0$*.*

Moreover, for every $E_i|F_i$*, there exists an index* α *such that* $Pl_\beta(F_i) = 0$ *for all* $\alpha > \beta$*,* $Pl_\alpha(F_i) > 0$ *and*

$$Bel(E_i|F_i) = 1 - \frac{Pl_\alpha(E_i^c F_i)}{Pl_\alpha(F_i)}, \qquad (2)$$

(c) all the following systems (S^α)*, with* $\alpha = 0, 1, 2, ..., k \leq n$*, admit a solution* $\mathbf{X}^\alpha = x_k^\alpha = m_\alpha(H_k)$*:*

$$(S^\alpha) = \begin{cases} \sum\limits_{H_k F_i \neq \emptyset} x_k^\alpha \cdot [1 - Bel(E_i|F_i)] = \sum\limits_{H_k E_i^c F_i \neq \emptyset} x_k^\alpha, & \forall F_i \subseteq H_0^\alpha \\ \sum\limits_{H_k \in H_0^\alpha} x_k^\alpha = 1 \\ x_k^\alpha \geq 0, & \forall H_k \subseteq H_0^\alpha \end{cases}$$

where H_0^α *is the greatest element of* $\mathcal{K}$ *such that* $\sum\limits_{H_i H_0^\alpha \neq \emptyset} m_{(\alpha-1)}(H_i) = 0$*.*

The above characterization result holds for coherent conditional belief functions as well as for coherent conditional plausibility. In particular condition *(c)* stresses that this measure can be written in terms of a suitable class of basic assignments, instead of just one as in the classical case where all the conditioning events have positive plausibility.

Note that every class $\mathcal{L}$ (condition (b) of Theorem 3) is said to be agreeing with both the conditional belief *Bel* and its dual conditional plausibility *Pl*. Whenever there are events in $\mathcal{K}$ with zero plausibility the class of unconditional plausibilities is formed by more

than one element and we can say that Pl_1 gives a refinement of those events judged with zero plausibility under Pl_0.

The following example shows the construction of the class $\mathcal{L}$ characterizing (in the sense of the above result) a conditional belief.

EXAMPLE **1** *Let $\{C_1, ..., C_5\}$ be a partition of Ω, $\mathcal{E}$ the corresponding algebra and $\mathcal{K} = \{C_1 \vee C_5, C_2 \vee C_3 \vee C_4, C_1 \vee C_2 \vee C_5, \Omega\}$.*

Consider the following function f defined as follows on $\mathcal{E} \times \mathcal{H}$:
for $K \in \{\Omega, C_2 \vee C_3 \vee C_4\}$ and $H \subseteq C_1 \vee C_5$
$f(C_i|K) = f(H|K) = f(H \vee C_i|K) = 0$ *for* $i = 3, 4$
$f(C_2|K) = f(C_2 \vee H|K) = f(C_2 \vee C_4|K) =$
$f(C_2 \vee C_4 \vee H|K) = 0.5$,
$f(C_3 \vee C_4|K) = f(C_3 \vee C_4 \vee H|K) = 0.2$,
$f(C_2 \vee C_3|K) = f(C_2 \vee C_3 \vee H|K) = 0.8$
$f(C_2 \vee C_3 \vee C_4|K) = f(C_2 \vee C_3 \vee C_4 \vee H|K) = 1$;

moreover (for $i = 1, 5$)
$f(C_i|C_1 \vee C_2 \vee C_5) = f(C_1 \vee C_5|C_1 \vee C_2 \vee C_5) = 0$,
$f(C_2|C_1 \vee C_2 \vee C_5) = f(C_2 \vee C_i|C_1 \vee C_2 \vee C_5) =$
$f(C_1 \vee C_2 \vee C_5|C_1 \vee C_2 \vee C_5) = 1$;

and $f(C_1|C_1 \vee C_5) = 0.2$, $f(C_5|C_1 \vee C_5) = 0.3$,
$f(C_1 \vee C_5|C_1 \vee C_5) = 1$.

We can prove that the above function is a conditional belief since there exists a suitable class $\mathcal{L} = \{Pl_0, Pl_1\}$ of plausibilities such that, for any $E|F \in \mathcal{A} \times \mathcal{K}$, one has $f(E|F) = 1 - \frac{Pl_\alpha(E^c \wedge F)}{Pl_\alpha(F)}$. The function Pl_0 is defined on $\mathcal{A}$ as follows: for any $H \subseteq C_1 \vee C_5$ $Pl_0(H) =$ 0, $Pl_0(C_2) = Pl_0(C_2 \vee H) = 0.8$, $Pl_0(C_4) = Pl_0(C_4 \vee$ $H) = 0.2, Pl_0(C_3) = Pl_0(C_3 \vee H) = Pl_0(C_3 \vee C_4) =$ $Pl_0(C_3 \vee C_4 \vee H) = 0.5$,
$Pl_0(C_2 \vee C_3) = Pl_0(C_2 \vee C_3 \vee H) = Pl_0(C_2 \vee C_4) =$
$Pl_0(C_2 \vee C_4 \vee H) = Pl_0(C_2 \vee C_3 \vee C_4) =$
$Pl_0(C_2 \vee C_3 \vee C_4 \vee H) = 1$.
Note that Pl_0 is associated to the following basic assignment $m(C_2) = 0.5, m(C_2 \vee C_3) = 0.3, m(C_3 \vee C_4) = 0.2$ and it is zero otherwise.
Then, $H_1^0 = C_1 \vee C_5$, and Pl_1 is defined as follows $Pl_1(C_1) = 0.7, Pl_1(C_5) = 0.8$, $Pl_1(C_1 \vee C_5) = 1$.

Results similar to the above one, characterizing conditional possibility and necessity in terms of a class of unconditional possibilities, have been given in [4, 11, 24], and for conditional probability see e.g. [9].

2.2 Zero–layers

The characterization of conditional plausibility (and conditional belief function) in terms of a suitable class of plausibilities gives rise to the following notion of zero-layers.

Definition 3 *Let Pl be a coherent conditional plausibility on $\mathcal{F}$, and $\mathcal{L}$ a class agreeing with Pl, then, for every event $H \in \mathcal{E}$, the* zero–layer *of H (denoted as $\circ(H)$) related to $\mathcal{L}$ is defined as the minimum number α such that $Pl_\alpha(H) > 0$.*
Moreover, define $\circ(\emptyset) = +\infty$.

Zero-layers single-out a partition of the algebra, in particular it follows that the zero-layer of any event E with positive plausibility is zero. Then, if the class $\mathcal{L}$ contains only an everywhere positive plausibility Pl_o, there is only one (trivial) zero-layer.

Remark 1 *It is immediate to prove that the zero-layers, related to $\mathcal{L}$, satisfy the following formal proprieties*

$$\circ(A \vee B) = \min\{\circ(A), \circ(B)\},$$
$$\circ(A \wedge B) \geq \max\{\circ(A), \circ(B)\}.$$

Note that zero-layers (which are obviously significant for events of zero plausibility) are a tool to detect "how much" a null event is ... null. In fact, if $\circ(A) > \circ(B)$ (that is, roughly speaking, the plausibility of A is a "stronger" zero than the plausibility of B), then by Theorem 3 (b) $Pl(A|A \vee B) = 0$ and so $Pl(B|A \vee B) = 1$. On the other hand $\circ(A) = \circ(B)$ iff $Pl(A|A \vee B)Pl(B|A \vee B) > 0$; this formula recalls the probabilistic notion of *commensurable* given by de Finetti in [17].

Definition 4 *Let Pl be a coherent conditional plausibility on $\mathcal{F}$, and $\mathcal{L}$ a class agreeing with Pl, then, for every event $E|H \in \mathcal{E} \times \mathcal{K}$, the* zero–layer *of $E|H$ (denoted as $\circ(E|H)$) related to $\mathcal{L}$ is defined as the (positive) number*

$$\circ(E|H) = \circ(E \wedge H) - \circ(H).$$

Since $\circ(\emptyset) = \infty$ it results $\circ(E|H) = \infty$ iff $E \wedge H = \emptyset$.

Remark 2 *More precisely, $Pl(A|B) > 0$ if and only if $\circ(A|B) = 0$ (i.e. $\circ(A \wedge B) = \circ(B)$).*

Moreover, from the properties of conditional plausibilities, for any conditioning event H, there is at least an atom $C \subseteq H$ such that $\circ(C|H) = 0$.

EXAMPLE 1 (continued) Let us consider again the conditional plausibility in Example 1, which admits a unique agreeing class and note that $\circ(C_1 \vee C_5) = 1$ and $\circ(C_1|C_1 \vee C_5) = \circ(C_5|C_1 \vee C_5) = 0$.

The above properties recall those related to the notion of zero-layer [9] arising in de Finetti conditional probability framework and they satisfy the same properties of k-functions of Spohn [30], so suggest relevant connections with the results shown in [11, 22, 24].

3 Independence

The background is now ready to introduce a definition of independence for coherent conditional plausibilities (i.e. the measure can be assessed on arbitrary set of conditional events without requiring any algebraic structure).

Definition 5 *Given a coherent conditional plausibility Pl on a set of conditional events $\mathcal{F}$ containing $\mathcal{D} = \{A^*|B^*, A^*\}$ - where A^* (analogously B^*) stands for either A or A^c-, A is independent of B under Pl (in symbol $A \perp\!\!\!\perp B[Pl]$), if both the following condition holds:*

(a) $Pl(A|B) = Pl(A|B^c) = Pl(A)$
$Pl(A^c|B) = Pl(A^c|B^c) = Pl(A^c)$,

(aa) *there exists an agreeing class $\mathcal{L} = \{Pl_\alpha\}$ for the restriction of Pl to $\mathcal{D}$ such that*

$$\circ(A|B) = \circ(A|B^c) \text{ and } \circ(A^c|B) = \circ(A^c|B^c).$$

Remark 3 *Definition 5 requires for the statement "A independent of B under $[Pl]$" that $B \neq \Omega$ and $B \neq \emptyset$ (since conditioning events cannot be impossible).*

This syntactical constraint has also a semantical counterpart: Ω and $\emptyset$ correspond to a situation of complete information (since the former is always true and the latter always false), and so it does not make sense to ask whether they could influence the plausibility of another event.

Conversely, by definition it follows that, under any coherent conditional plausibility, the events Ω and $\emptyset$ are independent of every possible (i.e. different from Ω and $\emptyset$) event B. In fact, condition (i) holds and for any agreeing class $\circ(\Omega|B) = \circ(\Omega|B^c) = 0$ and $\circ(\emptyset|B) = \circ(\emptyset|B^c) = +\infty$.

This conclusion is natural, since the plausibility (1 and 0, respectively) of Ω and $\emptyset$ cannot be changed by assuming the occurrence of any other possible event B.

In condition *(a)* of Definition 5 we require equalities that could seem very strong at the first light, this is due to remove situations such as those arising in the following examples:

EXAMPLE **2** *Let consider a basic assignment on the algebra generated by two possible events A and B, with focal elements*

$$m(A \wedge B) = m(A \wedge B^c) = m(A^c \wedge B) =$$

$$m(A^c \wedge B^c) = m(A \vee B) = \frac{1}{5}$$

(i.e. on all the other events of the algebra $m(\cdot)$ is equal to zero). This basic probability assignment implies $Pl(A) = Pl(A^c) = Pl(B) = Pl(B^c) = \frac{3}{5}$, $Pl(A \wedge B) = Pl(A^c \wedge B) = Pl(A \wedge B^c) = \frac{2}{5}$ but $Pl(A^c \wedge B^c) = \frac{1}{5}$. By applying the conditioning rule (Definition 1) it follows that $Pl(A|B) = Pl(A|B^c) = \frac{2}{3} \neq Pl(A)$. Moreover, $Pl(A^c|B) = \frac{2}{3} \neq \frac{1}{3} = Pl(A^c|B^c)$.

The above example shows that $Pl(A|B) = Pl(A|B^c)$ does not imply neither $Pl(A|B) = Pl(A)$ nor $Pl(A^c|B) = Pl(A^c|B^c)$, furthermore from the next example it arises the necessity of requiring all the equalities in condition *(a)*.

EXAMPLE **3** *Consider the following basic assignment*

$$m(A \wedge B) = m(A \wedge B^c) = m(A^c \wedge B) =$$

$$m(A^c \wedge B^c) = m(\Omega) = \frac{1}{5}.$$

Then, $Pl(A^* \wedge B^*) = \frac{2}{5}, Pl(A^*) = Pl(B^*) = \frac{3}{5}$ and $Pl(A^*|B^*) = \frac{2}{3}$. This implies that

$$Pl(A|B) = Pl(A|B^c)$$

and

$$Pl(A^c|B) = Pl(A^c|B^c),$$

but $Pl(A|B) \neq Pl(A)$.

When both $Pl(A)$ and $Pl(A^c)$ are greater than zero condition *(a)* of Definition 5 assures that $A \perp\!\!\!\perp B[Pl]$, in fact in this case all the zero-layers in condition *(aa)* are equal to 0 and so condition *(aa)* is trivially satisfied.

If condition *(a)* holds and $Pl(A) = 0$ $[Pl(A^c) = 0]$, then the second [first] equality under *(aa)* is trivially satisfied, so that the statement $A \perp\!\!\!\perp B[Pl]$ is ruled by the first [second] one. In other words equality *(a)* is not enough to assure independence in this situation: *it needs to be reinforced by the requirement that also their zero-layers must be equal.*

We finally note that the statement $A \perp\!\!\!\perp B\,[Pl]$ depends only on the restriction of the assessment Pl on $\mathcal{D}$, hence the statement is not effected by the values of the assessment Pl on $\mathcal{F} \setminus \mathcal{D}$ (actually the influence, e.g. of $Pl(B|A)$ is related to condition *(aa)*, as it will be clear from the next result). Since *(aa)* depends on a class agreeing with the coherent conditional plausibility, and since this class is in general not unique, it is necessary to prove that independence is well-defined by Definition 5, that means that is invariant with respect to the choice of any agreeing class.

Theorem 4 *Given two events A and B such that $B \neq \emptyset, \Omega$ and a coherent conditional plausibility Pl defined on $\mathcal{F}$, containing $\mathcal{D} = \{A^*|B^*, A^*\}$, such that*

$$Pl(A|B) = Pl(A|B^c) = Pl(A)$$
$$Pl(A^c|B) = Pl(A^c|B^c) = Pl(A^c).$$

If there exists a class agreeing with $Pl|_{\mathcal{D}}$ such that

$$\circ(A|B) = \circ(A|B^c) \text{ and } \circ(A^c|B) = \circ(A^c|B^c),$$

then this holds for any other class agreeing with $Pl|_{\mathcal{D}}$.

Proof: This theorem can be decomposed in three main cases:

1. $Pl(A) \cdot Pl(A^c) > 0$,
2. $Pl(A) = 0$,
3. $Pl(A^c) = 0$.

1. If $Pl(A) \cdot Pl(A^c) > 0$ the theorem is true since $\circ(A^*|B^*) = 0$ for all agreeing class.

2. If $Pl(A) = 0$ then $Pl(A^c) = 1$ and the only masses which can be greater than zero (i.e. the focal elements) are $m(A^c \wedge B)$, $m(A^c \wedge B^c)$, $m(A^c)$. If an agreeing class is such that $Pl(B) \cdot Pl(B^c) > 0$ (i.e. $m(A^c) > 0$ or $m(A^c \wedge B) \cdot m(A^c \wedge B^c) > 0$) then (in both the cases) $\circ(A^c|B) = \circ(A^c|B^c) = 0$. Now, we need to look at $B|A$ and $B^c|A$, through the system (S^1) (of Theorem 3), that can be written in a compact form by referring to Pl^1 and m^1, i.e.

$$(S^1) = \begin{cases} Pl^1(A \wedge B) = Pl(B|A) \cdot Pl^1(A), \\ Pl^1(A \wedge B^c) = Pl(B^c|A) \cdot Pl^1(A), \\ m^1(A \wedge B) + m^1(A \wedge B^c) + m^1(A) = 1, \\ m^1(\cdot) \geq 0. \end{cases}$$

To second equality of condition *(aa)* of Definition 5 holds if and only if $\circ(A \wedge B) = \circ(A \wedge B^c)$, that means $Pl(B|A) \cdot Pl(B^c|A) > 0$. Then, if $Pl(B|A) \cdot Pl(B^c|A) > 0$ all the agreeing class with $Pl_{|\mathcal{D}}$ are such that $\circ(A|B) = \circ(A|B^c) = 1$ and $\circ(A^c|B) = \circ(A^c|B^c) = 0$; otherwise none agreeing class satisfies condition *(aa)*.

If $Pl(B) = 0$ (i.e. $m(A^c \wedge B^c) = 1$) then $\circ(A^c|B^c) = \circ(B^c) = 0$ and (S^1) is

$$(S^1) = \begin{cases} Pl^1(A \wedge B) = 0 \cdot Pl^1(B), \\ Pl^1(A^c \wedge B) = 1 \cdot Pl^1(B), \\ Pl^1(A \wedge B) = Pl(B|A) \cdot Pl^1(A), \\ Pl^1(A \wedge B^c) = Pl(B^c|A) \cdot Pl^1(A), \\ m^1(\cdot) \geq 0. \end{cases}$$

A solution of (S^1) is such that $m(D) = 0$ for any $D \wedge (A \wedge B) \neq \emptyset$; then when $Pl(B|A) = 0$ we need to take in consideration the following cases

- $Pl^1(A) \cdot Pl^1(B) > 0$, then it follows
$Pl^1(A \wedge B^c) \cdot Pl^1(A^c \wedge B) > 0$ and
$\circ(A|B) = \circ(A \wedge B) - 1 = 1, \circ(A|B^c) = 1$ and
$\circ(A^c|B) = 1 - 1 = 0 = \circ(A^c|B^c)$.
- $Pl^1(A) > 0$ and $Pl^1(B) = 0$, then it follows
$\circ(A|B) = \circ(A \wedge B) - 2 = 1, \circ(A|B^c) = 1$ and
$\circ(A^c|B) = 2 - 2 = 0 = \circ(A^c|B^c)$.
- $Pl^1(A) = 0$ and $Pl^1(B) > 0$, then it follows
$\circ(A|B) = \circ(A \wedge B) - 1 = 2, \circ(A|B^c) = 2$ and
$\circ(A^c|B) = 1 - 1 = 0 = \circ(A^c|B^c)$.
On the other hand, when $Pl(B|A) > 0$, it follows from the above system $Pl^1(A) = 0$, so $Pl^1(B) = 1$ and $\circ(A^c \wedge B) = 1, \circ(A \wedge B) = 2$, while $\circ(A \wedge B^c) \geq 2$. It implies $\circ(A|B) = 1$ while $\circ(A|B^c) \geq 2$.

We can conclude this case: if $Pl(B|A) = 0$ any agreeing class of $Pl_{\mathcal{D}}$ satisfies the two equalities; while if $Pl(B|A) > 0$ no agreeing class satisfies the two equalities among the relevant zero-layers.

If $Pl(B^c) = 0$ is analogous to the previous one, just exchange B with B^c.

3. If $Pl(A^c) = 0$ is the same as 2., with A^c playing the role of A.

From the above result we get

Corollary 1 *Given a coherent conditional plausibility Pl defined on $\mathcal{F}$. If A is independent of B under Pl, then*

$$Pl(A \wedge B) = Pl(A)Pl(B).$$

It follows that the proposed notion of independence implies *cognitive independence* of Shafer [29], called also *weak independence* by Kong [27].

We have also the converse implication under suitable hypothesis, as shown in the next result.

Proposition 1 *Given a coherent conditional plausibility Pl defined on $\mathcal{F}$. If $Pl(B)$, $Pl(B^c)$, $Pl(A), Pl(A^c)$ are greater than 0, and*

$$Pl(A \wedge B) = Pl(A)Pl(B)$$
$$Pl(A \wedge B^c) = Pl(A)Pl(B^c)$$
$$Pl(A^c \wedge B) = Pl(A^c)Pl(B)$$
$$Pl(A^c \wedge B^c) = Pl(A^c)Pl(B^c)$$

then A is independent of B under $[Pl]$.

Proof: It follows directly from the definition of conditional plausibility and the properties of zero-layers.

The following example shows that the positivity condition cannot be avoided.

EXAMPLE **4** *Let A, B be two possible events and consider the assessment*
$Pl(B) = 0, Pl(B^c) = 1, Pl(A \wedge B^c) = Pl(A^c \wedge B^c) = Pl(A) = Pl(A^c) = Pl(A|B^c) = Pl(A^c|B^c) = \frac{2}{3}$,
$Pl(A|B) = Pl(A^c|B) = \frac{1}{2}$.
It is easy to show that Pl is a coherent conditional plausibility and for any atom generated by A and B, e.g. $A \wedge B$, its plausibility is equal to the product of the plausibilities of A and B, i.e.

$$Pl(A \wedge B) = Pl(A)Pl(B).$$

But, under Pl, we have that A is not independent of B.

Proposition 2 *Under any coherent conditional plausibility Pl, for any event A the statement "A is independent of itself" does not hold.*

Proof: Since by the axioms of conditional plausibilities we have that $Pl(A|A) = 1$, while $Pl(A|A^c) = Pl(\emptyset|A^c) = 0$, if follows that the statement does not hold.

The previous property (irreflexivity) is natural and essential, in fact *any event must be dependent on itself.*

Moreover, *independence implies logical independence*, as proved below. Recall that two events A and B are logically independent if all the events of the form $A^* \wedge B^*$ (where A^* stands for A or A^c) are possible, i.e. the number of relevant atoms is maximal.

Theorem 5 *Let Pl be a coherent conditional plausibility defined on $\mathcal{F}$. Given two possible events $A, B \in \mathcal{F}$, if A is independent of B under Pl, then A and B are logically independent.*

Proof: If there is a logical constraint between A and B we show that there is no agreeing class satisfying condition *(aa)*. If, for example, $A \wedge B = \emptyset$, then $Pl(A|B) = 0$ and $\circ(\emptyset) = \circ(A|B) = +\infty$; while being $A \wedge B^c = A$ a possible event $\circ(A|B^c) \leq \circ(A \wedge B^c) < +\infty$. The proof for other logical constraints follows similarly.

This is an intuitive implication: in fact *if an event is "logically" related to another, the two events must be not independent* under *any* uncertainty measure. Handling logical constraints is interesting also from a practical point of view, since in many real applications variables are suitably linked.

Remark 4 *Actually, independence under a measure assures the logical independence and this implication is guaranteed by the requirement (aa) of Definition 5. We recall that logical independence is taken into account also in [29] (as well in [2]), and it looks natural looking on Dempster rule. However, the independence notion introduced in [2] do not respect the above implication when events with degree of belief 0 are involved.*

We recall that the main difference between the approaches of [29] and [2] is that the first is referred to belief or plausibility function that are normalized (as in this paper) while in the second approach are taken in consideration also not normalized measures.

The following result characterizes independence in terms of the conditional plausibility (avoiding zero-layers).

Theorem 6 *Let A and B be two logically independent events. If a coherent conditional plausibility Pl is such that*

$$Pl(A|B) = Pl(A|B^c) = Pl(A)$$

and

$$Pl(A^c|B) = Pl(A^c|B^c) = Pl(A^c)$$

then $A \perp\!\!\!\perp B[Pl]$ if and only if one (and only one) of the following conditions holds:

1. $Pl(A) \cdot Pl(A^c) > 0$;
2. $Pl(A) = 0$ *and the coherent extension of Pl to $Pl(B)$, $Pl(B^c)$, $Pl(B|A)$, $Pl(B^c|A)$ satisfies one of the following:*
   a) $Pl(B) \cdot Pl(B^c) > 0$ *and* $Pl(B|A) \cdot Pl(B^c|A) > 0$,
   b) $Pl(B) = 0$ *and* $Pl(B|A) = 0$,
   c) $Pl(B^c) = 0$ *and* $Pl(B^c|A) = 0$;
3. $Pl(A^c) = 0$ *and the coherent extension of Pl to $Pl(B)$, $Pl(B^c)$, $Pl(B|A^c)$, $Pl(B^c|A^c)$ satisfies one of the following:*
   a) $Pl(B) \cdot Pl(B^c) > 0$ *and* $Pl(B|A^c) \cdot Pl(B^c|A^c) > 0$,
   b) $Pl(B) = 0$ *and* $Pl(B|A^c) = 0$,
   c) $Pl(B^c) = 0$ *and* $Pl(B^c|A^c) = 0$;

Proof: The items highlighted in the theorem statement follow directly by the proof of Theorem 4. In particular when $Pl(A) \cdot Pl(A^c) > 1$ is obvious because $\circ(A^*|B^*) = 0$ for all agreeing class. Moreover cases $Pl(A) = 0$ and $Pl(A^c) = 0$ correspond to the case 2. and 3. of Theorem 4 respectively: this result follows along the same proof of the previous result.

From the above result it comes out that the provided independence notion is not symmetric, and this happens when events with zero plausibility are involved. If $Pl(A), Pl(A^c), Pl(B), Pl(B^c)$ takes positive values and $A \perp\!\!\!\perp B$ under Pl, then

$$Pl(B|A) = \frac{Pl(A|B)Pl(B)}{Pl(A)} = Pl(B),$$

so going along the same computations $Pl(B|A) = Pl(B|A^c)$ and $Pl(B^c|A) = Pl(B^c|A^c) = Pl(B^c)$, which implies that also the statement $B \perp\!\!\!\perp A$ holds.

Since coherent conditional probability are a particular plausibility, and since the provided conditional independence for conditional plausibility is just a generalization of that given for conditional probability [9, 32] the fact that symmetry can fail when possible events of zero plausibility are involved is not a surprise, different examples have been given in the quoted papers to show that the lack of symmetry can be intuitive.

3.1 Independence for belief functions

By means of duality we obtain that if Pl is a coherent conditional plausibility on a set of conditional events $\mathcal{C}$ and Bel is its dual function on

$$\mathcal{C}^* = \{E|H \,:\, E^c|H \in \mathcal{C}\},$$

then Bel is a coherent conditional belief function (see Theorem 3).

Moreover, for $A|B, A|B^c, A \in \mathcal{C}$

$$Pl(A|B) = Pl(A|B^c) = Pl(A)$$

if and only if

$$Bel(A^c|B) = Bel(A^c|B^c) = Bel(A^c)$$

for $A^c|B, A^c|B^c, A^c \in \mathcal{C}$.

Then, it could seem reasonable to take A independent of B under Bel if and only if A is independent of B under the dual conditional plausibility Pl.

Recall that as shown in Section 2 a class $\mathcal{L}$ is agreeing for Bel if and only if it is agreeing also for the dual conditional plausibility Pl.

Note that this means that many properties of independence under plausibilities continue to be valid under belief functions, as for example independence implies logical independence. Moreover, also several results can be reformulated, as e.g. the characterization of independence of two possible events in terms of their belief (as done for plausibilities in Theorem 6).

Nevertheless, this notion of independence need to be studied more deeply: we need to detect better the role of zero-layers, and to exploit the relationship with the factorization property, i.e.

$$Bel(A^* \wedge B^*) = Bel(A^*)Bel(B^*).$$

Actually, the factorization has been adopt (as notion of independence) in [28] to prove under some technical hypothesis a strong law of large numbers for belief functions.

In the following example we propose a situation where, under a plausibility Pl, the event A is independent of B, but the factorization fails under the dual function of Pl.

EXAMPLE **5** *Consider the following basic assignment with focal elements*

$$m_{A\wedge B} = \frac{1}{2}, m_{A\vee B} = m_\Omega = \frac{1}{4};$$

which gives rice to the following belief function $Bel(A \wedge B) = Bel(A) = Bel(B) = Bel(A \vee B^c) = Bel(A^c \vee B) = \frac{1}{2}$, $Bel(A^c \wedge B^c) = Be(A^c) = Bel(B^c) = Bel(A^c \vee B^c) = 0$, *and so to plausibility* $Pl(A \wedge B) = Pl(A) = Pl(B) = 1$ $Pl(A \wedge B^c) = Pl(A^c \wedge B) = \frac{1}{2}$ $Pl(A^c \wedge B^c) = \frac{1}{4}$.

Then, the induced conditional plausibility is such that $A \perp\!\!\!\perp B[Pl]$ *(and* $B \perp\!\!\!\perp A[Pl]$*), but* $Bel(A \wedge B) = \frac{1}{2} \neq Bel(A)Bel(B) = \frac{1}{4}$.

Thus, our notion of independence under a plausibility is stronger than *cognitive independence* [27, 29]. However, in the case of positive events it does not imply *evidential independence*, called also strong independence [29, 27], which coincides with the requirement of factorization of the belief function and its dual. However, by adding to our independence notion the factorization property with respect to the belief function, we obtain a notion stronger than evidential independence. These considerations are useful also for comparing our notion with some concepts of independence, irrelevance and non-interactivity, given in [2] (also for non necessarily normalized measures) since the notion of doxastic independence and non-interactivity coincide with evidential independence in the case of interest, i.e. for normalized measures.

3.2 Conditional independence

The notion of independence between two events given in Definition 5 can be generalized to that of conditional independence:

Definition 6 *Given a coherent conditional plausibility Pl on a set of conditional events $\mathcal{F}$ containing $\mathcal{D} = \{A^*|B^* \wedge C, A^*|C\}$, A is independent of B conditionally to C under Pl (in symbol $A \perp\!\!\!\perp B|C$ $[Pl]$), if both the following conditions hold:*

i. $Pl(A|B \wedge C) = Pl(A|B^c \wedge C) = Pl(A|C)$
$Pl(A^c|B \wedge C) = Pl(A^c|B^c \wedge C) = Pl(A^c \wedge C)$,

ii. there exists an agreeing class $\mathcal{L} = \{Pl_\alpha\}$ *for the restriction of* Pl *to* $\mathcal{D}$ *such that*

$$\circ(A|B \wedge C) = \circ(A|B^c \wedge C) \text{ and}$$
$$\circ(A^c|B \wedge C) = \circ(A^c|B^c \wedge C).$$

Considerations similar to the unconditional case can be done: when $Pl(A|B \wedge C)$ and $Pl(A^c|B \wedge C)$ are both positive, then both equalities in condition *ii.* are trivially (as 0=0) satisfied. While in the other two cases (i.e. $Pl(A|B \wedge C) = 0$ or $Pl(A^c|B \wedge C) = 0$) the equality *i.* is not enough to assure independence, so it is "reinforced" by the requirement that also their relevant zero-layers must be equal.

Remark 5 *If the events* A *and* C *(or* A^c *and* C*) are incompatible, then* A *is independent of any event* B *given* C *whenever* $\emptyset \neq B \wedge C \neq C$*. This conclusion is natural since the plausibility* 0 *(or* 1*) of* $A|C$ *cannot be changed by assuming the occurrence of* B*.*

Actually, even if the restriction of Pl to $\mathcal{D}$ admits more than one agreeing class, we can prove along the line of Theorem 4, that condition *ii.* of Definition 6 holds either for all agreeing classes or for none of them.

Going on the same line of the proofs given for Theorem 4 and Theorem 6, we can characterize conditional independence in terms of plausibilities: it would be a simply generalization of Theorem 6.

4 Summary and Conclusions

In this paper, we look to conditional plausibility and belief from a more general point of view. In particular we are able to handle events with null measure. In this framework we provide a definition of independence and we give a characterization of it, we study its main properties, which allow us to compare our definition with other given in literature (in particular with respect to the notions introduced in [2, 3, 27, 29]). We recall that our notion of independence for plausibility is in the same line of that studied in [9, 11, 24, 32] for probability and possibility.

Through different examples we explain also the reason for taking exactly the provided definition, our choice has been guided mainly by two main reasons: to get a natural condition overcoming critical aspects and to get a suitable factorization of the joint plausibility distribution.

We show that the provided independence notion is not necessarily symmetric, then to represent such statements we need to refer to some not necessarily symmetric separation criterion such as that proposed in [33]. An open problem consists into looking for the representability of the set of independence statements induced by a conditional plausibility (belief) by means of a directed or undirected graph by testing which properties among the graphoid ones are satisfied: this would allow to compare our definition also with other independence notions given in the context of other uncertainty formalisms.

References

[1] N. Ben Amor, K. Mellouli, S. Benferhat, D. Dubois, H. Prade. A theoretical framework for possibilistic independence in a weakly ordered setting. Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(2), 117–155, 2002.

[2] B. Ben Yaghlane, P. Smets, K. Mellouli. Belief function independence: I. The marginal case *Int. Journal of Approximate Reasoning*, 29, 47-70, 2002.

[3] B. Ben Yaghlane, P. Smets, K. Mellouli. Belief function independence: I. The conditional case *Int. Journal of Approximate Reasoning*, 31, 31-75, 2002.

[4] B. Bouchon-Meunier, G. Coletti, C. Marsala. Independence and Possibilistic Conditioning. *Annals of Mathematics and Artificial Intelligence*, 35, . 107-124, 2002.

[5] L.M. Campos, J.F. Huete. Independence concepts in upper and lower probabilities. In: *Uncertainty in Intelligent Systems* (Eds. Bouchon-Meunier et al.), 85-96, 1993.

[6] G. Coletti, M. Mastroleo Conditional belief functions: a comparison among different definitions. *Proceeding of* 7^{th} *Workshop on Uncertainty Processing (WUPES 2007)*, 2006.

[7] G. Coletti, R. Scozzafava. From conditional events to conditional measures: a new axiomatic approach. *Annals of Mathematics and Artificial Intelligence* 32, 373–392, 2001.

[8] G. Coletti, R. Scozzafava. Stochastic Independence in a Coherent Setting. *Annals of Mathematics and Artificial Intelligence* 35, 151176, 2002.

[9] G. Coletti, R. Scozzafava. Probabilistic Logic in a Coherent Setting (Trends in Logic, n.15). Dordrecht: Kluwer: 2002.

[10] G. Coletti, R. Scozzafava. Toward a General Theory of Conditional Beliefs. *International Journal of Intelligent Systems* 21, 229-259, 2006.

[11] G. Coletti, B. Vantaggi. Possibility theory: conditional independence. *Fuzzy Sets and Systems*, 157(11), 1491–1513, 2006.

[12] G. Coletti, B. Vantaggi. A view on conditional measures through local representability of binary relations. Submitted in *International Journal of Approximate Reasoning*.

[13] F.G. Cozman. Irrelevance and independence axioms in quasi-Bayesian theory, in: T. Hunter, S. Parsons (Eds.), *Proc. of ECSQARU99, Lecture Notes in AI 1638*, Springer-Verlag, Berlin, 1999, pp. 128136, 2006.

[14] R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14 (1), 1–13, 1946.

[15] A.P. Dawid Conditional independence in statistical theory. *Journal of Royal Statistical Society B*, 41, 15-31, 1979.

[16] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Univ. Trieste*, 19, 3–55 1949 - *Engl. transl.*: Ch.5 in *Probability, Induction, Statistics*, Wiley, London, 1972.

[17] B. de Finetti. Les probabilitès nulles. *Bull. Sci. Math.*, 60, 175–288, 1936.

[18] A.P. Dempster. A generalizatin of Bayesian Inference. *The Royal Stat. Soc. B*, 50, 205–247, 1968.

[19] T. Denoeux, P. Smets, Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches, To appear in *IEEE Transactions on Systems, Man and Cybernetics B*, 2006.

[20] F.T. de Dombal, F. Gremy. *Decision Making and Medical Care*. North-Holland, 1976.

[21] J.R. Hill. Comment on Graphical models. *Statistical Science* 8, 258-261, 1993.

[22] D. Dubois, L. Fariñas del Cerro, A. Herzig, H. Prade. An ordinal view of independence with application to plausible reasoning. *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence* (R. Lopez de Mantaras, D. Poole, eds.), Seattle, WA, 195-203 1994.

[23] Dubois, H. Prade, Possibility Theory New York, Plenum Press, 1988.

[24] L. Ferracuti, B. Vantaggi. Independence and conditional possibilities for strictly monotone triangular norms. *International Journal of Intelligent Systems*, 21(3), 299-323, 2006.

[25] J.Y. Halpern. A counterexample to theorems of Cox and Fine. *J. of Artificial Intelligence Research*, 10, 67–85, 1999.

[26] J.Y. Jaffray. Bayesian updatingand Belief functions. *IEEE Trans. on Systems, man and Cybernetic*, 5, 1144–1152, 1992.

[27] C.T.A. Kong. A belief function generalization of Gibbs Ensemble. Tec. Report S-122 Harvard University, 1988.

[28] F. Maccheroni, M. Marinacci. A strong law of large numbers for capacities. *Annals of Probabilities*, 33(3), 1171–1178, 2005.

[29] G. Shafer. *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, 1976.

[30] W. Spohn. On the Properties of Conditional Independence, in: Humphreys, Suppes, (Eds), Scientific Philosopher 1: Probability and Probabilistic Causality, Kluwer, Dordrecht, 173–194.

[31] M. Studeny. Formal properties of conditional independence in different calculi of AI, in: Clarke, Kruse, Moral (Eds), Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Berlin, Springer-Verlag, 341–348, 1993.

[32] B. Vantaggi. Conditional independence in a finite coherent setting. *Annals of Mathematics and Artificial Intelligence*, 32, 287-314, 2001.

[33] Vantaggi B. The L-separation criterion for description of cs-independence models. *International Journal of Approximate Reasoning*, 29, 291-316, 2002.

[34] B. Vantaggi. Conditional independence structures and graphical models. *Int. Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 11(5), 545-571, 2003.

[35] J. Vejnarova. Conditional Independence Relations in Possibility Theory. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 8(3), 253-269, 2000.

[36] P. Walley. Belief function representations of statistical evidence. *Annals of Statistics*, 4, 1439–1465, 1987.

Coherence graphs

Enrique Miranda
Rey Juan Carlos University
Dept. of Statistics and Operations Research
C-Tulipán, s/n, 28933 Móstoles,
Spain
enrique.miranda@urjc.es

Marco Zaffalon
IDSIA
Galleria 2, CH-6928 Manno (Lugano),
Switzerland
zaffalon@idsia.ch

Abstract

We consider the task of proving Walley's (*joint* or *strong*) *coherence* of a number of probabilistic assessments, when these assessments are represented as a collection of conditional lower previsions. In order to maintain generality in the analysis, we assume to be given nearly no information about the numbers that make up the lower previsions in the collection. Under this condition, we investigate the extent to which the above global task can be decomposed into simpler and more local ones. This is done by introducing a graphical representation of the conditional lower previsions, that we call the *coherence graph*: we show that the coherence graph allows one to isolate some subsets of the collection whose coherence is sufficient for the coherence of all the assessments. The situation is shown to be completely analogous in the case of Walley's notion of *weak coherence*, for which we prove in addition that the subsets found are optimal, in the sense that they embody the maximal degree to which the task of checking weak coherence can be decomposed. In doing all of this, we obtain a number of related results: we give a new characterisation of weak coherence; we characterise, by means of a special kind of coherence graph, when the local notion of *separate coherence* is sufficient for coherence; and we provide an envelope theorem for collections of lower previsions whose graph is of the latter type.

Keywords. Walley's strong and weak coherence, coherent lower previsions, graphical models, coherence graph.

1 Introduction

Suppose we plan to carry out a statistical analysis about a certain domain modelled by the following *lower previsions*:

$$\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_3|X_2), \underline{P}_4(X_4|X_3),$$
$$\underline{P}_5(X_5, X_6|X_1), \underline{P}_6(X_2|X_3), \underline{P}_7(X_7|X_4), \underline{P}_8(X_8|X_5),$$
$$\underline{P}_9(X_8|X_6), \underline{P}_{10}(X_9, X_{10}|X_6, X_7), \underline{P}_{11}(X_{11}|X_9, X_{10}).$$

Each of them represents a real functional interpreted as a subject's lower prevision (i.e., lower expectation) for every bounded real-valued function of the random variables on the l.h.s. of the bar, conditional on given values of the variables on the r.h.s. of the bar.

In order to carry out the analysis, we should first verify that the assessments represented by the lower previsions are self-consistent, or *coherent*. Indeed coherence is a (minimal) requirement of rationality, and it is the key that enables one to use a number of powerful theoretical tools to do statistical inference.

Yet, checking coherence can be particularly difficult even in the simple setting illustrated above. In fact, this is a common problem. The power of coherence comes with a price: the technical complications that arise when dealing explicitly with it.

This is the case of the coherence notion that is the focus of this paper, i.e., Walley's definition of coherence [4, Section 7.1.4(b)], which we also call *joint* or *strong coherence*, so as to distinguish it from a weaker notion, also developed by Walley, and called *weak coherence*. Weak and strong coherence are reviewed in Section 2 of this paper, along with other introductory material about Walley's theory of coherent lower previsions.

We argue that the mentioned difficulty is strictly related to the fact that coherence, by its very nature, is a *global* notion: as such, it seems to resist being represented and verified in a local fashion. This is enforced by our initial results in Section 3: we show that a number of (conditional) lower previsions, such as $\underline{P}_1$, $\ldots$, $\underline{P}_{11}$ in the above example, are weakly coherent if and only if there is an extension, i.e., a lower prevision $\underline{P}(X_1, \ldots, X_{11})$ in the example, that is *pairwise* coherent with each of them; and they are strongly co-

herent if and only if they are *globally* coherent with such an extension. In other words, strong coherence seems to be much less amenable to local considerations than other, weaker notions of coherence.

Still, locality is an important property: it is the basis for having compact and efficient models of uncertainty, as well as models that are easier to understand, as it is widely acknowledged after the lesson given by *graphical models* in statistics and artificial intelligence.

The question, at this point, is the following: can we preserve both locality and (strong) coherence?

We regard the present paper as a first positive answer to this question; such an answer is made possible by a new graphical model that we propose in Section 5, and that we call *coherence graph*. Coherence graphs are graphical representations of the structural connections of the lower previsions in a given collection. For example, the coherence graph for the lower previsions $\underline{P}_1, \ldots, \underline{P}_{11}$ is shown in Figure 1. Its semantics should be obvious once we identify the lower previsions with the black solid circles in the graph.

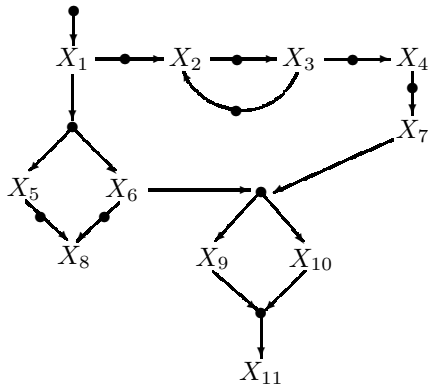


Figure 1: The coherence graph for $\underline{P}_1, \ldots, \underline{P}_{11}$.

We talk of structural connections, or of *collection template*, as defined in Section 4, because we do not focus on the numbers that make up the lower previsions; by coherence graphs we rather aim at revealing the *structure* behind the notion of coherence. This structure tells us to what extent the task of checking coherence can be made local. For instance, from the graph in Figure 1 we shall deduce that $\underline{P}_1, \ldots, \underline{P}_{11}$ are coherent if so are the lower previsions

$$\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_3|X_2), \underline{P}_5(X_5, X_6|X_1),$$
$$\underline{P}_6(X_2|X_3), \underline{P}_8(X_8|X_5), \underline{P}_9(X_8|X_6).$$

More generally speaking, our main result, stated in Section 6, is that coherence graphs allow us to graphically find out a so-called *minimal* partition of the collection of lower previsions, such that the coherence of the lower previsions in each set of the partition is sufficient for the coherence of the overall collection.

We show that the situation is completely analogous if we focus on weak coherence: proving weak coherence within each set of the minimal partition is sufficient for the weak coherence of the overall collection. In addition, in the case of weak coherence we can show that the partition found using the coherence graph is indeed minimal, in the sense that it is not possible to use non-coarser[1] alternative partitions to the same extent.

We should mention that these results are fully general with respect to the kind of admissible possibility spaces: they hold irrespective of the fact that we are dealing with finite, countable, or continuous spaces, possibly at the same time. Hence, our results can be used in fields as diverse as expert systems and statistics, just to name a few. Moreover, they are also valid for collections of *linear* previsions, i.e., they do not depend on the precise or imprecise character of the assessments.

We see two major consequences of our results. The first is directly related to proving coherence. We believe that coherence graphs, by making the structure behind coherence explicit, have the potential to give a boost to the theoretical advances in probability, and especially in imprecise probability. Similarly, there seems to be substantial hope for coherence graphs to enhance also the state-of-the-art algorithms for proving coherence. In the case of finite spaces of possibilities, this task is typically addressed by linear programming problems [1, 2, 5] that tend to grow very large as a consequence of the underlying *NP-hardness* of the task itself. But, by exploiting the structure of coherence graphs, it will often be possible to decompose the overall linear programming problem in a number of smaller ones, thus speeding up computation.

The second consequence is more of a principled kind, and is related to a subset of coherence graphs, called of *type A1*, that leads to partitions entirely made up of singletons. This implies that the related collections are immediately known to be coherent, irrespective of their numerical values, provided each of their elements satisfies a local property, called *separate coherence*. We should like to give a special perspective on these collections, by making an analogy with propositional logic. In propositional logic, the formulas that hold irrespective of the values their Boolean variables take, are called *tautologies*, and are regarded as the rules of

[1] Proving weak coherence within the sets of a coarser partition would immediately imply weak coherence within the sets of the minimal partition.

logic. We think that collections of lower previsions that have an A1-representation have a special role, and may embody a kind of 'compositional' rules that deliver jointly coherent collections by local considerations alone.

2 Coherent lower previsions

Let us give a short introduction to the concepts and results from the behavioural theory of imprecise probabilities that we shall use in the rest of the paper. We refer to [4] for an in-depth study of these and other properties.

Given a possibility space Ω, a *gamble* is a bounded real-valued function on Ω. This function represents a random reward $f(\omega)$, which depends on the a priori unknown value ω of Ω. We shall denote by $\mathcal{L}(\Omega)$ the set of all gambles on Ω. A *lower prevision* $\underline{P}$ is a real functional defined on some set of gambles $\mathcal{K} \subseteq \mathcal{L}(\Omega)$. It is used to represent a subject's supremum acceptable buying prices for these gambles, in the sense that for any $\epsilon > 0$ and any f in $\mathcal{K}$ the subject is disposed to accept the uncertain reward $f - \underline{P}(f) + \epsilon$.

We can also consider the supremum buying prices for a gamble, *conditional* on a subset of Ω. Given such a set B and a gamble f on Ω, the lower prevision $\underline{P}(f|B)$ represents the subject's supremum acceptable buying price for the gamble f, updated after coming to know that the unknown value ω belongs to B, and nothing else. If we consider a partition $\mathcal{B}$ of Ω (for instance a set of categories), then we shall represent by $\underline{P}(f|\mathcal{B})$ the gamble on Ω that takes the value $\underline{P}(f|B)$ if and only if $\omega \in B$. The functional $\underline{P}(\cdot|\mathcal{B})$ that maps any gamble f on its domain into the gamble $\underline{P}(f|\mathcal{B})$ is called a *conditional lower prevision*.

Let us now re-formulate the above concepts in terms of random variables, which are the focus of our attention in this paper. Consider random variables $X_1, \ldots, X_n$, taking values in respective sets $\mathcal{X}_1, \ldots, \mathcal{X}_n$. For any subset $J \subseteq \{1, \ldots, n\}$ we shall denote by X_J the (new) random variable

$$X_J := (X_j)_{j \in J},$$

which takes values in the product space

$$\mathcal{X}_J := \times_{j \in J} \mathcal{X}_j.$$

We shall also use the notation $\mathcal{X}^n := \mathcal{X}_{\{1,\ldots,n\}}$. This will be our possibility space in the rest of the paper.

Definition 1. Let J be a subset of $\{1, \ldots, n\}$, and let $\pi_J : \mathcal{X}^n \to \mathcal{X}_J$ be the so-called *projection operator*, i.e., the operator that drops the elements of a vector in $\mathcal{X}^n$ that do not correspond to indexes in J. A gamble f on $\mathcal{X}^n$ is called $\mathcal{X}_J$*-measurable* when for any $x, y \in \mathcal{X}^n$, $\pi_J(x) = \pi_J(y)$ implies that $f(x) = f(y)$.

There exists a one-to-one correspondence between the gambles on $\mathcal{X}^n$ that are $\mathcal{X}_J$-measurable and the gambles on $\mathcal{X}_J$: given an $\mathcal{X}_J$-measurable gamble f on $\mathcal{X}^n$, we can define f' on $\mathcal{X}_J$ by $f'(x) := f(x')$, where x' is any element in $\pi_J^{-1}(x)$; conversely, given a gamble g on $\mathcal{X}_J$, the gamble g' on $\mathcal{X}^n$ given by $g'(x) := g(\pi_J(x))$ is $\mathcal{X}_J$-measurable.

Consider two disjoint subsets O, I of $\{1, \ldots, n\}$. Then, $\underline{P}(X_O|X_I)$ represents a subject's behavioural dispositions about the gambles that depend on the outcome of the variables $\{X_k, k \in O\}$, after coming to know the outcome of the variables $\{X_k, k \in I\}$. As such, it is defined on the set of gambles that depend on the values of the variables in $O \cup I$ only, i.e., in the set $\mathcal{K}_{O \cup I}$ of the $\mathcal{X}_{O \cup I}$-measurable gambles on $\mathcal{X}^n$. Given such a gamble f and $x \in \mathcal{X}_I$, $\underline{P}(f|X_I = x)$ represents his supremum acceptable buying price for the gamble f, if he came to know that the variable X_I took the value x (and nothing else). Under the notation we gave above for lower previsions conditional on events and partitions, this would be $\underline{P}(f|B)$, where $B := \pi_I^{-1}(x)$. When there is no possible confusion about the variables involved in the lower prevision, we shall use the notation $\underline{P}(f|x)$ for $\underline{P}(f|X_I = x)$. The sets $\{\pi_I^{-1}(x) : x \in \mathcal{X}_I\}$ form a partition of Ω. Hence, we can define the gamble $\underline{P}(f|X_I)$, which takes the value $\underline{P}(f|x)$ on $x \in \mathcal{X}_I$. This is a conditional lower prevision.

The $\mathcal{X}_I$*-support* $S(f)$ of a gamble f in $\mathcal{K}_{O \cup I}$ is given by $S(f) := \{\pi_I^{-1}(x) : x \in \mathcal{X}_I, f\mathbb{I}_{\pi_I^{-1}(x)} \neq 0\}$, i.e., it is the set of conditioning events for which the restriction of f is not identically zero. Here, and in the rest of the paper, $\mathbb{I}_A$ will be used to denote the indicator function of the set A, i.e., the function whose value is 1 in the elements of A and 0 elsewhere. Also, for any gamble f in the domain $\mathcal{K}_{O \cup I}$ of the conditional lower prevision $\underline{P}(X_O|X_I)$, and any $x \in \mathcal{X}_I$, we shall denote by $G(f|x)$ the gamble $\mathbb{I}_{\pi_I^{-1}(x)}(f - \underline{P}(f|x))$, and by $G(f|X_I)$ the gamble that takes the value $G(f|\pi_I(y))$ in all $y \in \mathcal{X}^n$.

These assessments can be made for any disjoint subsets O, I of $\{1, \ldots, n\}$, and therefore it is not uncommon to model a subject's beliefs using a finite number of different conditional previsions. Then, we should verify that all the assessments modelled by these conditional previsions are coherent with each other. The first requirement we make is that for any disjoint $O, I \subseteq \{1, \ldots, n\}$, the conditional lower prevision $\underline{P}(X_O|X_I)$ defined on $\mathcal{K}_{O \cup I}$ should be *separately coherent*.[2] In this case, where the domain is a linear set of gambles, separate coherence holds if and

[2] We refer to [4] for more general definitions of the following notions in this section in terms of partitions, and for domains that are not necessarily (these) linear sets of gambles.

only if the following conditions are satisfied for any $x \in \mathcal{X}_I, f, g \in \mathcal{K}_{O\cup I}$, and $\lambda > 0$:

1. $\underline{P}(f|x) \geq \inf_{y\in\pi_I^{-1}(x)} f(y)$.
2. $\underline{P}(\lambda f|x) = \lambda \underline{P}(f|x)$.
3. $\underline{P}(f+g|x) \geq \underline{P}(f|x) + \underline{P}(g|x)$.

Separate coherence means on the one hand that, if a subject knows that the variable X_I has taken the value x, he cannot raise the (conditional) lower prevision of a gamble by considering the acceptable buying transactions that are implied by other gambles in the domain, and on the other hand that he should bet at any odds on the event that $X_I = x$ after having observed it. In general, separate coherence is not enough to guarantee the consistency of the lower previsions: conditional lower previsions can be conditional on the values of many different variables, and still we should verify that the assessments they provide are consistent not only separately, but also with each other. Formally, we are going to consider what we shall call *collections* of conditional lower previsions.

Definition 2. Consider a set of conditional lower previsions $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ with respective domains $\mathcal{K}^1, \ldots, \mathcal{K}^m \subseteq \mathcal{L}(\mathcal{X}^n)$, where $\mathcal{K}^i$ is the set of $\mathcal{X}_{O_i\cup I_i}$-measurable gambles,[3] for $i = 1, \ldots, m$. Then, this is called a *collection* on X^n when for each $i \neq j$ in $\{1, \ldots, m\}$, either $O_i \neq O_j$ or $I_i \neq I_j$.

This means that we do not have two different conditional lower previsions giving information about the same set of variables X_O, conditional on the same set of variables X_I.

Let $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ be a collection of conditional lower previsions, and let us see the different ways in which we can guarantee their consistency.

Definition 3. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are *weakly coherent* when for any $f_i \in \mathcal{K}^i$, $i = 1, \ldots, m$, $j \in \{1, \ldots, m\}$, $f_0 \in \mathcal{K}^j, z \in \mathcal{X}_{I_j}$,

$$\sup_{x\in\mathcal{X}^n} [\sum_{i=1}^{m} G_i(f_i|X_{I_i}) - G(f_0|z)](x) \geq 0.$$

Although this condition already assures that each of the conditional lower previsions is separately coherent, it does not prevent some inconsistencies from appearing: see [4, Example 7.3.5] for an example. This is the reason why we consider a stronger notion, called (*joint* or *strong*) coherence:

[3] We use $\mathcal{K}^i$ instead of $\mathcal{K}_{O_i\cup I_i}$ in order to alleviate the notation.

Definition 4. $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are *coherent* when for every $f_i \in \mathcal{K}^i$, $i = 1, \ldots, m$, $f_0 \in \mathcal{K}^j, z \in \mathcal{X}_{I_j}$, there exists some $B \in \{\pi_J^{-1}(z)\} \cup \bigcup_{i=1}^m S_i(f_i)$ such that

$$\sup_{x\in B} [\sum_{i=1}^{m} G_i(f_i|X_{I_i}) - G(f_0|z)](x) \geq 0,$$

where $S_i(f_i)$ is the $\mathcal{X}_{I_i}$-support of f_i.

In the next section, we prove a number of results that will help to understand better the differences between weak and strong coherence. But before we do that, we introduce a special case that will be of special interest for us: that of *conditional linear previsions*. We say that a conditional lower prevision $\underline{P}(X_O|X_I)$ on the set $\mathcal{K}_{O\cup I}$ is linear if and only if it is separately coherent and moreover $\underline{P}(f+g|x) = \underline{P}(f|x) + \underline{P}(g|x)$ for any $x \in \mathcal{X}_I$ and $f, g \in \mathcal{K}_{O\cup I}$. Conditional linear previsions correspond to the case where a subject's supremum acceptable buying price (lower prevision) coincides with his infimum acceptable selling price (or *upper prevision*) for any gamble on the domain. When a separately coherent conditional lower prevision $\underline{P}(X_O|X_I)$ is linear we shall denote it by $P(X_O|X_I)$.

If we consider a collection of conditional linear previsions $P_1(X_{O_1}|X_{I_1}), \ldots, P_m(X_{O_m}|X_{I_m})$ with domains $\mathcal{K}^1, \ldots, \mathcal{K}^m$, then they are coherent if and only if they *avoid partial loss*: for every $f_i \in \mathcal{K}^i$, $i = 1, \ldots, m$, there is some $B \in \bigcup_{i=1}^m S_i(f_i)$ such that

$$\sup_{x\in B} [\sum_{i=1}^{m} G_i(f_i|X_{I_i})](x) \geq 0,$$

where, again, $S_i(f_i) = \{\pi_{I_i}^{-1}(x) : x \in \mathcal{X}_{I_i}, f_i \mathbb{I}_{\pi_{I_i}^{-1}(x)} \neq 0\}$.

One interesting feature of conditional linear previsions allows to easily characterise separate coherence: a conditional lower prevision $\underline{P}(X_O|X_I)$ is separately coherent if and only if it is the lower envelope of a closed (in the *weak-* topology)* convex set of dominating conditional linear previsions, where $P(X_O|X_I)$ is said to *dominate* $\underline{P}(X_O|X_I)$ when for every $\mathcal{X}_{O\cup I}$-measurable gamble f, $P(f|x) \geq \underline{P}(f|x)$ for every $x \in \mathcal{X}_I$. Note, however, that in general a collection of coherent conditional lower previsions is not necessarily the lower envelope of a collection of coherent (i.e., avoiding partial loss) conditional linear previsions.

Finally, one interesting particular case is that where we are given only an unconditional lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ and a conditional lower prevision $\underline{P}(X_O|X_I)$ on $\mathcal{K}_{O\cup I}$. Then, weak and strong coherence are equivalent, and they both hold if and only if, for any $\mathcal{X}_{O\cup I}$-measurable f and any $x \in \mathcal{X}_I$,

(C1) $\underline{P}(G(f|X_I)) \geq 0$

(C2) $\underline{P}(G(f|x)) = 0$.

If both P and $P(X_O|X_I)$ are linear previsions, they are coherent if and only if for any $\mathcal{X}_{O\cup I}$-measurable f it holds that $P(f) = P(P(f|X_I))$.

3 Weak and strong coherence

The following theorem gives a new characterisation of the weak coherence of the conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$.

Theorem 1. *$\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are weakly coherent if and only if there is some coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ such that*

$$\begin{cases} \underline{P}(G_i(f|X_{I_i})) \geq 0 & \text{for any } f \text{ in } \mathcal{K}^i \\ \underline{P}(G_i(f|x)) = 0 & \text{for any } f \text{ in } \mathcal{K}^i, x \text{ in } \mathcal{X}_{I_i}. \end{cases}$$

Remark 1. When all the conditional previsions are linear, then weak coherence is equivalent to the existence of a *linear* prevision that is coherent with each of the conditionals: we can deduce from Theorem 1 and [4, Section 6.5.5] that any linear prevision P dominating $\underline{P}$ will satisfy $P(G_j(f|X_{I_j})) = 0$ for any $f \in \mathcal{K}^j$, and this implies that P is coherent with $P_j(X_{O_j}|X_{I_j})$.

When moreover all the spaces $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are finite, we deduce from Theorem 1 that the weak coherence of the conditional previsions $P_j(X_{O_j}|X_{I_j})$, $j = 1, \ldots, m$, is equivalent to the existence of a linear prevision (a finitely additive probability) on $\mathcal{X}^n$ inducing the conditional previsions by means of Bayes rule. This is not enough, however, for the conditional previsions to be coherent. For a counterexample, see [4, Example 7.3.5]. ♦

From this theorem, we can easily deduce the following two results, that relate (weak or strong) coherence to the existence of an unconditional lower prevision that is (weakly or strongly) coherent with the collection.

Proposition 1. *The conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are coherent if and only if there is some coherent unconditional lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ such that $\underline{P}$, $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are coherent.*

Corollary 1. *The conditional lower previsions $\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are weakly coherent if and only if there is some coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}^n)$ such that $\underline{P}, \underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})$ are weakly coherent.*

These results allow us to understand a bit better the conceptual difference between weak coherence and (strong) coherence: weak coherence amounts to the existence of a joint that is pairwise coherent with each of the conditional lower previsions; coherence means that there is a joint that is coherent with all the conditional lower previsions, *taken together*.

4 Collection templates

In this paper we are interested in proving coherence properties of lower previsions without assuming to be given information about the numbers that make up the lower previsions themselves, other than they produce separately coherent assessments. For this we need at least to focus on the 'form' of the lower previsions, which we call *template*.

Definition 5. Let $\underline{P}_{j'}(X_{O_{j'}}|X_{I_{j'}})$ and $\underline{P}_{j''}(X_{O_{j''}}|X_{I_{j''}})$ be two lower previsions on X^n. We say that they *have the same template* if $O_{j'} = O_{j''}$ and $I_{j'} = I_{j''}$. The class of all the lower previsions on X^n with the same template is just called *lower prevision template on X^n* (of the generic lower previsions in the class). We denote a lower prevision template in the same way as we denote a lower prevision (the distinction should be clear from the context): i.e., by $\underline{P}_j(X_{O_j}|X_{I_j})$.

Definition 6. Similarly, we say that *two collections of lower previsions on X^n have the same template* if they contain the same number m of lower previsions, and if it is possible to order the elements in each collection in such a way that for all j in $\{1, \ldots, m\}$ the two respective j-th lower previsions have the same template. The class of all the collections on X^n with the same template is just called *collection template on X^n* (of the generic collection in the class). We denote a collection template in the same way as we denote a collection of lower previsions (again, the distinction should be clear from the context): i.e., by $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$.

The notion of a collection template should be regarded a special kind of assessment about a collection of lower previsions, in the sense that knowing the template of a collection means knowing that the collection belongs to a certain set. An equivalent way to look at a collection template is as a collection of lower prevision templates. For this reason, we shall sometimes refer to the lower previsions templates in a collection template.

5 Coherence graphs

In this section, we introduce a graphical representation of collection templates based on directed graphs. For this, we start by recalling some terminology from graph theory.

A *directed graph* is a structure made up of a set of *nodes* and a set of *directed arcs* between nodes. Two nodes connected by an arc are also called its *endpoints*. A sequence of at least two nodes for which each pair of adjacent nodes is an arc in the graph, is called *directed path* between the first and the last node in the sequence (also called *origin* and *destination nodes*, respectively). When the origin and destination nodes coincide, and this is the only case of repeated node in the sequence, we say that the path is a *directed cycle*, or just a cycle, for short. Note that a path uniquely identifies a sequence of arcs; for this reason, by an abuse of terminology, we shall sometimes refer to the arcs of a path.

The *predecessors* of a node are all the nodes that have a directed path towards the given node. The predecessors for which there is a directed path made up of a single arc, are called *parents*. The *indegree* of a node is the number of its parents. A node with indegree equal to zero is called a *root*. Similarly, the *successors* of a node are all the nodes that can be reached from the given node following directed paths. The successors for which there is a directed path made up of a single arc, are called *children*. The *outdegree* of a node is the number of its children. A node with outdegree equal to zero is called a *leaf*.

The union of the set of parents and children of a node is called the set of its *neighbors*. The *union of two graphs* is a graph created by taking the union of their nodes and their arcs, respectively.

Now we are ready to define the most important graphical notion used in this paper.

Definition 7. Consider two finite sets $\mathcal{Z} = \{X_1, \ldots, X_n\}$ and $\mathcal{D} = \{D_1, \ldots, D_m\}$ of so-called *actual* and *dummy nodes*, respectively. Call $\mathcal{N} := \mathcal{Z} \cup \mathcal{D}$ the set of *nodes*, and a given $\mathcal{A} \subseteq \mathcal{N} \times \mathcal{N}$ the set of *arcs*. The triple $< \mathcal{Z}, \mathcal{D}, \mathcal{A} >$ is called a *coherence graph* on $\mathcal{Z}$ if the following properties hold:

(CG1) $\mathcal{Z}$ is non-empty.

(CG2) All neighbors of dummy nodes are actual nodes, and vice versa.

(CG3) The set of the parents and that of the children of any dummy node have an empty intersection.

(CG4) Dummy nodes are not leaves.

(CG5) Different dummy nodes do not have both the same parents and the same children.

Figure 1 used in the Introduction is just an example of a coherence graph, with actual nodes X_1, ..., X_{11}. Note that to make graphs easier to see, we represent dummy nodes in a simplified way: we do not show their labels and rather represent each of them simply as a black solid circle (this does not pose a problem since each dummy node is univocally identified by its neighbors); moreover, when a dummy node has exactly one parent and one child, we do not represent the arrow entering the dummy node (this is not going to cause ambiguity either).

Next, we show that there is a one-to-one relationship between coherence graphs on $\mathcal{Z} = \{X_1, \ldots, X_n\}$ and collection templates on X^n. To this extent, it is useful to isolate the notion of a D-structure in a coherence graph.

Definition 8. Given a dummy node D of a coherence graph, we call *D-structure* the subgraph whose nodes are D and its neighbors, and whose arcs are those connecting D to its neighbors.

In the graph of Figure 1 there are 11 D-structures, one per dummy node. For example, a D-structure is the subgraph made by the actual nodes X_9, X_{10}, X_{11}, by the dummy node in the middle, and by the arcs that connect them; another D-structure is the subgraph made by X_1, X_2, the dummy node in between, and the arc(s) connecting them.

At this point we consider two functions: the first one, that we shall denote by Γ, maps a collection template $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$, related to the variables $\{X_1, \ldots, X_n\} =: \mathcal{Z}$, into a coherence graph on $\mathcal{Z}$, with dummy nodes $\{D_1, \ldots, D_m\}$. This mapping is determined by the following procedure:

(Γ1) Let $\mathcal{Z} := \{X_1, \ldots, X_n\}$ be the set of actual nodes.

(Γ2) Let $\mathcal{D} := \{D_1, \ldots, D_m\}$ be the set of dummy nodes.

(Γ3) Let $\mathcal{A} := \emptyset$.

(Γ4) For all $j \in \{1, \ldots, m\}$, all $i' \in I_j$, all $i'' \in O_j$, add the arcs $(X_{i'}, D_j)$ and $(D_j, X_{i''})$ to $\mathcal{A}$.

The second function, that we denote by Γ^{-1}, maps a coherence graph on $\mathcal{Z} = \{X_1, \ldots, X_n\}$, with dummy nodes $\{D_1, \ldots, D_m\}$, into the collection template $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$, related to the variables $\{X_1, \ldots, X_n\}$. This mapping is determined by the following procedure:

(Γ^{-1}1) Set the collection of lower prevision templates equal to the empty set.

(Γ^{-1}2) For all $j \in \{1, \ldots, m\}$, add $\underline{P}_j(X_{O_j}|X_{I_j})$ to the collection template, where O_j and I_j are the set of indexes of the children and the parents of D_j, respectively.

The idea behind the two functions is very simple: identifying lower prevision templates in a collection with D-structures in the related coherence graph, and vice versa.

Consider the graph of Figure 1 once again. By applying Function Γ^{-1} we, unsurprisingly, obtain the collection of lower prevision templates used as starting example in the Introduction:

$\{\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_3|X_2), \underline{P}_4(X_4|X_3),$
$\underline{P}_5(X_5, X_6|X_1), \underline{P}_6(X_2|X_3), \underline{P}_7(X_7|X_4), \underline{P}_8(X_8|X_5),$
$\underline{P}_9(X_8|X_6), \underline{P}_{10}(X_9, X_{10}|X_6, X_7), \underline{P}_{11}(X_{11}|X_9, X_{10})\}.$

It is easy then to see that Function Γ gives back the original graph once it is applied to such a collection template. The reason is that the two functions turn out to be each other's inverses. This is shown by the next theorem, which also allows us to prove the wanted one-to-one relationship between coherence graphs and collection templates.

Theorem 2. *There is one-to-one relationship between coherence graphs and collection templates.*

Next, we introduce some graph-based terminology that is more directly relevant to our subsequent results.

Definition 9. We say that an actual node of a coherence graph is a (potential) *source of contradiction* (or *conflict*) if it has more than one parent or if it belongs to a cycle.

Definition 10. A coherence graph without sources of contradiction is said to be of type *A1*: i.e., acyclic and with maximum indegree for actual nodes equal to one. The corresponding collection template is said to be *representable as a graph of type A1*, or simply *A1-representable*.

The graph in Figure 2 is clearly not of type A1, as there are three sources of contradiction: X_8, given its two parents; X_2, because it has two parents and also because it is part of a cycle; and X_3, because it is in such a cycle, too.

Definition 11. Given a source of contradiction Z, call *block for* Z, or B_Z, the subgraph obtained by taking the union of the D-structures of the dummy nodes that are predecessors of Z.

Definition 12. Call *superblock* of a coherence graph, any union of all the blocks that share at least one actual node.

Figure 2 displays the only two different blocks of the coherence graph under consideration: the block for X_8 and that for X_3 (note that the latter coincides with the block for X_2). Those blocks have the node X_1 in common (besides its dummy parent); their union is thus a superblock, which is also the only one in the graph.

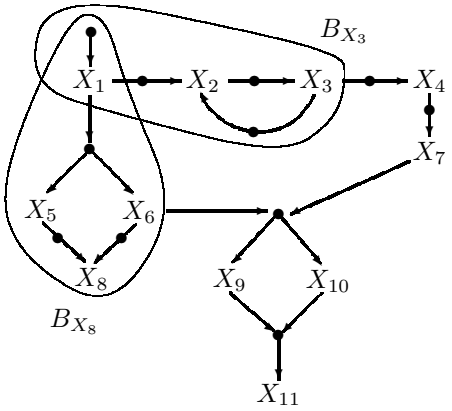


Figure 2: The areas delimited by closed lines contain two blocks of the coherence graph: B_{X_8} and B_{X_3}. Their union is a superblock.

Observe that there can be many configurations of blocks in a superblock: a superblock can be made up of a single block; if it is made up of more than one block, it may be the case that some blocks coincide (as B_{X_2} and B_{X_3} in Figure 2), that one of them is included in another, or that two of them share only some nodes (as B_{X_3} and B_{X_8} in the same figure).

We use the notion of superblock in order to build a partition of the dummy nodes.

Definition 13. Call *minimal partition of the dummy nodes* in a coherence graph, the partition whose elements are the sets of dummy nodes in each superblock, and the singletons made up of the remaining dummy nodes. The corresponding partition of $\{1, \ldots, m\}$ is denoted by $\mathcal{B}$ and is simply called the *minimal partition*.

Note that $\mathcal{B}$ refers also to a partition of the related collection template, given the one-to-one correspondence between dummy nodes and lower prevision templates. With respect to the graph in Figure 2, we obtain the following partition of the related collection template:

$\{\{\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_3|X_2), \underline{P}_5(X_5, X_6|X_1),$
$\underline{P}_6(X_2|X_3), \underline{P}_8(X_8|X_5), \underline{P}_9(X_8|X_6)\},$
$\{\underline{P}_4(X_4|X_3)\}, \{\underline{P}_7(X_7|X_4)\}, \{\underline{P}_{10}(X_9, X_{10}|X_6, X_7)\},$
$\{\underline{P}_{11}(X_{11}|X_9, X_{10})\}\}.$

Moreover, note that for A1-representable collection templates, the minimal partition is entirely made up of singletons, because their coherence graph has no sources of contradiction.

6 Coherence graphs as tools to prove coherence

The following theorem gives us conditions under which the coherence of some subsets of a collection of conditional lower previsions implies the coherence of all the elements in the collection. It shows that it is sufficient that the conditional lower previsions whose indices belong to the same element in $\mathcal{B}$ are coherent.

Theorem 3. *Consider a collection* $\{\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})\}$ *of separately coherent conditional lower previsions with known templates. Then, if for any* $B \in \mathcal{B}, \{\underline{P}_j(X_{O_j}|X_{I_j})\}_{j\in B}$ *are coherent, then* $\{\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})\}$ *are coherent.*

The intuition behind the proof of the theorem is the following. We exploit the properties of the coherence graph to create a total order on a set of coherent lower previsions strongly related to $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$. That order allows us to use the generalisation of the *Marginal Extension Theorem* (MET, in short) established in [3] to show that the lower previsions in that set are coherent, and from this to derive the coherence of $\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})$.

It is easy to see a similar result holds when we work with weak coherence instead of coherence:

Theorem 4. *Consider a collection* $\{\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})\}$ *of separately coherent conditional lower previsions with known templates. Then, if for any* $B \in \mathcal{B}, \{\underline{P}_j(X_{O_j}|X_{I_j})\}_{j\in B}$ *are weakly coherent, then* $\{\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})\}$ *are weakly coherent.*

Next, we investigate in which sense the partition $\mathcal{B}$ given by Definition 13 is minimal. For this, we should like to know if there are other partitions of $\{1,\ldots,m\}$ that we can use for the same end, meaning that the coherence of the conditional lower previsions within each of the elements of the partition guarantees the coherence of the collection template.

A first positive result in this regard is that the partition $\mathcal{B}$ is indeed minimal when we are studying the problem for weak coherence:

Proposition 2. *Let* $\mathcal{B}'$ *be a partition of* $\{1,\ldots,m\}$*, and assume that, for any* B' *in* $\mathcal{B}'$*,* $\{\underline{P}_j(X_{O_j}|X_{I_j})\}_{j\in B'}$ *are weakly coherent. Then, this implies the weak coherence of* $\{\underline{P}_1(X_{O_1}|X_{I_1}),\ldots,\underline{P}_m(X_{O_m}|X_{I_m})\}$ *if and only if* $\mathcal{B}$ *is finer than* $\mathcal{B}'$*.*

The sufficiency part in this proposition is actually Theorem 4, which can be proven in a similar way as Theorem 3. The idea for the necessity part is to show that, when the necessary condition fails, we can create conditional *linear* previsions $P_1(X_{O_1}|X_{I_1}),\ldots,P_m(X_{O_m}|X_{I_m})$ that are not weakly coherent and yet for all B' in $\mathcal{B}'$, $\{P_j(X_{O_j}|X_{I_j})\}_{j\in B'}$ are weakly coherent.

A basic step in the construction of such lower previsions is to prove that for any given $j \in \{1,\ldots,m\}$ and any $x \in \mathcal{X}_{O_j}$, we can define weakly coherent conditional previsions $P_1(X_{O_1}|X_{I_1}),\ldots,P_m(X_{O_m}|X_{I_m})$ such that any compatible joint P satisfies $P(\pi_{O_j}^{-1}(x)) = 1$. Even stronger, we can show that any compatible joint with *some* of these conditional previsions satisfies $P(\pi_{O_j}^{-1}(x)) = 1$. This is proven using the following lemmas:[4]

Lemma 1. *For any* $i = 1,\ldots,n$*, let us consider* $x_1^i, x_2^i \in \mathcal{X}_i$*. Define the conditional previsions* $P_1(X_{O_1}|X_{I_1}),\ldots,P_m(X_{O_m}|X_{I_m})$ *with respective domains* $\mathcal{K}^1,\ldots,\mathcal{K}^m$ *by*[5]

$$P_j(f|y) := \begin{cases} f((x_1^i)_{i\in O_j}, y) & \text{if } y = (x_1^i)_{i\in I_j} \\ f((x_2^i)_{i\in O_j}, y) & \text{otherwise,} \end{cases}$$

for any $j = 1,\ldots,m$*,* $y \in \mathcal{X}_{I_j}$ *and* $f \in \mathcal{K}^j$*. Then,* $P_1(X_{O_1}|X_{I_1}),\ldots,P_m(X_{O_m}|X_{I_m})$ *are coherent.*

Lemma 2. *For any* $i = 1,\ldots,n$*, let us consider* $x_1^i, x_2^i \in \mathcal{X}_i$*. Define the conditional previsions* $P_1(X_{O_1}|X_{I_1}),\ldots,P_{m-1}(X_{O_{m-1}}|X_{I_{m-1}})$ *with respective domains* $\mathcal{K}^1,\ldots,\mathcal{K}^{m-1}$ *by*

$$P_j(f|y) := \begin{cases} f((x_1^i)_{i\in O_j}, y) & \text{if } y = (x_1^i)_{i\in I_j} \\ f((x_2^i)_{i\in O_j}, y) & \text{otherwise,} \end{cases}$$

for any $y \in \mathcal{X}_{I_j}, f \in \mathcal{K}^j$*, and define* $P_m(X_{O_m}|X_{I_m})$ *by*

$$P_m(f|y) := f((x_2^i)_{i\in O_m}, y)$$

for any $y \in \mathcal{X}_{I_m}$ *and* $f \in \mathcal{K}^m$*. Then,* $P_1(X_{O_1}|X_{I_1}),\ldots,P_m(X_{O_m}|X_{I_m})$ *are weakly coherent.*

However, a similar result to Proposition 2 does not apply for coherence, due, among other things, to the fact that the previsions in Lemma 2 are weakly coherent but not coherent. As a consequence, there exist instances of collection templates where the coherence within the elements of a partition which is not coarser than $\mathcal{B}$ guarantees the coherence of all of them. One such case is given in the following example.

[4] Although the previsions in these lemmas correspond to 0-1 valued probabilities, this is not essential for the developments made in the proof of the theorem; it is possible to obtain similar results using probabilities that are not 0-1 valued.

[5] We are using there the one-to-one correspondence between gambles on $\mathcal{X}^j$ and gambles in $\mathcal{K}^j$.

Example 1. Consider the collection template $\{\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_2, X_3|X_1)\}$. Then, the minimal partition $\mathcal{B}$ associated to its coherence graph is $\{1, 2, 3\}$. However, we can deduce the coherence of the collection template using a smaller subset. For this, we must prove first that the coherence of $\underline{P}_2(X_2|X_1), \underline{P}_3(X_2, X_3|X_1)$ holds if and only if for any $\mathcal{X}_1 \times \mathcal{X}_2$-measurable gamble f and for any $x_1 \in \mathcal{X}_1$,

$$\underline{P}_2(f|x_1) = \underline{P}_3(f|x_1).$$

Using this property, we deduce that, when $\underline{P}_2(X_2|X_1), \underline{P}_3(X_2, X_3|X_1)$ are coherent, then $\{\underline{P}_1(X_1), \underline{P}_2(X_2|X_1), \underline{P}_3(X_2, X_3|X_1)\}$ are coherent if and only if $\underline{P}_1(X_1), \underline{P}_3(X_2, X_3|X_1)$ are. But since $\underline{P}_1(X_1), \underline{P}_3(X_2, X_3|X_1)$ are always coherent because of the marginal extension theorem in [4, Theorem 6.7.2], we deduce that the coherence of $\underline{P}_2(X_2|X_1), \underline{P}_3(X_2, X_3|X_1)$ implies the coherence of the collection template. ♦

It remains an open problem at this stage to determine a minimal partition with the property that the coherence within each of the elements of the partition guarantees the coherence of the collection template, and that is minimal in the sense that it is finer than any other partition with the same property.

In this respect, we can deduce from Theorem 3 that the separate coherence of the conditional lower previsions $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ implies their joint coherence when their associated coherence graph is of type A1. Using Lemma 1, we can prove that being of type A1 is also necessary for this property.

Proposition 3. *Consider a collection* $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ *of separately coherent conditional lower previsions with known templates. Then the separate coherence of* $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ *implies their coherence if and only if their coherence graph is of type A1.*

Note on the other hand that, with respect to weak coherence, we also have a necessary and sufficient condition for the separate coherence to imply the weak coherence, because of Proposition 2:

Corollary 2. *Consider a collection* $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ *of separately coherent conditional lower previsions with known templates. Then the separate coherence of* $\{\underline{P}_1(X_{O_1}|X_{I_1}), \ldots, \underline{P}_m(X_{O_m}|X_{I_m})\}$ *implies their weak coherence if and only if their coherence graph is of type A1.*

We should like to conclude this section remarking that if the collection template is A1, then we can give the following Bayesian sensitivity analysis interpretation:

Theorem 5. *Consider a collection of separately coherent conditional lower previsions. If their coherence graph is A1, then these lower previsions are lower envelopes of a family of coherent linear previsions.*

The interest of this result lies in the fact that the lower envelopes of a family of coherent conditional linear previsions are coherent conditional lower previsions, but the converse does not hold in general: there exist instances of coherent conditional lower previsions that are not even dominated by any family of coherent conditional linear previsions. A sufficient condition for the converse to hold is that the spaces $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are finite. This theorem shows that, if the coherence graph is A1, then the coherent conditional lower previsions are also lower envelopes of coherent conditional linear previsions, no matter the cardinality of the spaces.

7 Discussion

Coherence can be regarded as the very essence of a theory of personal probability. But working directly with coherence can be particularly onerous.

This paper is an attempt to deal with this difficulty, and to deliver tools that make checking coherence easier. We have been inspired in this by the lesson of graphical models, and have indeed defined a new graphical model called a coherence graph.

Coherence graphs are means to render explicit the structure behind the notion of coherence. We have shown that such a structure induces a partition of the available collection of lower previsions, with the characteristic that the coherence within each set of the partition implies the coherence of the overall collection.

This result is very general: it holds for lower previsions and for any cardinality of the possibility spaces involved. In particular, since it holds for lower previsions, it is also applicable to determine the coherence of a collection of conditional *linear* previsions, and therefore is also useful in the precise context.

More generally speaking, we expect the results in this paper to have substantial theoretical as well as practical consequences, whenever the focus is on the task of proving coherence. They already appear to shed light on specific aspects of coherence, thanks especially to coherence graphs of type A1. These graphs correspond to collections of separately coherent lower previsions that are coherent irrespective of the numerical values that make them up.

Remember that we have shown that there are important conceptual differences between the notions of weak and strong coherence proposed by Walley. Weak coherence is equivalent to the existence of a joint lower prevision that is coherent with each of the assessments. In the particular case of conditional linear previsions and finite spaces, this is equivalent to the existence of a joint mass function inducing each of the conditionals by means of Bayes rule. The introduction of the notion of strong coherence is needed because some conditional lower previsions can have a common joint and still be clearly incoherent with one another. Remarkably, this happens even in the linear and finite case mentioned above.

Taking this into account, we find it noteworthy that, for the problem tackled here, weak and strong coherence exhibit a similar behaviour: if we have a number of assessments and all we know about them is that each of them is separately coherent, we can guarantee that they are weakly coherent exactly under the same conditions for which we can deduce their joint coherence: we just need the graph representing the collection template to be A1. More generally, we have established a partition of the graph for which weak coherence inside implies weak coherence of them all, and we have proven that strong coherence inside this partition also implies the strong coherence of all the assessments. It is worth pointing out that there are also differences: we have shown that the minimal partition obtained using a coherence graph is indeed minimal in the case of weak coherence and not necessarily so for strong coherence.

Another point worth emphasising is the connection, used repeatedly in the proofs of this paper, between the A1 condition and the generalisation of the MET established in [3]: the relationship arises as from the A1 condition we can establish a total order on the conditional lower previsions in our collection template, and such an order is just what allows us to use the generalised MET. In this way, we have also given an easy graphical characterisation of the extent to which the theorem can be applied: to A1-representable collection templates.

Finally, we have proven that if the separate coherence of the lower previsions in a collection template implies their joint coherence (that is, if the associated coherence graph is A1), then the conditional lower previsions in the template are lower envelopes of coherent linear previsions. This does not hold for all collections of coherent conditional lower previsions, as is shown in [4, Section 6.6]. So it is remarkable that our results lead naturally to a Bayesian sensitivity analysis interpretation of the collection of conditional lower previsions.

As a topic for future research, we should like to mention the study of the coherence of collection templates when we have some additional structural assessments, such as considerations of irrelevance or independence.

Acknowledgements

We are grateful to Gert de Cooman for encouraging us to study the problems presented in this paper, and for many helpful comments. We acknowledge financial support by the MCYT projects MTM2004-01269, TSI2004-06801-C04-01, and by the Swiss NSF grants 200020-109295/1, 200021-113820/1.

References

[1] V. Biazzo, A. Gilio, T. Lukasiewicz, and G. Sanfilippo. Probabilistic logic under coherence: complexity and algorithms. *Annals of Mathematics and Artificial Intelligence*, 45(1–2):35–81, 2005.

[2] B. Jaumard, H. Hansen, and Poggi de Arag ao. Column generation methods for probabilistic logic. *ORSA Journal on Computing*, 3:135–148, 1991.

[3] E. Miranda and G. de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 2006. Accepted for publication.

[4] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[5] P. Walley, R. Pelessoni, and P. Vicig. Direct algorithms for checking consistecy and making inferences for conditional probability assessments. *Journal of Statistical Planning and Inference*, 126:119–151, 2004.

Scoring Rules, Entropy, and Imprecise Probabilities

Robert Nau, Victor Richmond Jose, and Robert Winkler
Fuqua School of Business, Duke University
robert.nau@duke.edu

Abstract

Suppose that a risk-averse expected utility maximizer with a precise probability distribution **p** bets optimally against a risk neutral opponent (or equivalently invests in an incomplete market for contingent claims) whose beliefs (or prices) are described by a convex set $\mathcal{Q}$ of probability distributions. This utility-maximization problem is the dual of the problem of finding the distribution **q** in $\mathcal{Q}$ that minimizes a generalized divergence (relative entropy) with respect to **p**. A special case is that of logarithmic utility, in which the corresponding divergence is the Kullback-Leibler divergence, but we present a closed-form solution for the entire family of linear-risk-tolerance (a.k.a. HARA) utility functions and show that this corresponds to a particular parametric family of generalized divergences, which is derived from an entropy measure originally proposed by Arimoto and which is also related to a generalization of pseudospherical scoring rule originally proposed by I.J. Good. A variant of this decision problem, in which the decision maker has quasilinear utility for consumption over two periods, leads to the family of power divergences, which is related to a generalization of the power family of scoring rules.

Keywords. entropy, divergence, scoring rules, portfolio optimization, incomplete markets

1 Introduction

There are many applications in which it is of interest to measure the difference between a precise probability distribution **p** and another precise probability **q**, or between a precise probability and the nearest or farthest point in some set of imprecise probabilities, in terms of the gain or loss that a decision maker experiences as a result of that difference. For example, **q** might be a prior probability distribution over some set of events, which is later updated to a posterior distribution **p** based on new information, and the magnitude of the difference between **p** and **q** might determine the quantity or value of that information for purposes of signal transmission or decision making. Or, **p** might be the precise probability of a decision maker who has the opportunity to bet or trade against an opponent whose beliefs are described by a precise probability **q** or by a set of imprecise probabilities $\mathcal{Q}$, in which case the decision maker can obtain a greater expected payoff or expected utility the farther that **p** is from **q** or from the nearest point in $\mathcal{Q}$. Or, the decision maker's probability **p** might itself be imprecise, known only to lie within some set $\mathcal{P}$, and it might be of interest to find the distribution that is nearest to the center of $\mathcal{P}$ in the sense of minimizing the maximum loss that the decision maker could suffer by acting upon the wrong probability.

The considerable literature on this topic includes (at least) three distinct but intertwined strands: scoring rules, entropy, and decision analysis. Scoring rules are reward functions for eliciting and evaluating probability forecasts, and the expected score associated with a forecast can be interpreted as a measure of the value of the forecaster's information. Entropy is a measure of the channel capacity required to communicate a stream of signals generated by a stationary process, and relative entropy measures the reduction in channel capacity that is possible when new information yields an updated signal distribution. Decision analysis provides a general framework for measuring information in terms of gains in expected utility, as well for determining how to optimally use information to choose portfolios of financial assets.

These information-theoretic tools have been used for many decades, but new applications and theoretical developments have emerged during the last few years on several fronts, including experimental economics, robust Bayesian statistics, and financial engineering. The objective of this paper is to add to this recent stream of interdisciplinary literature by broadening the concept of a scoring rule to include a not-

necessarily-uniform baseline distribution and to show that this leads immediately to tight connections with some well-known measures of divergence (relative entropy) as well as with models of utility maximization in markets under uncertainty. In the setting where some probabilities are imprecise, we focus on the problem in which $\mathbf{p}$ is outside the set $\mathcal{Q}$ and the quantity of interest is the divergence between $\mathbf{p}$ and its nearest neighbor in $\mathcal{Q}$. More details and proofs of the main results are given in Jose et al. (2007)

2 Scoring rules

Scoring rules are reward functions for eliciting and evaluating probabilities, and they have played an important role in the foundations of subjective probability theory (de Finetti 1937 & 1974, Good 1952, Winkler 1967 & 1996, Savage 1971, Lindley 1982) as well as practical applications such as incentive schemes for paying weather forecasters (Brier 1950) and subjects in economic experiments (Selten 1998) and for evaluating the quality of forecasts used in risk analysis (Cooke 1991). Consider an individual (the "forecaster") who is asked to assess a probability distribution over a set of n mutually exclusive and collectively exhaustive events. Let $\mathbf{p}$ denote the forecaster's true distribution, let $\mathbf{r} = (r_1, ..., r_n)$ denote her reported distribution (if different from $\mathbf{p}$), and let $\mathbf{e}_i$ denote the probability distribution that assigns probability 1 to event i and zero to all other events, i.e., the indicator vector for event i. A scoring rule is conventionally expressed as a function $S(\mathbf{r}, \mathbf{p})$, linear in its second argument, such that the score obtained if event i occurs is $S(\mathbf{r}, \mathbf{e}_i)$, and the forecaster's *expected* score for reporting $\mathbf{r}$ when her true distribution is $\mathbf{p}$ is $S(\mathbf{r}, \mathbf{p}) = \sum_i p_i S(\mathbf{r}, \mathbf{e}_i)$. It is assumed that the forecaster's objective is to maximize her expected score, which means that either she is risk neutral and $S(\mathbf{r}, \mathbf{e}_i)$ is measured in units of money or else she is non-risk neutral and $S(\mathbf{r}, \mathbf{e}_i)$ is measured in units of utility.

The scoring rule is defined to be [*strictly*] *proper* if it encourages honest reporting in the sense that $S(\mathbf{p}, \mathbf{p}) \geq S(\mathbf{r}, \mathbf{p})$ for every $\mathbf{r}$ and $\mathbf{p}$ [with equality only when $\mathbf{r} = \mathbf{p}$], so that the forecaster whose true distribution is $\mathbf{p}$ maximizes her expected score by truthfully reporting $\mathbf{p}$ rather than some other distribution. The forecaster's optimal expected score that is obtained when her distribution is $\mathbf{p}$ will be denoted by merely suppressing the first argument: $S(\mathbf{p}) \equiv S(\mathbf{p}, \mathbf{p})$. A proper scoring rule has a canonical representation in terms of its optimal-expected-score function, as noted by McCarthy (1956) and further elaborated by Hendrickson and Buehler (1971) and Savage (1971). In particular, if $S(.)$ is a differentiable function, then $S(\cdot, \cdot)$ is uniquely determined by the formula

$$S(\mathbf{r}, \mathbf{p}) = S(\mathbf{r}) + \nabla S(\mathbf{r}) \cdot (\mathbf{p} - \mathbf{r}). \qquad (1)$$

where $\nabla S(\mathbf{r})$ denotes the gradient of $S(\cdot)$ evaluated at $\mathbf{r}$, and conversely every function S that is [strictly] convex and differentiable uniquely defines a [strictly] proper scoring rule. Written in this form, the expected score yielded by a proper scoring rule is seen to be closely related to a particular measure of divergence between probability distributions that is known as a *Brègman divergence* (Brègman 1967), a connection that has been discussed by Grünwald and Dawid (2004), Dawid (2006), and Gneiting and Raftery (2007). Any strictly convex function F defines a Brègman divergence $B_F(\mathbf{p} \| \mathbf{r})$ as follows:

$$B_F(\mathbf{p} \| \mathbf{r}) = F(\mathbf{p}) - F(\mathbf{r}) - \nabla F(\mathbf{r}) \cdot (\mathbf{p} - \mathbf{r}).$$

Letting $F(\mathbf{p}) = S(\mathbf{p})$, it follows that for any strictly proper scoring rule, the function $S(\mathbf{p}) - S(\mathbf{r}, \mathbf{p})$, which represents the forecaster's expected *loss* for reporting $\mathbf{r}$ when her true distribution is $\mathbf{p}$, is a Brègman divergence, and vice versa. Thus, there is a one-to-one correspondence between strictly proper scoring rules and Brègman divergences.

The literature of scoring rules has mainly focused on a few strictly proper rules with particularly convenient parametric forms, axiomatic representations, and/or geometrical interpretations, namely the *quadratic*, *logarithmic*, and *spherical* scoring rules. The quadratic rule (a.k.a. "Brier score") is $S(\mathbf{p}, \mathbf{e}_i) = -\left(\|\mathbf{e}_i - \mathbf{p}\|_2\right)^2$. Thus, under the quadratic rule, the forecast $\mathbf{p}$ is treated as an estimate of the indicator vector of the uncertain event $\mathbf{e}_i$, and the forecaster is ultimately penalized in proportion to the squared Euclidean distance between $\mathbf{p}$ and the realized value of $\mathbf{e}_i$, in the tradition of least squares estimation. The logarithmic scoring rule is $S(\mathbf{p}, \mathbf{e}_i) = \ln(p_i)$, whose optimal expected score function is the negative entropy of the forecaster's true distribution, an issue to which we return below. (Some prescient comments on the potential connection between scoring rules and entropy were made by Good (1971).) The spherical scoring rule is $S(\mathbf{p}, \mathbf{e}_i) = p_i / \|\mathbf{p}\|_2$, and it is obtained by letting the set of feasible score vectors be the simplest strictly convex object in $\mathbb{R}^n$, namely the unit sphere.

The quadratic and spherical rules can be generalized into parametric families by replacing the 2-norm with the vector β-norm, $\|\mathbf{p}\|_\beta \equiv \left(\sum_{j=1}^n p_j^\beta\right)^{1/\beta}$. The generalized spherical rule is the *pseudospherical scoring rule*, $p_i / (\|\mathbf{p}\|_\beta)^{\beta-1}$, which was first proposed by Good (1971). The generalized quadratic rule is the *power scoring rule*, $\beta p_i^{\beta-1} - (\beta - 1)\left(\|\mathbf{p}\|_\beta\right)^\beta$. Written in this

conventional fashion, these families of rules are well-defined and proper only for $\beta > 1$ and the corresponding optimal-expected-score functions that generate them via McCarthy's formula are simply $\left(\|\mathbf{p}\|_\beta\right)^\beta$ and $\|\mathbf{p}\|_\beta$, respectively. The logarithmic scoring rule is the limiting case of both the pseudospherical and power scores as $\beta \longrightarrow 1$, but otherwise the two families do not intersect.

3 Weighted score rules and divergence measures

A key property of the aforementioned scoring rules is that they treat events *symmetrically* in the sense that if $p_i = [>]\ p_j$, then the score in event i is equal to [greater than] the score in event j, regardless of the descriptions of the events, and the forecaster's expected score is smallest when $\mathbf{p}$ is the uniform distribution. Thus, they implicitly reward the forecaster in proportion to some measure of the difference of $\mathbf{p}$ from a uniform distribution. However, in most real (and even hypothetical) applications, the relevant reference point is not a uniform distribution. For example, in weather forecasting the events that are of interest are often known to have widely varying a priori probabilities, and "baseline" values for those probabilities, upon which the forecaster is supposed to improve, are obtainable from historical records (Winkler 1994) or alternative forecasting models. In predicting the outcomes of sporting events or movements of financial markets, there are public betting lines or posted prices for contingent claims that implicitly assign probabilities to events. Therefore, we propose that scoring rules should be generalized so as to reward the forecaster in proportion to some measure of the difference between $\mathbf{p}$ and an appropriate baseline distribution $\mathbf{q}$. Such a scoring rule will be henceforth referred to as *weighted* scoring rule; it will be expressed as a function of three arguments, $S(\mathbf{r}, \mathbf{p}\|\mathbf{q})$, and its associated optimal expected score will be expressed as a function of two arguments, $S(\mathbf{p}\|\mathbf{q})$.

There are various functional forms through which the dependence of the score on the baseline distribution could be modeled, and the one we that we find most compelling, for both practical and theoretical reasons, is that for fixed $\mathbf{p}$ and $\mathbf{q}$ the score in state i should depend on the ratio p_i/q_i, so that if $p_i/q_i = [>]\ p_j/q_j$, then the score in event i should be equal to [greater than] the score in event j. One simple rationale for this desideratum is that when bets may be placed on outcomes of events, *relative* rather than absolute differences in probabilities are what matter, insofar as a \$1 bet on state i has an expected payoff of $\$p_i/q_i$ when the bettor's probability is p_i and the posted odds are based on q_i. Another rationale can be illustrated by a simple example: suppose that the state space consists of 4 states formed by the Cartesian product of two binary events E and F, and suppose it happens that the forecaster and client both agree on the probability of F and they also agree that E and F are probabilistically independent. Then it seems reasonable that the forecaster's payment should depend only on the outcome of E, not F, and this requires the payoff in each of the four states to depend only on the ratio of p to q, which is the relative change in the evaluation of the probability of E.

The measurement of differences between probabilities in terms of ratios has a long history in statistics and information theory. It was noted above that under a strictly proper scoring rule, the forecaster's expected *loss* for reporting a distribution $\mathbf{r}$ that is other than her true distribution $\mathbf{p}$ is a particular kind of divergence between $\mathbf{r}$ and $\mathbf{p}$, namely a Brègman divergence. Under a weighted strictly proper scoring rule that bases the score on the ratio p_i/q_i the forecaster's expected *gain* for possessing a distribution $\mathbf{p}$ that differs from $\mathbf{q}$ is a second kind of divergence, which is not a Brègman divergence. Rather, it turns out to be a special case (or a simple transformation) of another kind of generalized divergence known as an f-divergence (Csiszár 1967). If f is a strictly convex function, the corresponding f-divergence is defined as

$$D_f(\mathbf{p}\|\mathbf{q}) = E_{\mathbf{p}}[f(\mathbf{p}/\mathbf{q})]. \tag{2}$$

Divergences of this general form have been widely used in statistics for many years as "utility-free" measures of the value of the information - e.g., Goel (1983) uses f-divergence to define a "conditional amount of sample information" for measuring prior-to-posterior information gains in Bayesian hierarchical models. More recently it has been recognized that f-divergences are interpretable as measures of expected utility gains that are available to decision makers who have opportunities to bet against less-well-informed opponents or to invest in financial markets, as will be more fully discussed in later sections of this paper.

When the ratio p_i/q_i is substituted for p_i in the pseudospherical and power scoring rules, and they are affinely transformed so as to yield scores of zero when $\mathbf{p} = \mathbf{q}$, we obtain the *weighted pseudospherical score*, denoted $S^{\mathbf{S}}_\beta$, and the *weighted power score*, denoted

by $S_\beta^{\mathbf{P}}$, with the following parametric forms:

$$S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_i \| \mathbf{q}) \equiv \frac{1}{\beta - 1}\left(\left(\frac{p_i/q_i}{\left(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]\right)^{1/\beta}}\right)^{\beta-1} - 1\right), \quad (3)$$

$$S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_i \| \mathbf{q}) \equiv \frac{(p_i/q_i)^{\beta-1} - 1}{\beta - 1} - \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}] - 1}{\beta}. \quad (4)$$

Note that for any fixed values of $\mathbf{p}$, $\mathbf{q}$, and β, the pseudospherical score vector $(S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_1 \| \mathbf{q}), ..., S_\beta^{\mathbf{S}}(\mathbf{p}, \mathbf{e}_n \| \mathbf{q}))$ is a positive affine transformation of the power score vector $(S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_1 \| \mathbf{q}), ..., S_\beta^{\mathbf{P}}(\mathbf{p}, \mathbf{e}_n \| \mathbf{q}))$, since both vectors are affine transformations of $(\mathbf{p}/\mathbf{q})^{\beta-1}$, although the origins and scale factors of the transformations vary with $\mathbf{p}$, $\mathbf{q}$, and β. Thus, although the two rules yield different expected payoffs as a function of $\mathbf{p}$ (for the same $\mathbf{q}$ and β), and they create different incentives for information-gathering and different penalties for dishonest reporting, they nevertheless present the same relative risk profile to a truthful forecaster whose $\mathbf{p}$ is already fixed. At $\beta = 1$ both rules converge to the weighted logarithmic score $\ln(p_i/q_i)$. At $\beta = 2$, weighted forms of the quadratic and spherical scoring rules are obtained. The cases $\beta = 0$ and $\beta = \frac{1}{2}$ have not received much (if any) attention in the antecedent literature, but it will be shown later that $\beta = 0$ corresponds to a decision model involving exponential utility, which is the utility function most commonly used in applied decision analysis, while $\beta = \frac{1}{2}$ arises from a decision model involving reciprocal utility, which has some appealing symmetry properties and is closely related to the Hellinger distance between $\mathbf{p}$ and $\mathbf{q}$. These special cases will be further explored in the next two sections.

The corresponding optimal-expected-score functions for the two families of weighted scoring rules are:

$$S_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q}) = \frac{\left(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]\right)^{1/\beta} - 1}{\beta - 1}, \quad (5)$$

$$S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}] - 1}{\beta(\beta - 1)}, \quad (6)$$

and one is a monotonically increasing function of the other for any fixed β. Our first result is to point out that these expected score functions correspond exactly to two parametric families of generalized *divergence* (cross-entropy) between probability distributions. In particular the weighted power expected score $S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q})$ is precisely the *directed divergence of order β between $\mathbf{p}$ and $\mathbf{q}$* proposed by Havrda and Chavrát (1967), variants of which have been discussed by Rathie and Kannappan (1972), Cressie and Read (1984), and Haussler and Opper (1997). Cressie and Read refer to this quantity as the *power divergence*, which we shall also do here.

The weighted pseudospherical score $S_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$ is the cross-entropy measure that arises from a generalized entropy introduced by Arimoto (1971) and further elaborated by Sharma and Mittal (1975), Boekee and Van der Lubbe (1980) and Lavenda and Dunning-Davies (2003). Arimoto's generalized entropy of order β is defined for $\beta > 0$ by $\beta/(\beta - 1)\left(E_{\mathbf{p}}[\mathbf{p}^{\beta-1}]^{1/\beta} - 1\right)$. The factor of β in the numerator plays no essential role when β is restricted to be positive, and without it the measure is actually valid for all real β, and when $\mathbf{p}^{\beta-1}$ is replaced by $(\mathbf{p}/\mathbf{q})^{\beta-1}$ so as to define a cross-entropy, the weighted pseudospherical expected score is obtained. It is therefore appropriate to refer to the latter quantity as the *pseudospherical divergence of order β* between $\mathbf{p}$ and $\mathbf{q}$. Both of these generalized divergences reduce to the Kullback-Leibler divergence $\mathbf{E}_{\mathbf{p}}[\ln(\mathbf{p}/\mathbf{q})]$ at $\beta = 1$, and for other special cases of β they are related to two other well known divergences, namely the Chi-square divergence $\chi^2(\mathbf{q}\|\mathbf{p}) = E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}] - 1$ and the Hellinger distance $D_H(\mathbf{p}\|\mathbf{q}) \equiv \left(\sum_{j=1}^n \left(\sqrt{p_j} - \sqrt{q_j}\right)^2\right)^{1/2}$ as shown in the following table:

Table 1. Power & pseudospherical divergences

β	$S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q})$	$S_\beta^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$
-1	$\frac{1}{2}\chi^2(\mathbf{q}\|\mathbf{p})$	$\frac{1}{2}\left(1 - (\chi^2(\mathbf{q}\|\mathbf{p}) + 1)^{-1}\right)$
0	$D_{KL}(\mathbf{q}\|\mathbf{p})$	$1 - \exp(-D_{KL}(\mathbf{q}\|\mathbf{p}))$
$\frac{1}{2}$	$2D_H(\mathbf{p}\|\mathbf{q})^2$	$2\left(1 - \left(1 - \frac{1}{2}D_H(\mathbf{p}\|\mathbf{q})^2\right)^2\right)$
1	$D_{KL}(\mathbf{p}\|\mathbf{q})$	$D_{KL}(\mathbf{p}\|\mathbf{q})$
2	$\frac{1}{2}\chi^2(\mathbf{p}\|\mathbf{q})$	$\sqrt{\chi^2(\mathbf{p}\|\mathbf{q}) + 1} - 1$

Note that the power divergence is symmetric around $\beta = \frac{1}{2}$ in the sense that $S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) = S_{1-\beta}^{\mathbf{P}}(\mathbf{q}\|\mathbf{p})$, i.e., the roles of $\mathbf{p}$ and $\mathbf{q}$ are merely reversed when β is replaced by $1 - \beta$.

4 Decision models and information measures

Our second result is to show that the same two families of generalized divergence arise naturally as the solutions of two canonical expected-utility maximization problems, involving the most widely-used parametric family of utility functions, in which a risk averse decision maker with subjective probability dis-

tribution $\mathbf{p}$ bets against a non-strategic risk-neutral opponent with distribution $\mathbf{q}$, or equivalently, invests in a contingent claims market where prices are determined by taking expectations with respect to $\mathbf{q}$. A contingent claim is a claim to monetary payments that are contingent on states of the world, and it can be represented as an n-vector of payoffs $\mathbf{y}$ that has some market price $p(\mathbf{y})$ at which it can be purchased in arbitrary positive multiples. (In a financial market, the relevant states of the world might be possible values of a stock price or stock index on a particular future date, and a contingent claim might be a share of stock or an option to buy a share of that stock at a pre-specified strike price.). A decision maker who buys α units of $\mathbf{y}$ at its market price receives a net payoff of $\alpha(y_i - p(\mathbf{y}))$ in state i, hence the vector $\alpha(\mathbf{y} - p(\mathbf{y})\mathbf{1})$ is a feasible net payoff vector for the decision maker for all positive α. The market is *complete* if every contingent claim has a unique price at which it can be both bought *and* sold, in which case $\alpha(\mathbf{y} - p(\mathbf{y})\mathbf{1})$ is a feasible payoff vector for all real α, positive or negative. If the market prices are also *arbitrage-free* ("coherent"), then there exists a unique probability distribution $\mathbf{q}$ that prices all contingent claims according to their expected payoffs, so that $p(\mathbf{y}) = E_{\mathbf{q}}[\mathbf{y}]$ for all $\mathbf{y} \in \mathbb{R}^n$, and any $\mathbf{x} \in \mathbb{R}^n$ that satisfies $E_{\mathbf{q}}[\mathbf{x}] = 0$ is a feasible net payoff vector. In Bayesian theory this existence result is known as de Finetti's "fundamental theorem of probability," with $p(\mathbf{y})$ referred to as the "prevision" of $\mathbf{y}$, and in finance theory it is known as the "fundamental theorem of asset pricing," with $\mathbf{q}$ referred to as a "risk neutral distribution" because assets are priced "as if" by a risk neutral opponent whose probability distribution is $\mathbf{q}$.

In the first canonical problem (**"S"**), there is a single time period in which consumption occurs and the decision maker has a single-attribute vNM utility function $u(x)$. The decision maker's optimal expected utility, denoted $U^{\mathbf{S}}(\mathbf{p}\|\mathbf{q})$, is determined by:

$$\begin{aligned} &\textbf{Problem S} \quad : \qquad (7)\\ &U^{\mathbf{S}}(\mathbf{p}\|\mathbf{q}) \quad \equiv \quad \max_{\mathbf{x}\in\mathbf{R}^n} \; E_{\mathbf{p}}[u(\mathbf{x})] \;\; \text{s.t. } E_{\mathbf{q}}[\mathbf{x}] = 0, \end{aligned}$$

where $u(\mathbf{x}) \equiv (u(x_1), ..., u(x_n))$ denotes the vector of utilities that u yields when applied to $\mathbf{x}$. In the second problem (**"P"**), there are two periods in which consumption occurs and the decision maker with probability distribution $\mathbf{p}$ has a quasilinear vNM utility function $u(a, b) = a + u(b)$ where a is money consumed at time 0 and b is money consumed at time 1. The decision maker's objective is to choose a vector $\mathbf{x}$ of time-1 payoffs to be purchased from time-0 funds at market prices so as to maximize the expected utility of consumption in both periods. The time-0 cost of purchasing $\mathbf{x}$ is $E_{\mathbf{q}}[\mathbf{x}]$, so the optimal expected utility, denoted $U^{\mathbf{P}}(\mathbf{p}\|\mathbf{q})$, is the solution of:

$$\begin{aligned} &\textbf{Problem P} \quad : \qquad (8)\\ &U^{\mathbf{P}}(\mathbf{p}\|\mathbf{q}) \quad \equiv \quad \max_{\mathbf{x}\in\mathbf{R}^n} \; E_{\mathbf{p}}[u(\mathbf{x})] - E_{\mathbf{q}}[\mathbf{x}]. \end{aligned}$$

Next, let u be a utility function from the general exponential/logarithmic/power family, which will be parameterized here as:

$$u_\beta(x) \equiv \frac{1}{\beta - 1}((1 + \beta x)^{(\beta-1)/\beta} - 1) \qquad (9)$$

for $\beta x > -1$. This parameterization has two key properties. First, $u_\beta(0) = 0$ and $u'_\beta(0) = 1$, so that for any β the marginal rate of substitution between time-0 consumption and time-1 consumption is unity at $x = 0$ for the decision maker in Problem **P**. Second, the corresponding *risk tolerance function* $\tau_\beta(x)$, which is the reciprocal of the Pratt-Arrow risk aversion measure, is the following linear function of wealth: $\tau_\beta(x) \equiv -u'_\beta(x)/u''_\beta(x) = 1 + \beta x$. Thus, the risk tolerance as well as the marginal utility is normalized to a value of 1 at $x = 0$, and β is the coefficient of risk tolerance, i.e., the increase in risk tolerance per unit of increase in wealth. The linear-risk-tolerance utility functions are also known as hyperbolic-absolute-risk-aversion (HARA) utility functions in the literature of financial economics, although parameterizing them in terms of their risk tolerance coefficients rather than their risk aversion coefficients is more useful for our purposes. Some important special cases of u_β are given in Table 2:

Table 2. Linear-risk-tolerance utility functions

β	$u_\beta(x)$	*Functional form*
-1	$u_{-1}(x) = -\frac{1}{2}((1-x)^2 - 1)$	Quadratic
0	$u_0(x) = 1 - \exp(-x)$	Exponential
$\frac{1}{2}$	$u_{1/2}(x) = 2\left(1 - \frac{1}{1+x/2}\right)$	Reciprocal
1	$u_1(x) = \ln(1 + x)$	Logarithmic
2	$u_2(x) = \sqrt{1 + 2x} - 1$	Square-root

The utility functions $\{u_\beta\}$ also exhibit a symmetry around $\beta = \frac{1}{2}$, namely that $u_{1-\beta}(x) = -u_\beta(-x)$, or equivalently $u_\beta(-u_{1-\beta}(-x)) = x$. In other words, the graph of $u_{1-\beta}$ is obtained from the graph of u_β by merely reflecting it around the line $y = -x$. Note that the power (exponent) in u_β is the term $(\beta-1)/\beta$, which has the property that $((\beta-1)/\beta)^{-1} = ((1-\beta)-1)/(1-\beta)$, so that swapping β for $1-\beta$ results in another power utility function whose power is the reciprocal of the original. Thus, up to affine scaling, the reciprocal utility function ($\beta = \frac{1}{2}$) is self-symmetric, the exponential and logarithmic utility

functions ($\beta = 0$ and $\beta = 1$) are symmetric to each other, and the power utility function with exponent δ is symmetric to the power utility function with exponent $1/\delta$ for any positive or negative δ other than 0 or 1.

Let $\mathbf{x}^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q})$ and $\mathbf{x}^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q})$ denote the solutions of Problems **S** and **P** when $u = u_\beta$, with i^{th} elements $x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\|\mathbf{q})$ and $x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\|\mathbf{q})$, respectively, and let $U^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q})$ and $U^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q})$ denote their corresponding expected utilities. In these terms, we have:

THEOREM 1:

(a) $S^{\mathbf{S}}_{\beta}(\mathbf{p}, \mathbf{e}_i\|\mathbf{q}) = u_\beta(x^{\mathbf{S}}_{\beta,i}(\mathbf{p}\|\mathbf{q}))$,

and $\quad S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q}) = U^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q})$

(b) $S^{\mathbf{P}}_{\beta}(\mathbf{p}, \mathbf{e}_i\|\mathbf{q}) = u_\beta(x^{\mathbf{P}}_{\beta,i}(\mathbf{p}\|\mathbf{q})) - E_{\mathbf{q}}[\mathbf{x}^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q})]$,

and $\quad S^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q}) = U^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q})$

(c) $S^{\mathbf{P}}_{\beta}(\mathbf{p}\|\mathbf{q}) \geq S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{q})$ for all $\mathbf{p}$, $\mathbf{q}$, and β.

Thus, the statewise utility gains to the decision maker under problems **S** and **P** are precisely the pseudospherical and power scores for the same $\mathbf{p}$, $\mathbf{q}$, and β, and the expected utilities are the corresponding divergences.

5 Utility/entropy duality in incomplete markets

We now extend the preceding results to a setting in which the decision maker's risk neutral betting opponent has imprecise probabilities, which is equivalent to an incomplete market where a contingent claim may have a "bid-ask spread" rather than a single price at which it can be both bought and sold. The bid-ask spreads generally do not suffice to determine a unique risk neutral distribution; rather, they only determine a convex set $\mathcal{Q}$ of risk-neutral distributions such that $\mathbf{x}$ is a feasible net payoff vector for the decision maker if and only if $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ for all $\mathbf{q} \in \mathcal{Q}$. (The payoff to the opponent is $-\mathbf{x}$, hence the constraint $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ for all $\mathbf{q} \in \mathcal{Q}$ means that the opponent with imprecise probabilities $\mathcal{Q}$ will accept only those bets yielding non-negative expected payoffs for all $\mathbf{q} \in \mathcal{Q}$.) The problem of expected-utility maximization in incomplete markets has been widely studied in the mathematical finance literature in recent years, and it has been shown that there is a duality relationship between maximization of expected utility and minimization of an appropriate measure of relative entropy or divergence (e.g., Frittelli 2000, Rouge and El Karoui 2000, Goll and Rüschendorf 2001, Delbaen et al. 2002, Slomczyński and Zastawniak 2004, Ilhan et al. 2004, Samperi 2005). Most of this literature has focused on the case of exponential utility, for which the dual problem is the minimization of the reverse KL divergence $D_{KL}(\mathbf{q}, \mathbf{p})$, as well as on issues that arise in multi-period or continuous-time markets. In this section we will show that in a single-period or two-period market, the duality relationship applies to the entire spectrum of linear-risk-tolerance utility and pseudospherical divergence or power divergence.

An incomplete, single-period market can be parameterized in either of two ways. One is in terms of an $m \times n$ matrix $\mathbf{A}$ whose rows are feasible net payoff vectors, i.e., $\mathbf{A} = \{a_{ij}\}$ where a_{ij} is the net payoff that the decision maker receives in the j^{th} state of the world for purchasing one unit of the i^{th}contingent claim at its asking price. (It suffices to consider only purchases at asking prices, rather than sales at bid prices, since a bid price of p for a contingent claim $\mathbf{y}$ is equivalent to an asking price of $-p$ for $-\mathbf{y}$.) Alternatively, the market can be parameterized in terms of a $k \times n$ matrix $\mathbf{Q}$ whose rows are risk neutral probability distributions that support the contingent claim prices, i.e., $\mathbf{Q} = \{q_{ij}\}$ where q_{ij} is the probability of the j^{th} state of the world under the i^{th} risk neutral distribution. The rows of $\mathbf{Q}$ are the extremal risk-neutral probability distributions assigning non-positive expectation to all the rows of $\mathbf{A}$, i.e., the rows of $-\mathbf{Q}$ are the dual cone of the rows of $\mathbf{A}$. The second parameterization will be adopted here, in terms of which $\mathcal{Q}$ is the convex hull of the rows of $\mathbf{Q}$, so that a generic element of $\mathcal{Q}$ can be expressed as $\mathbf{q} = \mathbf{z}^T\mathbf{Q}$ where $\mathbf{z}$ is an element of $\triangle^k$, the unit simplex in $\mathbb{R}^k$, and the feasibility requirement that $E_{\mathbf{q}}[\mathbf{x}] \leq 0$ for all $\mathbf{q} \in \mathcal{Q}$ can be expressed as $\mathbf{Qx} \leq \mathbf{0}$.

In the incomplete-market generalization of Problem **S**, the problem of finding the maximum expected utility, which will be denoted as $U^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{Q})$, is dual to the problem of finding the minimum pseudospherical divergence of order β between $\mathbf{p}$ and all $\mathbf{q}$ in the convex hull of the rows of $\mathbf{Q}$, which will be denoted as $S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{Q})$:

$$\begin{aligned} &\textbf{Primal Problem S} \quad : \\ &U^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{Q}) \quad \equiv \quad \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})],\ \mathbf{Qx} \leq \mathbf{0} \\ &\textbf{Dual Problem S} \quad : \\ &S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{Q}) \quad \equiv \quad \min_{\mathbf{z}\in\triangle^k} S^{\mathbf{S}}_{\beta}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}). \end{aligned}$$

In the incomplete-market generalization of Problem **P**, the decision maker's objective is to determine an amount w to be spent at time 0 to finance consumption in period 1. For the period-1 payoff vector $\mathbf{x}$ that the decision maker wishes to purchase, the risk-neutral expected value of $\mathbf{x}$ needs to be less than or equal to w for all the extremal risk neutral distributions. The corresponding primal and dual problems

are:

$$\begin{aligned}
&\textbf{Primal Problem P} \quad : \\
&\qquad U_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{Q}) \quad \equiv \quad \max_{\mathbf{x}\in\mathbb{R}^n} E_{\mathbf{p}}[u_\beta(\mathbf{x})] - w, \ \mathbf{Qx} \le w\mathbf{1} \\
&\textbf{Dual Problem P} \quad : \\
&\qquad S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{Q}) \quad \equiv \quad \min_{\mathbf{z}\in\triangle^k} S_\beta^{\mathbf{P}}(\mathbf{p}\|\mathbf{z}^T\mathbf{Q}).
\end{aligned}$$

The special case $\beta = 1$ corresponds to logarithmic utility in the primal problem and KL divergence in the dual problem, while $\beta = 0$ corresponds to exponential utility in the primal problem and reverse KL divergence in the dual problem, and the cases $\beta = 1/2$ and $\beta = \pm 2$ are related to the squared Hellinger distance and the Chi-square divergence as shown in Table 1. These duality relationships are formalized in:

THEOREM 2:
(a) In an incomplete, single-period market, maximization of expected linear-risk-tolerance utility with risk tolerance coefficient β (Primal Problem **S**) is equivalent to minimization of the pseudospherical divergence of order β between the decision maker's subjective distribution $\mathbf{p}$ and a risk neutral distribution $\mathbf{q}$ consistent with contingent claim prices (Dual Problem **S**). Their optimal objective values are the same and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other.
(b) In an incomplete, two-period market, maximization of quasi-linear expected linear-risk-tolerance utility with second-period risk tolerance coefficient β (Primal Problem **P**) is equivalent to minimization of the power divergence of order β between the decision maker's subjective distribution $\mathbf{p}$ and a risk neutral distribution $\mathbf{q}$ consistent with contingent claim prices (Dual Problem **P**). Their optimal objective values are the same and the optimal values of the decision variables in one problem are equal to the normalized optimal values of the Lagrange multipliers in the other.

Note that because the pseudospherical divergence is a monotonic transformation of the power divergence, the distribution $\mathbf{q}$ ($= \mathbf{z}^T\mathbf{Q}$) that solves Dual Problem **S** is the same one that solves Dual Problem **P**, although the objective values and the primal payoff vectors are generally different. The geometry of the dual solutions is illustrated in Figure 1.

Grünwald and Dawid (2004) have explored duality relationships among strictly proper scoring rules, generalized entropies and divergences, and expected-utility-maximization (or in their terms, expected-loss-minimization) in the context of robust Bayesian inference, where the decision maker does not know the true probability distribution and her opponent is "Nature" who chooses the true distribution $\mathbf{p}$ from some

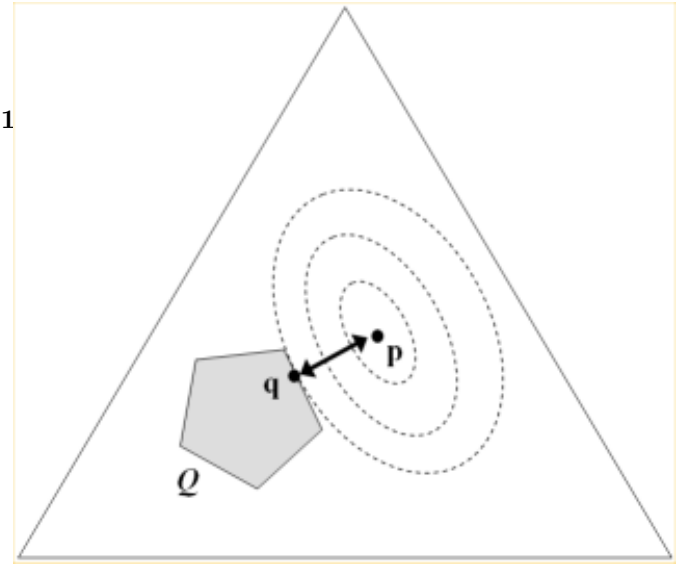


Figure 1: Geometry of minimizing the divergence between $\mathbf{p}$ and the nearest element of $\mathcal{Q}$ ($n = 3$)

convex set $\mathcal{P}$, such as the set of distributions satisfying a mean-value constraint. The robust Bayes problem for the decision maker is to determine the distribution $\mathbf{r}$ that minimizes her maximum expected loss over all $\mathbf{p} \in \mathcal{P}$, where the expected loss (in our terms) is the negative expected score $-S(\mathbf{r}, \mathbf{p})$. Grünwald and Dawid show that the optimal-expected-loss function, $-S(\mathbf{p})$, is interpretable as a generalized entropy, and minimizing the maximum expected loss is equivalent to maximizing this entropy on the set $\mathcal{P}$. This scoring-rule entropy uniquely determines a corresponding Brègman divergence $B_S(\mathbf{p}\|\mathbf{r}) \equiv S(\mathbf{p}) - S(\mathbf{r}, \mathbf{p})$, as noted earlier, and Grünwald and Dawid go on to show that the distribution $\mathbf{r}$ that minimizes the maximum expected loss on $\mathcal{P}$ is also the distribution that minimizes this divergence with respect to an uninformative "reference" distribution $\mathbf{p}_0$ at which the entropy $-S(\mathbf{p})$ is maximized. For typical symmetric scoring rules, the reference distribution is the uniform distribution, but any scoring rule entropy can be transformed so as to shift the reference point to any other distribution $\mathbf{p}_0^*$ by the addition of a linear function of $\mathbf{p}$, namely $S(\mathbf{p}_0^*, \mathbf{p})$. The reference distribution $\mathbf{p}_0$ in their model therefore plays an analogous role to the baseline distribution $\mathbf{q}$ in our model, insofar as $-S(\mathbf{p}\|\mathbf{r})$ is maximized in the uninformative case where $\mathbf{p} = \mathbf{q}$. Grünwald and Dawid also discuss scoring rules for continuous probability distributions drawn from the generalized exponential family, focusing in particular on the logarithmic and quadratic scoring rules.

6 Summary and Conclusions

We have shown that when a risk averse decision maker with a precise probability distribution $\mathbf{p}$ bets against a risk neutral opponent with a convex set $\mathcal{Q}$ of imprecise probabilities, or equivalently invests in an incomplete market for contingent claims where $\mathcal{Q}$ is the set of risk neutral distributions determined by market prices, there is a natural duality between maximizing LRT utility and minimizing pseudospherical or power divergence with the same value of β. In particular, maximization of logarithmic utility ($\beta = 1$) corresponds to finding the distribution $\mathbf{q}$ in $\mathcal{Q}$ that minimizes the KL divergence $D_{KL}(\mathbf{p}\|\mathbf{q})$, maximization of exponential utility ($\beta = 0$) corresponds to minimizing the reverse KL divergence $D_{KL}(\mathbf{q}\|\mathbf{p})$, and maximization of reciprocal utility ($\beta = \frac{1}{2}$) or square-root utility ($\beta = 2$) correspond to minimization of the Hellinger distance $D_H(\mathbf{p}\|\mathbf{q})$ or the Chi-square divergence $\chi^2(\mathbf{p}\|\mathbf{q})$, respectively.

References

[1] Arimoto, S. 1971. Information-theoretical considerations on estimation problems. *Infor. Contr.* **19**:181-194.

[2] Boekee, D. E., J. C. A. van der Lubbe. 1980. The R-norm information measure. *Infor. Contr.* **45**:136-155.

[3] Brègman, L. 1967. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. and Math. Phys.*, **7**:200-217.

[4] Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**(1):1-3.

[5] Cooke, R. 1991. Experts in uncertainty: Opnion and subjective probability in science. Oxford University Press, Oxford.

[6] Cressie, N., T. R. C. Read. 1984. Multinomial goodness of fit. *J. Royal Stat. Soc. B* **46**(3):440-464.

[7] I. Csiszár. 1967 Information type measures of differences of probability distribution and indirect observations. Studia Math. Hungarica **2**:299-318

[8] Dawid, A.P. 2006. The geometry of proper scoring rules. Research Report No. 268, Department of Statistical Science, University College, London. (April 2006)

[9] De Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Ann. Inst. Henri Poincaré* 7, 1—68. Translation reprinted in H.E. Kyburg and H.E. Smokler, Eds. (1980), *Studies in Subjective Probability*, 2nd ed., Robert Krieger, New York, 53-118.

[10] De Finetti, B. 1974. *Theory of probability, Vol. 1.* Wiley, New York.

[11] Delbaen, F., P. Grandits, T. Rheinländer, D. Samperi, M. Schweizer, C. Stricker. 2002. Exponential hedging and entropy penalties. *Math. Finance.* **12:**99-123.

[12] Frittelli, M. 2000. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Math. Finance.* **10:**39-52.

[13] Gneiting, T., Raftery, A. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**:359-378

[14] Goel, P. 1983. Information measures and Bayesian hierarchical models. *J. Am. Stat. Assoc.* **78**:408-410.

[15] Goll, T., Rüschendorf, L. 2001. Minimax and minimal distance martingale measures and their relationship to portfolio optimization. Finance Stoch. **5**:557-581

[16] Good, I.J. 1952 Rational decisions. *J. Royal Statist. Soc. B.* **14**: 107-114.

[17] Good, I.J. 1971. Comment on paper by Buehler. In *Foundations of Statistical Inference* (Godambe and Sprott, eds., Holt, Reinhart, & Winston, Toronto) 337-339.

[18] Grünwald, P.D., A.P. Dawid. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Ann. Statist.* **32**: 1367-1433.

[19] Havrda, J., F. Chavrát. 1967. Quantification method of classification processes: the concept of structural α-entropy. *Kybernetika* **3**:30-35.

[20] Haussler, D., M. Opper. 1997. Mutual information, metric entropy, and cumulative relative entropy risk. *Ann. Statist.* **25**: 2451-2492.

[21] Hendrickson, A. D., R. J. Buehler 1971. Proper scores for probability forecasters. *Ann. Math. Stat.* **42**:1916-1921.

[22] Ilhan, A., M. Jonsson, R. Sircar. 2004. Portfolio optimization with derivatives and indifference pricing. Working Paper, Princeton University.

[23] Jose, V., R. Nau, R. Winkler. 2007. Scoring Rules, Entropy, and Utility Maximization. Working Paper, Duke Univerity

[24] Lavenda, B. H., J. Dunning-Davies. 2003. Qualms concerning Tsallis's condition of pseudo-additivity as a definition of non-extensivity. http://arxiv.org/abs/cond-mat/0312132.

[25] Lindley, D. 1982. Scoring rules and the inevitability of probability. *Int. Stat. Rev.* **50**(1):1-26.

[26] McCarthy, J. 1956. Measures of the value of information. *Proc. Nat. Acad. Sci. USA* **42**:654-655.

[27] Rathie, P. N., P. Kannappan. 1972. A directed-divergence function of type β. *Infor. Contr.* **20**:38-45.

[28] Rouge, R., N. El Karoui. 2000. Pricing via utility maximization and entropy. *Math. Finance.* **10:**259-276.

[29] Samperi, D. 2005. Model selection using entropy and geometry: complements to the six-author paper. Working paper, Decision Synergy.

[30] Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *J. Am. Stat. Assoc.* **66**:783-801.

[31] Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Exper. Econ.* **1**:43-62.

[32] Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* **27**:379-423.

[33] Sharma, B. D., D. P. Mittal. 1975. New non-additive measures of entropy for discrete probability distributions *J. Math. Sci.* **10**:28-40.

[34] Slomczyński, W., T Zastawniak. 2004. Utility maximizing entropy and the second law of thermodynamics. *Ann. Prob.* **32**:2261-2285.

[35] Winkler, R. L. 1967. The quantification of judgment: some methodological suggestions. *J. Am. Stat. Assoc.* **62**:1105-1120.

[36] Winkler, R. L. 1994. Evaluating probabilities: asymmetric scoring rules. *Manage. Sci.* **40**(11):1395-1405.

[37] Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1):1-60.

Imprecise probability methods for sensitivity analysis in engineering

Michael Oberguggenberger
Unit for Engineering Mathematics
University of Innsbruck, Austria
michael@mat1.uibk.ac.at

Julian King
Department of Mathematics
University of Innsbruck, Austria
csae2209@uibk.ac.at

Bernhard Schmelzer
Department of Mathematics
University of Innsbruck, Austria
csae1209@uibk.ac.at

Abstract

This article addresses questions of sensitivity of output values in engineering models with respect to variations in the input parameters. Such an analysis is an important ingredient in the assessment of the safety and reliability of structures. A major challenge in engineering applications lies in the fact that high computational costs have to be faced. Methods have to be developed that admit assertions about the sensitivity of the output with as few computations as possible. This article serves to explore various techniques from imprecise probability that may contribute to achieving this goal.

Keywords. Reliability of structures, sensitivity analysis, random sets, fuzzy sets, simulation methods, aerospace engineering.

1 Introduction

The goal of this article is to demonstrate how various methods from imprecise probability theory can be employed in sensitivity analysis of engineering structures. We are motivated by a research project in aerospace engineering[1] which involves the determination of the buckling load of the frontskirt of the ARIANE 5 launcher under various loading and flight scenarios. The frontskirt is a reinforced light weight shell structure. The computation of the decisive parameter indicating failure, the load proportionality factor (LPF), is based on a finite element model[2]. Part of the project is to determine the most influential input parameters (loads, material constants, geometry) on the load proportionality factor in a sensitivity analysis. The goal is to evaluate the design and to assess the safety of the structure. The calculation of the output variable LPF – under a given single set of input parameters – takes about 32 hours on a high performance computer. In addition to the extremely high computational cost, the LPF may depend in a non-differentiable manner on some of the input parameters, especially variations in the geometry. A classical sensitivity analysis of the complete structure is currently out of reach.

Engineering information on the variability of the input parameters usually consists of a central value and a coefficient or range of variation. The basic strategy for arriving at a sensitivity assessment will be to successively freeze the input parameters and study the effect on the variability of the output. We wish to do this without artificial parametric assumptions and with as few calls of the finite element program as possible. We will explore the usability of methods from imprecise probability theory for this purpose. In particular, we shall model the input variability by means of

– random sets and Tchebycheff's inequality;

– fuzzy sets and Hartley-like measures;

– intervals and sampling from a Cauchy distribution;

– standard Monte-Carlo simulation and resampling.

A detailed description of the respective methods will follow in four sections, with a final section devoted to a comparison of the methods. The question of modelling correlations between the input variables will be addressed in the appropriate sections. We shall exemplify the results with the aid of a simplified finite element model simulating part of a space craft launcher (Figure 1). The computational cost for the simplified model is one hour per call of the program.

In the sensitivity analysis, up to 17 input parameters were taken into account. A tentative description of the meaning of the parameters as well as their nominal values can be read off from Table 1.

[1]ICONA-project, Intales GmbH Engineering Solutions and University of Innsbruck, supported by TransIT Innsbruck and by EADS Astrium ST.

[2]The load proportionality factor is defined as the limiting value in an incremental procedure, in which the dynamic loads during a flight scenario are increased stepwise until breakdown of the structure is reached.

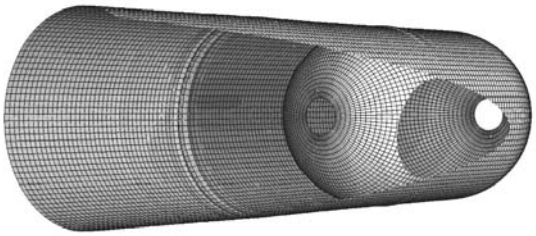


Figure 1: Simplified finite element model.

For background material on sensitivity analysis we refer to the Special Issue [9], in particular the survey article [10] and to [8], for random sets, to [18, 19], for random and fuzzy sets, to [5, 14], for probability boxes, to [3], for a review on probabilistic treatment of uncertainty in structural engineering as well as information on variability of typical input parameters, to [24].

i	Parameter X_i	Mean μ_i
1	Initial temperature	293 K
2	Step1 thermal loading cylinder1	450 K
3	Step1 thermal loading cylinder2	350 K
4	Step1 thermal loading cylinder3	150 K
5	Step1 thermal loading sphere1	150 K
6	Step1 thermal loading sphere2	110 K
7	Step2 hydrostatic pressure cylinder3	0.4 MPa
8	Step2 hydrostatic pressure sphere1	0.4 MPa
9	Step2 hydrostatic pressure sphere2	0.4 MPa
10	Step3 aerodynamic pressure	−0.05 MPa
11	Step4 booster loads y-direction node1	40000 N
12	Step4 booster loads y-direction node2	20000 N
13	Step4 booster loads z-direction node1	$3.e6$ N
14	Step4 booster loads z-direction node2	$1.e6$ N
15	Step4 mechanical loads x-direction	100 N
16	Step4 mechanical loads y-direction	50 N
17	Step4 mechanical loads z-direction	300 N

Table 1: Description of input parameters no. 1 – 17.

2 Random set methods

It has been argued in [20, 21] that random intervals constructed by Tchebycheff's inequality can serve as a non-parametric model of the variability of a parameter, given its mean value and variance as sole information. We begin with the univariate case of a real-valued random variable X. Let $\mu = \mathrm{E}(X)$ be its expectation and $\sigma^2 = \mathrm{V}(X)$ be its variance. Tchebycheff's inequality asserts that

$$P\big(|X-\mu| > d_\alpha\big) \le \alpha, \qquad d_\alpha = \sigma/\sqrt{\alpha}. \tag{1}$$

Equipping the unit interval $(0,1]$ with the uniform probability distribution, the non-parametric confidence intervals

$$I_\alpha = [\mu - d_\alpha, \mu + d_\alpha], \quad \alpha \in (0,1] \tag{2}$$

define a random set. By construction, the following formulas for the belief in the set I_α and the plausibility of its complement I_α^{c} hold:

$$\begin{aligned}\underline{P}(I_\alpha) &= \textstyle\int_{\{\beta\in(0,1]:I_\beta\subset I_\alpha\}} \mathrm{d}\beta = 1-\alpha \le P(I_\alpha),\\ \overline{P}(I_\alpha^{\mathsf{c}}) &= \textstyle\int_{\{\beta\in(0,1]:I_\beta\cap I_\alpha^{\mathsf{c}}\neq\emptyset\}} \mathrm{d}\beta = \alpha \ge P(I_\alpha^{\mathsf{c}}).\end{aligned}$$

This shows that the random set description provides a conservative assessment of the variability X. In applications, the range of the parameter X may be confined to a compact interval $[x_{\min}, x_{\max}]$. In this case, the random set will be truncated to

$$I_\alpha = [(\mu - d_\alpha) \vee x_{\min}, (\mu + d_\alpha) \wedge x_{\max}].$$

In the multivariate case $X = (X_1, \ldots, X_d)$ where each parameter X_i is modelled as a random set as in (2), we form the joint random set (assuming random set independence)

$$\alpha = (\alpha_1, \ldots, \alpha_d) \to A_\alpha = I^1_{\alpha_1} \times \ldots \times I^d_{\alpha_d}$$

again with the uniform distribution on the probability space $(0,1]^d$.

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. If the input variables $X = (X_1, \ldots, X_d)$ are modelled as a random set A_α, $\alpha \in (0,1]^d$ (equipped with the uniform probability distribution), the output variable is given by the random set $g(A_\alpha)$, $\alpha \in (0,1]^d$. A visualization of the output can be obtained by means of the upper and lower distribution functions (or *probability box*, [3])

$$\begin{aligned}\overline{F}(x) &= P\big(\alpha : g(A_\alpha) \cap (-\infty, x] \neq \emptyset\big)\\ \underline{F}(x) &= P\big(\alpha : g(A_\alpha) \subset (-\infty, x]\big).\end{aligned} \tag{3}$$

In the numerical evaluation, the joint random set is approximated by a finite random set with focal elements

$$I^1_{\alpha_1} \times \ldots \times I^d_{\alpha_d}, \qquad \alpha_j \in \{\tfrac{1}{n}, \tfrac{2}{n}, \ldots, 1\},$$

each with probability weight n^{-d}. The input-output function is evaluated as follows: First, an interval $Q \subset \mathbb{R}^d$ is determined that bounds the relevant range of the input variables X. Next, the values of the function g are computed at the m^d nodes of a uniform grid on Q. The output $g(Q)$ is approximated by a response surface $\hat{g}(Q)$ obtained by multilinear splines. More precisely, to compute the image of one of the sets A_α, $\hat{g}(Q)$ is evaluated at all grid points inside A_α and all points on its edges intersecting one of the grid lines. The interval $g(A_\alpha)$ is approximated by

the minimum and maximum value thus obtained. Finally, the probability box (3) is calculated by adding the weights when appropriate. The essential computational effort thus amounts to m^d calls of the finite element program.

Figure 2 shows the result of the calculation of the load proportionality factor (LPF) where the three input parameters X_3, X_{13}, X_{14} (temperature cylinder 2, booster load node 1 in z-direction, booster load node 2 in z-direction) were kept variable. The variance σ for the Tchebycheff model was adjusted such that the base intervals $[x_{\min}, x_{\max}]$ for each of the parameters was symmetric around the corresponding mean μ with spread $\pm 0.15\mu$. In this case, $d = 3$ and we chose $m = 5$ so that 125 calls to the FE-program were required.

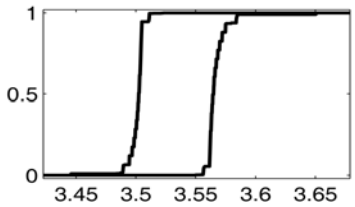


Figure 2: Probability box: LPF, 3 input variables.

Example 1 To assess the sensitivity of the load proportionality factor LPF with respect to the parameters X_3, X_{13}, X_{14} we again use the Tchebycheff model for each of the parameters with spread 0.15 times their mean values. Then we successively set one of the resulting $\sigma_3, \sigma_{13}, \sigma_{14}$ equal to zero (while keeping the others at their given value), go through the calculation indicated above and plot the resulting probability box (solid lines – the thin lines indicate the unperturbed result from Figure 2). This is displayed in Figure 3 and shows that setting $\sigma_{13} = 0$ produces the biggest reduction of the width of the probability box, while setting $\sigma_{14} = 0$ has little effect. We infer that the parameter X_{14} has the least influence on the variability of the response, while X_{13} exerts the biggest influence.

The pinching strategy in the case of probability boxes is further explicated in [4] and applied in [21]. Questions of dependence or interactivity of the input variables are left aside in this section. Dependence could be modelled by copulas on the underlying probability space $(0,1]^d$ or by restrictions on the set of probability measures on $\mathbb{R}^d$ defined by the random set.

3 Fuzzy sets

In this section, one-dimensional input variables will be modelled as *normalized fuzzy numbers*, that is as fuzzy subsets B of the real line with upper semi-continuous

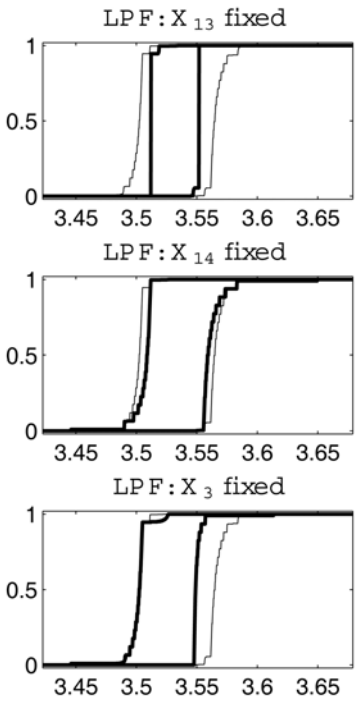


Figure 3: Probability box: LPF, frozen variables.

membership function $\pi_B(x)$ that attains the value 1. The α-level set of B is the set

$$B_\alpha = \{x \in \mathbb{R} : \pi_B(x) \geq \alpha\}, \quad \alpha \in (0,1].$$

In the multivariate case, the non-interactive joint fuzzy set is defined as follows. Given d univariate fuzzy sets $B^1, \ldots, B^d$, the joint fuzzy set has the α-level sets

$$B_\alpha = B^1_\alpha \times \ldots \times B^d_\alpha, \quad \alpha \in (0,1].$$

Interactivity will be modelled by certain parametric restrictions on the α-level sets. To avoid combinatorial complications, we shall treat interactivity of at most two out of the d variables. Since an α-level set of the form $B^i_\alpha \times B^j_\alpha$ is a homothetic image of the unit square, it suffices to give the definitions for $B^1_\alpha = B^2_\alpha = [0,1]$. Following [27], interactivity will be modelled by replacing the unit square by a diamond-shaped region, symmetric around one of the diagonals. Let $0 \leq \rho \leq 1$ and define the points $P_1, \ldots, P_4$ by

$$\begin{array}{ll} P_1 = (\rho/2, \rho/2), & P_2 = (1-\rho/2, \rho/2), \\ P_3 = (1-\rho/2, 1-\rho/2), & P_4 = (\rho/2, 1-\rho/2). \end{array}$$

Interactivity of *positive degree* ρ is modelled by taking the rhombus with corners $\{(0,0), P_2, (1,1), P_4\}$ as joint level set, while interactivity of *negative degree* $-\rho$ is modelled by the rhombus with corners $\{(0,1), P_1, (1,0), P_3\}$ as joint level set (Figure 4).

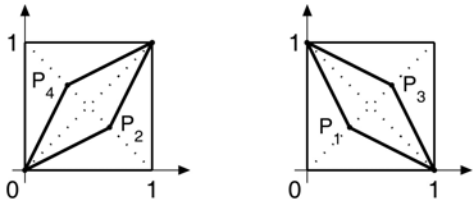


Figure 4: Positive/negative interactivity.

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. If the input variables $X = (X_1, \ldots, X_d)$ are modelled as a non-interactive or interactive fuzzy set with α-level sets B_α as above, Zadeh's extension principle yields the output variable as the fuzzy number with level sets $g(B_\alpha)$, $\alpha \in (0, 1]$.

While a fuzzy set can be interpreted as a random set (cf. e. g. [5]) and the procedure appears similar to the one of Section 2, there is a fundamental difference in the multivariate case: in fuzzy set theory, only α-level sets of the same level are combined to produce the joint fuzzy set, while for random sets, the focal elements are obtained as products with respect to any combination and thus are indexed by the product space $(0, 1]^d$.

Example 2 In the assessment of the sensitivity of the load proportionality factor LPF with respect to the input parameters X_3, X_{13}, X_{14}, these parameters were modelled as symmetric triangular fuzzy numbers, with central values μ_i from Table 1 and spread $\pm 0.15\mu_i$ as before. The numerical calculation is based on the response surface method explained in Example 1. The images of the α-level sets are again computed by piecewise multilinear combination. To handle possible lack of monotonicity of the function g, we start with level $\alpha = 1$ and go the way down to $\alpha = 0$, insuring at each step that the approximations satisfy $g(A_\beta) \subset g(A_\alpha)$ for $\alpha < \beta$.

In the non-interactive case, the procedure for determining the sensitivity of the output with respect to the input variables is the same as in Example 1. The initial calculation is performed with proportional spreads $\pm 0.15\mu_i$. Then we successively replace one of the triangular fuzzy numbers by its crisp central value μ_i, and compute the output as a fuzzy number. The result gives a good visual representation of the change of variability. This can be quantified using e. g. the Hartley-like measure

$$\mathrm{HL}(B) = \int_0^1 \log\big(1 + \lambda(B_\alpha)\big)\, \mathrm{d}\alpha$$

of fuzzy sets B as proposed by [14] (see also [1] for further implementation of this idea in sensitivity analysis and [6] for interval-valued indices). The result is

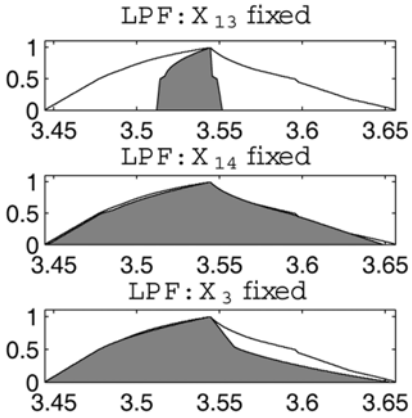


Figure 5: Fuzzy sets: LPF, frozen variables, noninteractive case.

depicted in Figure 5, where the outer contour is the membership function of the fuzzy LPF with all input parameters fuzzy, while the shaded region is bounded by the membership function of the fuzzy LPF with successively frozen input parameters. It confirms the observations obtained by the random set method: X_{13} is the most influential parameter, followed by X_3 and then X_{14}. This can be explained by the model set-up: X_{13} refers to a large booster load on one side of the frontskirt, while X_{14} signifies a much smaller booster load on the opposite side. The Hartley-like measures displayed in Table 2, though, show that some, albeit small, influence of parameter X_{14} is detectable.

Fuzzy set	HL-measure
no fixing	0.1481
X_{13} fixed	0.0398
X_{14} fixed	0.1430
X_3 fixed	0.1268

Table 2: Hartley-like measures of outputs, non-interactive input.

Example 3 This example serves to show how the effect of possible correlations between two of the input parameters on the sensitivity can be assessed. *Correlation* will be interpreted here as degree of interactivity as described above. In this example, we assume a degree of interactivity $\rho = 0.98$ between parameters X_{13} and X_{14}. The remaining parameters are treated as non-interactive. The α-level sets are of cylindrical shape with a rhombic base R_α, say. Their images are again computed by piecewise multilinear combination. Otherwise, the procedure of successively freezing variables is similar: For example, when X_{13} is frozen at its central value μ_{13}, the interactivity restricts X_{14} to vary along the intersection of R_α with the line through

μ_{13} parallel to the x_{14}-axis, while X_3 varies in its original α-level interval.

The result is shown in Figure 6; the meaning of the contour and the shaded region is the same as in Figure 5. The outcome confirms the prominence of pa-

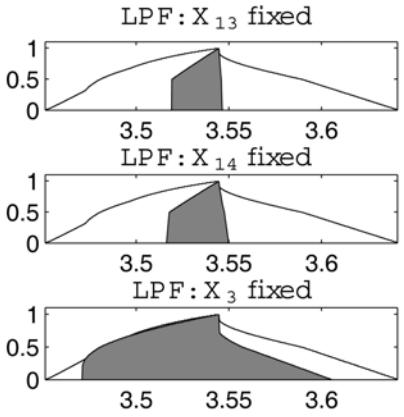


Figure 6: Fuzzy sets: LPF, frozen variables, interactive case.

rameter X_{13}; as a consequence of the correlation, parameter X_{14} is seen to exert a comparable influence. The result also demonstrates that the correlation changes the sensitivity of the output with respect to parameter X_3. Table 3 shows the Hartley-like measures of the fuzzy output under successive freezing of input variables. One may note that the study of the influence of correlations can be implemented in the fuzzy approach with ease.

Fuzzy set	HL-measure
no fixing	0.1357
X_{13} fixed	0.0287
X_{14} fixed	0.0329
X_3 fixed	0.1011

Table 3: Hartley-like measures of outputs, interactive input.

As in Example 1, the computational effort using the response surface consisted in 125 calls of the finite element program. The vertical jumps of the membership function in Figure 6 indicate that the output does not depend monotonically on the input variables. Closer inspection (done by producing an array of two-dimensional plots of the partial maps $X_i \to$ LPF) showed that this is indeed the case. Therefore, the accuracy of the method using just 125 grid values is in question. A number of additional explicit evaluations showed that the accuracy of the boundaries of the α-level sets for the LPF is in the range of ± 0.02 in absolute value.

4 Interval bounds

This section is devoted to interval estimates of input and output parameters. Suppose that the variability of each input parameter X_i is described by an interval $[\mu_i - \Delta_i, \mu_i + \Delta_i]$ of spread Δ_i around a central value μ_i. It has been argued in [16], that an estimate of the output interval can be obtained by Monte Carlo simulation using the Cauchy distribution.

The underlying theory from [16] is as follows. Suppose we wish to estimate the difference

$$\Delta y = g(x_1, \ldots, x_d) - g(\mu_1, \ldots, \mu_d)$$

where $|\Delta x_i| = |x_i - \mu_i| \le \Delta_i$. Linearization around the mean value gives

$$|\Delta y| \le \Delta = \sum_{i=1}^{d} |c_i| \Delta_i, \qquad c_i = \frac{\partial g}{\partial x_i}\big(\mu_1, \ldots, \mu_d\big).$$

If the X_i are independent random variables following a Cauchy distribution with scale parameter Δ_i, then $Y = c_1 X_1 + \ldots + c_d X_d$ obeys a Cauchy distribution with scale parameter Δ. This offers the possibility of computing the bound Δ on the output spread by Monte Carlo simulation.

The algorithm runs along the following lines. To produce a single realization, a d-dimensional sample $(z_1, \ldots, z_d)$ of Cauchy distributed variables with scale parameters 1 is taken. Setting $K = \max_{1 \le i \le d} |z_i|$, one has that $\delta_i = \Delta_i z_i / K$ has a Cauchy distribution with scale parameter Δ_i / K. Putting $x_i = \mu_i + \delta_i$ it follows that

$$Z = K\big(g(x_1, \ldots, x_d) - g(\mu_1, \ldots, \mu_d)\big)$$

is a realization of a Cauchy distributed variable with desired scale parameter Δ (this is true exactly when g is linear and otherwise approximately). An n-fold repetition yields the Monte Carlo sample of size n of the variable Z. Fitting a Cauchy distribution – e. g. by the maximum likelihood method – produces an estimate of the spread Δ of the output interval $[g(\mu_1, \ldots, \mu_d) - \Delta, g(\mu_1, \ldots, \mu_d) + \Delta]$. The computational effort for this estimate is n calls of the finite element program and thus independent of the dimension d. This offers the possibility to include a larger number of input variables in the analysis.

Example 4 In this calculation, 17 input parameters were included with nominal values displayed in Table 1. The spreads Δ_i were taken as 0.15-times the nominal values μ_i. We used a direct Monte Carlo method to produce a sample of size $n = 100$. The value of the load proportionality factor LPF was obtained as $\mu = g(\mu_1, \ldots, \mu_d) = 3.5443$. The simulation resulted in an estimate for its spread of $\hat{\Delta} = 0.2924$.

In the next step, the distribution of the resulting spread Δ was estimated by resampling. We employed 10000 random subsamples of size 100 (with repetition), following the suggestions in [23]. This resulted in a 95%-confidence interval for Δ of $CI_{0.95}(\hat{\Delta}) = [0.2281, 0.3685]$. The essential computational effort consisted in $n = 100$ calls of the finite element program.

Remark 5 A sensitivity analysis could be based on this method, again by freezing variables successively. It is possible to reduce computational cost by using the same Monte Carlo sample and approximating the frozen variables by a truncated Cauchy distribution. More precisely, instead of setting $\Delta_1 = 0$, say, we select the random numbers $(x_2, \ldots, x_d)$ computed above from the part of the population $(x_1, x_2, \ldots, x_d)$ which satisfies $|\delta_1| < \varepsilon$ for a suitably chosen small ε. This is justified, because the resulting truncated $(d-1)$-dimensional random variables converge in distribution to the ones with Δ_1 frozen at the value 0 as $\varepsilon \to 0$. However, successive simultaneous freezing of two or more variables requires repeated Monte Carlo simulation because the sample size would be too small for repeated truncation.

A more troublesome observation concerns the accuracy of the Cauchy method in our situation where the output function g is a nonlinear finite element computation resulting in the LPF. It turned out that the simulations of the auxiliary variable Z actually failed the KS-test for being Cauchy distributed. This means that our output function g is too far away from linearity and thus puts the accuracy of the Cauchy method into question in this context.

5 Monte Carlo simulation

To complete the analysis, we have a glimpse at direct Monte Carlo simulation in sensitivity analysis. Methods like scatterplots (input – output) and computing the weighted contribution of each input variable to the variance of the output are commonplace and will not be discussed here. These methods suffer the problem that hidden interactions may have a significant effect on the decomposition of the variance (see, however, [2]). We therefore turn to a method which intends to remove the influence of co-variates on the correlation between a given input variable X_i and the output variable Y. This method is based on the partial rank correlation coefficient (PRCC).

We recall that partial correlation between two random variables X_i and Y given a set of co-variates $X_{\setminus i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d\}$ is defined as the correlation between the two residuals $e_{X_i \cdot X_{\setminus i}}$ and $e_{Y \cdot X_{\setminus i}}$ obtained by regressing X_i on $X_{\setminus i}$ and Y on $X_{\setminus i}$, respectively. More precisely, one first constructs the two regression models

$$\widehat{X}_i = \alpha_0 + \sum_{j \neq i} \alpha_j X_j, \quad \widehat{Y} = \beta_0 + \sum_{j \neq i} \beta_j X_j,$$

obtaining the residuals

$$e_{X_i \cdot X_{\setminus i}} = X_i - \widehat{X}_i, \quad e_{Y \cdot X_{\setminus i}} = Y - \widehat{Y}.$$

Since $e_{X_i \cdot X_{\setminus i}}$ and $e_{Y \cdot X_{\setminus i}}$ are those parts of X_i and Y that remain after subtraction of the best linear estimates in terms of $X_{\setminus i}$, the partial correlation coefficient

$$\rho_{X_i, Y \cdot X_{\setminus i}} = \rho(e_{X_i \cdot X_{\setminus i}}, e_{Y \cdot X_{\setminus i}})$$

quantifies the linear relationship between X_i and Y after removal of any part of the variation due to the linear influence of $X_{\setminus i}$. Applying a rank transformation to the variables X_i and Y leads to the partial rank correlation coefficient (PRCC). For further background on PCCs and PRCCs, see [7, 11, 22].

Example 6 To estimate the influence of each of the 17 input parameters from Table 1 on the output LPF, we performed a Monte Carlo simulation of size $n = 100$ with uniformly distributed input variables (on the intervals as in Example 4), using Latin hypercube sampling, an efficient stratified sampling strategy.

To obtain a sample of size n, the Latin hypercube sampling plan divides the range of each variable X_i into n disjoint subintervals of equal probability. First, n values of each variable X_i, $i = 1, \ldots, d$, belonging to the respective subintervals are randomly selected. Then the n values for X_1 are randomly paired without replacement with the n values for X_2. The resulting pairs are then randomly combined with the n values of X_3 and so on, until a set of n d-tuples is obtained. This set forms the Latin hypercube sample. The advantage of Latin hypercube sampling is that sampled points are evenly distributed through design space, thereby covering regions possibly important for the input-output map which might be missed by direct Monte Carlo simulation. It can be shown that the variance of an estimator based on Latin hypercube sampling is asymptotically smaller than the variance of the direct Monte Carlo estimator, and possibly markedly smaller when the input-output map is partially monotonic [8, 17, 26].

For additional accuracy in view of the rather small sample size we subjected the simulated variables to correlation control (see [12, 13]). This procedure consists in a rearrangement of the originally simulated values such that the resulting empirical rank correlation matrix is close to diagonal.

The resulting PRCCs can be seen in Figure 7. For further statistical confirmation, we performed a resam-

pling procedure as in Example 4, producing bootstrap confidence intervals for the partial rank correlation coefficients as displayed in Figure 8. Accordingly, only the PRCCs of the parameters X_1, X_3, X_9, X_{13} and X_{14} test to be nonzero.

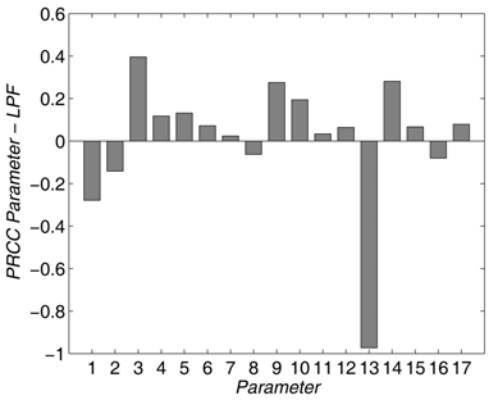


Figure 7: Partial rank correlation coefficients.

The outcome confirms the results of the sensitivity analysis in the previous sections: Among the parameters X_3, X_{13} and X_{14}, the one with the biggest influence is X_{13}, followed by X_3 and X_{14}.

We also ran various tests with correlated input as in Example 3 which confirmed the observed sensitivities. However, each test required a new Monte Carlo simulation with sample size $n = 100$. In addition, we computed Sobol indices [25] for groups of variables; this, however, again requires additional Monte Carlo simulations.

6 Summary and Conclusions

Starting from a research project in aerospace engineering one of whose goals was to determine the sensitivity of the buckling load of the frontskirt of the ARIANE 5 launcher with respect to certain input parameters, we explored various methods from probability and imprecise probability theory. In view of the excessive computational costs of a single run of the finite element program, the major challenge was to develop methods with as few calls of the program as possible. We used a simplified model of the launcher for the numerical tests of the methods.

The methods under scrutiny were random sets and Tchebycheff's inequality, fuzzy sets and Hartley-like measures, intervals and sampling from a Cauchy distribution, standard Monte-Carlo simulation and resampling. Criteria for the evaluation are

– computational effort

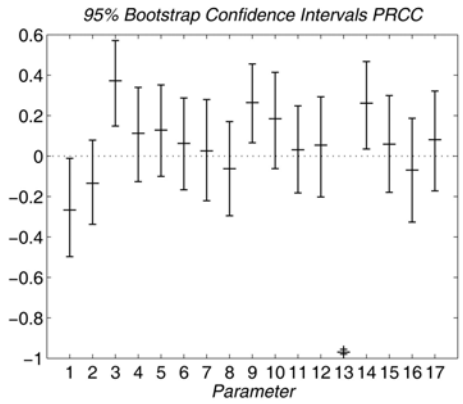


Figure 8: Partial rank correlation coefficients, confidence intervals.

– applicability to large scale problems

– accuracy

– avoidance of tacit assumptions

– reliability and clarity of interpretation

– possibility of analyzing correlated input.

Generally speaking, the Monte Carlo simulation methods are computationally least expensive. For our sensitivity study, a sample size of $n = 100$ appeared sufficient. In addition – as is well known – the sample size can be chosen independently of the number of input variables, so that we could include all 17 variables in our study. These methods are clearly applicable to large scale problems. Disadvantages are that parametric assumptions on the input variables have to be made and that freezing of variables requires repetition of the $n = 100$ simulations. Thus computing PRCCs plus resampling is possible irrespective of the problem scale, but variance decomposition by freezing variables is not. The same applies to analyzing sensitivity with respect to input correlations, which requires repetition of the simulation as well. The numerical accuracy of the Monte Carlo simulation is well known to be of order $1/\sqrt{n}$ times the standard deviation of the simulated variable. In view of the coefficients of variation which were in the range of 10% this appeared sufficient for the sensitivity study.

We emphasize that the results of a Monte Carlo simulation are amenable to resampling, which introduces little additional computational effort (no further evaluations of the costly input-output map are needed). In this way, bootstrap confidence intervals can be obtained that may serve as statistical estimates of the accuracy of the results. For example, we estimated the bias of each partial rank correlation coefficient, that is, the absolute value of the difference of the mean

of the resampled data and the initial estimate. The estimated bias resulted to be less than 2% of the initial estimate. Further, the significance of the resulting ranking of the influence of the respective input parameters can be assessed by comparing the bootstrap confidence intervals.

The Cauchy method is a simulation method for estimating the spread of the output interval. The resulting estimate is non-parametric in as much as only the spreads of the input variables enter. As a subcase of Monte Carlo simulation, everything that has been said above applies here as well. A problematic point is that the method is derived under the assumption that the output function is approximately linear. In our case, the output function is substantially nonlinear. By means of repeated simulations we observed a quite substantial lack of accuracy of the estimate of the output spread in our case. Namely, direct Monte Carlo simulations of size $n = 100$ of the output variable LPF, with uniformly distributed input variables, produced an output range of $[3.45, 3.65]$. This indicates that the range was largely overestimated by the Cauchy method (see Example 4). This could possibly be overcome by the suggestion of [16] of repeated bisection of the input interval, though at an increase in computational cost.

Both in the fuzzy set and random set methods, the output α-level sets and focal sets, respectively, are computed by searching for the maximum and minimum of the corresponding output range. Sufficient accuracy can only be obtained by a larger number of calls of the output function, evaluated on a grid of input data. In addition, the grid size increases exponentially with the number of input variables. These methods appear feasible only in the case of medium size problems and a small number of input variables. Monotonicity or partial monotonicity of the output function increases accuracy and helps reducing the number of computations required.

Test runs with finer grids showed that the numerical error of the interpolation (i. e. replacing the true output function by a piecewise bilinear response surface) was less than 1%, thus definitely satisfactory. However, the optimization error introduced when calculating the boundaries of the output level sets turned out to be about ± 0.02 in absolute value, which is around 10 - 20% of the spread of the base level (see end of Section 3).

The numerical error in the boundaries of the output level sets appears less influential in the random set method. This is due to a certain averaging effect. Indeed, in the fuzzy model the computation of ℓ output level sets corresponds to ℓ input level sets, whereas in the random set model – at least when using random set independence – a combination of ℓ^d input focal sets enters (d the number of variables).

Both methods are essentially non-parametric. The random set model we used is generated by Tchebycheff's inequality and hence non-parametric by definition. In the fuzzy set model, we used triangular fuzzy numbers as input. These can be seen as a collection of intervals of linearly changing length. The α-level sets resulting from the computation determine the output range when the input varies over d-dimensional intervals of length proportional to $1 - \alpha$.

The fuzzy model in combination with the response surface technique has an additional advantage: it allows the a-posteriori introduction of interactivity between the input variables without the need for new calls of the output function. The effect of interactive input can simply be evaluated by interpolation in the response surface.

We finally comment on the practicality of upscaling to the full problem. This remains a major challenge. The computational structure of the given problem consists in a nonlinear, incremental procedure. The LPF is obtained as the ultimate load value beyond which the computed solution cannot be prolonged. This may be either due to a bifurcation point or to a breakdown of the structure. We currently pursue two strategies. One strategy is a perturbation method that replaces the full model by a quadratic approximation when a bifurcation point is reached. This is based on Koiter's asymptotic analysis of post-buckling of shells, see e. g. [15]. The sensitivity analysis would be done with the asymptotic model in place of the full model. The second strategy is to start the sensitivity analysis at a later stage of the iterative procedure. Both methods require to access the finite element code at a deeper level. A certain difficulty which we expect to encounter stems from the fact that the incremental procedure is path dependent. Thus varying the input parameters late in the process could be misleading, as initial variations might result in a quite different path to breakdown.

Acknowledgements

We are grateful to Herbert Haller for providing the finite element models and advice on engineering questions. Thanks are also due to Christof Neuhauser and Alexander Ostermann for discussions on numerical problems, to Hermann Starmans for the project management and to Robert Winkler for help with the finite element codes.

References

[1] D. A. Alvarez. Nonspecifity for infinite random sets. Preprint, Unit for Engineering Mathematics, University of Innsbruck, 2006.

[2] G. E. B. Archer, A Saltelli, I. M. Sobol. Sensitivity measures, ANOVA-like techniques and the use of the bootstrap. *Journal of Statistical Computation and Simulation*, 58:99–120, 1997.

[3] S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. *SANDIA Report* SAND2002-4015, Sandia National Laboratories, Albuquerque, 2003.

[4] S. Ferson and W. T. Tucker. Sensitivity analysis using probability bounding. *Reliability Engineering & System Safety*, 91:1435-1442, 2006.

[5] I. R. Goodman, H. T. Nguyen. Fuzziness and randomness. In: C. Bertoluzza, M. Á. Gil, D. A. Ralescu, editors. *Statistical Modeling Analysis and Management of Fuzzy Data*. Physica-Verlag, Heidelberg, 2002.

[6] J. W. Hall. Uncertainty-based sensitivity indices for imprecise probability distributions. *Reliability Engineering & System Safety*, 91:1443-1451, 2006.

[7] J. C. Helton, F. J. Davis. Sampling-based methods for uncertainty and sensitivity analysis. *SANDIA Report* SAND99-2240, Sandia National Laboratories, Albuquerque, 2000.

[8] J. C. Helton, F. J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *SANDIA Report* SAND2001-0417, Sandia National Laboratories, Albuquerque, 2002.

[9] J. C. Helton, R. M. Cooke, M. D. McKay, A. Saltelli, editors. *The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004), Santa Fe, New Mexico*. Special Issue, *Reliability Engineering & System Safety*, 91(10-11):1105–1474, 2006.

[10] J. C. Helton, J. D. Johnson, C. J. Sallaberry, C. B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91:1175-1209, 2006.

[11] R. L. Iman, W. J. Conover. The use of the rank transformation in regression. *Technometrics*, 21:499-509, 1979.

[12] R. L. Iman, W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11:311–334, 1982.

[13] R. L. Iman, M. J. Shortencarier, J. D. Johnson. A FORTRAN 77 program and user's guide for the calculation of partial correlation and standardized regression coefficients. *SANDIA Report* SAND85-0044, Sandia National Laboratories, Albuquerque, 1985.

[14] G. Klir, M. J. Wiermann. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Physica-Verlag, Heidelberg, 1998.

[15] W. T. Koiter. Current trends in the theory of buckling. In: B. Budiansky, editor. *Buckling of Structures. Proceedings IUTAM Symposium, Cambridge 1974*. Springer-Verlag, Berlin 1976.

[16] V. Kreinovich, S. A. Ferson. A new Cauchy-based black-box technique for uncertainty in risk analysis. *Reliability Engineering & and System Safety*, 85:267–279, 2004.

[17] M. D. McKay, R. J. Beckman, W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239-245, 1979.

[18] I. Molchanov. *Theory of Random Sets*. Springer-Verlag, Berlin, 2005.

[19] H. T. Nguyen. *An Introduction to Random Sets*. Chapman & Hall, Boca Raton, 2006.

[20] M. Oberguggenberger, W. Fellin. Assessing the sensitivity of failure probabilities: a random set approach. In: G. Augusti, G. I. Schuëller, M. Ciampoli, editors. *Safety and Reliability of Engineering Systems and Structures. ICOSSAR 2005 - Rome*. Millpress, Rotterdam 2005, 1755–1760.

[21] M. Oberguggenberger, W. Fellin. Reliability bounds through random sets: nonparametric methods and geotechnical applications. *Computers & Structures*, to appear.

[22] A. Saltelli, I. M. Sobol. About the use of rank transformation in sensitivity analysis of model output. *Reliability Engineering & System Safety*, 50:225-239, 1995.

[23] J. Shao, D.-S. Tu. *The Jackknife and Bootstrap*. Springer Verlag, New York, 1995.

[24] G. I. Schuëller. On the treatment of uncertainties in structural mechanics and analysis. *Computers & Structures*, 85:235–243, 2007.

[25] I. M. Sobol. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling & Computational Experiment*, 1:407-414, 1993.

[26] M. Stein. Large sample properies of simulations using Latin hypercube sampling. *Technometrics*, 29:143-151, 1987.

[27] F. Tonon, C. L. Pettit. Toward a definition and understanding of correlation for variables constrained by random relations. In: G. Augusti, G. I. Schuëller, M. Ciampoli, editors. *Safety and Reliability of Engineering Systems and Structures. ICOSSAR 2005 - Rome*. Millpress, Rotterdam 2005, 1741–1745.

Luceños' Discretization Method and its Application in Decision Making under Ambiguity

Michael Obermeier
Department of Statistics
University of Munich
Germany
obermeierm@web.de

Thomas Augustin
Department of Statistics
University of Munich
Germany
Thomas@stat.uni-muenchen.de

Abstract

When extending classical statistical models to imprecise probabilities, one fundamental difficulty, which may have hindered some powerful practical applications, is the following gap: While classical statistical models are typically based on absolutely continuous probability distributions, most computational methods developed for handling imprecise probability models rely on finite sample spaces. A natural way to close this gap is discretization of the underlying continuous probability distribution. This, however, is far from straightforward, because naïve discretization by mere rounding may cause a substantial bias; even moments of very low order would be distorted. The present paper discusses the application of Luceños' ([10]) so-to-say adaptive discretization method in imprecise probability models. We firstly recall two theorems, showing, for any fixed natural number r, how to construct a discrete random variable such that its first, second, ... r-th moment coincides with the corresponding moment of the underlying continuous distribution. (In addition, also coincidence of the distribution functions in a fixed number of points can be enforced.) Then we illustrate the power of the method by utilizing it in decision problems under ambiguity.

Keyword: Decision making under ambiguity, discretization, Gaussian quadrature, imprecise probabilities, interval probability, linear programming, Luceño, numerical integration.

1 Introduction

Classical statistical models typically are based on parametric, absolutely continuous probability distributions on the real line. Handling extensions of these models in the imprecise probability framework, quite often becomes very demanding from the computational point of view, and then approximative techniques are the best one can hope for, the more as also in classical statistics many integrals of less smooth functions can be only obtained numerically. A natural idea in this context is discretization, in order to make available powerful algorithms (for instance for handling graphical models ([4]) or decision making ([9], [17]) that explicitly rely on finite spaces to obtain approximate solutions in this generalized setting. However, such discretizations need some care; for more than hundred years, since the work of Sheppard ([13]) at the end of the nineteenth century, statisticians have been well aware that analysis based on rounded data may be severely biased, and so discretization by mere rounding or other ad-hoc techniques is a bad advice. Sheppard also developed a simple correction formula, the applicability of which, however, is restricted to the r-th moment of a normal distribution and further regularity conditions. Applying this "correction" to other distributions or more complex functions of random variables can even increase the bias ([18], [1]).

In this paper we study a more sophisticated discretization procedure, which discretizes the underlying distribution in an adaptive way such that, for any r, the r-th first moments of the new, discretized distribution agree with the ones from the original distribution. The method is based on Gaussian quadrature; its power in statistical applications is advocated by Luceño ([10]), to which we refer when describing the essentials of the technique. We claim that these results are even more important in the area of imprecise probabilities, for the reasons outlined above. To corroborate this thesis, we illustrate the discretization technique by applying it to two types of decision problems under ambiguity.

In more detail, the paper is organized as follows: In Section 2 we prepare the ground by recalling Luceños ([10]) two main theorems and briefly discussing issues of concrete calculation of the discretization. Our application to decision theory is divided in four sections. In Section 3 we recall basic notions, which then are used to explain in Section 4 two different general types of discretizations. In Section 5 it is shown how to use

one discretization technique in parametric situations with non-elementary utility functions: prior information is assumed to be described by a set of parametric distributions with varying parameters. We calculate E-admissible actions as well as the Γ-maximin solutions. While these results may be understood as a direct extension of numerical integration procedures from classical probability to the imprecise probability framework, the importance of the second type of discretization, the construction of an approximately equivalent decision problem becomes unambiguously evident in Section 6, where envelopes of parametric sets are considered. After choosing a reference distribution which determines the concrete discretization, we transform the whole decision problem on an infinite state of natures to an approximately equivalent decision problem based on a finite set of states of nature, and then give E-admissible and Γ-maximin solutions by adapting the algorithms from [9] and [17].

2 Luceños Discretization Method

2.1 Two Fundamental Theorems

One can understand Luceños article *discrete approximations to continuous univariate distributions* ([10]) as a kind of "translation" of the Gaussian quadrature method for integration into probability theory.

The central idea of the Gaussian quadrature method is to replace an integral over a function $h(x)$ and its weight function $w(x)$ by a sum, i.e. to take

$$\int_a^b h(x)w(x)d(x) \approx \sum_{j=1}^{N} h(x_j)w_j, \tag{1}$$

where w_j and x_j are chosen in a sophisticated way such that one can approximate the value of the integral numerically with rather high accuracy. To find the nodes x_j and the weights w_j, $j = 1, \ldots, N$, the roots of recursively defined orthogonal polynomials of degree N are used, which depend on $w(x)$. Unlike some other numerical integration methods, e.g. the Newton-Cotes formulas, the abscissae and weights here are found dynamically, which means that they are adapted to the shape of the original function and therefore the approximation of the integral is very accurate (for Gaussian quadrature and other numerical integration methods see for example [15]).

In the following statistical application of this method the weight function $w(x)$ is the probability density function (PDF) of a univariate continuous random variable X, whereas w_i, $i = 1, \ldots, N$ constitute the probability mass function (PMF) of the corresponding discrete random variable Y. The detailed proceeding and some fundamental properties of the approximations are described in two theorems, which we recall here from [10].

Proposition 1 (Luceño) *Consider two univariate random variables X and Y on a domain $[a, b]$ with $-\infty \leq a \leq x \leq b \leq \infty$.[1] Let X be (absolutely) continuous with probability density function (PDF) $w(x)$, having finite moments of any order, and let Y be discrete on N atoms $x_1, \ldots, x_N$ with probability mass function (PMF) $w_1, \ldots, w_N$. Then*

$$\mathbb{E}(X^r) = \mathbb{E}(Y^r), \forall r \in \{0, \ldots, 2N-1\} \tag{2}$$

if and only if the nodes $x_1, \ldots, x_N$ *and the* weights $w_1, \ldots, w_N$ *satisfy the following two conditions:*

i) $x_1, \ldots, x_N$ are the roots of the polynomial $Q_N(x)$ of degree N defined by the three-term recursion

$$Q_{i+1}(x) = (x - \delta_{i+1})Q_i(x) - \gamma_{i+1}^2 Q_{i-1}(x), \tag{3}$$

$$i \geq 0, \text{ where } Q_{-1}(x) \equiv 0,\ Q_0(x) \equiv 1 \text{ and}$$

$$\delta_{i+1} = \mathbb{E}\{XQ_i^2(X)\}/\mathbb{E}\{Q_i^2(X)\},\ i \geq 0, \tag{4}$$

$$\gamma_{i+1}^2 = \begin{cases} 0, & i = 0 \\ \mathbb{E}\{Q_i^2(X)\}/\mathbb{E}\{Q_{i-1}^2(X)\}, & i \geq 1. \end{cases} \tag{5}$$

ii) the probabilities $w_1, \ldots, w_N$ are the solution of the linear system

$$\sum_{j=1}^{N} Q_k(x_j)w_j = \begin{cases} 1, & k = 0 \\ 0, & k = 1, \ldots, N-1. \end{cases} \tag{6}$$

In addition, the cumulative distribution functions (CDFs) of X and Y can be forced to agree at least in a given set of points:

Proposition 2 (Luceño) *Consider the situation of Proposition 1. Let $c_0 = a < c_1 < \ldots < c_{M-1} < c_M = b$ such that $I_i = \int_{c_{i-1}}^{c_i} w(x)dx > 0$ for all $i = 1, \ldots, M$, and consider the discrete random variables Y_i, $i = 1, \ldots, M$ with atoms $x_{i1}, \ldots, x_{iN}$ and weights $w_{i1}, \ldots, w_{iN}$ arising from applying Proposition 1 to the random variables $X_i := X \cdot 1_{[c_{i-1}, c_i]}$.*

Then, for the random variable Z with atoms $x_{11}, \ldots, x_{1N}, x_{21}, \ldots, x_{2N}, \ldots, x_{M1}, \ldots, x_{MN}$ and weights $I_1 \cdot w_{11}, \ldots, I_1 \cdot w_{1N}, I_2 \cdot w_{21}, \ldots, I_2 \cdot w_{2N}, \ldots, I_M \cdot w_{M1}, \ldots, I_M \cdot w_{MN}$,

$$\mathbb{E}(X^r) = \mathbb{E}(Z^r), \quad \forall r \in \{0, \ldots, 2N-1\}, \tag{7}$$

and the CDFs of X and Y coincide at least at the abscissae $c_0, c_1, \ldots, c_{M-1}, c_M$.

[1] To include the statistical standard distributions, without the need to distinguish between the domain $\mathbb{R}$ and some bounded domain, we allow for $a = \infty$ and $b = \infty$, but implicitly assume $f(-\infty) = f(\infty) = 0$

2.2 Easier Calculations in Standard Cases

In the case where the weight function belongs to a standard family, there are well known polynomials, which can be used instead of the three-term recursion just described to find the abscissae and corresponding weights. For example, for the normal distribution with mean μ and variance σ^2 one can use the Gauss-Hermite polynomials with the weight function $w(x) = e^{-x^2}$ on the interval $-\infty < x < \infty$:[2]

$$H_{j+1}(x) = 2xH_j(x) - 2jH_{j-1}(x) \qquad \text{with} \quad H_1(x) = 1$$

Given the roots $x_i^{(GH)}$ and weights $w_i^{(GH)}$ of the polynomial with degree N, one can obtain the random Variable Y and its PMF through (see [10], p.347):

$$Y_j = \mu + \sigma x_j^{(GH)}\sqrt{2}, \quad w_j = \frac{1}{\sqrt{\pi}} w_j^{(GH)}, j = 1, \ldots, N.$$

However there are only very few, classical families where one can easily use well known polynomials to find the new variable.

2.3 Nonstandard Case

For the partial intervals used in Proposition 2 the weights do not generally belong to any of these classical families. As a consequence, there are no well known polynomials which can be used to find the new variable, and the three term recursion described above has to be used, leading to another numerical problem: how to solve the inner products in Part i) of Proposition 1 to determine the γ_is and δ_is, if $w(x)$ is no classical weight?

Based on the knowledge of the so called "modified moments" $\nu_j = \int_a^b \pi_j(x)w(x)dx$ of orthogonal polynomials π_j, Sack and Donavan ([12]) offer a numerically stable algorithm to find the coefficients γ_j^2 and δ_j of the recursion. Wheeler ([22]) improved this method to an $O(N^2)$ algorithm. Other solutions are presented by Gautschi ([6]). One simple and heuristic way is to approximate the inner products with an adequate quadrature rule. In the further calculations presented here this simple method is used, because the focus of this paper is mostly on the construction of the new discrete random variable, not directly on the exact calculation of an integral.

In a last step the weights w_i and nodes x_i of the new variable have to be found. They result from the eigenvalues and eigenvectors of a tridiagonal matrix consisting of the γ_i^2 and δ_i from (5) and (4) (for details see [15], p.179f.).

2.4 Accuracy of the Approximation

If the focus of the approximation is on the shape of a continuous CDF, one will use the method described in Proposition 2. The accuracy of this discrete approximation depends on the way the partition of $[a, b]$ is chosen. One first possibility is to split the support in inner intervals $[c_{i-1}, c_i]$, $i = 2, \ldots, M-1$ of the same size and two possibly larger outer intervals if the domain is infinite. The obvious problem is that then one has the same number of interpolation points in areas where the PDF is high (which means that the CDF has a big increase) as in areas with low PDF. A more satisfying method is to use the PDF (if it is numerically manageable) to find a more appropriate partition. One can split the support in M quantiles and use them as the c_j's. In this way the intervals are adjusted to the shape of the original distribution: they are small where the PDF is high and wide where the PDF is low, and so finally in the important areas of the support there are more nodes than in the less important ones.[3]

3 Decision Making under Ambiguity, Basic Notions

To illustrate and exemplify the power of Luceños method in the area of imprecise probabilities, we apply it to some general decision problems under ambiguity. To prepare the ground we briefly recall the basic setting of decision theory, where one has to choose an optimal *action* from a non-empty, finite set $\mathbb{A} = \{a_1, ..., a_n\}$ of possible actions. The consequences of every action depend on the true, but unknown *state of nature* ϑ being an element of a space Θ. The corresponding outcome is evaluated by the *utility function* $u : (\mathbb{A} \times \Theta) \to \mathbb{R}$ and by the associated random variable $\mathbf{u}(a)$ on Θ. Often it makes sense to study *randomized actions* in addition, which can be understood as a classical probability measure $\lambda = (\lambda_1, ..., \lambda_n)$ on $(\mathbb{A}, \mathcal{P}o(\mathbb{A}))$, where λ_i is interpreted as the probability with which action a_i is taken. Then $u(\cdot)$ and $\mathbf{u}(\cdot)$ are extended to randomized actions by defining $u(\lambda, \vartheta) := \sum_{s=1}^n u(a_s, \vartheta)\lambda_s$. (Next to simplifying calculations, under some criteria the optimal randomized action may be superior to the optimal unrandomized one.)

[2] For some other distributions there are also well known polynomials, which can be used directly to find the discretization: For the gamma distribution the Gauss-Laguerre polynomials can be used, for the beta distribution the Gauss-Jacobi polynomials and for the uniform distribution the Gauss-Legendre polynomials (for details see [10]).

[3] Several tests in [11], chapter 3 for this approximation to the standard normal distribution show empirically that, for a sufficiently large M, samples of the new discrete variable cannot be distinguished any more from the original variable.

This model contains the essentials of every (formalized) decision situation under uncertainty and is applied in a huge variety of disciplines. If the states of nature are produced by a perfect random mechanism, and if the corresponding probability measure $\pi(\cdot)$ on $(\Theta, \mathcal{P}o(\Theta))$ is completely known, the Bernoulli principle is nearly unanimously favored. One chooses then the unrandomized action a^* or the randomized action λ^* maximizing the expected utility

$$\mathbb{E}_\pi \mathbf{u}(a) := \int u(a, \vartheta) d\pi(\vartheta) \tag{8}$$

and $\mathbb{E}_\pi \mathbf{u}(\lambda) := \int u(\lambda, \vartheta) d\pi(\vartheta)$ among all a and all λ, respectively. a^* and λ^*, respectively, is called *Bayes action with respect to* π. In many applications, however, it is not possible to describe the prior knowledge on the stochastic behavior of the states of nature by a classical probability measure, and a more general description of ambiguity is needed, as provided by imprecise probabilities and related approaches (see, in particular, [19] and [21]).

From the technical point of view, the usual concepts of imprecise probability lead to convex sets $\mathcal{M}$ of classical probabilities. Every distribution π from $\mathcal{M}$ produces a classical expected utility $\mathbb{E}_\pi \mathbf{u}(\lambda)$. Assuming $\mathcal{M} \neq \varnothing$, all possible expected utilities $\mathbb{E}_\pi \mathbf{u}(\lambda)$ range within the interval

$$\left[\underline{\mathbb{E}}_{\mathcal{M}} \mathbf{u}(\lambda) \, , \, \overline{\mathbb{E}}_{\mathcal{M}} \mathbf{u}(\lambda) \right] , \tag{9}$$

and this interval-valued quantity is called *generalized expected utility*. Based on this notion of generalized expected utility several optimality criteria are common. An overview is given in [16] where also further references are provided. Here two of them are considered: the Γ-maximin criterion and the criterion of E-admissibility.

The Γ-*maximin* criterion considers a worst case scenario, which means $\underline{\mathbb{E}}_{\mathcal{M}} u(\lambda)$ is evaluated only. Then an action λ^* is optimal iff for all λ

$$\underline{\mathbb{E}}_{\mathcal{M}} \mathbf{u}(\lambda^*) \geq \underline{\mathbb{E}}_{\mathcal{M}} \mathbf{u}(\lambda). \quad {}^{4} \tag{10}$$

The concept of *E-admissibility* on the other hand typically does not offer a unique action as the one to choose, but a set of optimal actions: An action[5] a^* is said to be E-admissible in $\mathcal{M}$ with respect to a set of prior probabilities $\mathcal{M}$, iff there exists a classical prior $\pi(\cdot) \in \mathcal{M}$ such that a^* is Bayes action with respect to $\pi(\cdot)$ for all actions a under consideration.[6]

4 Two Types of Discretization

In the case of continuous distributions of the states of nature practical handling these criteria may encounter severe difficulties. Except in special cases, where the distributions are stochastically ordered and or the expected utility is easily expressed by a underlying parameter, it is hard or even impossible to determine optimal actions by evaluating the integrals (8). Since for finite set of states of nature powerful algorithms exist, decision making provides an area where Luceños discretization techniques is quite welcome.

To implement the discretization, we apply Proposition 1 and 2 by assuming there is a random variable X. In the background, that takes values in Θ producing the states of natures. Applying the general techniques to this variable X setting in (1) as well as to $w(\cdot) = \pi(\cdot)$ and $h(\cdot) = u(a, \cdot)$ and $h(\cdot) = u(\lambda, \cdot)$, respectively, note that not only the weights but also the nodes depend on the underlying probability distribution. Therefore, for a given set $\mathcal{M}$ of continuous distributions on $\Theta \subset \mathbb{R}$ (equipped with the corresponding Borel $\sigma-$field $\mathcal{B}$), two different types of discretization have to be distinguished, depending whether a separate or a common discretization scheme is used:

Type-I discretization: The first possibility is to apply Proposition 1 or 2 to every element $\pi(\cdot) \in \mathcal{M}$ separately. This means that to every $\pi(\cdot) \in \mathcal{M}$ a corresponding discrete distribution $\nu_\pi(\cdot)$ is constructed with atoms $\vartheta_{1,\pi}, \ldots, \vartheta_{N,\pi}$, and $\mathbb{E}_\pi \mathbf{u}(\lambda)$ is replaced by its approximative equivalent

$$\mathbb{E}_{\nu_\pi} \mathbf{u}(\lambda) = \sum_{j=1}^{N} u(\lambda, \vartheta_{j,\pi}) \nu_\pi \left(\vartheta_{j,\pi} \right). \tag{11}$$

Type-II discretization: Here Θ itself is discretized. For this, a certain *reference distribution* $\pi_0(\cdot) \in \mathcal{M}$ is selected, to which then Proposition 1 or 2 is applied.[7] The resulting nodes[8] $x_1, \ldots, x_N$ are used to define a new discrete space $\Theta_d = \{\theta_1, \ldots, \theta_N\}$ with $\vartheta_1 = \big[a, (x_1 + x_2)/2\big]$, $\vartheta_2 = \big((x_1 + x_2)/2, (x_2 + x_3)/2\big]$, $\ldots$, $\vartheta_N = \big((x_{N-1} + x_N)/2, b\big]$, which then is used to replace Θ. The utility function is then extended to Θ_d

[4] This concept is a very conservative decision rule, similarly to the maximin rule in the classical decision theory: in the case of complete ambiguity both criteria coincide.

[5] Under the criterion of E-admissibility usually consideration is confined to the unrandomized actions. If needed, the algorithms used later on can be extended to randomized actions (cf. [17, p. 357]).

[6] E-admissibility can be considered in a broader sense as a generalization of the criterion of admissibility in classical decision theory.

[7] This method is described later on in Section 6. A glance at Figure 4 may therefore be helpful.

[8] For the sake of readability, the dependence on $\pi_0(\cdot)$ is suppressed in the notation here throughout the following definitions.

by assigning the values at the nodes, i.e. by defining

$$u(a, \vartheta_j) := u(a, x_j), \quad \forall j \in \{1, \ldots, N\}, \ \forall a \in \mathbb{A}\,.$$

The nodes of the reference distribution are used to discretize all elements of $\mathcal{M}$. More precisely, the set $\mathcal{M}$ of continuous probability distributions on $(\Theta, \mathcal{B})$ is replaced by the set $\mathcal{P}$ of discrete probability distributions, being the set of all classical probabilities in accordance with $P(\cdot) = [L(\cdot), U(\cdot)]$ on $(\Theta_d, \mathcal{P}o(\Theta_d))$[9] defined via:[10]

$$L\Big(\bigcup_{j\in J}\{\vartheta_j\}\Big) = \inf_{\pi\in\mathcal{M}} \pi\Big(\bigcup_{j\in J}\vartheta_j\Big) \tag{12}$$

and

$$U\Big(\bigcup_{j\in J}\{\vartheta_j\}\Big) = \sup_{\pi\in\mathcal{M}} \pi\Big(\bigcup_{j\in J}\vartheta_j\Big) \tag{13}$$

$$\forall J \subset \{1, \ldots, N+1\}\,. \tag{14}$$

Since N may be quite large, in many applications the computational effort may be substantially reduced by a further approximation in which not all elements of $\mathcal{P}o(\Theta_d)$ are used in the assignment of $P(\cdot)$. Then the power set of $\{1, \ldots, N+1\}$ in (14) is replaced by some subset $\mathcal{J}_{\max}$, and natural extension is applied to obtain the remaining interval limits $L(\cdot)$ and $U(\cdot)$. A natural choice for $\mathcal{J}_{\max}$, that is also applied in Section 6, is to consider only connected intervals in the assignment procedure (cf. also Figure 4).

Both types of discretization have different types of applications. Type-I discretization necessarily requires that the density functions of all elements of $\mathcal{M}$ have to be known, in order to able to apply Luceño's theorems to each of them. In particular, the set $\mathcal{M}$ must be dominated by the Lebesgue measure (in the measure-theoretic sense) to guarantee the existence of appropriate densities. These conditions do not apply for the second option, which therefore is more general. There it is sufficient that the reference distribution has a known density, the set $\mathcal{M}$ itself may even be undominated, which is usually for instance the case when considering neighborhood models from robust statistics (e.g., [7], [3], [20], [14]).

We will discuss both methods, from the general point of view as well as with the help of examples. For ease of illustration we will use in both examples a set of normal distributions. In Case i) it is used immediately as the credal prior information $\mathcal{M}$, in Case ii) it serves as a building block to define an appropriate interval-valued assignment.

[9] $\mathcal{P}o(\Theta_d)$ denotes the power set of Θ_d.

[10] By conjugacy ($U(\cdot) = 1 - L(\cdot^C)$) either (12) or (13) would be sufficient to describe $P(\cdot)$.

5 Applications of Type-I Discretization

We start with Type-I discretization, where $\mathcal{M}$ is a set of absolutely continuous probability distributions, to each of which the discretization procedure is applied. We assume that $\mathcal{M}$ can be described by a set $(f(\cdot)_\psi))_{\psi\in\Psi}$ of densities with the parameter space Ψ being a compact subset of $\mathbb{R}^k$ for some finite k. In this situation, for every $\pi(\cdot) \in \mathcal{M}$, the expected utility of an action can be (approximately) calculated by Equation (11), relying on the new discrete distribution ν_π. The optimal action (the action with the largest expected utility) can be found with linear optimization. It can be seen as the value of a function depending on the unknown parameter ψ. When n, the number of competing actions is small, as in the following example, E-admissible actions as well as the Γ-maximin action can be extracted graphically by plotting these functions. Section 5.2 then sketches general computational tools for complexer situations with larger n.

5.1 Numerical example

In the following the procedure should be firstly explained with the help of a numerical example. Consider the actions a_1, a_2 and a_3 with their utility functions

$$\begin{aligned} u(a_1, \vartheta) &= \exp(-\exp(\vartheta)) \\ u(a_2, \vartheta) &= \exp(-\exp(\vartheta^2)) \\ u(a_3, \vartheta) &= 0.1\,. \end{aligned}$$

The associated state of nature follow a normal distribution with $\mu = 1$ and σ varying between 0.5 and 1.5. This is an example where it is impossible to solve the corresponding integrals of the expected utility analytically, and we have to rely on Luceños method.[11]

Relying on the criterion of E-admissibility, the results can be found in Figure 1. As we are interested in the expected utility of an action in dependence on σ, which means in the value of an integral, it is reasonable to use here the simple Gaussian quadrature rule from Proposition 1 for the discretization.[12] On the left hand side one can identify the optimal action in dependence on σ, while on the right hand side the corresponding expected utility of the optimal action is shown. Two of the three actions are optimal for special values of σ; the set of E-admissible actions is $\{a_1, a_2\}$.

[11] Such integrals for instance occur when handling frailty or measurement error in survival models. Note further that with the first utility function even such common techniques like Taylor series expansion fail to calculate the integral.

[12] For Figure 1 a discretization with $N = 10$ points is used for the normal distribution with $\mu = 1$ and any fixed σ.

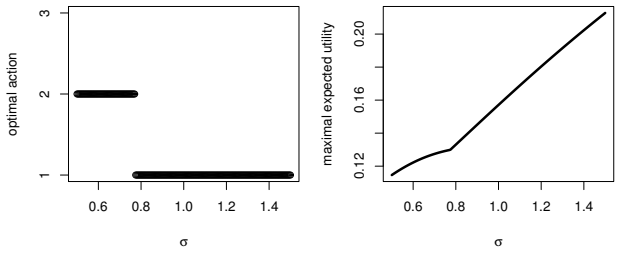


Figure 1: Action with maximal expected utility depending on σ (left), maximal possible utility depending on σ (right)

If additionally the mean μ is uncertain, the two-dimensional Figure 1 becomes three-dimensional (cf. Figure 2). The parameter μ now also varies in

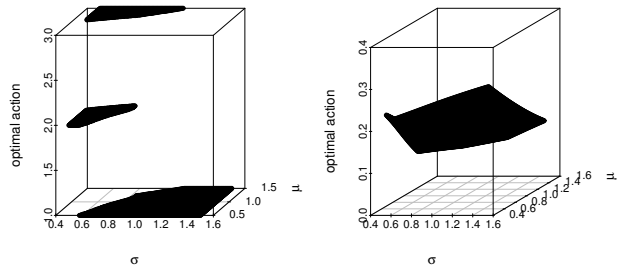


Figure 2: Action with maximal expected utility depending on σ and μ (left), maximal possible utility depending on σ and μ (right)

the interval $[0.5, 1.5]$. Like in the picture before, one can identify on the left hand side for a special μ-σ combination the optimal action. On the right there are again the corresponding expected utilities. For some μ-σ combinations now also action a_3 is optimal, which means that the set of E-admissible actions consists of all three actions. The degree of the polynomial was again chosen as 10, so that the continuous prior distribution was substituted by 10 nodes.

Also the Γ-maximin action can be found graphically. Figure 3 shows the values of the expected utility of the actions from the example in dependence on σ, which is calculated with the help of discretizations with 10 nodes. Action a_2 has the highest minimal expected utility, so it is Γ-maximin action.

5.2 General Algorithms

The method exemplified here is quite general. In more complex situations, with less smooth utility functions, or when the utility functions are very similar to each other, the number of nodes can be enlarged to obtain calculations of sufficient accuracy. This would increase the computational effort, but does not make a substantial difference. Of course, especially when the set of actions is large, graphical solutions may be insufficient, and general algorithms are needed. For that purpose, also discretize Ψ, resulting in a grid $\psi_1, \ldots, \psi_s, \ldots, \psi_Q$ of different values.[13]

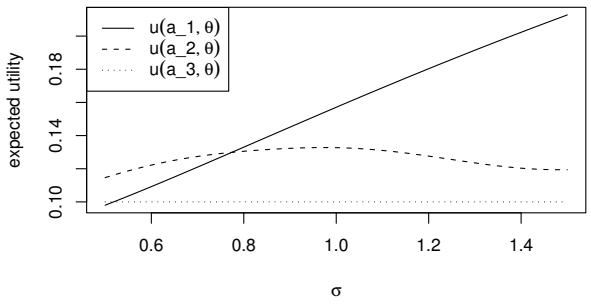


Figure 3: Expected utility depending on σ

That way a *finite* number of probability distributions from $\mathcal{M}$ is processed, each of which is discretized by one of Luceño's theorems, and so eventually a *finite set* of probability distributions with a *finite domain* is considered. In such a setting the algorithm to determine Γ−Maximin solutions described in detail in [2, Theorem 1] can be used mutatis mutandis.[14]

Also an algorithm to determine E-admissible actions can be obtained. For its construction, consider the elements $\pi_{\psi_1}(\cdot), \ldots, \pi_{\psi_Q}(\cdot)$ of $\mathcal{M}$ corresponding to the parameter values $\psi_1, \ldots, \psi_s, \ldots, \psi_Q$. Note that, with defining for all $l = 1, \ldots, n$ and $s = 1, \ldots, Q$,

$$z_{sl} := \mathbb{E}_{\pi_{\psi_s}} u(a^*, \theta) - \mathbb{E}_{\pi_{\psi_s}} u(a_l, \theta), \tag{15}$$

an action a^* is E-admissible, if

$$\exists s \; \forall l : z_{sl} \geq 0\,, \tag{16}$$

or equivalently if

$$\exists s : z_s := \min_{l=1,\ldots,n} a_l \geq 0\,, \tag{17}$$

which is the case iff the optimum $(z_1^*, \ldots, z_Q^*)$ of the following optimization problem

$$\begin{aligned} \sum_{s=1}^{Q} z_s \quad &\rightarrow \quad \max \\ z_{jl} \quad &\geq \quad z_j \quad \forall l = 1, \ldots, n, \;\; \forall j = 1, \ldots, Q, \end{aligned}$$

[13] If the parametrization of the elements of $\mathcal{M}$ is continuous, as is the case in the commonly used statistical models, no substantial loss of information should occur as long as Q is sufficiently large.

[14] Only the set $\mathcal{E}(\mathcal{M})$ arising there in Equation (16) has to be replaced by $\mathcal{Q}$, and the states of natures have to be redefined appropriately.

has a component $z^*_{s_0}$ which is non-negative. The fact that the expectations in (15) can be approximated according to (11) yields directly an algorithm based on linear optimization.

6 Application of Type-II discretization

6.1 Construction of the Discretized Prior Information

Now we turn to Type-II discretization, which produces finally an interval probability on a *finite* state of natures Θ_d with corresponding structure $\mathcal{P}$. In essence, we are now directly in a situation where we can apply algorithms from [9] and [17] to determine the E-admissible actions and the Γ-maximin action(s). To illustrate the general procedure we discuss in some detail the case where $\mathcal{M}$ is a set of parametric distributions, just as before, but now $\mathcal{M}$ is understood as the prestructure of an interval probability, i.e. our prior information consists of the lower and upper envelopes of $\mathcal{M}$, and therefore we explicitly take, for instance, also convex mixtures of elements of $\mathcal{M}$ into consideration allowing for some ambiguity in the shape of the distributions.

Firstly, in order to define the nodes, a reference distribution $\pi_0(\cdot) \in \mathcal{M}$ is chosen, which should be located in the "middle" of $\mathcal{M}$; in its neighborhood there are all possible prior distributions. As described in Section 4 this reference distribution is discretized with N nodes $x_1, \ldots, x_N$, and based on this the new space Θ_d with the elements $\varphi_1, \ldots, \varphi_N$ is obtained.[15] When constructing the interval probability $P(\cdot)$ on $(\Theta_d, \mathcal{P}o(\Theta))$, the next step is to determine the infima and suprema in (12) and (13) from $\mathcal{M}$.[16] For this purpose, for every element $\pi(\cdot) \in \mathcal{M}$ the distribution on the discretized space Θ_d has to be determined. Taking the lower envelope over all these distributions, we confine ourselves for complexity reasons to a support consisting of connected intervals (and then apply natural extension). This means we take for every $\pi(\cdot) \in \mathcal{M}$[17] the lower and upper envelope, resulting in

$$\underline{b}_{s,t} := L\Big(\bigcup_{l=s}^{t-1}\{\vartheta_l\}\Big) \quad \text{and} \quad \overline{b}_{s,t} := U\Big(\bigcup_{l=s}^{t-1}\{\vartheta_l\}\Big). \quad (18)$$

An essential building block in the whole discretization procedure are the nodes obtained by discretizing the reference distribution: Its weights are in this context much less important than the location of the nodes. The location determines the intervals ϑ_j, constituting the states, and so finally the bounds $\underline{b}_{s,t}$ and $\overline{b}_{s,t}$. So a discrete variable is aimed at, the distribution of which approximates the continuous distribution function as exactly as possible. This means, Proposition 2 should be used here.

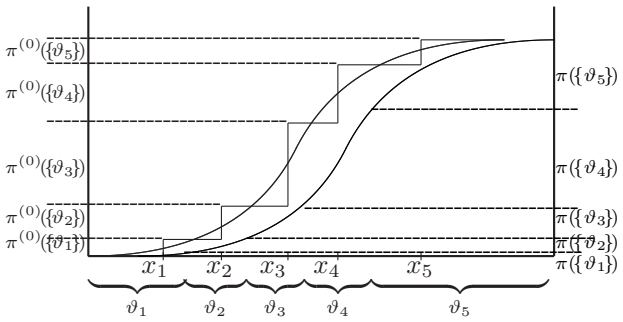


Figure 4: Finding of lower and upper probabilities with the neighborhood of a reference distribution (sketch). For details see footnote 15.

6.2 E-admissibility

If one now considers again the optimization problem determining the expected utility, one has to respect that the considered distributions are not known explicitly, they just have to satisfy the condition

$$L\Big(\bigcup_{l=s}^{t-1}\{\vartheta_l\}\Big) \leq \pi\Big(\bigcup_{l=s}^{t-1}\{\vartheta_l\}\Big) \leq U\Big(\bigcup_{l=s}^{t-1}\{\vartheta_l\}\Big).$$

Next to the auxiliary conditions for λ, i.e., $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0 \quad \forall i$, now therefore also the constraints on $\pi_d(\cdot) \in \mathcal{P}$ have to be considered.

This can be solved by adopting the algorithm developed independently by ([9]) and ([17]): With the help of linear optimization it is possible to decide in the situation with a discrete, but ambiguous state distribution whether an action a_i is E-admissible or not. For this purpose, for every action a_i, the set of all probability measures from the structure $\mathcal{P}$, for which

[15] It may be helpful to look at Figure 4, which sketches graphically several steps of the described procedure to find the lower and upper bounds. The reference distribution chosen is depicted here together with one other distribution from $\mathcal{M}$.
The curve in the middle shows the reference distribution π_0, while the steps represent its discretization with nodes $x_{j,\pi_0}, j = 1 \ldots 5$. The corresponding states ϑ_j are denoted at the abscissa; the probability masses can be seen on the left. As an example, one other distribution is discretized, the corresponding values of the $\pi(\{\vartheta_j\})$ can be read at the right side.

[16] Note that when directly an F-probability on Θ is given, for instance, by one of the neighborhood models used in robust statistics, this step can be skipped. Moreover the methods seem quite attractive to provide a further discretization when p-boxes [5] are considered.

[17] In practical calculations most often a further discretization by considering a grid analogous to 5.2 has to be used.

a_i is optimal, is considered:

$$\Pi_i = \Big\{\pi_d(\cdot) \in \mathcal{P} \Big| \sum_{j=1}^{N} u(a_i, \vartheta_j)\pi_d(\{\vartheta_j\}) \geq \sum_{j=1}^{N} u(a_l, \vartheta_j)\pi_d(\{\vartheta_j\}), \quad \forall l = 1, \ldots, n\Big\}$$

If Π_i is not empty, then there is a classical probability measure in $\mathcal{P}$ under which a_i is optimal and consequently a_i is an E-admissible action.

6.3 Γ-maximin Criterion

Linear programming can be used also for the Γ-maximin criterion. A straightforward, but inefficient possibility is, to find the Γ-maximin action with the help of $n = |\mathbb{A}|$ linear programming procedures, where the expected utility of each action is minimized, and then the action with the highest minimal value has to be found. But it is also possible (and more efficient) to find the optimal action, by considering a single optimization problem. As described in ([17]) the optimization problem:

$$\min_{\pi_d \in \mathcal{P}} \sum_{i=1}^{n} \left(\sum_{j=1}^{N} u(a_i, \theta_j)\pi_d(\{\vartheta_j\}) \right) \lambda_i \longrightarrow \max_{\lambda},$$

subject to the additional constraint $\sum_i \lambda_i = 1$, can be transformed into a single linear programming problem, either by introducing the vertices of the corresponding structure $\mathcal{P}$ or by dualization.[18] Straightforward implementations of the method with dualization and the algorithm described before for the lower and upper bounds have been used in the example below.

6.4 Numerical Example

In the following these algorithms are applied to a numerical example. Let the utility functions $u(a_i, \vartheta)$ of the actions $a_1, \ldots, a_5$ have the form:

$$\begin{aligned} u(a_1, \vartheta) &= 1 \\ u(a_2, \vartheta) &= -(\vartheta - 0.5)^2 + 2.3 \\ u(a_3, \vartheta) &= -(\vartheta + 0.75)^2 + 4.5 \\ u(a_4, \vartheta) &= -|\vartheta - 1| + 2.1 \qquad \text{and} \\ u(a_5, \vartheta) &= -\frac{(\vartheta - 1)^2}{4} + 1.5. \end{aligned}$$

Again we assume that $\mathcal{M}$ consists of all normal distributions with $\mu \in [0.75, 1.25]$ and $\sigma \in [0.75, 1.25]$, but, as discussed above, we explicitly want to allow for ambiguity concerning the type of the distribution and therefore take $\mathcal{M}$ only as a prestructure (cf. [21]), i.e. as a building block to construct an interval-valued assignment - and a corresponding structure (set of compatible distributions) - by passing over to the lower and upper envelope. Firstly the lower and upper bounds have to be found. A normal distribution with $\mu = 1$ and $\sigma = 1$ seems to be a natural choice for the reference distribution. This distribution is now discretized with a fixed number N of nodes. Accordingly normal distributions, which are inside the given bounds for μ and σ^2 are used to find the interval limits $\underline{b}_{s,t}$ and $\overline{b}_{s,t}$ in (18). In the first part of this example the discretization method based on Proposition 2 with $N = 3$ and $M = 10$, i.e. together 30 nodes, has been chosen. To find the bounds, the normal distributions with $\mu \in [0.75, 0.76, \ldots, 1.25]$ and $\sigma \in [0.75, 0.76, \ldots, 1.25]$ have been considered. Implementation of the algorithms from ([17]) yields for the criterion of E-admissibility the vector $(0, 1, 1, 1, 1)$: actions a_2, a_3, a_4 and a_5 are E-admissible under these constraints, action a_1 is not E-admissible. The optimal action under the Γ-maximin criterion is a_5 ($\lambda = (0, 0, 0, 0, 1)$) with a minimal expected utility of 1.106. For comparison, the same calculation has been made with the discretization method based on Proposition 1 with $N = 30$ nodes. The resulting values differ: just a_4 and a_5 are E-admissible actions, while a_5 is also here Γ-maximin action with a minimal expected utility of 1.087.

6.5 Notes on the Accuracy of the Results

It is certainly better to use the method of Proposition 2: it produces a new random variable with a distribution function which is more similar to the shape of the original distribution than the function of a variable produced with the ordinary Gaussian quadrature rule. Indeed the results, as seen above, are different. For explanation see the following Figure 5. It shows the differences between the application of both propositions and their consequences for finding the lower and upper bounds: The new variable produced with the method in Proposition 1 shows big differences to the distribution function of the original distribution, while the curve of the second method can hardly be distinguished from the continuous distribution (picture on the left). The number of nodes was in the first theorem $N = 60$, while in the second one with $N = 3$ and $M = 20$ was used, leading altogether again to 60.

The relatively bad approximation of the original CDF by Proposition 1 follows from the fact that in the simple Gaussian quadrature the nodes for the whole sup-

[18] A second advantage of this algorithm is the fact, that it considers also the mixed extension of the set of actions: the Γ-maximin action does not necessarily have to be a pure i.e. non-randomized action (see [2]).

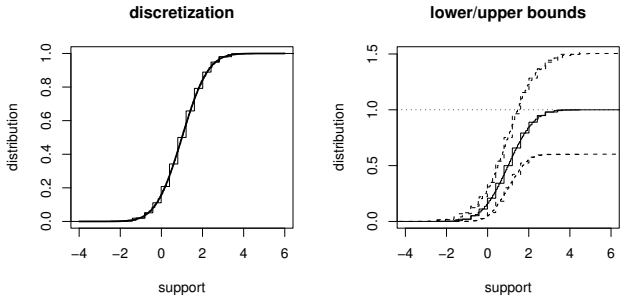


Figure 5: Discretization with Proposition 1 respectively 2 (left). Lower/upper bounds for the probabilities $\pi(\{\vartheta_j\})$ calculated with both methods (right).

port are chosen with one single polynomial. For this reason a lot of nodes are located in the less interesting outer areas of the support. And, as explained above, the locations of the nodes are more important to find the bounds $\underline{b}_{s,t}$ and $\overline{b}_{s,t}$ than the weights (which would be no problem for the method in Proposition 1). The approximation in the inner areas of the support is much exacter with the second method: with a clever choice of the intervals there are a lot of nodes in the important areas. At the end both methods lead to different values of $\overline{b}_{s,t}$'s and $\underline{b}_{s,t}$'s, whereas the values obtained by applying Proposition 2 are to be preferred as explained above. On the right hand side of Figure 5 one can see these differences in the displayed $\underline{b}_{j-1,j}$ and $\overline{b}_{j-1,j}$. The curves in the middle are the discretizations of the reference distribution, below there are the $\underline{b}_{j-1,j}$, above the $\overline{b}_{j-1,j}$.

For the differences between both methods concerning the evaluation of the Γ-maximin action watch Figure 6 which shows the results of the linear optimization for obtaining the Γ-maximin solution with both methods.

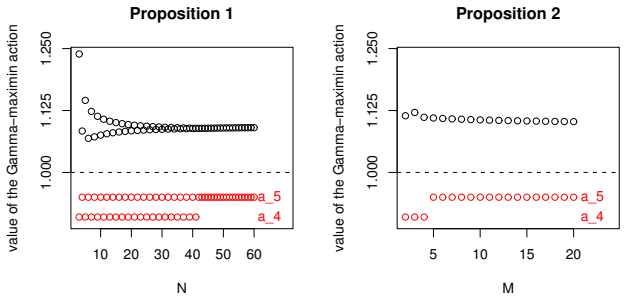


Figure 6: Γ-maximin action, found with Theorem 1 respectively Theorem 2. Upper part of the figure: expected utility value of the Γ-maximin action, lower part: which action is Γ-maximin?

In the upper part of each graph one can see the calculated minimal expected utility value of the Γ-maximin action. The lower part shows the corresponding action. The results based on Proposition 1 oscillate at the beginning between a_5 and a_4 and stay finally stable at action a_5. Also with the method of Proposition 2 (with $N = 3$) the Γ-maximin action at the beginning is a_4. But relatively quickly, from $M = 4$ on, which is discretization with 12 nodes, a_5 stays the optimal solution.

The expected values oscillate with both methods in the same way. But with a high number of nodes the results become more stable. As explained above, the results of Proposition 2 are better in a case, where the results differ.

7 Concluding Remarks

We have discussed a sophisticated method for discretizing a continuous random variable. In contrast to straightforward ad-hoc discretizations, for instance by rounding, one is able to enforce important relations between both variables: the discrete variable and the continuous variable have a given number of moments in common, and also their distribution functions can be ensured to coincide in a certain set of points.

In our view this makes the method quite attractive in imprecise probability theory far beyond decision theory, where we have exemplified the power of the method in two typical applications. Further fruitful areas of application include the calculation of posterior probabilities from the generalized Bayes rule for continuous distributions and the extension of graphical models based on continuous distributions to imprecise probabilities.

Of course, the presentation given here is mainly an exploratory sketch of some basic ideas. Deeper investigations are urgently needed in order to find general recommendations on the trade-off between complexity and accuracy of the approximation. In this respect, also special attention has to be paid to the utility function, in particular when it is not smooth.[19] Another important detail is the sensitivity of the results with respect to the choice of the reference distribution. In rare case only, like the application in neighborhood models (see the survey in [3, Section 4] as well as [7], [20], [14]), there is a unambiguously natural candidate, and canonical examples providing well-accepted recommendations have still to be developed.[20]

[19] One referee suggested to utilize duality of utility and probability for this purpose and to take discreteness of the utility function explicitly into account as well.

[20] One general idea in this direction we owe a referee, who suggested to choose that distribution in $\mathcal{M}$ which minimizes

8 Acknowledgement

We wish to thank all three referees for their helpful and very stimulating remarks.

References

[1] A. S. Ahmad. *Statistical Analysis of Heaping and Rounding Effects*. Dissertation. Department of Statistics, Univ. of Munich, 2006.

[2] T. Augustin. Expected Utility within a Generalized Concept of Probability - a Comprehensive Framework for Decision Making under Ambiguity. *Stat. Pap.*, 43: 5–22, 2002a.

[3] T. Augustin. Neyman-Pearson Testing under Interval Probability by Globally Least Favorable Pairs. Reviewing Huber-Strassen Theory and Extending it to General Interval Probability. *J. Stat. Plan. Inf.*, 105: 149–173, 2002b.

[4] F. G. Cozman. Credal Networks. *Artif. Intell.*. 120: 199–233, 2000.

[5] S. Ferson et al. *Constructing probability boxes and Dempster-Shafer structures*. Technical Report (SAND 2002-4015), Sandia National Laboratories, 2003.

[6] W. Gautschi. Algorithm 726: ORTHPOL – A Package of Routines for Generating Orthogonal Polynomials and Gauss-Type Quadrature Rules. *ACM Trans. Math. Software*, 20: 21–62, 1994.

[7] P. J. Huber and V. Strassen. Minimax Tests and the Neyman-Pearson Lemma for Capacities. *Ann. Stat.*, 1: 251–263, Correction, 2: 223–224, 1973.

[8] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics, Part 2*, 2nd edition, D. Van Nostrand Company, INC, 1951.

[9] D. Kikuti, F. G. Cozman and C. P. de Campos. Partially Ordered Preferences in Decision Trees: Computing Strategies with Imprecision in Probabilities, In: R. Brafman, U. Junker (eds.): Multidisciplinary IJCAI-05 Workshop on Advances in Preference Handling, Edinburgh, Scotland, 118–123, 2005. see also `http://wikix.ilog.fr/wiki/pub/Preference05/WsProceedings/Pref05.pdf`

[10] A. Luceño. Discrete Approximations to Continous Univariate Distributions – an Alternative to Simulation. *J. Roy. Stat. Soc.*, B 61: 345–352, 1999.

[11] M. Obermeier. *Luceños Diskretisierungstechnik für stetige Verteilungen — Implementierung und Anwendung*. Diploma thesis, Dep. of Statistics, Univ. of Munich, 2006.

[12] R.A. Sack and A.F. Donavan. An Algorithm for Gaussian Quadrature given Modified Moments. *Numer. Math.*, 18: 465–478, 1971/72.

[13] W. F. Sheppard. On the Calculation of the most Probable Values of Frequency Constants for Data Arranged according to Equidistant Division of a Scale. *Proc. London Math. Soc.*, 29: 353–380, 1898.

[14] D. Skulj. Jeffrey's Conditioning Rule in Neighbourhood Models. *Int. J. Approx. Reasoning*, 42: 192–211, 2006.

[15] J. Stoer and R. Bulirsch. *Numerische Mathematik I*. Springer, 1999.

[16] M. C. M. Troffaes. Decision Making under Uncertainty using Imprecise Probabilities. *Int. J. Approx. Reasoning*, in press, 2007.

[17] L. V. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In: F.G. Cozman, R. Nau and T. Seidenfeld (eds.) *Proc. of the 4th Int. Symposium on Imprecise Probabilities and their Applications* (ISIPTA '05), 349–358, 2005.

[18] S. B. Vardeman. Sheppard's Correction for Variances and the "Quantization Noise Model". *IEEE Trans. Instr. Meas.*, 54: 2117–2119, 2005.

[19] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[20] A. Wallner. Bi-elastic Neighbourhood Models. In: J.-M. Bernard, T. Seidenfeld, and M. Zaffalon (eds) *Proc. of the 3rd Int. Symposium on Imprecise Probabilities and their Applications (ISIPTA '03)*, 593–607, 2003.

[21] K. Weichselberger. *Elementare Grundbegriffe einer allgemeinen Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physica, 2001.

[22] J. C. Wheeler. Rocky Mountain. *J. Math.*, 4: 287–296, 1974.

the maximal Kullback-Leibler distance over $\mathcal{M}$.

Some Bounds for Conditional Lower Previsions

Renato Pelessoni
University of Trieste
renato.pelessoni@econ.units.it

Paolo Vicig
University of Trieste
paolo.vicig@econ.units.it

Abstract

In this paper we consider some bounds for lower previsions that are either coherent or centered convex. As for coherent conditional previsions, we adopt a structure-free version of Williams' coherence, which we compare with Williams' original version and with other coherence concepts. We then focus on bounds concerning the classical product and Bayes' rules. After discussing some implications of product rule bounds, we generalise a well-known lower bound, which is a (weak) version for coherent lower probabilities of Bayes' theorem, to the case of (centered) convex previsions. We obtain a family of bounds and show that one of them is undominated in all cases.

Keywords. Conditional lower previsions, product rule, Bayes' theorem, Williams' coherence, centered convex previsions.

1 Introduction

Quite recently, P.M. Williams' 1975 seminal paper *Notes on conditional previsions* was published in a slightly revised version [21], preceded by an introductory paper discussing basic aspects and historical motivations for his work [14]. This fact confirms that Williams' ideas on coherence still play a very important role in the theory of conditional imprecise previsions.

One of the aims of this paper is to show that Williams' coherence, while being more general than other coherence concepts that have been developed, may be quite simple to work with in several problems. Precisely, we shall use a variant of Williams' original coherence which does not impose any structural constraint on the set of conditional (bounded) random variables forming the domain of the lower prevision $\underline{P}$ and which is a generalisation of Walley's coherence for unconditional (bounded) random variables (or gambles) [16].

After recalling some preliminary notions in Section 2, we discuss this variant in Section 3, comparing it firstly with Williams' original version and then with other generalisations of Walley's unconditional coherence, either potential or proposed in [16]. When being equivalent to the notion of coherence mainly adopted by Walley in [16], as is the case in the sequel of the paper, Williams' coherence may be conveniently used to prove certain results, which therefore hold in Walley's approach too.

We shall use Williams' coherence to study some bounds for conditional lower previsions. Actually we prove that several results hold also for previsions that are (centered) convex, i.e. satisfy a consistency notion (introduced in [10]) which is more general than Williams' coherence.

We focus on generalisations of product rule and Bayes' rule bounds together with other bounds which we termed sign rules. A motivation for investigating all these bounds is that they may give us some guidance for extending coherent or convex lower previsions. This is particularly relevant when conditioning, given that many rules or standard procedures for inferences or anyway for getting unconditional coherent evaluations do not apply in a conditional framework (for instance, convex combinations of coherent conditional lower previsions are not necessarily coherent).

In Section 4 we discuss some inequalities (product and sign rules), which are essentially known, exploring some of the implications they have for extending $\underline{P}$ under an epistemic irrelevance assumption. It appears here that when the product rule may hold with equality, the lower prevision obtained from this equality is not necessarily the natural extension as in the case of events, but may also coincide with the opposite concept of upper extension. In Section 5 we generalise the well-known lower bound $\underline{P}(A|B) \geq \frac{\underline{P}(A \wedge B)}{\underline{P}(A \wedge B) + \overline{P}(\overline{A} \wedge B)}$ to the case of conditional random variables and of lower previsions that are Williams' coherent or, more

generally, centered convex. We derive a family of bounds, proving that one of them, given by equation (11), is the best in all cases.

Section 6 contains some further comments and conclusions.

2 Preliminaries

In the sequel, $\mathcal{D}$ is an *arbitrary* (non-empty) set of *bounded* random variables (also termed gambles [16] or random quantities [21]), or more generally of bounded conditional random variables.

In the conditional case, if $X|B \in \mathcal{D}$, X is a random variable and B a non-impossible event. When $B = \Omega$, we obtain the (unconditional) random variable $X = X|\Omega$.

The supremum $\sup(X|B)$ of $X|B$ may be computed as $\sup_{\omega \Rightarrow B} X(\omega)$ ($\sup_{\omega \in B} X(\omega)$ in the set-theoretic interpretation of events), where all ω belong to a large enough partition (possibility space) $I\!P$. It will be also denoted as $\sup_B X$. Analogously, $\inf(X|B) = \inf_B X$.

We write B for both an event B and its indicator function $|B|$ (de Finetti's convention), appearing from the context which of the two meanings is intended.

A *lower prevision* $\underline{P}$ on $\mathcal{D}$ is a map $\underline{P} : \mathcal{D} \to \mathbb{R}$. An upper prevision $\overline{P}$ may be defined through the equality $\overline{P}(-X) = -\underline{P}(X)$, which always lets us refer to either lower or upper previsions only. A *precise* prevision P is the special case $\overline{P}(X) = \underline{P}(X) = P(X)$.

The consistency notions we shall consider for $\underline{P}$ are those of *coherence* or *(centered) convexity*. More specifically, when $\mathcal{D}$ is made of unconditional random variables, $\underline{P}$ is said to be *coherent* when satisfying the definition in [16], sec. 2.5.4 (a):

Definition 1 *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a* coherent lower prevision *on $\mathcal{D}$ iff, for all $n \in \mathbb{N}^+$, $\forall\ X_0, X_1, \ldots, X_n \in \mathcal{D}$, $\forall\ s_0, s_1, \ldots, s_n$ real and non-negative, defining $\underline{G} = \sum_{i=1}^{n} s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$, $\sup \underline{G} \geq 0$.*

This definition has a well-known behavioural interpretation: $\underline{P}(X)$ is an agent's supremum buying price for X, and $\underline{G}$ is the agent's *gain* resulting from her/his buying $s_i X_i$, for $i = 1, \ldots, n$, and selling $s_0 X_0$. We shall use this terminology too, saying that the agent *bets* on $X_0, \ldots, X_n$ with *stakes* $s_0, \ldots, s_n$ respectively.

In a conditional environment, we adopt the following generalisation of Definition 1 to define a *coherent* $\underline{P}(\cdot|\cdot)$:

Definition 2 *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a* coherent conditional lower prevision *on $\mathcal{D}$ iff, for all $n \in \mathbb{N}^+$, $\forall X_0|B_0, \ldots, X_n|B_n \in \mathcal{D}$, $\forall\ s_0, s_1, \ldots, s_n$ real and non-negative, defining $B = \bigvee_{i=0}^{n} B_i$ and $\underline{G} = \sum_{i=1}^{n} s_i B_i(X_i - \underline{P}(X_i|B_i)) - s_0 B_0(X_0 - \underline{P}(X_0|B_0))$, $\sup(\underline{G}|B) \geq 0$.*

Here the gain is $\underline{G}|B$, a conditional random variable itself. Conditioning on B has the meaning of considering only those values for $\underline{G}$ when at least one of $B_0, \ldots, B_n$ is true. It is easy to realise that we would get an equivalent definition (adopted in [18]) by replacing $\underline{G}|B$ with $\underline{G}|S$, where the *support* S is defined as $S = \vee\{B_i : s_i \neq 0, i = 0, \ldots, n\}$.

Throughout the paper, Definition 2 will be referred to as Williams' coherence, or *W-coherence* or simply coherence, but as we will explain in Section 3, it is actually a structure-free version of the original Williams' coherence.

A weaker notion than W-coherence is that of lower prevision that *avoids uniform loss* [16, 18]. It may be obtained from Definition 2 by ruling out the bet on $X_0|B_0$ and modifying B and $\underline{G}$ accordingly. In the unconditional environment it is termed condition of *avoiding sure loss* and is defined in [16], Sec. 2.4.4 a).

The consistency notion of *centered convexity* [10, 11] is weaker than coherence, but sufficiently stronger than the conditions of avoiding sure or uniform loss to allow for interesting properties and applications (for instance, in risk measurement [10]). In fact, several of the results in the next sections apply to centered convex previsions too.

Formally, the definition of *convex lower prevision* is obtained from Definition 1 and Definition 2 by introducing just the extra *convexity constraint* $\sum_{i=1}^{n} s_i = s_0$ (> 0) and eventually by further imposing (this is not restrictive) that $s_0 = 1$ [9, 10]. Again, we could condition $\underline{G}$ on its support S rather than on B, getting an equivalent definition of convex conditional lower prevision. This is done in [10, 11]. *Centered* convexity requires in addition that ($0 \in \mathcal{D}$ and) $\underline{P}(0) = 0$ in the unconditional case, and further that $\forall X|B \in \mathcal{D}$, $0|B \in \mathcal{D}$ and $\underline{P}(0|B) = 0$ in the conditional case.

Centering is quite a natural requirement: non-centered convex previsions have rather weak consistency properties, but special instances of them may be found in the risk literature (cf. [10]).

Proposition 1 *If $\underline{P} : \mathcal{D} \to \mathbb{R}$ is centered convex, then necessarily [10]:*

P1) $\inf X \leq \underline{P}(X) \leq \sup X$ *(internality);*

P2) $Y \leq X \Rightarrow \underline{P}(Y) \leq \underline{P}(X)$, $\forall X, Y \in \mathcal{D}$ *(monotonicity);*

P3) $\underline{P}(\lambda X + (1-\lambda)Y) \geq \lambda \underline{P}(X) + (1-\lambda)\underline{P}(Y)$, $\forall X, Y \in \mathcal{D}$, $\forall \lambda \in [0, 1]$.

These properties obviously hold for coherent lower previsions too, while P1) might fail for non-centered convex previsions.

Let $\underline{P}$ be a lower prevision defined on an arbitrary set $\mathcal{D}$. Any consistency condition satisfied by $\underline{P}$ should guarantee that there exists an extension of $\underline{P}$ on any $\mathcal{D}' \supset \mathcal{D}$ which satisfies the same consistency condition. If such an extension is not unique, its vaguest or least-committal one, if existing, has a special importance. This peculiar extension is the *natural extension* $\underline{E}$ in the case of coherent or, when conditioning, W-coherent previsions [14, 16, 21], the *convex* natural extension $\underline{E}_c$ for centered convex (unconditional or conditional) previsions [9, 10]. The natural or convex natural extensions always exist for these consistency notions, not necessarily with other ones, like Walley-coherence in [16], Section 7.1.4 (b), or non-centered convexity. Hence, the consistency notions we shall be working with always allow for extensions of the same kind on any superset: we shall often use this fact in the proofs of the results, without always mentioning explicitly that we are performing an extension.

When working with conditional random variables, like $\underline{G}|B$, we shall employ the equality

$$f(X_1, \ldots, X_n)|B = f(X_1|B, \ldots, X_n|B) \qquad (1)$$

where f is any real function [3].

3 Two or Three Things on Williams' Coherence

3.1 About Williams' definition

Williams' original definition ([21], Definition 1) differs formally from our definition of W-coherence. One reason is that it refers to upper rather than lower previsions, but this is unimportant, since using the conjugacy relation $\overline{P}(-X) = -\underline{P}(X)$ our condition $\sup \underline{G}|B \geq 0$ corresponds exactly to his inequality in (A^*) of [21]. The true difference is that his notion is not completely structure-free, as it asks that for every $X|B$ in $\mathcal{D}$, $\underline{P}(X|B)$ is assigned for any X in a linear space $\mathcal{X}_B$. It follows for instance that Williams' definition does not formally generalise Walley's coherence for unconditional previsions (our Definition 1), which is structure-free: when $B = \Omega$ for all $X|B \in \mathcal{D}$, the set of all X is constrained to form a linear space $\mathcal{X}_\Omega$. On the contrary, Definition 2 is in particular a generalisation of Walley's unconditional coherence and appears to be, in general, nimbler. For instance, the bounds in Section 4 involve just a few random variables and no structure is actually needed for proving them. The fundamental link between the two versions of Williams' coherence is ensured by the following *extension theorem*.

Proposition 2 *If $\underline{P} : \mathcal{D} \to \mathbb{R}$ is W-coherent on $\mathcal{D}$ (according to Definition 2), it has a W-coherent extension on* any *$\mathcal{D}' \supset \mathcal{D}$.*

Although we are not aware of any published proof for this proposition, nevertheless it should be regarded as essentially known. In fact, it can be proven by adapting the proofs concerning the convex natural extension in [10], thus proving that there *always* exists the natural extension of a W-coherent lower prevision on any $\mathcal{D}' \supset \mathcal{D}$. Alternatively, the scheme of de Finetti's extension theorem can be followed, with suitable (but basically minor) modifications. After de-Finetti's path-breaking proof concerning precise (unconditional) previsions in [6], this scheme was employed in several generalisations (see e.g. [1, 4]). Its two-step proof shows in the first step that there exist W-coherent extensions on $\mathcal{D}' = \mathcal{D} \cup \{X|B\}$, $\forall X|B$, while the second step generalises the proof to any $\mathcal{D}'$ using Zorn's lemma or equivalent results. A by-product of the first step is that the set of admissible W-coherent extensions on $X|B$ is proved to be a closed interval. Its lower endpoint is the *natural extension* $\underline{E}(X|B)$, while the upper endpoint is the *upper extension* $\underline{U}(X|B)$ of $\underline{P}$. We shall meet again upper extensions in Section 4.

As an important implication of Proposition 2 in our framework, when $\mathcal{D}$ in Definition 2 does not meet Williams' structure requirements in his definition it is always possible to coherently extend $\underline{P}$ on a set $\mathcal{D}'$ such that these requirements hold, and there the two notions of coherence coincide. It follows that our W-coherent lower previsions have all the properties established for Williams' coherence in [21], including the important *envelope theorem*, stating that $\underline{P}$ is coherent on $\mathcal{D}$ if and only if

$$\underline{P}(X|B) = \inf_{P \in \mathcal{M}} P(X|B), \forall X|B \in \mathcal{D}$$

where $\mathcal{M}$ is the set of the coherent precise previsions $P(\cdot|\cdot)$ dominating $\underline{P}(\cdot|\cdot)$ on $\mathcal{D}$ ($P(X|B) \geq \underline{P}(X|B), \forall X|B \in \mathcal{D}$).

3.2 Alternative concepts of coherence

Another issue concerning Definition 2 of W-coherence is: equivalent formulations of Definition 1 are known, so why not rather generalise them in a conditional environment? An answer is that Definition 2 seems more appropriate for further generalisations. In fact, an equivalent version of coherence in Definition 1 is obtained by restricting the stakes $s_0, \ldots, s_n$ to be integer (this is Walley's Definition 2.5.1 in [16]), and

this can be done in a conditional environment too. However, considering integer combinations only is not sufficient when the random numbers are unbounded, even in the unconditional case, as shown in [12].

Another definition, less used [1] but equivalent to Definition 1, is:

Definition 3 *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a coherent lower prevision on $\mathcal{D}$ iff, for all $n \in \mathbb{N}^+$, $\forall\ X_0, X_1, \ldots, X_n \in \mathcal{D}$, $\forall\ s_1, \ldots, s_n \geq 0$, $\forall\ \lambda_0 \in \mathbb{R}$ such that $X_0 \geq \sum_{i=1}^n s_i X_i + \lambda_0$, it holds that $\underline{P}(X_0) \geq \sum_{i=1}^n s_i \underline{P}(X_i) + \lambda_0$.*

To the best of our knowledge, no generalisation of this definition to a conditional environment is available in the literature, nor does the problem of generalising it to an equivalent version of W-coherence seem to have a straightforward solution.

A further issue is that a number of different generalisations of coherence (Definition 1 or equivalent) to a conditional framework have been proposed in [16]: how do they relate to W-coherence? We observe some basic facts about it.

a) The generalisations in [16] are not structure-free: the conditioning events have some special features. *When being comparable*, W-coherence in Definition 2 is *equivalent* to the following two of them:

   - the concept of *coherence* defined in Sec. 7.1.4 (b) (referred to as *Walley-coherence* here), with the extra assumption that all partitions $\mathcal{B}_i$ in that definition are finite (this equivalence is stated (without proof) in [16]);
   - the concept of *separate coherence* defined in Sec. 6.2.2, *without* any other extra assumption (this equivalence is proved in the Appendix).

b) In general, W-coherence may be weaker than Walley-coherence (when at least one $\mathcal{B}_i$ is infinite). This fact may lead to the disadvantages discussed in [16], but has the non-negligible advantages over Walley-coherence that the natural extension always exists and that the envelope theorem characterises W-coherence. A weaker notion than Walley-coherence, *weak coherence* defined in Sec. 7.1.4 (a), is sometimes stronger and sometimes weaker than W-coherence. This notion is anyway rather counterintuitive, and Walley-coherence is in fact the major conditional coherence condition in [16].

c) At any rate, properties of W-coherence involving only finitely many distinct conditioning events hold for Walley-coherence too (a W-coherent assessment or possibly one of its W-coherent extensions, cf. Proposition 2, may be referred in this case to a finite set of finite partitions $\mathcal{B}_i$). In particular, the bounds we investigate later on hold in Walley's framework too.

Last but not least, we note that the notion of conditional random variable (and of conditional event) is often left at an informal level in the literature, including [16, 21]. A formal approach to these and other descriptive tools of uncertainty is developed in [3, 4].

Although this issue is seemingly not particularly relevant in many matters, a greater formalisation turns out to be useful with other ones. For an example, consider Lemma 6.2.4 in [16]: this lemma states that, if $BX = BY$ and other coherence conditions hold for a lower prevision $\underline{P}$, then $\underline{P}(X|B) = \underline{P}(Y|B)$. But using (1), $BX|B = (B|B) \cdot (X|B) = X|B$, thus condition $BX = BY$ alone implies $X|B = Y|B$. Consequently $\mu(X|B) = \mu(Y|B)$ whatever the uncertainty measure μ is, not because of coherence (μ could even be incoherent), but merely because we are evaluating the same thing.

4 Product and Sign Rules

The *product rule* is among the basic inferential rules in Bayesian statistics. In its simplest version for probabilities, it requires that $P(A \wedge B) = P(A) \cdot P(B|A)$; in a more general version involving a precise prevision P, events A and B and a random variable X, we have $P(AX|B) = P(A|B) \cdot P(X|A \wedge B)$.

We investigate now some generalisations of this rule, and related properties, for coherent lower previsions.

Proposition 3 *Let $\underline{P}$ be coherent on $\mathcal{D} \supset \{AX|B, A|B, X|A \wedge B\}$. Then, necessarily:*

a) (product rule) *if $\underline{P}(X|A \wedge B) > 0$, then*

$$\underline{P}(AX|B) \geq \underline{P}(A|B) \cdot \underline{P}(X|A \wedge B) \tag{2}$$

b) (product rule) *if $\underline{P}(X|A \wedge B) < 0$, then*

$$\underline{P}(AX|B) \leq \underline{P}(A|B) \cdot \underline{P}(X|A \wedge B) \tag{3}$$

c) $\underline{P}(AX|B) = 0$ iff $\underline{P}(A|B) \cdot \underline{P}(X|A \wedge B) = 0$

[1] Definition 3 has a curios story: not mentioned explicitly in Walley's book [16], although following directly from results established there, it appears in [2], but without being related to coherence for imprecise previsions. It was then discussed extensively in [7].

d) (sign rules)

$\underline{P}(AX|B) > 0 \Rightarrow \underline{P}(X|A \wedge B) > 0;$

$\underline{P}(AX|B) < 0 \Rightarrow \underline{P}(X|A \wedge B) < 0;$

Proof. Put $p_1 = \underline{P}(AX|B)$, $p_2 = \underline{P}(A|B)$, $p_3 = \underline{P}(X|A \wedge B)$, and consider a gain $\underline{G}$ in Definition 2 arising from betting on $AX|B$, $A|B$, $X|A \wedge B$: $\underline{G} = s_1B(AX - p_1) + s_2B(A - p_2) + s_3AB(X - p_3) = ((s_1 + s_3)AX + (s_2 - s_3p_3)A - s_1p_1 - s_2p_2)B$. Now choose s_1, s_2, s_3 such that

$$s_1 = -s_3, s_2 = s_3p_3 \tag{4}$$

and $\underline{G}$ specialises into

$$\underline{G} = (p_1 - p_2p_3)s_3B. \tag{5}$$

Proof of a). We have $p_3 > 0$. Choose $s_3 > 0$. Then from (4) $s_1 < 0$, $s_2 > 0$. Since only one of the stakes s_1, s_2, s_3 is negative, we have an admissible bet according to Definition 2. To ensure $\sup \underline{G}|B \geq 0$ it is necessary from (5) that $p_1 - p_2p_3 \geq 0$, which is (2).

Proof of b). Analogous to a), after choosing $s_3 < 0$.

Proof of c). To prove the implication $\underline{P}(X|A \wedge B) \cdot \underline{P}(A|B) = 0 \Rightarrow \underline{P}(AX|B) = 0$, note that when $\underline{P}(X|A \wedge B) \cdot \underline{P}(A|B) = p_2p_3 = 0$ the gain $\underline{G}$ in (5) reduces to $\underline{G} = s_3p_1B$, and $\underline{G}|B = s_3p_1$. To ensure $\sup \underline{G}|B \geq 0$, whatever the sign of s_3 may be, it is necessary that $p_1 = \underline{P}(AX|B) = 0$. The proof of the converse implication is similar.

Proof of d). For the first implication, suppose $\underline{P}(AX|B) > 0$. Then $\underline{P}(X|A \wedge B)$ can be neither negative (since then b) would contradictorily imply $\underline{P}(AX|B) \leq 0$), nor zero (c) would imply $\underline{P}(AX|B) = 0$). Hence $\underline{P}(X|A \wedge B) > 0$. The other implication is proven similarly. ■

4.1 Comments

The sign rules are obtained here from the product rule. A simpler version of the first rule holds for convex previsions too, and may be derived from Lemma 1 (cf. Section 5.1). Sign rules introduce some rough inferential constraints. For instance, let $B = \Omega$. Then knowing or assuming that $\underline{P}(AX) > 0$ implies necessarily $\underline{P}(X|A) > 0$ (no matter what sign $\underline{P}(X)$ has).

The product rule has interesting implications, involving the natural and upper extension. To outline this point, let $B = \Omega$, and suppose that A is *epistemically irrelevant* for X, so $\underline{P}(X|A) = \underline{P}(X)$. If we have assessed $\underline{P}(A)$ and $\underline{P}(X)$, but not $\underline{P}(AX)$, it is tempting to extend $\underline{P}$ on AX putting $\underline{P}(AX) = \underline{P}(A) \cdot \underline{P}(X)$ (multiplicative rule). There are instances when this is possible: if X is an event too, under an additional assumption (logical independence of A and X); in this case $\underline{P}(AX)$ is the natural extension $\underline{E}(AX)$ [13]. These properties do not necessarily hold if we further introduce some constraints on $\underline{P}$. For instance, it was shown in [8] that the multiplicative rule holds only in very special cases if we require $\underline{P}$ to be a necessity measure. However, as long as only events are involved, we can hope to simultaneously apply $\underline{P}(AX) = \underline{P}(A) \cdot \underline{P}(X)$ and obtain the natural extension $\underline{E}(AX) = \underline{P}(AX)$. Proposition 3 informs us that in the realm of random variables the situation is more complex: even assuming that $\underline{P}(AX) = \underline{P}(A) \cdot \underline{P}(X)$ is a coherent extension of $\underline{P}$ on AX, there are instances (cf. a)) when $\underline{P}(AX) = \underline{E}(AX)$, but other conditions (cf. b)) imply that $\underline{P}(AX)$ is just the opposite, i.e. the upper extension of $\underline{P}$.[2] This happens in particular when $\sup X < 0$ (hence $\underline{P}(X) < 0$ by P1) of Proposition 1), or also $X \leq 0$ if $\underline{P}(X) \neq 0$.

Finally, note that some sign constraints arise as a joint consequence of *a*), *b*), *c*), depending on the sign of $\underline{P}(A|B)$: if $\underline{P}(A|B) > 0$ then $\underline{P}(AX|B)$ and $\underline{P}(X|A \wedge B)$ must take the same sign (both positive, both negative, or both null), while if $\underline{P}(A|B) = 0$ then $\underline{P}(AX|B) = 0$, but $\underline{P}(X|A \wedge B)$ is unconstrained.

5 Bayes' Rule Bounds for Centered Convex Previsions

The following inequality, which holds if $\underline{P}(B) > 0$ and its terms are well-defined, is well-known in the theory of coherent imprecise probabilities [15, 16, 17]:

$$\underline{P}(A|B) \geq \frac{\underline{P}(A \wedge B)}{\underline{P}(A \wedge B) + \overline{P}(\overline{A} \wedge B)} \tag{6}$$

Together with an analogous bound, eq. (6) generalises Bayes' theorem for precise probabilities (when $\underline{P} = \overline{P} = P$ it reduces to $P(A|B) \geq P(A \wedge B)/P(B)$. The reverse inequality may be obtained from $\overline{P}(A|B) \leq \overline{P}(A \wedge B)/(\overline{P}(A \wedge B) + \underline{P}(\overline{A} \wedge B)))$. In fact, an immediate, inferential way of interpreting (6) is to suppose that an unconditional coherent $\underline{P}$ is assigned: then (6) gives a lower bound for extending $\underline{P}$ on $A|B$, hence also a lower bound for its natural extension $\underline{E}(A|B)$. It is well-known [15] that when $\underline{P}$ is defined on an algebra $\mathcal{A}$ and is 2-monotone there (i.e. $\underline{P}(A \vee B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \wedge B)$, $\forall A, B \in \mathcal{A}$), the bound in (6) is precisely equal to $\underline{E}(A|B)$, which under these assumptions may be written in terms of Choquet integrals. Inequality (6) was also studied in various other papers, including [17], where it is also compared with Dempster's rule of conditioning, and [19].

[2] Upper extensions received little attention in [16], while they were investigated in [20].

Our main purpose in this section is to generalise eq. (6) introducing a more general bound, holding for random variables with corresponding lower previsions that are centered convex. For this, we need a preliminary Lemma, which has also other implications, commented below.

Lemma 1 *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$. Whenever the lower previsions below are defined,*

a) if $\underline{P}$ is convex on $\mathcal{D}$, then for $\lambda \in \mathbb{R}$

$$\underline{P}(B(X-\lambda)) > 0 \Rightarrow \underline{P}(X|B) > \lambda;$$
$$\underline{P}(X|B) > \lambda \Rightarrow \underline{P}(B(X-\lambda)) \geq 0;$$

b) if $\underline{P}$ avoids uniform loss on $\mathcal{D}$, then for $\lambda \in \mathbb{R}$,

$$\underline{P}(B(\lambda - X)) > 0 \Rightarrow \underline{P}(X|B) < \lambda.$$

Proof. To prove a), write the gain $\underline{G}|\Omega = \underline{G}$ for a bet on $B(X-\lambda)$, $X|B$ with stakes $s_1 = s_0$ (note that $s_1 = s_0$ is the convexity condition in this case): $\underline{G} = s_1(B(X-\lambda) - \underline{P}(B(X-\lambda))) - s_1 B(X - \underline{P}(X|B)) = s_1(B(\underline{P}(X|B) - \lambda) - \underline{P}(B(X-\lambda)))$.

To prove the first inequality, put $s_1 = 1$. To ensure $\sup \underline{G} \geq 0$ (note that $\underline{G}$ varies only with B), the following inequality must be false for at least one value of B: $B(\underline{P}(X|B) - \lambda) < \underline{P}(B(X-\lambda))$. If $\underline{P}(B(X-\lambda)) > 0$, then necessarily $\underline{P}(X|B) - \lambda > 0$.

To prove the second inequality, put $s_1 = -1$. To guarantee now that $\sup \underline{G} \geq 0$, the reversed inequality $B(P(X|B) - \lambda) > \underline{P}(B(X-\lambda))$ must be false. If $\underline{P}(X|B) > \lambda$, it is necessary for this that $\underline{P}(B(X - \lambda)) \geq 0$.

To prove b), consider the bet on $B(\lambda - X)$, $X|B$ with gain $\underline{G} = B(\lambda - X) - \underline{P}(B(\lambda - X)) + B(X - \underline{P}(X|B))$, and argue similarly to the preceding cases. ■

Corollary 1 *Under the assumptions of Lemma 1, b), $\overline{P}(B(X-\lambda)) < 0 \Rightarrow \underline{P}(X|B) < \lambda$.*

Proof. Follows from Lemma 1, b) and $\underline{P}(B(\lambda - X)) = -\overline{P}(B(X-\lambda))$. ■

5.1 Comments

Only the first inequality in a) will be actually used to generalise (6), but the three inequalities deserve some comments. The inequalities in a) imply when $\lambda = 0$ a simpler version of the first inequality in Proposition 3 d) (sign rules), but holding under the weaker assumption that $\underline{P}$ is convex. As for the inequality in b), it holds also for centered convex previsions, since these previsions avoid uniform loss [10].

5.2 A Generalised Lower Bound

We obtain now a generalisation of the lower bound (6).

Proposition 4 *Let $\underline{P}$ be an unconditional centered convex lower prevision on $\mathcal{D} \supset \{B, B(X - \sup(X|B)), B(X - \inf(X|B))\}$ and $\underline{P}(B) > 0$. If $\underline{P}(B(X - \inf(X|B))) - \underline{P}(B(X - \sup(X|B))) \neq 0$, any (centered) convex extension of $\underline{P}$ on $X|B$ is such that, $\forall h \leq \inf(X|B)$, $\forall k \geq \sup(X|B)$,*

$$\underline{P}(X|B) \geq \phi(h,k) = \frac{k\underline{P}(B(X-h)) - h\underline{P}(B(X-k))}{\underline{P}(B(X-h)) - \underline{P}(B(X-k))} \qquad (7)$$

Proof. We preliminarily observe that the denominator in (7) is positive. This follows from the assumptions and internality and monotonicity of $\underline{P}$ (Proposition 1, P1) and P2)), which imply: $B(X-h) \geq 0 \Rightarrow \underline{P}(B(X-h)) \geq 0$, $B(X-k) \leq 0 \Rightarrow \underline{P}(B(X-k)) \leq 0$, and then $0 < \underline{P}(B(X - \inf(X|B))) - \underline{P}(B(X - \sup(X|B))) \leq \underline{P}(B(X-h)) - \underline{P}(B(X-k))$.

To start now the proof, note that for any $\lambda \in \mathbb{R}$, $B(X - ((1-\lambda)h + \lambda k)) = (1-\lambda)B(X-h) + \lambda B(X-k)$. From this equality, we get for any $\lambda \in [0,1]$ (use P3) of Proposition 1) $\underline{P}(B(X - ((1-\lambda)h + \lambda k))) = \underline{P}((1-\lambda)B(X-h) + \lambda B(X-k)) \geq (1-\lambda)\underline{P}(B(X-h)) + \lambda \underline{P}(B(X-k)) = \underline{P}(B(X-h)) - \lambda(\underline{P}(B(X-h)) - \underline{P}(B(X-k)))$. Defining $\overline{\lambda} = \frac{\underline{P}(B(X-h))}{\underline{P}(B(X-h)) - \underline{P}(B(X-k))}$, $\overline{\lambda} \in [0,1]$. We can therefore replace λ with $\overline{\lambda}$ in the above derivation, getting $\underline{P}(B(X - ((1-\overline{\lambda})h + \overline{\lambda}k)) \geq \underline{P}(B(X-h)) - \overline{\lambda}[\underline{P}(B(X-h)) - \underline{P}(B(X-k))] = 0$.

If $\underline{P}(B(X - ((1-\overline{\lambda})h + \overline{\lambda}k))) > 0$, use Lemma 1, a) to obtain $\underline{P}(X|B) > (1-\overline{\lambda})h + \overline{\lambda}k = \phi(h,k)$.

If $\underline{P}(B(X - ((1-\overline{\lambda})h + \overline{\lambda}k))) = 0$, then $\underline{P}(X|B) = (1-\overline{\lambda})h + \overline{\lambda}k = \phi(h,k)$. We apply here Proposition 9 in [10], which generalises to convex lower previsions a result known for coherent lower previsions [16], ensuring that $r = \underline{P}(X|B)$ is the unique solution of $\underline{P}(B(X-r)) = 0$, if $\underline{P}$ is convex and $\underline{P}(B) > 0$. ■

Notation. When unambiguous we write $S_B = \sup(X|B)$, $I_B = \inf(X|B)$.

Remark 1 *When $\underline{P}$ is coherent, the assumptions in Proposition 4 ensuring that the denominators are non-zero simplify as follows: it is sufficient to ask that*

i) $X|B$ is non-constant;

ii) $\underline{P}(B) > 0$.

In fact, i) and ii) imply $\underline{P}(B(X - I_B)) - \underline{P}(B(X - S_B)) > 0$. To see this, consider a bet on B, $B(X - S_B)$, $B(X - I_B)$ with stakes $S_B - I_B$, 1, -1 respectively. Then $\underline{G} = (S_B - I_B)(B - \underline{P}(B)) + B(X -$

$S_B) - \underline{P}(B(X-S_B)) - B(X-I_B) + \underline{P}(B(X-I_B)) = \underline{P}(B(X-I_B)) - \underline{P}(B(X-S_B)) - (S_B-I_B)\underline{P}(B) = \sup \underline{G}$. *Thus* $\sup \underline{G} \geq 0$ *iff* $\underline{P}(B(X-I_B)) - \underline{P}(B(X-S_B)) \geq (S_B-I_B)\underline{P}(B) > 0$.

As a further remark, note that $\underline{P}(B) = 0$ *(*$\underline{P}$ *coherent) implies* $\underline{P}(B(X-I_B)) - \underline{P}(B(X-S_B)) = 0$*, by Proposition 3 c).*

The lower bound (7) is as a matter of fact a family of lower bounds, indexed on h and k. The immediate question is therefore: which h, k should be chosen? It is not clear a priori that there should be a unique couple (h,k) preferable in all cases, but the following proposition solves the problem in favour of the remarkable couple $h = \inf(X|B)$, $k = \sup(X|B)$.

Proposition 5 *Under the assumptions of Proposition 4,* $\phi(I_B, S_B) \geq \phi(h,k)$, $\forall h \leq I_B$, $\forall k \geq S_B$.

Proof. The proof is made up of two steps. In the first step we prove that for any fixed $h \leq I_B$, $\phi(h,k) \leq \phi(h,S_B)$; in the second that $\phi(h,S_B) \leq \phi(I_B,S_B)$.

To shorten notation, we define $f(r) = \underline{P}(B(X-r))$, so that for instance $f(h) = \underline{P}(B(X-h))$ and $\phi(h,k) = \frac{kf(h)-hf(k)}{f(h)-f(k)}$.

First step. Fix h and define $\delta = \delta(k) = k-h$. We have $\delta \geq S_B - I_B > 0$ (the last inequality is implied by the assumption $\underline{P}(B(X-I_B)) - \underline{P}(B(X-S_B)) \neq 0$ in Proposition 4, which rules out the trivial case that $X|B$ is constant).

We write now $\phi(h,k)$ as a function $u(\delta)$ of δ: $u(\delta) = \frac{(h+\delta)f(h)-hf(h+\delta)}{f(h)-f(h+\delta)}$, or also

$$u(\delta) = \phi(h, h+\delta) = h + \delta \frac{f(h)}{f(h)-f(h+\delta)}. \tag{8}$$

We now consider the function of δ in (8), $\delta/[f(h) - f(h+\delta)]$, proving that:

$$\delta_1 > \delta_2\ (>0) \Rightarrow \tfrac{\delta_1}{f(h)-f(h+\delta_1)} \leq \tfrac{\delta_2}{f(h)-f(h+\delta_2)}. \tag{9}$$

To prove (9), we first verify that $f(r)$ is concave on $\mathbb{R}$. In fact, for $\lambda \in [0,1]$ and using also P3) of Proposition 1, $f(\lambda r_1 + (1-\lambda)r_2) = \underline{P}(B(X - \lambda r_1 - (1-\lambda)r_2)) = \underline{P}(\lambda B(X-r_1) + (1-\lambda)B(X-r_2)) \geq \lambda \underline{P}(B(X-r_1)) + (1-\lambda)\underline{P}(B(X-r_2)) = \lambda f(r_1) + (1-\lambda) f(r_2)$.

For a standard property of concave real functions, $F(\delta) = \frac{f(h+\delta)-f(h)}{\delta}$ is monotone non-increasing for $\delta \in \mathbb{R}$, hence in particular for $\delta \in I = [S_B - I_B, +\infty[$. Interval I is the domain of δ in our case; here $\delta > 0$ and (cf. the beginning of the proof of Proposition 4) $f(h+\delta) - f(h)$ is negative, thus $F(\delta) < 0$, $\forall \delta \in I$. Recalling this, we easily get (9) from $\delta_1 > \delta_2 \Rightarrow F(\delta_1) \leq F(\delta_2)$.

Using (9), and recalling that $f(h) \geq 0$, $u(\delta)$ is maximised, for a given h, by minimising δ, putting hence $\delta = S_B - h$. This is equivalent to choosing $k = S_B$ in $\phi(h,k)$. Thus $\phi(h,k) \leq \phi(h,S_B)$, $\forall k \geq S_B$.

Second step. Define $\delta = \delta(h) = h - S_B < 0$ and write $\phi(h,S_B)$ as a function $v(\delta)$ of δ:

$$v(\delta) = \phi(S_B + \delta, S_B) = S_B - \delta \frac{f(S_B)}{f(S_B+\delta) - f(S_B)}.$$

We prove now that

$$\delta_1 < \delta_2\ (<0) \Rightarrow v(\delta_1) \leq v(\delta_2). \tag{10}$$

For this, we can follow a scheme similar to the proof of the first step (alternatively, a longer proof essentially exploiting the definition of convex prevision is possible). As before, the function $F(\delta) = \frac{f(S_B+\delta)-f(S_B)}{\delta}$ is monotone non-increasing, and negative for $\delta \in]-\infty, I_B - S_B]$. From this and recalling that $f(S_B) \leq 0$, (10) follows straightforwardly.

We conclude that $\phi(h,k) \leq \phi(h,S_B) \leq \phi(I_B,S_B)$, $\forall h \leq I_B$, $\forall k \geq S_B$, where the first inequality follows from step 1, whilst the second is a consequence of step 2. ■

The most notable consequence of Proposition 5 is that we get the following lower bound for $\underline{P}(X|B)$:

$$\underline{P}(X|B) \geq \tfrac{S_B \underline{P}(B(X-I_B)) - I_B \underline{P}(B(X-S_B))}{\underline{P}(B(X-I_B)) - \underline{P}(B(X-S_B))}. \tag{11}$$

When X is an event, $X = A$, (11) reduces to

$$\underline{P}(A|B) \geq \frac{\underline{P}(A \wedge B)}{\underline{P}(A \wedge B) - \underline{P}(B(A-1))}$$

and then to (6), with simple manipulations ($B(A-1) = -B\overline{A}$).

Thus the lower bound in (11) generalises (6) to random variables and to lower previsions that are centered convex (in particular, W-coherent).

An upper bound for $\overline{P}(X|B)$ can be derived from (11):

Corollary 2 *In the assumptions of Proposition 4 and whenever the relevant previsions are defined* [3]

$$\overline{P}(X|B) \leq \tfrac{I_B \underline{P}(B(S_B-X)) - S_B \underline{P}(B(I_B-X))}{\underline{P}(B(S_B-X)) - \underline{P}(B(I_B-X))} \tag{12}$$

Proof. Write (11) for $-X|B$:

$$\underline{P}(-X|B) \geq \tfrac{-I_B \underline{P}(B(S_B-X)) + S_B \underline{P}(B(I_B-X))}{\underline{P}(B(S_B-X)) - \underline{P}(B(I_B-X))}.$$

[3] When X is an event A, (12) reduces to $\overline{P}(A|B) \leq \frac{\overline{P}(A\wedge B)}{\overline{P}(A\wedge B) + \underline{P}(\overline{A}\wedge B)}$. We already met this bound in the paragraph following eq. (6).

Eq. (12) follows, reversing signs in the above inequality and since $-\underline{P}(-X|B) = \overline{P}(X|B)$. ■

An issue which remains to be investigated is under what conditions the bound in (11) is sharp, i.e. it is actually equal to the natural extension $\underline{E}(X|B)$ if $\underline{P}$ is coherent, or to the convex natural extension $\underline{E}_c(X|B)$, when $\underline{P}$ is centered convex. The following example illustrates the case of coherence.

Example *Given the partition $\mathbb{P} = \{e_1, e_2, e_3, e_4\}$, define X such that $X(e_1) = 1$, $X(e_2) = -1$, $X(e_3) = 0$, $X(e_4) = 2$. Given the precise probabilities P_1, P_2, having the following values on $\mathbb{P}$: $P_1(e_1) = 0.2$; $P_1(e_2) = 0.3$; $P_1(e_3) = 0.2$; $P_1(e_4) = 0.3$; $P_2(e_1) = 0.5$; $P_2(e_2) = 0.1$; $P_2(e_3) = 0$; $P_2(e_4) = 0.4$, and calling $\mathcal{A}(\mathbb{P})$ the powerset of $\mathbb{P}$, each of P_1, P_2 has a unique coherent extension to a precise prevision on $U = \mathcal{A}(\mathbb{P}) \cup \{X\} \cup \{B(X - r) : r \in \mathbb{R}\}$, where B is a given event in $\mathcal{A}(\mathbb{P})$. A coherent lower prevision $\underline{P}$ may be defined on any subset $\mathcal{D}$ of U as $\underline{P}(Y) = \min\{P_1(Y), P_2(Y)\}$, $\forall Y \in \mathcal{D}$ (lower envelope theorem). We choose $\mathcal{D} = \mathcal{A}(\mathbb{P}) \cup \{B(X - \inf(X|B)), B(X - \sup(X|B))\}$. Thus in particular $\underline{P}(e_1) = 0.2$, $\underline{P}(e_1 \vee e_2 \vee e_3) = 0.6$, etc. Note that the restriction of $\underline{P}$ on $\mathcal{A}(\mathbb{P})$ is a lower probability which is not 2-monotone (for instance $\underline{P}(e_1 \vee e_3 \vee e_4) = 0.7 < \underline{P}(e_1 \vee e_3) + \underline{P}(e_3 \vee e_4) - \underline{P}(e_3) = 0.8$). We have 10 non-trivial different choices for the conditioning event B in $\mathcal{A}(\mathbb{P})$. It may be verified that the bound is sharp in all of these but one.*

a) For instance, let $B = e_1 \vee e_2 \vee e_3$. This is one of the 9 choices for B giving a sharp bound (11). In fact, $S_B = 1$, $I_B = -1$. Since $P_1(B(X-r)) = P_1(BX) - rP_1(B) = -0.1 - 0.7r$ and $P_2(B(X - r)) = 0.4 - 0.6r$, we obtain $\underline{P}(B(X - r)) = \min\{-0.1 - 0.7r, 0.4 - 0.6r\} = -0.1 - 0.7r$ iff $r \geq -5$. Then the bound (11) is $\phi(I_B, S_B) = \phi(-1, 1) = -\frac{1}{7} = \underline{E}(X|B)$, because $P_1(X|B) = \frac{P_1(BX)}{P_1(B)} = -\frac{1}{7}$. Note that $\phi(-1, 1)$ is not the only sharp bound in the $\phi(h, k)$ family: $\phi(h, k) = -\frac{1}{7}$ for $h \in [-5, -1]$, $k \geq 1$.

b) Let now $B = e_1 \vee e_4$. This choice corresponds to the unique non-exact bound (11). In fact, now $P_1(B(X - r)) = 0.8 - 0.5r$, $P_2(B(X - r)) = 1.3 - 0.9r$, $\underline{P}(B(X-r)) = 1.3 - 0.9r$ iff $r \geq \frac{5}{4}$ and the bound is $\phi(I_B, S_B) = \phi(1, 2) = \frac{11}{8}$. To see that the bound cannot be reached, note that [15] if it were sharp, there would be a precise prevision P, in the set $\mathcal{M}_{\mathcal{D}}(\underline{P})$ of precise previsions dominating $\underline{P}$ on $\mathcal{D}$, such that its extension on $X|B$ ensures that $P(X|B) = \frac{P(BX)}{P(B)} = \frac{P(e_1)+2P(e_4)}{P(e_1)+P(e_4)} = \frac{11}{8}$, which means that

$$P(e_4) = \frac{3}{5} P(e_1) \tag{13}$$

It is then easy to check that no such P may be found in $\mathcal{M}_{\mathcal{D}}(\underline{P})$: just verify that there is no real solution for the system of linear inequalities formed by (13), the dominance constraints $P \geq \underline{P}$ on $\mathcal{D}$, and the non-negativity and normalisation constraints for P on $\mathbb{P}$.

c) Let us introduce partition $\mathbb{P}' = \{\omega_1, \omega_{2,3}, \omega_4\}$ which is a coarsening of $\mathbb{P}$: $\omega_1 = e_1$, $\omega_{2,3} = e_2 \vee e_3$, $\omega_4 = e_4$. We do not modify the uncertainty evaluations of b), defining P_1, P_2 on $\mathbb{P}'$ as the restrictions of the previously defined P_1, P_2 respectively (thus $P_1(\omega_1) = 0.2$, $P_1(\omega_{2,3}) = 0.5$, $P_1(\omega_4) = 0.3$, $P_2(\omega_1) = 0.5$, $P_2(\omega_{2,3}) = 0.1$, $P_2(\omega_4) = 0.4$) and $\underline{P}$ as their lower envelope on $\mathcal{D}' = \mathcal{A}(\mathbb{P}') \cup \{B(X - \sup(X|B)), B(X - \inf(X|B))\}$, $B \in \mathcal{A}(\mathbb{P}')$. Here $B = \omega_1 \vee \omega_4$, and $X(\omega_1) = 1$, $X(\omega_4) = 2$, while $X(\omega_{2,3})$ may take any value, it does not influence the following computations. Note that now $\underline{P}$ is 2-monotone on $\mathcal{A}(\mathbb{P}')$, since $\mathbb{P}'$ is a three-atom partition [15]. Obviously, the bound (11) is again $\frac{11}{8}$ as in b), since, when passing from b) to c), we essentially only grouped together e_2 and e_3, which are irrelevant in the computation of $\phi(1, 2)$. However, there is now a prevision P in $\mathcal{M}_{\mathcal{D}'}(\underline{P})$ which reaches the bound, i.e. such that $P(X|B) = \frac{11}{8}$: its values on $\mathbb{P}'$ are $P(\omega_1) = 0.5$, $P(\omega_{2,3}) = 0.2$, $P(\omega_4) = 0.3$.

6 Conclusions

In a first part of the paper (Section 3) we related W-coherence with Williams' original definition, and also with other notions of coherence in a conditional framework, especially Walley-coherence. Our main purpose here was to show that W-coherence can be profitably employed to obtain results which hold for Walley-coherence too (Section 4). A more extended comparison between these two coherence concepts is beyond the aims of the present paper, but is an undoubtedly interesting question. It requires analysing further issues, like the role of the conglomerative property or the interpretation of Walley's updating principle.

In the sequel of the paper, we have discussed some implications of product rule bounds and generalised a Bayes' theorem bound to either W-coherent or centered convex lower previsions. Although we did not consider them here, other similar bounds or simple generalisations may be found (for instance, for upper

previsions), with analogous properties. A less immediate question is that of investigating further the relationships of these bounds with important concepts in the theory of imprecise previsions: epistemic irrelevance and natural and upper extension for (more general) product rule bounds, 2-monotonicity and possibly Choquet integration for the bound (11). Concerning the latter issue, a generalisation to lower previsions of 2-monotonicity with related results was recently proposed in [5]. There remain anyway two features in our approach which, while ensuring generality, make it difficult to apply pre-existing results to sufficiently general situations, for instance in the problem of establishing when the bound (11) is sharp. One feature is that we are working in a structure-free environment, while 2-monotonicity is customarily referred to algebras of events [15] or (linear) lattices of random variables [5]. With respect to this feature, our example is still rather peculiar: there is a partition $\mathbb{P}$ there such that $\mathcal{A}(\mathbb{P}) \subset \mathcal{D}$, but this inclusion is generally not required. A second issue is that we consider also the centered convexity condition, and relationships of 2-monotonicity (for previsions) with convexity are still largely to be explored.

Appendix. W-coherence and separate coherence

Let $\mathbb{P}$ be an *arbitrary* (finite or not) partition of non-impossible events. We recall the definition of *separate coherence* in [16]:[4]

Definition 4 *The conditional lower previsions $\underline{P}_B(X|B)$, defined for any $B \in \mathbb{P}$ and $X \in \mathcal{H}(B)$, where $\mathcal{H}(B)$ is an arbitrary set of gambles containing B, are* separately coherent *iff, for every $B \in \mathbb{P}$,*

i) $\underline{P}_B(B|B) = 1$

ii) $\forall s_0, \ldots, s_n \geq 0$, $\forall X_0, \ldots, X_n \in \mathcal{H}(B)$, *defining* $\underline{G} = \sum_{i=1}^n (X_i - \underline{P}(X_i|B)) - s_0(X_0 - \underline{P}(X_0|B))$, *it holds that* $\sup \underline{G} \geq 0$.

When defined on the same domain, separate coherence and W-coherence are equivalent, as we now prove. Define for this the conditional lower prevision $\underline{P}$ such that $\underline{P}(X|B) = \underline{P}_B(X|B)$, $\forall B \in \mathbb{P}$, $\forall X \in \mathcal{H}(B)$ ($\underline{P}$ is the collection of all $\underline{P}_B$).

Proposition 6 *The lower previsions $\underline{P}_B$ ($B \in \mathbb{P}$) in Definition 4 are separately coherent iff $\underline{P}$ is W-coherent on $\mathcal{D} = \cup_{B \in \mathbb{P}} \mathcal{D}_B$, where $\mathcal{D}_B = \{X|B : X \in \mathcal{H}(B)\}$.*

[4] The integer stakes in [16] may be equivalently replaced by real non-negative ones, as we do here.

Proof. We prove first that W-coherence implies separate coherence. If $\underline{P}$ is W-coherent, i) necessarily holds. As for ii), it follows from $\sup \underline{G} = \max\{\sup_B \underline{G}, \sup_{\overline{B}} \underline{G}\} \geq \sup_B \underline{G} = \sup \underline{G}|B = \sup(B\underline{G}|B) \geq 0$, the last equality holding by (1), the inequality by W-coherence.

To prove the converse implication, suppose that separate coherence holds. Betting on B, $X_0, \ldots, X_n$, it follows then $\sup(s(B - \underline{P}(B|B)) + \sum_{i=1}^n s_i(X_i - \underline{P}(X_i|B)) - s_0(X_0 - \underline{P}(X_0|B))) = \sup(s(B-1)+\underline{G}) = \max(\sup_B(s(B-1)+\underline{G}), \sup_{\overline{B}}(s(B-1)+\underline{G})) \geq 0$. The last inequality implies $\sup_B(s(B-1)+\underline{G}) \geq 0$, if we choose $s > \max(\sup_{\overline{B}} \underline{G}, 0)$, since then $\sup_{\overline{B}}(s(B-1)+\underline{G}) = -s + \sup_{\overline{B}} \underline{G} < 0$. Using also (1), $\sup_B(s(B-1)+\underline{G}) = \sup(\underline{G}|B) = \sup(B\underline{G}|B) = \sup(\sum_{i=1}^n s_i B(X_i - \underline{P}(X_i|B)) - s_0 B(X_0 - \underline{P}(X_0|B))|B) \geq 0$, which means, given the arbitrariness of n, $X_0, \ldots, X_n$ and $s_0, \ldots, s_n \geq 0$, that $\underline{P}$ is W-coherent on $\mathcal{D}_B$. It is then a simple exercise to prove that W-coherence of $\underline{P}$ on each $\mathcal{D}_B$ implies W-coherence of $\underline{P}$ on $\mathcal{D}$, because of the special structure of $\mathcal{D}$. ■

Acknowledgements

We are grateful to the referees for their constructive suggestions.

References

[1] T.E. Armstrong, W.D. Sudderth. Locally coherent rates of exchange. *Annals of Statistics*, 17:1394-1408, 1989.

[2] P. Artzner, F. Delbaen, S. Eber and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203-228, 1999.

[3] L. Crisma. Events and conditional events: a subjectivistic approach. In *The notion of event in probabilistic epistemology*, Pubbl. n. 2, Dip. Mat. Appl. 'B. de Finetti', Ediz. Lint, Trieste, 43–89, 1996.

[4] L. Crisma. *Introduzione alla teoria delle probabilità coerenti.* Ediz. EUT, Trieste, 2006.

[5] G. de Cooman, M. C. M. Troffaes and E. Miranda. n-Monotone lower previsions. *Journal of Intelligent & Fuzzy Systems*, 16: 253–263, 2005.

[6] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali triestini dell'Università di Trieste*, XIX 2, 29–81, 1949.

[7] S. Maaß. Continuous linear representations of coherent lower previsions. In *Proc. ISIPTA'03*, Lugano (CH), 372–382, 2003.

[8] E. Miranda and G. de Cooman. Epistemic independence in numerical possibility theory. *International Journal of Approximate Reasoning*, 32:23–42, 2003.

[9] R. Pelessoni and P. Vicig. Convex imprecise previsions. *Reliable Computing*, 9(6):465–485, 2003.

[10] R. Pelessoni and P. Vicig. Uncertainty modelling and conditioning with convex imprecise previsions. *International Journal of Approximate Reasoning*, 39(2–3):297–319, 2005.

[11] R. Pelessoni and P. Vicig. Envelope theorems and dilation with convex conditional previsions. In *Proc. ISIPTA'05*, Pittsburgh, PA, 266–275, 2005.

[12] M. C. M. Troffaes and G. de Cooman. Lower previsions for unbounded random variables. In *Soft Methods in Probability, Statistics and Data Analysis*, Advances in Soft Computing, 146–155, Physica-Verlag, New York, 2002.

[13] P. Vicig. Epistemic independence for imprecise probabilities. *International Journal of Approximate Reasoning*, 24:235–250, 2000.

[14] P. Vicig, M. Zaffalon and F. G. Cozman. Notes on 'Notes on conditional previsions'. *International Journal of Approximate Reasoning*, 44:358–365, 2007.

[15] P. Walley. Coherent lower (and upper) probabilities. *Research Report*, University of Warwick, Coventry, 1981.

[16] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[17] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83:1–58, 1996.

[18] P. Walley, R. Pelessoni and P. Vicig. Direct algorithms for checking consistency and making inferences from conditional probability assessments. *Journal of Statistical Planning and Inference*, 126(1):119-151, 2004.

[19] L.A. Wasserman and J.B. Kadane. Bayes' theorem for Choquet capacities. *Annals of Statistics*, 18:1328-1339, 1990.

[20] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I – Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.

[21] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007 (revised version of: Notes on conditional previsions. *Research Report*, School of Math. and Phys. Science, University of Sussex, 1975).

Human reasoning with imprecise probabilities: Modus ponens and denying the antecedent

Niki Pfeifer
Department of Psychology,
University of Salzburg,
Austria
niki.pfeifer@sbg.ac.at

Gernot D. Kleiter
Department of Psychology,
University of Salzburg,
Austria
gernot.kleiter@sbg.ac.at

Abstract

The MODUS PONENS ($A \to B, A \therefore B$) is, along with MODUS TOLLENS and the two logically not valid counterparts DENYING THE ANTECEDENT ($A \to B, \neg A \therefore \neg B$) and AFFIRMING THE CONSEQUENT, the argument form that was most often investigated in the psychology of human reasoning. The present contribution reports the results of three experiments on the probabilistic versions of MODUS PONENS and DENYING THE ANTECEDENT. In probability logic these arguments lead to conclusions with imprecise probabilities.

In the MODUS PONENS tasks the participants inferred probabilities that agreed much better with the coherent normative values than in the DENYING THE ANTECEDENT tasks, a result that mirrors results found with the classical argument versions. For MODUS PONENS a surprisingly high number of lower and upper probabilities agreed perfectly with the conjugacy property (upper probabilities equal one complements of the lower probabilities). When the probabilities of the premises are imprecise the participants do not ignore irrelevant ("silent") boundary probabilities. The results show that human *mental probability logic* is close to predictions derived from *probability logic* for the most elementary argument form, but has considerable difficulties with the more complex forms involving negations.

Keywords. Mental probability logic, modus ponens, coherence, imprecise probabilities

1 Introduction

While there is a long tradition of probabilistic approaches in human judgment and decision making [10], only recently probabilistic approaches are adopted in the psychology of reasoning [16, 17, 7, 19, 12, 14, 18, 27, 26, 29]. Traditionally, classical logic dominated the psychology of human reasoning [6, 28, 2]. Classical logic was the normative standard of reference and used as a criterion for the rationality of human inferences. See Table 1 for often investigated argument forms in psychology.

	MP	CMP	DA	CDA
P_1:	$A \to B$	$A \to B$	$A \to B$	$A \to B$
P_2:	A	A	$\neg A$	$\neg A$
Concl.:	B	$\neg B$	$\neg B$	B
L-valid:	yes	no	no	no
$V_i(\mathfrak{C})$	t	f	?	?

Table 1: Non-probabilistic version of the MODUS PONENS (MP), DENYING THE ANTECEDENT (DA), and their respective complementary versions (CMP, and CDA). A and B denote propositions. $\to$ and $\neg$ denote the material implication and negation, respectively, and are defined as usual. *L-valid* denotes logical validity, and V_i denotes the logical valuation-function V of the conclusion $\mathfrak{C}$ under the interpretation i that assigns t ("true"), to all premises (P_1 and P_2). If the antecedents, A, of the conditional premise is false, then the truth value of the conclusion is not determined (denoted by the question mark).

Traditional psychological research on human reasoning designates human inference as rational/not rational if it corresponds to logically valid/not valid argument forms. Everyday life situations, though, are inherently uncertain. The uncertainty cannot be captured by classical logic. Reasoning about uncertainty and uncertain knowledge is a fundamental human competence. Thus, classical logic cannot be an adequate normative standard of reference for the psychology of reasoning.

We proposed a psychological theory of human reasoning, called "*mental probability logic*" [20, 22, 21, 23, 24, 25], which evaluates the rationality of human reasoning not by means of logical validity but by means of *coherence* [9, 8, 4]. *Mental probability logic* is a psychological *competence theory* about how humans

interpret common sense conditionals, represent the premises of everyday life arguments and draw inferences by coherent manipulations of mental representations.

Why a *competence* theory? Many investigations on cognitive processes report errors, fallacies, or biases. Well known are perceptual illusions, biases in judgment under uncertainty, or errors in deductive reasoning. While these phenomena may be startling and stimulating in the scientific process, they do not lead to theories that explain human performance in a systematic way. Collecting slips of the tongue does not lead to a theory of speaking. Psycholinguistics distinguishes performance and competence. Competence describes what functions a cognitive system can compute. Human reasoning can solve complex problems and perform sophisticated inferences. While developing a theory of reasoning one should have the explanation of these processes in mind. One should strive for a competence theory. The distinction between competence and performance was introduced by Noam Chomsky [3]. The analogy to deductive reasoning is obvious. The emphasis on the function a cognitive system should compute is due to David Marr [15].

Mental probability logic claims that the common sense conditionals are represented as subjective conditional probabilities. Based on the available information, the premises are evaluated and represented by coherent precise (point) probabilities, coherent imprecise probabilities, or logical information. Coherent imprecise probabilities can be represented by coherent interval probabilities or second order probability distributions. Human reasoning is a mental process that forms new representations from old ones by using probabilistic versions of formal inference rules. We assume that a certain core set of probabilistic inference rules are hard wired in the human inference engine.

The normative standard of *mental probability logic* is based on *coherence*. Coherence is the key concept in the tradition of subjective probability theory. It was originally developed by de Finetti [5]. More recent work includes [31, 13, 4, 8]. A probability assessment is coherent[1] if it does not admit one or more bets with sure loss. Coherence provides an adequate normative foundation for the *mental probability logic* and has many psychologically plausible advantages compared with classical concepts of probability:

- In the framework of coherence a *complete Boolean algebra is is not required* for probabilistic inference. Full algebras are psychologically unrealistic as they can neither be unfolded in working memory nor be stored in the long term memory. *Mental probability logic* suggests that humans try to keep the memory load as small as possible and process only relevant informations (see also [30]).
- *Conditional probability,* $P(B|A)$, is a *primitive* notion. The probability values are assigned *directly*. Conditional probability is not "defined"—as in probability textbooks—via the fraction of the "joint", $P(A \wedge B)$, and the "marginal", $P(A)$, probabilities.[2] Conditional probabilities are *directly* encoded or just directly connected to the arguments of the if–then relation.
- Because lack of knowledge (time, effort) it may be impossible for a person to assign precise probabilities to an event. If a person is uncertain about probabilities, then *mental probability logic* supposes that human subjects make coherent *imprecise* probabilistic assessments (by interval-valued probabilities or by second order probability distributions).
- Coherence is in the tradition of subjective probability theory in which probabilities are conceived as *degrees of belief*. Degrees of belief are naturally affine to psychology.

Imprecise versions of the argument forms presented in Table 1 are formalized by interval probabilities [23] or by second order probability distributions [25]. In the present study we focus on interval probabilities only. We now present imprecise versions of the four argument forms of Table 1. While only the MODUS PONENS is logically valid, all four argument forms admit to infer coherent probability intervals for the conclusion.

The imprecise version of the MODUS PONENS has the form:

$$\begin{aligned} &P(B|A) \in [x', x''] \,,\; P(A) \in [y', y''] \\ &\quad \therefore \quad P(B) \in [x'y', 1 - y' + x''y'] \,. \end{aligned} \tag{1}$$

Since $P(\neg\mathfrak{C}) = 1 - P(\mathfrak{C})$ (conjugacy principle [31]) the complement of an interval $[l, u]$ is $[1-u, 1-l]$, it trivially follows that the imprecise COMPLEMENT MODUS PONENS has the form:

$$\begin{aligned} &P(B|A) \in [x', x''] \,,\; P(A) \in [y', y''] \\ &\quad \therefore \quad P(\neg B) \in [y' - x''y', 1 - x'y'] \,. \end{aligned} \tag{2}$$

The imprecise DENYING THE ANTECEDENT has the form:

$$\begin{aligned} &P(B|A) \in [x', x''], P(\neg A) \in [y', y''] \\ &\quad \therefore \quad P(\neg B) \in [(1-x'')(1-y''), 1-x'(1-y'')] \end{aligned} \tag{3}$$

[1] Throughout we use "coherent" as synonymous with "totally coherent" [9].

[2] The definition $P(B|A) =_{df.} P(A \wedge B)/P(A)$ is problematic if $P(A) = 0$.

The imprecise COMPLEMENT DENYING THE ANTECEDENT has the form:

$$P(B|A) \in [x', x''], P(\neg A) \in [y', y'']$$
$$\therefore \quad P(B) \in [x'(1-y''), x''+y''-x''y''] \quad (4)$$

Equations (1)-(4) may be obtained by natural extension [31], likewise, by de Finetti's Fundamental Theorem [5] or by Lad's generalized version [13]—or by elementary probability theory (for a demonstration see [23]).

In the psychological literature, the non-probabilistic versions of the MODUS PONENS and the DENYING THE ANTECEDENT were studied extensively. Meta-analytical results show that the MODUS PONENS is endorsed by 89-100% of human subjects [6]. The DENYING THE ANTECEDENT is endorsed by 17-73% of the subjects [6]. We do not judge the 17-73% of subjects as irrational. Rather, we propose to reinterpret the data in the light of *mental probability logic*. The question is not whether the human subjects endorse the non-probabilistic DENYING THE ANTECEDENT, but whether they infer coherent probabilities from the premises of the *imprecise* DENYING THE ANTECEDENT. In the next sections we present empirical data on the four imprecise argument forms (1)-(4).

2 Experiment 1

2.1 Method and Procedure

Thirty students of the University of Salzburg participated in Experiment 1. No students with special logical or mathematical education were included.

Each participant received a booklet containing a general introduction, one example explaining the response modality with point percentages, and one example explaining the response modality with interval percentages. Three target tasks were presented on separate pages. Eight additional target tasks were presented in tabular form. The first three MODUS PONENS target tasks had the following form:

> Please imagine the following situation. Several cars are parked on a parking lot. About these cars we know the following:
>
> *Exactly 80%* of the *red cars* on this parking lot are *two-door-cars*.
> *Exactly 90%* of the cars on this parking lot are *red cars*.
>
> Imagine all the cars that are on this parking lot. How many of these cars are *two-door-cars*?

Then, the participants were informed that the solution is either a point percentage or a percentage between two boundaries (*from at least ... to at most ...*). The booklet offered two response modalities where the participants had to choose one deliberately.

> *Response Modality 1:*
> If you think that the correct answer is a *point* percentage, please fill in your answer here:
> *Exactly* .....% of the cars on this parking lot are *two-door-cars*.
>
> **Point percentage:**
>
> 0 25 50 75 100 %

> *Response Modality 2:*
> If you think that the correct answer lies within two boundaries (*from at least ... to at most ...*), please mark the two values here:
> *At least* .....% and *at most* .....% of the cars on this parking lot are *two-door-cars*.
>
> **Within the bounds of:**
>
> 0 25 50 75 100 %

The subsequent two tasks were formulated accordingly. In the second task the numerical values in two premises were 20 and 40%, and in the third task 60 and 90%, respectively. Each task was on a separate page. After the third task the participants answered eight analogous tasks presented on one page in tabular form. Again, the cover story was kept constant, only the numerical values contained in the premises varied (see Table 2).

The thirty participants were divided into two groups, fifteen participants received the MODUS PONENS tasks, as just described, and fifteen participants received DENYING THE ANTECEDENT tasks. The DENYING THE ANTECEDENT tasks were formulated exactly as the MODUS PONENS tasks, with two differences. First, a negation was added in the second premise: "*Exactly 90%* of the cars on this parking lot are ***not*** *red cars*". Second, a negation was added to the question: "How many of these cars are ***not*** *two-door-cars*?". Both response modalities were adopted accordingly. Adding the negations to the second premise and to the conclusion as just described clearly reflects the form of the corresponding imprecise DENYING THE ANTECEDENT. In both conditions, we presented the same percentage numbers to the participants.

The booklets were mixed and assigned arbitrarily to

P_1	P_2	CLB	CUB	LBR	UBR
80	90	72	82	75.80 (5.16)	82.60 (8.33)
20	40	8	68	14.60 (11.15)	55.60 (30.24)
60	90	54	64	51.93 (10.13)	63.87 (14.21)
40	40	16	76	25.13 (12.65)	64.07 (35.43)
80	70	56	86	53.00 (21.91)	71.20 (27.26)
20	60	12	52	18.60 (13.37)	46.27 (23.59)
100	100	100	100	100.00 (0.00)	100.00 (0.00)
60	70	42	72	48.00 (15.08)	67.60 (22.04)
40	60	24	64	33.27 (13.88)	59.47 (23.55)
70	80	56	76	61.33 (13.80)	76.73 (12.40)
30	50	15	65	20.67 (12.80)	51.33 (30.32)

Table 2: Mean lower (LBR) and mean upper bound responses (UBR) of the MODUS PONENS tasks of Experiment 1 ($n_1 = 15$). The standard deviations are in parenthesis. P_1 and P_2 denote the percentages in the premises. CLB and CUB denote the normative/coherent lower and upper bounds, respectively.

P_1	P_2	CLB	CUB	LBR	UBR
80	90	2	92	27.13 (33.58)	64.13 (37.94)
20	40	48	88	36.07 (27.71)	62.80 (31.97)
60	90	4	94	22.47 (22.17)	73.00 (29.30)
40	40	36	76	33.87 (20.87)	62.67 (23.25)
80	70	6	76	22.93 (23.20)	55.27 (34.74)
20	60	32	92	29.07 (17.90)	66.53 (24.65)
100	100	0.00	100	10.00 (28.03)	36.67 (48.06)
60	70	12	82	28.40 (26.26)	60.93 (30.39)
40	60	24	84	22.53 (17.01)	59.07 (23.71)
70	80	6	86	19.93 (23.91)	54.93 (31.25)
30	50	35	85	27.87 (20.76)	62.67 (25.13)

Table 3: Mean lower LBR and mean upper bound responses UBR of the DENYING THE ANTECEDENT tasks of Experiment 1 ($n_2 = 15$). The standard deviations are in parenthesis. P_1 and P_2 denote the percentages in the premises. CLB and CUB denote the normative/coherent lower and upper bounds, respectively.

the participants. All participants were tested individually in a quiet test room in the department. They were told to take as much time as they wanted. In case of questions, the they were asked to reread the instructions carefully.

2.2 Results and Discussion

Table 2 lists the probabilities presented in the premises, the normative lower and upper bounds, and the participants' mean lower and upper bound responses for the MODUS PONENS tasks. In the task with certain premises (i.e., "100%" in both premises) all fifteen participants responded with point value of 100%, which is normatively correct.

Table 3 lists the respondents' mean lower and upper bound responses for the DENYING THE ANTECEDENT tasks. In the task with certain premises four of the fifteen participants responded with the unit interval 0-100%. These four participants clearly understood that the DENYING THE ANTECEDENT is probabilistically not informative if all premises have probabilities equal to 1. One participant responded with the point value 100 and one with the point value 50%. The majority (nine out of the fifteen participants) responded with the point value 0%.

In the MODUS PONENS tasks on the average 30% of the responses were point values (the task with the certain premises not included). In the DENYING THE ANTECEDENT tasks on the average 23% of the responses were point values (the task with the certain premises was not averaged).

In both conditions the standard deviations are high. This may be a consequence of the explicit presentation of the point value response modality. The participants could actually have some imprecise value in mind, but nevertheless respond with a representative point value just to reduce the complexity of the task. Such point value responses bias the mean lower and upper bound responses. To avoid constantly pointing explicitly to the possibility to give an interval value response we dropped the response point response modality in Experiment 2. Dropping the point response modality forces the participants to respond by intervals while still allowing point value responses by equating the lower and the upper bound responses.

Table 4 reports the frequencies of interval response categories of the MODUS PONENS condition in 3×3 tables. Each table contains the six possible interval responses together with the according empirical frequencies of the interval responses. The *columns* designate whether the participants' lower bounds are below (LB), within (LW), or above (LA) the normative intervals. The *rows* designate whether the upper bounds are above (UA), within (UW) or below (UB) the normative intervals.

Table 5 reports the frequencies of interval response categories of the DENYING THE ANTECEDENT condition in 3×3 tables. Figure 1 presents the averaged interval response frequencies in the MODUS PONENS tasks. The data of Task 2 and of Task 7 were not averaged. The normative lower bound of Task 2 is $\leq 10\%$. Both normative bounds of task 7 are equal to 100%. Figure 1 shows that in the MODUS PONENS tasks the majority of the participants gave coherent interval responses.

Figure 2 presents the averaged interval response fre-

quencies in the DENYING THE ANTECEDENT tasks. The data of Task 1, 3, 5, 6, 7, and 10 were not averaged. The normative upper bounds of Task 1, 3, 6, and 7 are $\geq$ 90%. The normative lower bounds of Task 2, 5, 7 and 10 are $\leq$ 10%. Figure 2 shows that in the DENYING THE ANTECEDENT tasks the participants responded with more incoherent intervals than in the MODUS PONENS tasks. More coherent interval responses were observed in the MODUS PONENS tasks (62.93% of the participants) compared with the DENYING THE ANTECEDENT tasks (41.33% of the participants).

In the MODUS PONENS condition, the mean responses agree very well with the normative lower ($r_{LBR,CLB}$ = .99) and upper ($r_{(UBR,CUB)}$ = .92) probabilities. The good agreement remains when the lower ($r_{(LBR,CLB).P1}$ = .91) and upper ($r_{(UBR,CUB).P2}$ = .95) percentages in the premises are partialled out. Partialling out the values contained in the tasks reduces the possible influence of anchoring and/or matching effects.

In the DENYING THE ANTECEDENT condition we observed a different pattern. While the mean responses still agree well with the normative lower ($r_{(LBR,CLB)}$ = .76) probabilities, the correlation is slightly negative for the upper ($r_{(UBR,CUB)} = -.20$) probabilities. Partialling out Premise 1 and Premise 2 reduces the correlations to $r_{(LBR,CLB).P_1}$ = .25 and $r_{(UBR,CUB).P_2}$ = .03. The results may be explained by assuming that the participants just respond with values close to those contained in the description of the tasks (known as "matching heuristic").

It is well known that logical tasks involving negations are difficult. In probabilistic inference tasks we consider the correlations between the probabilities of the premises and the normative lower and upper probabilities of the conclusions. For the set of our MODUS PONENS tasks the four correlations are all positive, $r_{(P_1,CLB)}$ = .97 and $r_{(P_2,CLB)}$ = .92 for the lower probabilities, and $r_{(P_1,CUB)}$ = .86 and $r_{(P_2,CUB)}$ = .51 for the upper ones.

For the DENYING THE ANTECEDENT tasks (with the identical numerical probabilities of the premises!) the lower bound correlations are highly negative, $r_{(P_1,CLB)} = -.92$ and $r_{(P_2,CLB)} = -.93$, and for the upper bounds positive, $r_{(P_1,CUB)}$ = .24 and $r_{(P_2,CUB)}$ = .62. The weighting and integration of affirmative and non-affirmative information makes tasks like the DENYING THE ANTECEDENT especially difficult. We note that linear regression predicts lower probabilities better than upper ones.

The overall conclusion of Experiment 1 is that the responses of the participants in the MODUS PONENS condition are very close to the normative values while in the DENYING THE ANTECEDENT condition the responses might be explained by matching based guessing.

	Schema			*Task 1*			*Task 2*		
UA	a	b	c	0	4	1	0	4	0
UW	d	$\boldsymbol{e}$	-	0	**10**	-	0	**11**	-
UB	f	-	-	0	-	-	0	-	-
	LB	LW	LA	LB	LW	LA	LB	LW	LA
	Task 3			*Task 4*			*Task 5*		
UA	0	5	0	0	5	0	0	2	0
UW	2	**7**	-	0	**9**	-	1	**10**	-
UB	1	-	-	1	-	-	2	-	-
	LB	LW	LA	LB	LW	LA	LB	LW	LA
	Task 6			*Task 7*			*Task 8*		
UA	0	4	1	-	-	-	0	4	1
UW	1	**9**	-	0	**15**	-	0	**9**	-
UB	0	-	-	0	-	-	1	-	-
	LB	LW	LA	LB	LW	LA	LB	LW	LA
	Task 9			*Task 10*			*Task 11*		
UA	0	5	0	0	3	1	0	4	0
UW	0	**10**	-	1	**10**	-	0	**11**	-
UB	0	-	-	0	-	-	0	-	-
	LB	LW	LA	LB	LW	LA	LB	LW	LA

Table 4: Frequencies of the interval responses in the MODUS PONENS condition of Experiment 1 ($n_1 = 15$). UA: the participants' upper bound response is *above* the normative upper bound, UW: upper bound response is *within* the normative interval, UB: upper bound response is *below* the normative lower bound; LA, LW, and LB: same for the participants' lower bound responses. a: too wide interval responses, b: lower bound responses coherent, c: both bound responses above, d: upper bound responses coherent, $\boldsymbol{e}$: both bound responses coherently within ±5% (bold), f: both bound responses below the normative lower bounds.

The presence of the negations in the DENYING THE ANTECEDENT is a possible explanation, why there were less coherent interval responses compared with the MODUS PONENS tasks. It is easier to cognitively represent an affirmed than a negated proposition. An affirmed proposition can be visualized, for example, more directly than a negated one. Classical MODUS PONENS was proposed as a basic and "hard wired" inference rule [28, 2]. Probabilistic MODUS PONENS is a similar candidate.

In addition to the MODUS PONENS and the DENYING THE ANTECEDENT, the respective complementary versions are investigated in Experiment 2. By investigating the complementary versions as well, the presence of the negation is more balanced.

	Schema			*Task 1*			*Task 2*		
UA	*a*	*b*	*c*	-	2	0	0	1	0
UW	*d*	***e***	-	-	**13**	-	3	**6**	-
UB	*f*	-	-	-	-	-	5	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task 3			*Task 4*			*Task 5*		
UA	0	1	0	0	3	0	0	4	0
UW	0	**14**	-	5	**6**	-	2	**9**	-
UB	0	-	-	1	-	-	0	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task 6			*Task 7*			*Task 8*		
UA	0	2	0	-	-	-	0	4	0
UW	6	**6**	-	-	**15**	-	4	**7**	-
UB	1	-	-	-	-	-	8	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*
	Task 9			*Task 10*			*Task 11*		
UA	0	2	0	0	2	0	0	2	0
UW	5	**7**	-	3	**10**	-	6	**5**	-
UB	1	-	-	0	-	-	2	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Table 5: Frequencies of the interval responses in the DENYING THE ANTECEDENT condition ($n_2 = 15$). For explanation of the schema see Table 4.

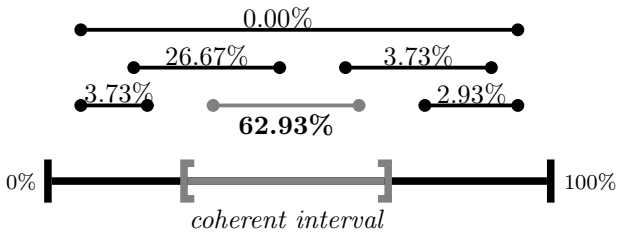


Figure 1: Averaged interval response frequencies over nine selected MODUS PONENS tasks (see text, $n_1 = 15$).

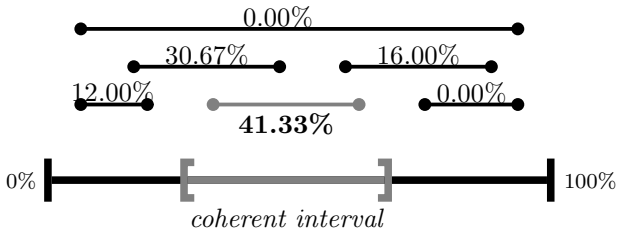


Figure 2: Averaged interval response frequencies over five selected DENYING THE ANTECEDENT tasks (see text, $n_2 = 15$).

3 Experiment 2

3.1 Method and Procedure

Method and procedure of Experiment 2 are analog to Experiment 1. Sixty students of the University of Salzburg participated in Experiment 2. No students with special logical or mathematical education were included. Thirty participants were assigned to the MODUS PONENS condition and thirty participants were assigned to the DENYING THE ANTECEDENT condition.

In the MODUS PONENS condition, each participant worked out three MODUS PONENS tasks and three COMPLEMENT MODUS PONENS tasks. To counterbalance position position effects, fifteen participants got the MODUS PONENS tasks at the beginning, and fifteen participants got the MODUS PONENS tasks at the end of the session. The MODUS PONENS tasks had the following form:

> Please imagine the following situation. Around Christmas time a certain ski-resort is very busy. This region is very popular among sportsmen, like skiers, snow-boarders, and sledge-rider. Every hour a cable-car brings the sportsmen to the top. About this cable-car we know:
>
> *Exactly 100%* of the *skiers* wear *red caps*.
> *Exactly 100%* of the sportsmen are *skiers*.
>
> Imagine all the sportsmen in this cable car. How many of these sportsmen wear a *red cap*?

Speaking about a closed room (cable-car) instead of an unspecified parking lot (Experiment 1) should help to represent and visualize the problems. As in Experiment 1, participants were free to respond either in terms of point percentages or in terms of interval percentages. In Experiment 2, however, the response modality 1 (point response) was dropped. The participants were informed by two examples at the beginning that point values can be given by equating the lower and the upper bounds.

All three MODUS PONENS tasks had the same structure. The percentages of the two premises in the first task were 100 and 100%, in the second task were 70 and 90%, and in the third task the percentages were 70 and 50%, respectively. The three COMPLEMENT MODUS PONENS tasks contained the same percentages and differed from the MODUS PONENS task only in one respect: a negation was added to the conclusion ("How many of these sportsman **do not** wear a *red cap*?").

The DENYING THE ANTECEDENT condition was analogue. Fifteen participants first received the three DENYING THE ANTECEDENT tasks and then the three complementary versions of the DENYING THE AN-

P_1	P_2	CLB	CUB	LBR	UBR
		MODUS PONENS			
100	100	100	100	100.00 (.00)	100.00 (.00)
70	90	63	73	62.43 (11.77)	69.17 (9.71)
70	50	35	85	42.5 (15.13)	54.83 (21.19)
		COMPLEMENT MODUS PONENS			
100	100	.00	.00	.00 (.00)	.00 (.00)
70	90	27	37	35.40 (16.73)	42.03 (17.45)
70	50	15	65	41.00 (18.82)	53.67 (17.71)
		DENYING THE ANTECEDENT			
100	100	.00	100	37.37 (47.53)	85.00 (35.11)
70	20	20	44	18.63 (15.25)	41.63 (15.97)
70	50	15	65	25.4 (21.12)	59.23 (20.73)
		COMPL. DENYING THE ANTECEDENT			
100	100	.00	100	0.83 (4.56)	53.33 (49.01)
70	20	56	80	51.9 (19.12)	75.87 (20.19)
70	50	35	85	32.70 (12.92)	65.17 (27.43)

Table 6: Mean lower LBR and mean upper bound responses UBR of the MODUS PONENS condition ($n_1 = 30$) and of the DENYING THE ANTECEDENT condition ($n_2 = 30$) of Experiment 2. The standard deviations are in parenthesis. P_1 and P_2 denote the percentages in the premises. CLB and CUB denote the normative/coherent bounds.

TECEDENT tasks. The order was reversed for the other fifteen participants. The premises and the conclusions were adopted accordingly. The only difference in the percentages between the MODUS PONENS condition and the DENYING THE ANTECEDENT condition was, that "90%" was replaced by "20%". The reason for this was to avoid non-informative assessments.

3.2 Results and Discussion

The results of t-tests indicate that there were no position effects. We therefore pooled the data in the MODUS PONENS condition ($n_1 = 30$) and in the DENYING THE ANTECEDENT condition ($n_2 = 30$).

Table 6 lists the probabilities presented in the premises, the normative lower and upper bounds, and the participants' mean lower and upper bound responses for the MODUS PONENS tasks and the COMPLEMENT MODUS PONENS tasks of Experiment 2.

In the MODUS PONENS tasks with certain premises ("100%" in both premises) all thirty participants responded with that point value 100%, which is normatively correct. In the according COMPLEMENT MODUS PONENS tasks all thirty participants responded correctly with the point value 0.00%. Thus, in Task 1 all participants inferred (correctly) point values. In the other tasks (both MODUS PONENS and COMPLEMENT MODUS PONENS), between 50 and 60% of the responses were point value responses, which is about double compared with Experiment 1.

This result is surprising, since dropping the explicit point value response modality should decrease and not increase the number of point value responses. A possible explanation is that, as for all participants the first task contained certain premises and as all participants responded by point values in the first task, they simply continued to give point value responses later on.

In the DENYING THE ANTECEDENT tasks with certain premises (i.e., "100%" in both premises) fourteen of the thirty participants inferred a unit interval, $[\leq 1, 100]\%$. Four participants inferred a point value equal to zero, and ten inferred a point value equal to 100%. One participant inferred a point value of 50% and one an interval between 70 and 100%. In the according COMPLEMENT DENYING THE ANTECEDENT tasks fifteen of the thirty participants inferred a unit interval, $[\leq 1, 100]\%$. Thirteen responded a point value equal to zero. One participant inferred a $[25, 50]\%$ interval and one inferred a $[0, 50]\%$ interval. Practically half of the participants understood that only a non-informative interval can be inferred if each premise is certain.

In the two DENYING THE ANTECEDENT tasks with 70 and 20%, and 70 and 50%, in the premises, 0 and 16.67% point value responses were observed, respectively. This amount of point value responses is smaller than in Experiment 1. In both according COMPLEMENT DENYING THE ANTECEDENT tasks 26.67% point value responses were observed, which is comparable to the results of Experiment 1.

Table 7 reports the frequencies of interval response categories of the MODUS PONENS condition and of the DENYING THE ANTECEDENT condition in 3×3 tables. 80.22% of the participants inferred coherent intervals in the MODUS PONENS condition on the average (the tasks with certain premises were not averaged). 55.00% of the participants inferred coherent intervals in the DENYING THE ANTECEDENT condition on the average (the tasks with certain premises were not averaged).

In the MODUS PONENS tasks 85.00% (Experiment 1: 62.93%) of the interval responses were coherent on the average. In the DENYING THE ANTECEDENT tasks 56.66% (Experiment 1: 41.33%) of the interval responses were coherent on the average. The improved cover-story explains why more coherent interval responses in Experiment 2 than in Experiment 1 were observed.

MODUS PONENS tasks							
	Task 2			*Task 3*			*Coh. Bounds*
UA	0	4	1	0	0	0	*Task 2:*
UW	2	**22**	-	1	**29**		63–73
UB	1	-	-	0	-	-	*Task 3:*
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	35–85
COMPLEMENT MODUS PONENS tasks							
	Task 2			*Task 3*			*Coh. Bounds*
UA	0	0	6	0	2	1	*Task 2:*
UW	5	**19**	-	0	**27**		27–37
UB	0	-	-	0	-	-	*Task 3:*
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	15–65
DENYING THE ANTECEDENT tasks							
	Task 2			*Task 3*			*Coh. Bounds*
UA	3	4	0	0	5	0	*Task 2:*
UW	6	**16**	-	7	**18**		20–44
UB	1	-	-	0	-	-	*Task 3:*
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	15–65
COMPLEMENT DENYING THE ANTECEDENT tasks							
	Task 2			*Task 3*			*Coh. Bounds*
UA	1	8	0	1	6	0	*Task 2:*
UW	6	**14**	-	4	**18**		56–80
UB	1	-	-	1	-	-	*Task 3:*
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	35–85

Table 7: Frequencies of the interval responses in the MODUS PONENS ($n_1 = 30$) and the DENYING THE ANTECEDENT condition ($n_2 = 30$) of Experiment 2. For explanation see Table 4.

All participants inferred a probability(interval) of a conclusion $\mathfrak{C}$, $P(\mathfrak{C}) \in [z'_{\mathfrak{C}}, z''_{\mathfrak{C}}]$, and the probability of the negated conclusion, $P(\neg\mathfrak{C}) \in [z'_{\neg\mathfrak{C}}, z''_{\neg\mathfrak{C}}]$. To test the conjugacy principle of the interval responses, we checked for each participant whether (i) $z'_{\mathfrak{C}} + z''_{\neg\mathfrak{C}} = 100\%$, and whether (ii) $z'_{\neg\mathfrak{C}} + z''_{\mathfrak{C}} = 100\%$.

In the MODUS PONENS tasks with certain premises, all participants satisfied both equalities, (i) and (ii). In the tasks with 70% and 90% in the premises sixteen of the thirty participants satisfied both, (i) and (ii). In the tasks with 70% and 50% in the premises fifteen of the thirty participants satisfied both, (i) and (ii). It is surprising that in the MODUS PONENS tasks more than half of the participants gave intervals that with lower/upper probabilities that exactly add up to 1. In the DENYING THE ANTECEDENT tasks with certain premises, twenty of the thirty participants satisfied both, (i) and (ii). In the tasks with 70 and 20% premises nobody satisfied both, (i) and (ii), eleven satisfied (i), and one satisfied (ii). In the tasks with 70 and 50% ten satisfied both, (i) and (ii).

In sum, more additive responses and more coherent interval responses were observed in MODUS PONENS tasks than in the DENYING THE ANTECEDENT tasks. It is reasonable that humans are better in argument forms that guarantee high probabilities of the conclusion if each premise is highly probable. If the premises of the MODUS PONENS are certain, then the conclusion is certain. However, if the premises of the DENYING THE ANTECEDENT are certain, then the probability of the conclusion is in the unit interval $[0, 1]$.

4 Experiment 3

This section reports data of an experiment with imprecise probabilities in the premises conducted by Florian Bauerecker [1]. Specifically, we focus on human understanding of what we call "silent bounds". We call a probability bound b of a premise *silent* if, and only if, b is *irrelevant* for the probability propagation from the premise(s) to the conclusion. E.g., in the probabilistic MODUS PONENS y'' is silent (y'' doesn't occur in the lower or upper probabilities of the conclusion, see (1)). Experiment 3 introduces an especially critical test of the claim that human subjects are capable to make coherent probabilistic inferences.

Method and procedure of Experiment 3 are analog to Experiment 1. Eighty participants were recruited for investigating questions going beyond the scope of the present study. Therefore, we report selected data on the imprecise MODUS PONENS only ($n = 40$). The MODUS PONENS tasks were formulated as follows:

> Please imagine the following situation. Claudia works at blood donation services. She investigates to which blood group the donated blood belongs and whether the donated blood is Rhesus-positive.
>
> > Claudia is 60% certain: If the donated blood belongs to the blood group 0, then the donated blood is Rhesus-positive.
> > Claudia knows as well that donated blood belongs with *more than* 75% certainty to the blood group 0.
>
> How certain should Claudia be that a recent donated blood is Rhesus-positive?

Contrary to Experiment 1 and Experiment 2, the conditional premise is here formulated in a if–then form. The cover-story remained constant, only the numbers in the premises varied. Table 8 lists the probabilities presented in the premises, the normative lower and upper bounds, and the participants' mean lower and upper bound responses for the MODUS PONENS tasks with and without silent bounds.

The participants inferred higher upper bounds in the MODUS PONENS task containing silent bounds ($M = 71.79$) compared with the according task not containing silent bounds ($M = 60.20$; t(39)=3.53, p=.001).

P_1/P_2	CI	LBR	UBR
60/[75,100*]	[45,70]	44.50 (21.57)	71.78 (20.07)
60/75	[45,70]	46.83 (23.76)	60.20 (16.86)
[75,100]/60	[45,100]	43.42 (22.00)	72.38 (22.98)
75/60	[45,85]	46.27 (21.73)	59.90 (17.19)

Table 8: Mean lower LBR and mean upper bound responses UBR of the MODUS PONENS tasks ($n = 40$) in [1]. * denotes the silent bound. The standard deviations are in parenthesis. P_1 and P_2 denote the percentages in the premises. CI denotes the normative/coherent interval.

	Task 1 (60/[75,100*])			*Task 2* (60/75)			*Coh. Bounds*
UA	1	14	0	0	2	4	*Task 1:*
UW	8	**16**	-	7	**25**		45–70
UB	1	-	-	2	-	-	*Task 2:*
	LB	LW	LA	LB	LW	LA	45–70
	Task 3 ([75,100]/60)			*Task 4* (75/60)			*Coh. Bounds*
UA	-	-	-	0	2	0	*Task 3:*
UW	7	**32**	-	8	**29**		45–100
UB	1	-	-	1	-	-	*Task 4:*
	LB	LW	LA	LB	LW	LA	45–85

Table 9: Frequencies of the interval responses in the MODUS PONENS ($n = 40$) tasks of Experiment 3. Coherent interval responses are bold (±5% tolerance interval). Further explanation in Table 4.

Thus the participants were sensitive to the silent bounds. They did not understand the irrelevance of the silent bound for the probability propagation from the premises to the conclusion.

[21] report data on a conjunction problem where in one condition interval-values in the premises were presented. All upper bounds were equal to 100%. In the other condition only corresponding point values were presented. The point values were equal to the lower bounds of the interval condition. Higher mean lower bounds were observed in the interval condition than in the point condition. An explanation for this finding is, that the participants reduced the processing load of the interval valued premises by representing only the means of the lower and upper bounds. Then, of course, the coherent lower bound must be higher.

This explanation of the [21] data on the conjunction problem is, however, not applicable to the data from the imprecise MODUS PONENS task. If the second premise (containing silent bounds) is represented as 88%, then the coherent interval of the conclusion is $[53\%, 65\%]$. Assuming that the participants represent "88" instead of the interval "$[75, 100]$", then the participants' mean upper bounds should be *lower* in the interval value condition than in the point value condition.

An alternative explanation is that higher explicit imprecision (by communicating interval-values in the premises) elicits larger interval responses. It could be that *conversational implicatures* [11] modulate the accumulation of imprecision. The participant assumes by conversational implicature that the experimenter communicates only relevant informations. Thus the silent bound is not understood as irrelevant, rather, the silent bound is understood by the participant as a hint from the experimenter to add imprecision to the conclusion: to infer wider intervals.

5 Concluding Remarks

We reported three psychological experiments on the probabilistic versions of two prominent argument forms in the framework of *mental probability logic*. Clearly more coherent responses were observed in MODUS PONENS than in DENYING THE ANTECEDENT tasks. Human subjects employ inference rules that guarantee high probability conclusions if the premises are highly probable. Practically no participant inferred "too wide" intervals such that the coherent intervals are subintervals. While most participants did not understand the irrelevance of the silent bounds in the MODUS PONENS task of Experiment 3 they are not completely "blind" for them. The close agreement of the mean responses and the normative values of the lower probabilities in Experiment 3 is stunning. One may speculate that human subjects are doing better in processing lower than upper probabilities. More than half of the participants responded with lower/upper probabilities that agreed perfectly with the conjugacy principle.

References

[1] F. Bauerecker. Konditionales Schließen mit Intervallen und Verständnis für das Komplement einer Schlussform. Master's thesis, Universität Salzburg, 2006.

[2] M. D. S. Braine and D. P. O'Brien, editors. *Mental logic*. Erlbaum, Mahwah, 1998.

[3] N. Chomsky. *Aspects of the theory of syntax*. MIT Press, Cambridge, 1965.

[4] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting*. Kluwer, Dordrecht, 2002.

[5] B. De Finetti. *Theory of probability*, volume 1, 2. John Wiley & Sons, Chichester, 1974. Original work published 1970.

[6] J. S. B. T. Evans, S. E. Newstead, and R. M. J. Byrne. *Human Reasoning*. Erlbaum, Hove, 1993.

[7] J. S. B. T. Evans, S. H. Handley, and D. E. Over. Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29:321–355, 2003.

[8] A. Gilio. Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence*, 34:5–34, 2002.

[9] A. Gilio and S. Ingrassia. Totally coherent set-valued probability assessments. *Kybernetika*, 34(1):3–15, 1998.

[10] T. Gilovich, D. Griffin, and D. Kahneman, editors. *Heuristics and biases. The psychology of intuitive judgment*. Cambridge University Press, Cambridge, 2002.

[11] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and semantics*, volume 3: Speech acts. Seminar Press, New York, 1975.

[12] P. N. Johnson-Laird, V. Girotto, P. Legrenzi, and M. S. Legrenzi. Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106(1):62–88, 1999.

[13] F. Lad. *Operational subjective statistical methods: A mathematical,philosophical, and historical introduction*. Wiley, New York, 1996.

[14] I.-M. Liu, K.-C. Lo, and J.-T. Wu. A probabilistic interpretation of 'If—Then'. *The Quarterly Journal of Experimental Psychology*, 49(A):828–844, 1996.

[15] D. Marr. *Vision. A computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco, 1982.

[16] M. Oaksford and N. Chater. Conditional probability and the cognitive science of conditional reasoning. *Mind & Language*, 18(4):359–379, 2003.

[17] M. Oaksford and N. Chater. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, Oxford, 2007.

[18] K. Oberauer and O. Wilhelm. The meaning(s) of conditionals: Conditional probabilities, mental models and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29:680–693, 2003.

[19] D. E. Over, C. Hadjichristidis, J. S. Evans, S. J. Handley, and S. A. Sloman. The probability of causal conditionals. *Cognitive Psychology*, 54:62–97, 2007.

[20] N. Pfeifer and G. D. Kleiter. Nonmonotonicity and human probabilistic reasoning. In *Proceedings of the 6th Workshop on Uncertainty Processing*, pages 221–234, Hejnice, 2003. September 24–27$^{\text{th}}$, 2003.

[21] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146(1-2):93–109, 2005.

[22] N. Pfeifer and G. D. Kleiter. Towards a mental probability logic. *Psychologica Belgica*, 45(1):71–99, 2005.

[23] N. Pfeifer and G. D. Kleiter. Inference in conditional probability logic. *Kybernetika*, 42:391–404, 2006.

[24] N. Pfeifer and G. D. Kleiter. Is human reasoning about nonmonotonic conditionals probabilistically coherent? In *Proceedings of the 7th Workshop on Uncertainty Processing*, pages 138–150, Mikulov, 2006. September 16–20$^{\text{th}}$, 2006.

[25] N. Pfeifer and G. D. Kleiter. Towards a probability logic based on statistical reasoning. In *Proceedings of the 11th IPMU Conference (Information Processing and Management of Uncertainty in Knowledge-Based Systems)*, pages 2308–2315, Paris, 2006. Edition E.D.K.

[26] G. Politzer. Uncertainty and the suppression of inferences. *Thinking & Reasoning*, 11(1):5–33, 2005.

[27] G. Politzer and G. Bourmaud. Deductive reasoning from uncertain conditionals. *British Journal of Psychology*, 93:345–381, 2002.

[28] L. J. Rips. *The psychology of proof: Deductive reasoning in human thinking*. MIT Press, Cambridge, 1994.

[29] G. Schurz. Non-monotonic reasoning from an evolution-theoretic perspective: Ontic, logical and cognitive foundations. *Synthese*, 1-2:37–51, 2005.

[30] M. J. Smithson. *Human judgment and imprecise probabilities*. `www.sipta.org/documentation`, 1997-2000.

[31] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Learning about a Categorical Latent Variable under Prior Near-Ignorance

Alberto Piatti
IDSIA
Switzerland
alberto.piatti@idsia.ch

Marco Zaffalon
IDSIA
Switzerland
zaffalon@idsia.ch

Fabio Trojani
University of S. Gallen
Switzerland
fabio.trojani@unisg.ch

Marcus Hutter
ANU&NICTA
Australia
marcus@hutter1.net

Abstract

It is well known that complete prior ignorance is not compatible with learning, at least in a coherent theory of (epistemic) uncertainty. What is less widely known, is that there is a state similar to full ignorance, that Walley calls *near-ignorance*, that permits learning to take place. In this paper we provide new and substantial evidence that also near-ignorance cannot be really regarded as a way out of the problem of starting statistical inference in conditions of very weak beliefs. The key to this result is focusing on a setting characterized by a variable of interest that is *latent*. We argue that such a setting is by far the most common case in practice, and we show, for the case of categorical latent variables (and general *manifest* variables) that there is a sufficient condition that, if satisfied, prevents learning to take place under prior near-ignorance. This condition is shown to be easily satisfied in the most common statistical problems.

Keywords. Prior near-ignorance, latent and manifest variables, observational processes, vacuous beliefs, imprecise probabilities.

1 Introduction

Epistemic theories of statistics are often concerned with the question of *prior ignorance*. Prior ignorance means that a subject, who is about to perform a statistical analysis, has not any substantial belief about the underlying data-generating process. Yet, the subject would like to exploit the available sample to draw some statistical inference, i.e., the subject would like to use the data to learn, moving away from the initial condition of ignorance. This situation is very important as it is often desirable to start a statistical analysis with weak assumptions about the problem of interest, thus trying to implement an objective-minded approach to statistics.

A fundamental question is if prior ignorance is compatible with learning. Walley gives a negative answer for the case of his self-consistent (or *coherent*) theory of statistics: he shows, in a very general sense, that *vacuous* prior beliefs lead to vacuous posterior beliefs, irrespective of the type and amount of observed data (Walley, 1991, Section 7.3.7). But, at the same time, he proposes focusing on a slighlty different state of beliefs, called *near-ignorance*, that does enable learning to take place (Walley, 1991, Section 4.6.9). Loosely speaking, near-ignorant beliefs are beliefs close but not equal to vacuous (see Section 3). The possibility to learn under prior near-ignorance is shown, for instance, in the special case of the near-ignorance prior defining the *imprecise Dirichlet model* (IDM). This is a popular model used in the case of inference from categorical data generated by a discrete process (Walley (1996); Bernard (2005)).

In this paper, we also focus on a categorical random variable X, expressing the outcomes of a multinomial process, but we assume that such a variable is *latent*. This means that we cannot observe the realizations of X, so we can learn about it only by means of another (not necessarily categorical) variable S, related to X in some known way. Variable S is assumed to be *manifest*, in the sense that its realizations can be observed (see Section 2).

In such a setting, we introduce a condition in Section 4, related to the likelihood of the observed data, that is shown to be sufficient to prevent learning about X under prior near-ignorance. The condition is very general as it is developed for any prior that models near-ignorance (not only the one used in the IDM), and for very general kinds of relation between X and S. We show then, by simple examples, that such a condition is easily satisfied, even in the most elementary and common statistical problems.

In order to appreciate this result, it is important to realize that latent variables are ubiquitous in problems of uncertainty. It can be argued, indeed, that there is a persistent distinction between (latent) facts (e.g., health, state of economy, color of a ball) and (manifest) observations of facts: one can regard them as being related by a so-called *observational process*; and the point is that these kinds of processes are imperfect in practice. Observational processes are often neglected in statistics, when their im-

perfection is deemed to be tiny. But a striking outcome of the present research is that, no matter how tiny the imperfection, provided it exists, learning is not possible under prior near-ignorance.

In our view, the present results raise serious doubts about the possibility to adopt a condition of prior near-ignorance in real, as opposed to idealized, applications of statistics. As a consequence, it may make sense to consider re-focusing the research about this subject on developing models of very weak states of belief that are, however, stronger than near-ignorance.

2 Categorical Latent Variables

In this paper, we follow the general definition of *latent* and *manifest variables* given by Skrondal and Rabe-Hesketh (2004): a *latent variable* is a random variable whose realizations are unobservable (hidden), while a *manifest variable* is a random variable whose realizations can be directly observed. The concept of latent variable is central in many sciences, like for example psychology and medicine. Skrondal and Rabe-Hesketh (2004) list several fields of application and several phenomena that can be modeled using latent variables, and conclude that latent variable modeling "*pervades modern mainstream statistics*," although "*this omni-presence of latent variables is commonly not recognized, perhaps because latent variables are given different names in different literatures, such as random effects, common factors and latent classes*," or hidden variables.

But what are latent variables in practice? According to Boorsbom et al. (2002), there may be different interpretations of latent variables. A latent variable can be regarded, for example, as an unobservable random variable that exists independently of the observation. An example is the unobservable health status of a patient that is subject to a medical test. Another possibility is to regard a latent variable as a product of the human mind, a construct that does not exist independent of the observation. For example the *unobservable state of the economy*, often used in economic models. In this paper, we assume the existence of a latent categorical random variable X, with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances $\theta \in \Theta := \{\theta = (\theta_1, \ldots, \theta_k) \mid \sum_{i=1}^{k} \theta_i = 1,\ 0 \leq \theta_i \leq 1\}$, without stressing any particular interpretation.

Suppose now that our aim is to predict, after N realizations of the variable X, the next outcome (or the next N' outcomes). Because the variable X is latent and therefore unobservable by definition, the only possible way to learn something about the probabilities of the next outcome is to observe the realizations of some manifest variable S related, in a known way, to the (unobservable) realizations of X. An example of known relationship between latent and manifest variables is the following.

Example 1 We consider a binary medical diagnostic test used to assess the health status of a patient with respect to a given disease. The accuracy of a diagnostic test[1] is determined by two probabilities: the *sensitivity* of a test is the probability of obtaining a positive result if the patient is diseased; the *specificity* is the probability of obtaining a negative result if the patient is healthy. Medical tests are assumed to be imperfect indicators of the unobservable true disease status of the patient. Therefore, we assume that the probability of obtaining a positive result when the patient is healthy, respectively of obtaining a negative result if the patient is diseased, are non-zero. Suppose, to make things simpler, that the sensitivity and the specificity of the test are known. In this example, the unobservable health status of the patient can be considered as a binary latent variable X with values in the set $\{\text{Healthy}, \text{Ill}\}$, while the result of the test can be considered as a binary manifest variable S with values in the set $\{\text{Negative result}, \text{Positive result}\}$. Because the sensitivity and the specificity of the test are known, we know how X and S are related. ♢

We continue discussion about this example later on, in the light of our results, in Example 2 of Section 4.

3 Near-Ignorance Priors

Consider a categorical random variable X with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances $\theta \in \Theta$. Suppose that we have no relevant prior information about θ and we are therefore in a situation of prior ignorance. How should we model our prior beliefs in order to reflect the initial lack of knowledge?

Let us give a brief overview of this topic in the case of coherent models of uncertainty, such as Bayesian probability and Walley's theory of *coherent lower previsions*.

In the traditional Bayesian setting, prior beliefs are modeled using a single prior probability distribution. The problem of defining a standard prior probability distribution modeling a situation of prior ignorance, a so-called *noninformative prior*, has been an important research topic in the last two centuries[2] and, despite the numerous contributions, it remains an open research issue, as illustrated by Kass and Wassermann (1996). See also Hutter (2006) for recent developments and complementary considerations. There are many principles and properties that are desirable to model a situation of prior ignorance and that have been used in past research to define noninformative priors. For example Laplace's *symmetry or indifference* principle has

[1] For further details about the modeling of diagnostic accuracy with latent variables see Yang and Becker (1997).

[2] Starting from the work of Laplace at the beginning of the 19th century (Laplace (1820)).

suggested, in case of finite possibility spaces, the use of the uniform distribution. Other principles, like for example the principle of *invariance under group transformations*, the *maximum entropy* principle, the *conjugate priors* principle, etc., have suggested the use of other noninformative priors, in particular for continuous possibility spaces, satisfying one or more of these principles. But, in general, it has proven to be difficult to define a standard noninformative prior satisfying, at the same time, all the desirable principles.

In the case of finite possibility spaces, we agree with De Cooman and Miranda (2006) when they say that there are at least two principles that should be satisfied to model a situation of prior ignorance: the *symmetry principle* and the *embedding principle*. The *symmetry principle* states that, if we are completely ignorant a priori about θ, then we have no reason to favour one possible outcome of X to another, and therefore our probability model on θ should be symmetric. This principle recalls Laplace's *symmetry or indifference* principle that, in the past decades, has suggested the use of the *uniform prior* as standard noninformative prior. The *embedding principle* states that, for each possible event A, the probability assigned to A should not depend on the possibility space $\mathcal{X}$ in which A is embedded. In particular, the probability assigned a priori to the event A should be invariant with respect to refinements and coarsenings of $\mathcal{X}$. It is easy to show that the embedding principle is not satisfied by the uniform distribution. How should we model our prior ignorance in order to satisfy these two principles? Walley (1991) gives a compelling answer to this question: he proves[3] that the only probability model consistent with coherence and with the two principles is the *vacuous probability model*, i.e., the model that assigns, for each non-trivial event A, lower probability $\underline{\mathrm{P}}(A) = 0$ and upper probability $\overline{\mathrm{P}}(A) = 1$. It is evident that this model cannot be expressed using a single probability distribution. It follows that, to model properly and in a coherent way a situation of prior ignorance, we need *imprecise probabilities*.[4]

Unfortunately, adopting the vacuous probability model for X is not a practical solution to our initial problem, because it produces only vacuous posterior probabilities. Walley (1991) suggests, as practical solution, the use of *near-ignorance priors*. A near-ignorance prior is a large closed convex set $\mathcal{M}_0$ of probability distributions for θ, very close to the vacuous probability model, which produces a priori *vacuous expectation*s for various functions f on Θ, i.e., such that $\underline{\mathbf{E}}(f) = \inf_{\theta\in\Theta} f(\theta)$ and $\overline{\mathbf{E}}(f) = \sup_{\theta\in\Theta} f(\theta)$.

An example of near-ignorance prior that is particularly instructive is the set of priors $\mathcal{M}_0$ used in the *imprecise Dirichlet model* (IDM). The IDM models a situation of prior ignorance about the chances θ of a categorical random variable X. The near-ignorance prior $\mathcal{M}_0$ used in the IDM consists in the set of all Dirichlet densities $p(\theta) = dir_{s,\mathbf{t}}(\theta)$ for a fixed $s > 0$ and all $\mathbf{t} \in \mathcal{T}$, where

$$dir_{s,\mathbf{t}}(\theta) := \frac{\Gamma(s)}{\prod_{i=1}^{k}\Gamma(st_i)} \prod_{i=1}^{k} \theta_i^{st_i-1}, \qquad (1)$$

and

$$\mathcal{T} := \{\mathbf{t} = (t_1, \ldots, t_k) \mid \sum_{j=1}^{k} t_k = 1,\ 0 < t_j < 1\}. \quad (2)$$

The particular choice of $\mathcal{M}_0$ in the IDM implies vacuous prior expectations for all functions $f(\theta) = \theta_i^{N'}$, for all $N' \geq 1$ and all $i \in \{1, \ldots, k\}$, i.e., $\underline{\mathbf{E}}(\theta_i^{N'}) = 0$ and $\overline{\mathbf{E}}(\theta_i^{N'}) = 1$. Choosing $N' = 1$, we have, a priori,

$$\underline{\mathrm{P}}(\mathrm{X} = x_i) = \underline{\mathbf{E}}(\theta_i) = 0, \quad \overline{\mathrm{P}}(\mathrm{X} = x_i) = \overline{\mathbf{E}}(\theta_i) = 1.$$

It follows that the particular near-ignorance prior $\mathcal{M}_0$ used in the IDM implies vacuous prior probabilities for each possible outcome of the variable X. It can be shown that this particular set of priors satisfies both the symmetry and embedding principles.

But what is the difference between the vacuous probability model and the the near-ignorance prior used in the IDM? In fact, although both models produce vacuous prior probabilities and both models satisfy the symmetry and embedding principles, the IDM yields posterior probabilities that are not vacuous, while the vacuous probability model produces only vacuous posterior probabilities. The answer to this question is the reason why we use the term *near-ignorance*: in the IDM, although we are completely ignorant about the possible outcomes of the variable X, we are not completely ignorant about the chances θ, because we assume a particular class of prior distributions, i.e., the Dirichlet distributions for a fixed value of s.

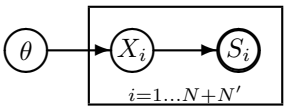
4 Limits of Learning under Prior Near-Ignorance

Consider a sequence of independent and identically distributed (IID) categorical latent variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}$ and unknown chances $\theta \in \Theta$, and a sequence of independent manifest variables $(\mathrm{S}_i)_{i\in\mathbf{N}}$. We assume that a realization of the manifest variable S_i can be observed only after an (unobservable) realization of the latent variable X_i and that the probability distribution of S_i given X_i is known for each $i \in \mathbf{N}$. Furthermore, we assume S_i to be independent of the chances θ of X_i given X_i. Define the random variables $\mathbf{X} := (\mathrm{X}_1, \ldots, \mathrm{X}_N)$, $\mathbf{S} := (\mathrm{S}_1, \ldots, \mathrm{S}_N)$ and $\mathbf{X'} := (\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'})$.

[3] In Note 7, p. 526. See also Section 5.5.

[4] For a complementary point of view, see Hutter (2006).

We focus on the problem of predictive inference.[5] Suppose that we observe a dataset $\mathbf{s}$ of realizations of manifest variables $\mathrm{S}_1, \ldots, \mathrm{S}_N$ related to the (unobservable) dataset $\mathbf{x} \in \mathcal{X}^N$ of realizations of the variables $\mathrm{X}_1, \ldots, \mathrm{X}_N$. Using the notation defined above we have $\mathbf{S} = \mathbf{s}$ and $\mathbf{X} = \mathbf{x}$. Our aim is to predict the outcomes of the next N' variables $\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'}$. In particular, given $\mathbf{x}' \in \mathcal{X}^{N'}$, our aim is to calculate $\underline{\mathrm{P}}(\mathbf{X}' = \mathbf{x}' \,|\, \mathbf{S} = \mathbf{s})$ and $\overline{\mathrm{P}}(\mathbf{X}' = \mathbf{x}' \,|\, \mathbf{S} = \mathbf{s})$. To simplify notation, when no confusion is possible, we denote in the rest of the paper $\mathbf{S} = \mathbf{s}$ with $\mathbf{s}$ and $\mathbf{X}' = \mathbf{x}'$ with $\mathbf{x}'$. The (in)dependence structure can be depicted graphically as follows:



Modelling our prior ignorance about the parameters θ with a near-ignorance prior $\mathcal{M}_0$ and denoting by $\mathbf{n}' := (n'_1, \ldots, n'_k)$ the frequencies of the dataset $\mathbf{x}'$, we have

$$\begin{aligned}
\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) &= \inf_{p \in \mathcal{M}_0} \mathrm{P}_p(\mathbf{x}' \,|\, \mathbf{s}) := \\
&= \inf_{p \in \mathcal{M}_0} \int_\Theta \prod_{i=1}^k \theta_i^{n'_i} p(\theta \,|\, \mathbf{s}) d\theta = \\
&=: \inf_{p \in \mathcal{M}_0} \mathbf{E}_p \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right) = \\
&= \underline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right),
\end{aligned}$$

where, according to Bayes theorem,

$$p(\theta \,|\, \mathbf{s}) = \frac{\mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta)}{\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta) d\theta},$$

provided that $\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta) d\theta \neq 0$. Analogously, substituting $\sup$ to $\inf$ in (3), we obtain

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right). \tag{3}$$

The central problem now is to choose $\mathcal{M}_0$ so as to be as ignorant as possible a priori and, at the same time, to be able to learn something from the observed dataset of manifest variables $\mathbf{s}$. Theorem 1 and the following corollaries yield a first partial solution to the above problem, stating several conditions for learning under prior near-ignorance.

Theorem 1 *Let $\mathbf{s}$ be given. Consider a bounded continuous function f defined on Θ and denote with $f_{\max}$ the Supremum of f on Θ. If the likelihood function $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is strictly positive*[6] *in each point in which f reaches its maximum value $f_{\max}$ and it is continuous in an arbitrary small neighborhood of these points, and $\mathcal{M}_0$ is such that a priori $\overline{\mathbf{E}}(f) = f_{\max}$, then*

$$\overline{\mathbf{E}}(f \,|\, \mathbf{s}) = \overline{\mathbf{E}}(f) = f_{\max}.$$

Many corollaries to Theorem 1 are listed in Section B of the Appendix. Here we discuss only the most important corollary. Consider, given a dataset $\mathbf{x}'$, the particular function $f(\theta) = \prod_{i=1}^k \theta_i^{n'_i}$. This function is particularly important for predictive inference, because its lower and upper expectations correspond to the lower and upper probabilities assigned to the dataset $\mathbf{x}'$. It is easy to show that, in this case, the minimum of f is 0 and is reached in all the points $\theta \in \Theta$ with $\theta_i = 0$ for some i such that $n'_i > 0$, while the maximum of f is reached in a single point of Θ corresponding to the relative frequencies $\mathbf{f}'$ of the sample $\mathbf{x}'$, i.e., at $\mathbf{f}' = \left(\frac{n'_1}{N'}, \ldots, \frac{n'_k}{N'} \right) \in \Theta$, and the maximum of f is given by $\prod_{i=1}^k \left(\frac{n'_i}{N'} \right)^{n'_i}$. It follows that vacuous probabilities regarding the dataset $\mathbf{x}'$ are given by

$$\underline{\mathrm{P}}(\mathbf{x}') = \underline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \right) = 0,$$

$$\overline{\mathrm{P}}(\mathbf{x}') = \overline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \right) = \prod_{i=1}^k \left(\frac{n'_i}{N'} \right)^{n'_i}.$$

Corollary 1 *Let $\mathbf{s}$ be given and let $\mathrm{P}(\mathbf{s} \,|\, \theta)$ be a continuous strictly positive function on Θ. Then, if $\mathcal{M}_0$ implies vacuous prior probabilities for a dataset $\mathbf{x}' \in \mathcal{X}^{N'}$, the predictive probabilities of $\mathbf{x}'$ are vacuous also a posteriori, after having observed $\mathbf{s}$, i.e.,*

$$\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \underline{\mathrm{P}}(\mathbf{x}') = 0,$$

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathrm{P}}(\mathbf{x}') = \prod_{i=1}^k \left(\frac{n'_i}{N'} \right)^{n'_i}.$$

In other words, Corollary 1 states a sufficient condition that prevents learning to take place under prior near-ignorance: if the likelihood function $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is continuous and strictly positive on Θ, then all the dataset $\mathbf{x}' \in \mathcal{X}^{N'}$ for which $\mathcal{M}_0$ implies vacuous probabilities have vacuous probabilities also a posteriori, after having observed $\mathbf{s}$. It follows that, if this sufficient condition is satisfied, we cannot use near-ignorance priors to model a state of prior ignorance for the same reason for which, in Section 3,

[5] For a general presentation of predictive inference see Geisser (1993); for a discussion of the imprecise probability approach to predictive inference see Walley et al. (1999).

[6] The Assumption about $\mathrm{P}(\mathbf{s} \,|\, \theta)$ in Theorem 1 can be substituted by the following weaker assumption. For a given arbitrary small $\delta > 0$, denote with Θ_δ the measurable set, $\Theta_\delta := \{\theta \in \Theta \,|\, f(\theta) \geq f_{\max} - \delta\}$. If $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is such that, $\lim_{\delta \to 0} \inf_{\theta \in \Theta_\delta} \mathrm{P}(\mathbf{s} \,|\, \theta) = c > 0$, then Theorem 1 holds.

we have excluded the vacuous probability model: because only vacuous posterior probabilities are produced.

The sufficient condition described above is satisfied very often in practice, as illustrated by the following striking examples.

Example 2 Consider the medical test introduced in Example 1 and an (ideally) infinite population of individuals. Denote with the binary variable $\mathrm{X}_i \in \{H, I\}$ the health status of the i-th individual of the population and with $\mathrm{S}_i \in \{+, -\}$ the results of the diagnostic test applied to the same individual. We assume that the variables in the sequence $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are IID with unknown chances $(\theta, 1-\theta)$, where θ corresponds to the (unknown) proportion of diseased individuals in the population. Denote with $1-\varepsilon_1$ the sensitivity and with $1-\varepsilon_2$ the specificity of the test. Then it holds that

$$\mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = H) = \varepsilon_1 > 0,$$

$$\mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = I) = \varepsilon_2 > 0,$$

where $(I, H, +, -)$ denote (patient ill, patient healthy, test positive, test negative).

Suppose that we observe the results of the test applied to N different individuals of the population; using our previous notation we have $\mathbf{S} = \mathbf{s}$. For each individual we have,

$$\begin{aligned}
&\mathrm{P}(\mathrm{S}_i = + \,|\, \theta) = \\
=&\mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = I)\mathrm{P}(\mathrm{X}_i = I \,|\, \theta) + \\
+&\mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = H)\mathrm{P}(\mathrm{X}_i = H \,|\, \theta) = \\
=&\underbrace{(1-\varepsilon_2)}_{>0}\cdot\theta + \underbrace{\varepsilon_1}_{>0}\cdot(1-\theta) > 0.
\end{aligned}$$

Analogously,

$$\begin{aligned}
&\mathrm{P}(\mathrm{S}_i = - \,|\, \theta) = \\
=&\mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = I)\mathrm{P}(\mathrm{X}_i = I \,|\, \theta) + \\
+&\mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = H)\mathrm{P}(\mathrm{X}_i = H \,|\, \theta) = \\
=&\underbrace{\varepsilon_2}_{>0}\cdot\theta + \underbrace{(1-\varepsilon_1)}_{>0}\cdot(1-\theta) > 0.
\end{aligned}$$

Denote with $n^{\mathbf{s}}$ the number of positive tests in the observed sample $\mathbf{s}$. Then, because the variables S_i are independent, we have

$$\begin{aligned}
&\mathrm{P}(\mathbf{S} = \mathbf{s} \,|\, \theta) = ((1-\varepsilon_2)\cdot\theta + \varepsilon_1\cdot(1-\theta))^{n^{\mathbf{s}}}\cdot \\
&\cdot(\varepsilon_2\cdot\theta + (1-\varepsilon_1)\cdot(1-\theta))^{N-n^{\mathbf{s}}} > 0
\end{aligned}$$

for each $\theta \in [0,1]$ and each $\mathbf{s} \in \mathcal{X}^N$. Therefore, according to Corollary 1, all the predictive probabilities that, according to $\mathcal{M}_0$, are vacuous a priori remain vacuous a posteriori. It follows that, if we want to avoid vacuous posterior predictive probabilities, then we cannot model our prior knowledge (ignorance) using a near-ignorance prior implying some vacuous prior predictive probabilities. This simple example shows that our previous theoretical results raise serious questions about the use of near-ignorance priors also in very simple, common, and important situations.

The situation presented in this example can be extended, in a straightforward way, to the general categorical case and has been studied, in the special case of the near-ignorance prior used in the imprecise Dirichlet model, in Piatti et al. (2005). ♢

Example 2 focuses on discrete latent and manifest variables. In the next example, we show that our theoretical results have important implications also in models with discrete latent variables and continuous manifest variables.

Example 3 Consider the sequence of IID categorical variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}^N$ and unknown chances $\theta \in \Theta$. Suppose that, for each $i \geq 1$, after a realization of the latent variable X_i, we can observe a realization of a continuous manifest variable S_i. Assume that $p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)$ is a continuous positive probability density, e.g., a normal $N(\mu_j, \sigma_j^2)$ density, for each $x_j \in \mathcal{X}$. We have

$$\begin{aligned}
p(\mathrm{S}_i \,|\, \theta) &= \sum_{x_j\in\mathcal{X}^N} p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)\cdot\mathrm{P}(\mathrm{X}_i = x_j \,|\, \theta) = \\
&= \sum_{x_j\in\mathcal{X}^N} \underbrace{p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)}_{>0}\cdot\theta_j > 0,
\end{aligned}$$

because θ_j is positive for at least one $j \in \{1, \ldots, N\}$ and we have assumed S_i to be independent of θ given X_i. Because we have assumed $(\mathrm{S}_i)_{i\in\mathbf{N}}$ to be a sequence of independent variables, we have,

$$p(\mathbf{S} = \mathbf{s} \,|\, \theta) = \prod_{i=1}^{N} \underbrace{p(\mathrm{S}_i = \mathbf{s}_i \,|\, \theta)}_{>0} > 0.$$

Therefore, according to Corollary 1, if we model our prior knowledge using a near-ignorance prior $\mathcal{M}_0$, the vacuous prior predictive probabilities implied by $\mathcal{M}_0$ remain vacuous a posteriori. It follows that, if we want to avoid vacuous posterior predictive probabilities, we cannot model our prior knowledge using a near-ignorance prior implying some vacuous prior predictive probabilities. ♢

Examples 2 and 3 raise, in general, serious criticisms about the use of near-ignorance priors in practical applications.

The only predictive model in the literature, of which we are aware, where a near-ignorance prior is used successfully to obtain non-vacuous posterior predictive probabilities is the IDM. In the next example, we explain how the IDM avoids our theoretical limitations.

Example 4 In the IDM, we assume that the IID categorical variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are observable. In other words, we have $\mathrm{S}_i = \mathrm{X}_i$ for each $i \geq 1$ and therefore the IDM is not a latent variable model. Having observed $\mathbf{S} = \mathbf{X} = \mathbf{x}$, we have

$$\mathrm{P}(\mathbf{S} = \mathbf{x} \,|\, \theta) = \mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) = \prod_{i=1}^{k} \theta_i^{n_i},$$

where n_i denotes the number of times that $x_i \in \mathcal{X}$ has been observed in $\mathbf{x}$. We have $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) = 0$ for all θ such that $\theta_j = 0$ for at least one j such that $n_j > 0$ and $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) > 0$ for all the other $\theta \in \Theta$, in particular for all θ in the interior of Θ.

The near-ignorance prior $\mathcal{M}_0$ used in the IDM consists in the set of all the Dirichlet densities $dir_{s,\mathbf{t}}(\theta)$ for a fixed $s > 0$ and all $\mathbf{t} \in \mathcal{T}$, where $dir_{s,\mathbf{t}}(\theta)$ and $\mathcal{T}$ have been defined in (1) and (2).

The particular choice of $\mathcal{M}_0$ in the IDM implies, for each $N' \geq 1$ and each $i \in \{1, \ldots, k\}$, that

$$\underline{\mathbf{E}}(\theta_i^{N'}) = 0, \quad \overline{\mathbf{E}}(\theta_i^{N'}) = 1.$$

Consequently, denoting with $\mathbf{d^i} \in \mathcal{X}^{N'}$ the dataset with $n'_i = N'$ and $n'_j = 0$ for each $j \neq i$, a priori we have,

$$\underline{\mathrm{P}}(\mathbf{X}' = \mathbf{d^i}) = 0, \quad \overline{\mathrm{P}}(\mathbf{X}' = \mathbf{d^i}) = 1,$$

and in particular

$$\underline{\mathrm{P}}(\mathrm{X}_1 = x_i) = 0, \quad \overline{\mathrm{P}}(\mathrm{X}_1 = x_i) = 1.$$

It can be shown that other prior predictive probabilities are not vacuous. For example, for $i \neq j$, we have

$$\overline{\mathbf{E}}(\theta_i\theta_j) = \frac{s}{4(s+1)} < \frac{1}{4} = \sup_{\theta\in\Theta} \theta_i\theta_j.$$

The IDM produces, for each possible observed data set $\mathbf{x}$, non-vacuous posterior predictive probabilities for each possible future data set (see Walley (1996)). This means that our previous theoretical limitations are avoided in some way. To explain this result we consider two cases. We consider firstly an observed data set $\mathbf{x}$ where we have observed at least two different outcomes. Secondly, we consider a data set $\mathbf{x}$ formed exclusively by outcomes of the same type, in other words, a data set of the type $\mathbf{d^i}$.

In the first case we have that $\mathrm{P}(\mathbf{x} \,|\, \theta) = \prod_{j=1}^{k} \theta_j^{n_j}$ is equal to zero for $\theta = \mathbf{e^i}$ for each $i \in \{1, \ldots, k\}$. In fact, $\theta_i = 1$ implies $\theta_j = 0$ for each $j \neq i$ and there is at least one j with $n_j > 0$. Therefore, the assumptions of Corollaries 4 and 5 are not satisfied. And in fact the IDM produces non-vacuous posterior predictive probabilities for each data set that, a priori, has vacuous predictive probabilities. On the other hand, all the datasets whose prior predictive probability reaches its maximum in a relative frequency $\mathbf{f} \in \Theta$ such that $\mathrm{P}(\mathbf{x} \,|\, \mathbf{f}) > 0$, are characterized by non-vacuous prior predictive probabilities.

The second case yields similar results. The only difference is that $\mathrm{P}(\mathbf{d^i} \,|\, \theta) = \theta_i^{N'}$ for a given $i \in \{1, \ldots, k\}$. In this case $\mathrm{P}(\mathbf{x} \,|\, \mathbf{e^i}) = 1 > 0$ and in fact, according to Corollaries 4 and 5, we obtain

$$\overline{\mathrm{P}}(x_i \,|\, \mathbf{x}) = \overline{\mathrm{P}}(x_i) = 1,$$

$$\overline{\mathrm{P}}(\mathrm{X}' = \mathbf{d^i} \,|\, \mathbf{x}) = \overline{\mathrm{P}}(\mathbf{d^i}) = 1,$$

and consequently, for each $j \neq i$ and each $\mathbf{y} \neq \mathbf{d^i}$,

$$\underline{\mathrm{P}}(x_j \,|\, \mathbf{x}) = \underline{\mathrm{P}}(x_j) = 0,$$

$$\underline{\mathrm{P}}(\mathrm{X}' = \mathbf{y} \,|\, \mathbf{x}) = \underline{\mathrm{P}}(\mathbf{y}) = 0.$$

But, on the other hand, we obtain

$$\underline{\mathrm{P}}(x_i \,|\, \mathbf{x}) > 0, \quad \underline{\mathrm{P}}(\mathrm{X}' = \mathbf{d^i} \,|\, \mathbf{x}) > 0,$$

$$\overline{\mathrm{P}}(x_j \,|\, \mathbf{x}) < 1, \quad \overline{\mathrm{P}}(\mathrm{X}' = \mathbf{y} \,|\, \mathbf{x}) < 1,$$

and therefore the posterior predictive probabilities are not vacuous for each possible future data set. $\diamondsuit$

Yet, since the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are assumed to be observable, the successful application of a near-ignorance prior in the IDM is not helpful in addressing the doubts raised by our theoretical results about the applicability of near-ignorance priors in situations where the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are latent.

5 Conclusions

In this paper we have proved a sufficient condition that prevents learning about a latent categorical variable to take place under prior near-ignorance about the data-generating process.

The condition holds as soon as the likelihood is strictly positive (and continuous), and so is satisfied frequently, even in the simplest settings. Taking into account that the considered framework is very general and pervasive of statistical practice, we regard this result as a form of substantial evidence against the possibility to use prior near-ignorance in real statistical problems. Given that complete prior ignorance is not compatible with learning, as it is well known, we deduce that there is little hope to use any form of prior ignorance to do objective-minded statistical inference in practice.

As a consequence, we suggest that future research efforts should be directed to study and develop new forms of knowledge that are close to near-ignorance but that do not coincide with it.

Acknowledgements

This work was partially supported by Swiss NSF grants 200021-113820/1 (Alberto Piatti), 200020-109295/1 (Marco Zaffalon) and 100012-105745/1 (Fabio Trojani).

A Technical preliminaries

In this appendix we provide some technical results that are used to prove the theorems in the paper. First of all, we introduce some notation used in this appendix. Consider a sequence of probability densities $(p_n)_{n\in\mathbf{N}}$ and a function f defined on a set Θ. Then, we use the notation,

$$\mathbf{E}_n(f) := \int_\Theta f(\theta)p_n(\theta)d\theta,$$

$$\mathrm{P}_n(\widetilde{\Theta}) := \int_{\widetilde{\Theta}} p_n(\theta)d\theta, \quad \widetilde{\Theta}\subseteq\Theta.$$

In addition, for a given probability density p on Θ,

$$\mathbf{E}_p(f) := \int_\Theta f(\theta)p(\theta)d\theta,$$

$$\mathrm{P}_p(\widetilde{\Theta}) := \int_{\widetilde{\Theta}} p(\theta)d\theta, \quad \widetilde{\Theta}\subseteq\Theta.$$

Finally, with $\to$ we denote $\lim_{n\to\infty}$.

Theorem 2 *Let $\Theta \subset \mathbf{R}^k$ be the closed k-dimensional simplex and let $(p_n)_{n\in\mathbf{N}}$ be a sequence of probability densities defined on Θ w.r.t. the Lebesgue measure. Let $f \geq 0$ be a bounded continuous function on Θ and denote with $f_{\max}$ the supremum of f on Θ. For this function define the measurable sets*

$$\Theta_\delta = \{\theta\in\Theta \,|\, f(\theta) \geq f_{\max}-\delta\}. \tag{4}$$

Assume that $(p_n)_{n\in\mathbf{N}}$ concentrates on a maximum of f for $n\to\infty$, in the sense that

$$\mathbf{E}_n(f) \to f_{\max}, \tag{5}$$

then, for all $\delta > 0$, it holds

$$\mathrm{P}_n(\Theta_\delta) \to 1.$$

Theorem 3 *Let $L(\theta)\geq 0$ be a bounded measurable function with*

$$\lim_{\delta\to 0}\inf_{\theta\in\Theta_\delta} L(\theta) =: c > 0, \tag{6}$$

under the same assumptions of Theorem 2. Then

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} = \frac{\int_\Theta f(\theta)L(\theta)p_n(\theta)d\theta}{\int_\Theta L(\theta)p_n(\theta)d\theta} \to f_{\max}. \tag{7}$$

Remark 1 *If f has a unique maximum in $\theta=\theta_0$ and L is a function, continuous in an arbitrary small neighborhood of $\theta=\theta_0$, such that $L(\theta_0)>0$, then (6) is satisfied.*

B Corollaries to Theorem 1

The following Corollaries to Theorem 1 are necessary to prove Corollary 1, and are useful to understand more deeply the limiting results implied by the use of near-ignorance priors with latent variables.

Corollary 2 *Let $\mathbf{x}'$ and $\mathbf{s}$ be given. Denote with $\mathbf{f}' := \left(\frac{n'_1}{N'},\ldots,\frac{n'_k}{N'}\right)\in\Theta$ the vector of relative frequencies of the dataset $\mathbf{x}'$. If $\mathrm{P}(\mathbf{s}\,|\,\theta)$ is continuous in an arbitrary small neighborhood of $\theta=\mathbf{f}'$, $\mathrm{P}(\mathbf{s}\,|\,\mathbf{f}')>0$ and $\mathcal{M}_0$ is such that*

$$\overline{\mathrm{P}}(\mathbf{x}') = \sup_{\theta\in\Theta}\left(\prod_{i=1}^k \theta_i^{n'_i}\right) = \prod_{i=1}^k\left(\frac{n'_i}{N'}\right)^{n'_i},$$

then

$$\overline{\mathrm{P}}(\mathbf{x}'\,|\,\mathbf{s}) = \overline{\mathrm{P}}(\mathbf{x}').$$

Corollary 3 *Let $\mathbf{x}'$ and $\mathbf{s}$ be given. If $\mathrm{P}(\mathbf{s}\,|\,\theta)>0$ for each $\theta\in\Theta$ with $\theta_i=0$ for at least one i with $n'_i>0$, and $\mathcal{M}_0$ is such that $\underline{\mathrm{P}}(\mathbf{x}')=0$, it follows that*

$$\underline{\mathrm{P}}(\mathbf{x}'\,|\,\mathbf{s}) = \underline{\mathrm{P}}(\mathbf{x}') = 0.$$

Corollary 4 *Let $\mathbf{s}$ be given. Consider an arbitrary $x_i\in\mathcal{X}$ and denote with $\mathbf{e^i}$ the particular vector of chances with $\theta_i=1$ and $\theta_j=0$ for each $j\neq i$. Suppose that $\mathcal{M}_0$ is such that, a priori, $\overline{\mathrm{P}}(\mathrm{X}_1=x_i) := \overline{\mathbf{E}}(\theta_i)=1$. Then, if $\mathrm{P}(\mathbf{s}\,|\,\mathbf{e^i})>0$ and $\mathrm{P}(\mathbf{s}\,|\,\theta)$ is continuous in a neighborhood of $\theta=\mathbf{e^i}$, we have*

$$\overline{\mathrm{P}}(\mathrm{X}_{N+1}=x_i\,|\,\mathbf{s}) = \overline{\mathrm{P}}(\mathrm{X}_1=x_i) = 1, \tag{8}$$

and consequently,

$$\underline{\mathrm{P}}(\mathrm{X}_{N+1}=x_j\,|\,\mathbf{s}) = \underline{\mathrm{P}}(\mathrm{X}_j=x_i) = 0, \tag{9}$$

for each $j\neq i$.

Corollary 5 *Let $\mathbf{s}$ and N' be given and consider an arbitrary $x_i\in\mathcal{X}$. Suppose that $\mathcal{M}_0$ is such that, a priori, $\overline{\mathrm{P}}(\mathrm{X}_1=x_i) := \overline{\mathbf{E}}(\theta_i)=1$. Denote with $\mathbf{d^i}\in\mathcal{X}^{N'}$ the data set with $n_i=N'$ and $n_j=0$ for each $j\neq i$. Then, if $\mathrm{P}(\mathbf{s}\,|\,\mathbf{e^i})>0$ and $\mathrm{P}(\mathbf{s}\,|\,\theta)$ is continuous in a neighborhood of $\theta=\mathbf{e^i}$, we have*

$$\overline{\mathrm{P}}(\mathbf{X}'=\mathbf{d^i}\,|\,\mathbf{s}) = 1,$$

and consequently,

$$\underline{\mathrm{P}}(\mathbf{X}'=\mathbf{y}\,|\,\mathbf{s}) = 0,$$

for each $\mathbf{y}\neq\mathbf{d^i}$.

References

Bernard J. M. (2005) An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39 (2–3), 123–150.

Boorsbom D., Mellenbergh G. J., van Heerden J. (2002) The theoretical status of latent variables. *Psychological Review*, 110 (2), 203–219.

De Cooman G., Miranda E. (2006) Symmetry of models versus models of symmetry. In *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr*. Eds. Hofer W. and Wheeler G., 82 pages. King's College Publications, London.

Geisser S. (1993) *Predictive Inference: An Introduction.* Monographs on Statistics and Applied Probability 55. Chapman and Hall, New York.

Hutter M. (2006) On the foundations of universal sequence prediction. In *Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC'06)*, 408–420, Beijing.

Kass R., Wassermann L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91: 1343–1370.

Laplace P. S. (1820), *Essai Philosophique sur Les Probabilités*. English translation: *Philosophical Essays on Probabilities* (1951), New York: Dover.

Piatti A., Zaffalon M., Trojani F. (2005) Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model, in: Cozman, F. G., Nau, B., Seidenfeld, T. (Eds), *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications.*, Manno (Switzerland), 276–286.

Skrondal A., Rabe-Hasketh S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman and Hall/CRC, Boca Raton.

Yang I., Becker M. P. (1997) Latent variable modeling of diagnostic accuracy. *Biometrics*, 53: 948–958.

Walley P. (1991) *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York.

Walley P. (1996) Inferences from multinomial data: learning about a bag of marbles. *J. R. Statistic. Soc. B* , 58(1): 3–57.

Walley P., Bernard J-M. (1999) Imprecise probabilistic prediction for categorical data. Tech. Rep. CAF-9901, Laboratoire Cognition et Activités Finalisées. Université Paris 8, Saint-Denis, France.

Conditioning in Chaotic Probabilities Interpreted as a Generalized Markov Chain

Leandro Chaves Rêgo
Statistics Department
Federal University of Pernambuco
Recife-PE Brazil
Email: leandro@de.ufpe.br

Abstract

We propose a new definition for conditioning in the Chaotic Probability framework. We show that the Conditional Chaotic Probability model that we propose can be given the interpretation of a generalized Markov chain. Chaotic Probabilities were introduced by Fine et al. as an attempt to model chance phenomena with a usual set of measures $\mathcal{M}$ endowed with an *objective, frequentist interpretation* instead of a compound hypothesis or behavioral subjective one. We follow the presentation of the univariate case chaotic probability model and provide an instrumental interpretation of random process measures consistent with a conditional chaotic probability source, which can be used as a tool for simulation of our model. Given a finite time series, we also present a universal method for estimation of conditional chaotic probability models that is based on the analysis of the relative frequencies taken along a set of subsequences chosen by a given set of rules.

Keywords: Imprecise Probabilities, Foundations of Probability, Church Place Selection Rules, Probabilistic Reasoning, Conditioning, Complexity

1 Introduction

1.1 What is Chaotic Probability About?

Unlike the standard theory of real valued probability which since its beginning was *Janus faced* having to deal with both objective and subjective phenomena, sets of measures are mainly used to model behavior and subjective beliefs. Chaotic Probabilities were developed by Fine et al. [2] [4] [12] as an attempt to make sense of an objective, frequentist interpretation of a usual set of probability measures $\mathcal{M}$. In this setting, $\mathcal{M}$ is intended to model stable (although not stationary in the standard stochastic sense) physical sources of finite time series data that are highly irregular. The work was in part inspired in the following quotation from Kolmogorov 1983 [7]:

"In everyday language we call random those phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking, there is no ground to believe that random phenomena should possess any definite probability. Therefore, we should distinguish between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges the problem of finding reasons for the applicability of the mathematical theory of probability to the real world."

Despite that fact pointed out by Kolmogorov, the idea of models of physical chance phenomena sharing the precision of real number system is so well-entrenched that identifications of chaotic probability phenomena are difficult to make and hard to defend.

1.2 Previous Work and Overview

A large portion of the literature on imprecise probabilities gives a behavioral, subjective interpretation of this model Walley 1991 [14]. But some work has been done on the development of a frequentist interpretation of imprecise probabilities. Fine et al. have worked on asymptotics or laws of large numbers for interval-valued probability models [9] [13] [11] [6].

The work of Cozman and Chrisman 1997 [1] studying estimation of credal sets by analyzing limiting relative frequencies along a set of subsequences of a time series is very similar to the approach taken by Fierens and Fine, except that the latter restrict themselves to studying finite time series data. Another quote from Kolmogorov 1963 [8] explains the reason for such a restriction:

"The frequency concept based on the notion of limiting frequency as the number of trials increases to infinity, does not contribute anything to substantiate the ap-

plicability of the results of probability theory to real practical problems where we have always to deal with a finite number of trials."

In their work on Chaotic Probability models [2] [4], Fierens and Fine provided an instrumental interpretation of the model, a method for simulation of a random sequence given the model, and a method for estimation of the model given a finite random sequence. They have worked both on the univariate case and in the conditional case. This paper will parallel their approach on a different conditional setting that can be interpreted as a generalized Markov chain. We discuss the differences between their approach to conditioning and ours in Section 3.5. Roughly speaking, in our setting we have that a conditional chaotic probability model $\mathcal{M}_{|K}$ is a function that associates for each possible sequence y of K-previous outcomes a univariate chaotic probability model $\mathcal{M}_{|K}(y)$, i.e., a set of probability measures.

Section 2.1 provides an instrumental interpretation of conditional chaotic probability models. Although we do not claim this interpretation for explaining real world data, it is useful to develop it because it provides a means to understand the behavior of conditional chaotic probabilities using standard well-known tools of probability theory, it also provides the basis for simulation of these models, and finally it extends the interpretation proposed by Fierens and Fine for univariate chaotic probabilities. With that interpretation in mind we also provide a method for simulation of a data sequence given the Conditional Chaotic Probability model in Section 2.2.

In Section 3, we analyze the problem of estimating conditional chaotic probabilities from data. As in the univariate setup, we do that by studying the relative frequency taken along selected subsequences. We define three properties of a set of subsequence selection rules: *Conditional Causal Faithfulness, Conditional Homogeneity and Conditional Visibility.* By Conditional Causally Faithful rules we mean rules that, for each fixed sequence of past K outcomes, select subsequences such that the empirical and theoretical time averages along the selected subsequence are sufficiently close together. A set of rules renders $\mathcal{M}_{|K}$ conditionally visible if, for each fixed sequence y of past K outcomes, all measures in $\mathcal{M}_{|K}(y)$ can be estimated by relative frequencies along the selected subsequences. Finally, a set of rules is conditionally homogeneous if, for each fixed sequence y of past K outcomes, it cannot expose more than a small neighborhood of a single measure contained in the convex hull of $\mathcal{M}_{|K}(y)$, intuitively a set of rules is conditionally homogenous if the relative frequencies taken along the terms selected by the rules and that have y as the previous K outcomes are all close to a single measure in the convex hull of $\mathcal{M}_{|K}(y)$. We then prove the existence of families of causal subsequence selection rules that can make $\mathcal{M}_{|K}$ conditionally visible. Following the steps of Rêgo and Fine 2005 [12], in Section 4 we describe a universal methodology for finding a family of causal subsequence selection rules that can make $\mathcal{M}_{|K}$ conditionally visible, and in Section 5, we strengthen this result by assuring that the relative frequency taken along every subsequence analyzed is close to some measure in $\cup_y \mathcal{M}_{|K}(y)$ with high probability. In Section 6, we give the interpretation of conditional chaotic probabilities as a Generalized Markov Chain that instead of a single transition probability measure has a set of transition probabilities. We conclude in Section 7.

2 From Model to Data

2.1 Instrumental Interpretation

Let $\mathcal{X} = \{z_1, z_2, ..., z_\xi\}$ be a finite sample space.[1] We denote by $\mathcal{X}^*$ the set of all finite sequences of elements taken from $\mathcal{X}$. A particular sequence of n samples from $\mathcal{X}$ is denoted by $x^n = \{x_1, x_2, ..., x_n\}$. $\mathcal{P}$ denotes the set of all measures on the power set of $\mathcal{X}$ and $x^{i:j} = \{x_i, x_{i+1}, ..., x_{j-1}, x_j\}$. A conditional chaotic probability model given the past K outcomes $\mathcal{M}_{|K} : \mathcal{X}^K \to 2^{\mathcal{P}}$ is a function associating for each sequence of past K outcomes a subset of $\mathcal{P}$. Intuitively, $\mathcal{M}_{|K}$ models the "marginals" of the next outcome of some process generating sequences in $\mathcal{X}^*$ given the previous K outcomes. This section provides an interpretation of such a process.

Let F be a conditional chaotic selection function, $F : \mathcal{X}^* \to \cup_{y \in \mathcal{X}^K} \mathcal{M}_{|K}(y)$. At each instant i, a measure $\nu_i = F(x^{i-1})$ is chosen according to this selection function F. We require that the complexity of F be neither too complex, so that $\mathcal{M}_{|K}$ can not be exposed on the basis of a finite time series, nor too simple so that a standard stochastic process can be used to model the phenomena. We also require that F satisfies the following restriction

$$F(x^{i-1}) \in \mathcal{M}_{|K}(x^{i-K:i-1}),\ \forall i > K. \qquad (1)$$

Let $\mu_F \in \mathcal{P}^K$ be the initial probability distribution over the first K symbols.

An actual data sequence x^n is assessed by the graded potential of the realization of a sequence of random

[1] Recently, Fierens 2007 [3] extended the univariate Chaotic Probability Model to be defined on any subset of the reals. For ease of exposition, we focus on the finite case here.

variables X^n described by:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) =$$
$$= \mu_F(X_1 = x_1, ..., X_K = x_K) \prod_{l=K+1}^{n} \nu_l(x_l) \qquad (2)$$

where $\nu_l \in M_{|K}(x^{l-K:l-1})$

We denote by $\mathcal{M}^*_{|K}$ the family of all such process measures P. This interpretation is usually considered instrumental (i.e., without commitment to reality), although theory also applies if there is empirical reality in the description of F, but F is simply unknown.

In the unconditional chaotic probability model, it is understood that all relevant information produced by the source is captured by the coarse-grained description provided by the set of measures $\mathcal{M}$ and the further information contained in the fine-grained description, F, has no empirical reality (Rêgo and Fine 2005 [12] emphasize that a similar situation occurs in quantum mechanics, see Gell-Mann 1994 [5], pg. 144–146). With this conditional model, we try to develop a model that does not discard all information provided by F, it considers the fact that for different previous K outcomes the source behaves differently. That is, it allows for the existence of a simple structure in the choice selection function.

Notice also, that in this account, for each sequence y of previous K outcomes, it is the whole set $\mathcal{M}_{|K}(y)$ that models the chance phenomena and not a "true" individual measure in $\mathcal{M}_{|K}(y)$ that is unknown to us, as in the usual compound hypothesis modeling.

Like in the unconditional case, no matter how complex the conditional selection function is, the process measure P is a standard stochastic process, the issue is whether it reflects the reality of the underlying phenomena. In the unconditional case, if the selection function is chaotic,[2] then all we can hope to learn and therefore predict for future terms in the sequence is the coarse-grained description of the model given by $\mathcal{M}$, a subset of $\mathcal{P}$. However, in the conditional case that we present here, the conditional selection function satisfies (1), and one can hope to learn $\mathcal{M}_{|K}(y)$ for each $y \in \mathcal{X}^K$. Therefore, in the conditional chaotic probability model, there is some structure in the chaotic behavior of the conditional selection function so that, as we will see in Section 3, the fact that the previous K outcomes were equal to some sequence y allow us to have a finer description of the model than just the coarse-grained description of the model given by $\mathcal{M}_{|K}$.

The next subsection digresses on a new statistical model that gives an application for the mathematical tools developed here. Fierens 2007 [3] also provides a motivation for the mathematical tools developed in the theory of chaotic probabilities using it for robust stochastic simulation.

2.1.1 Digression on a New Statistical Model

While our primary interest is not in a statistical compound hypothesis, the results of this paper do bear on a new statistical estimation model.

We can partially specify any stochastic process model $P \in \mathcal{P}$ by specifying the following set of conditional measures for the individual times for all possible $y \in \mathcal{X}^K$:

$$\mathcal{M}^P_{|K}(y) = \{\nu : (\exists j \geq K)(\exists x^j),\ x^{j-K+1:j} = y,$$
$$\nu(X_{j+1} \in A) = P(X_{j+1} \in A | X^j = x^j),\ \forall A \subset \mathcal{X}\}$$

Note that we do not keep track of the full conditioning event, only of the measure ν and of the previous K outcomes. We then wish to estimate $\mathcal{M}^P_{|K}(y),\ \forall y \in \mathcal{X}^K$, from data x^n. Also note that, in general, the process is not Markovian in the traditional sense, as the conditional selection function F depends on the whole history. Although, as we show in Section 6, it can be given the interpretation of a generalized Markov chain.

The model can be used in the following situation. Suppose we have an opponent in a game who can decide whether or not to act upon a trial t after examining the history of outcomes x^{t-1} prior to that trial. Certain distributions $P(X_t \in A | X^{t-1} = x^{t-1})$ for the trial at time t are favorable to us and others to our opponent. An assessment of the range of consequences to us from choices made by an intelligent opponent can be calculated from $\mathcal{M}^P_{|K}(y),\ \forall y \in \mathcal{X}^K$.

2.2 Simulation

In this section, we provide a method for sequence generation according to a source that is modeled by a conditional chaotic probability. First of all, we define a distance metric between probability measures as:

[2] As in the original work of Fierens and Fine [2] [4], the adjective "chaotic" is not used in the traditional technical sense of the mathematical literature on chaos, rather it is used in the sense of the selection function F being neither too simple nor too complex, where the complexity of the selection function can be measured, for example, in terms of Kolmogorov Complexity [10]. As well stated by an anonymous referee, *"the term 'chaotic probabilities' refers to viewpoint in which there is no true measure that is the model, because models for individual outcomes vary unpredictably while remaining in a given set $\mathcal{M}$"*.

$$(\forall \mu, \mu' \in \mathcal{P})\ d(\mu, \mu') \doteq \max_{z \in \mathcal{X}} |\mu(z) - \mu'(z)|$$

Note that $\mathcal{P}$ is compact with respect to d, so for all $\epsilon > 0$ we can find a minimal finite covering of it by Q_ϵ balls of radius ϵ, $\{B(\epsilon, \mu_i)\}$, where μ_i are computable measures. Let N_ϵ be the size of the smallest subset of the above covering of the simplex that covers the actual set of probabilities that can be selected by the conditional chaotic selection function, $\cup_{y \in \mathcal{X}^K} \mathcal{M}_{|K}(y)$, and denote this subset by $\mathcal{M}_\epsilon$. Let $\mathcal{M}_\epsilon(y)$ be the smallest subset of $\mathcal{M}_\epsilon$ that covers the set of probabilities that can be selected after the string of outcomes y, $\mathcal{M}_{|K}(y)$. Then, given an appropriate chaotic selection function $F : \mathcal{X}^* \to \mathcal{M}_\epsilon$, where $\mathcal{M}_\epsilon = \cup_{y \in \mathcal{X}^K} \mathcal{M}_\epsilon(y)$, satisfying $F(x^{i-1}) \in \mathcal{M}_\epsilon(x^{i-K:i-1})$, $\forall i > K$, and an appropriate initial probability distribution μ_F, the following algorithm can be used for simulation:

- Use a pseudo-random number generator to generate x^K according to μ_F
- For $i = K + 1$ to n
  - Choose $\nu_i = F(x^{i-1}) \in \mathcal{M}_\epsilon(x^{i-K:i-1})$
  - Choose any $\nu_i' \in B(\epsilon, \nu_i) \cap \mathcal{M}_{|K}(x^{i-K:i-1})$
  - Use a pseudo-random number generator to generate x_i according to ν_i'

Since we want to expose all of $\mathcal{M}_{|K}(y)$, $\forall y \in \mathcal{X}^K$, in a single but sufficiently long simulated sequence, we require F to visit several times each measure in $\mathcal{M}_\epsilon$. In the following sections, we study the problem of estimating a conditional chaotic probability model given a long enough but finite data sequence.

3 From Data to Model

3.1 Subsequence Analysis

The estimation process in the conditional chaotic probability framework uses a finite time series and analyzes it calculating ξ^K sets of relative frequencies taken along subsequences selected by causal subsequence selection rules (also known as Church place selection rules). These rules are called causal because the next choice is a function only of past values in the sequence and not, say, of the whole sequence. These rules satisfy the following:

Definition 3.1: An effectively computable function φ is a causal subsequence selection rule if:

$$\varphi : \mathcal{X}^* \to \{0, 1\}$$

and, for any $x^n \in \mathcal{X}^*$, x_k is the j-th term in the generated subsequence $x^{\varphi,n}$, of length $\lambda_{\varphi,n}$, if:

$$\varphi(x^{k-1}) = 1,\ \sum_{i=1}^{k} \varphi(x^{i-1}) = j,\ \lambda_{\varphi,n} = \sum_{k=1}^{n} \varphi(x^{k-1})$$

Given a set of causal subsequence selection rules, Ψ, for each $\varphi \in \Psi$ and $y \in \mathcal{X}^K$, define the empirical and theoretical conditional time averages along a chosen subsequence by:

$$(\forall \mathbf{A} \subset \mathcal{X}),$$

$$\overline{\mu}_{\varphi,n,y}(\mathbf{A}) \doteq \sum_{i=K+1}^{n} \frac{I_{\mathbf{A}}(x_i) I_{\{y\}}(x^{i-K:i-1}) \varphi(x^{i-1})}{\lambda_{\varphi,n,y}}$$

$$\overline{\nu}_{\varphi,n,y}(\mathbf{A}) \doteq \frac{1}{\lambda_{\varphi,n,y}} \sum_{i=K+1}^{n} E[I_{\mathbf{A}}(X_i) | X^{i-1} = x^{i-1}] \times$$

$$\times I_{\{y\}}(x^{i-K:i-1}) \varphi(x^{i-1})$$

where $I_{\mathbf{A}}$ is the $\{0,1\}$-valued indicator function of the event $\mathbf{A}$ and $\lambda_{\varphi,n,y} \doteq \sum_{i=K+1}^{n} I_{\{y\}}(x^{i-K:i-1}) \varphi(x^{i-1})$.

$\overline{\nu}_{\varphi,n}(.|y)$ can be rewritten in terms of the instrumental understanding as:

$$\overline{\nu}_{\varphi,n,y}(\mathbf{A}) \doteq \frac{1}{\lambda_{\varphi,n,y}} \sum_{i=K+1}^{n} \nu_i(\mathbf{A}) I_{\{y\}}(x^{i-K:i-1}) \varphi(x^{i-1})$$

A rule φ applied to x^n is said to be *conditionally causally faithful* if $\forall y \in \mathcal{X}^K$, $d(\overline{\nu}_{\varphi,n,y}, \overline{\mu}_{\varphi,n,y})$ is small. Essentially, φ is conditionally faithful if it does not extract an arbitrary pattern. The existence of such rules is shown by the following theorem.

Theorem 3.2: *Let ξ be the cardinality of $\mathcal{X}$ and denote the cardinality of Ψ by $||\Psi||$. Let $m \leq n$. If $||\Psi|| \leq t$, then for any process measure $P \in \mathcal{M}^*_{|K}$ and $y \in \mathcal{X}^K$:*

$$P(\max_{\varphi \in \Psi} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \lambda_{\varphi,n,y} \geq m\} \geq \epsilon) \leq$$

$$\leq 2\xi t \exp\{\frac{-\epsilon^2 m^2}{2(n-K)}\}$$

Proof: Follows immediately from Theorem 1 of Fierens and Fine 2003 [4], considering

$$I_{\{y\}}(x^{i-K:i-1}) \varphi(x^{i-1})$$

to be a selection rule, $\varphi'(x^{i-1})$, for the original sequence x^n. ∎

Note that, *as long as the size of the family of selection rules is not too big, conditional faithfulness is guaranteed with high probability if the subsequence selected is*

long enough. Note that the restriction on the size of Ψ is necessary, since if we allow all possible selection rules, we will get all the measures giving probability 1 to each one of the elements of the sample space $\mathcal{X}$.

Note also that, if we take $m = \alpha(n-K)$, for $\alpha \in (0,1)$, the size t of the family of selection rules can be as large as $e^{\rho(n-K)}$, for $\rho < \frac{\alpha^2\epsilon^2}{2}$; conditional faithfulness of the rules is guaranteed with high probability for large n.

3.2 Conditional Visibility and Estimation

The property that a set of rules, Ψ, must satisfy in order to expose all of $\mathcal{M}_{|K}(y)$, $\forall y \in \mathcal{X}^K$, is given by the following definition:

Definition 3.3: (Conditional Visibility) $\mathcal{M}_{|K}$ is **conditionally made visible** $(\Psi, \theta, \delta, m, n)$ by $P \in \mathcal{M}^*_{|K}$ if $\forall y \in \mathcal{X}^K$:

$$P(\bigcap_{\nu \in M_{|K}(y)} \bigcup_{\varphi \in \Psi} \{X^n : \lambda_{\varphi,n,y} \geq m, \\ d(\nu, \overline{\mu}_{\varphi,n,y}) \leq \theta\}) \geq 1 - \delta$$

Let $\hat{\mathcal{M}}^{\theta,\Psi,y}_{|K}$ be an estimator of $\mathcal{M}_{|K}(y)$ defined by:

$$\forall x^n \in \mathcal{X}^*,\ \hat{\mathcal{M}}^{\theta,\Psi,y}_{|K}(x^n) = \bigcup_{\varphi \in \Psi : \lambda_{\varphi,n,y} \geq m} B(\theta, \overline{\mu}_{\varphi,n,y})$$

where, $B(\theta, \overline{\mu}_{\varphi,n}) \doteq \{\mu \in \mathcal{P} : d(\mu, \overline{\mu}_{\varphi,n}) < \theta\}$.

Let $[\mathbf{A}]^\epsilon$ denote the ϵ-enlargement of a set $\mathbf{A}$ defined by:

$$(\forall \mathbf{A} \subseteq \mathcal{P})(\forall \epsilon > 0)[\mathbf{A}]^\epsilon \doteq \{\mu : (\exists \mu' \in \mathbf{A}) d(\mu, \mu') < \epsilon)\}$$

The next theorem shows that for an appropriate set of rules Ψ, it is possible to conditionally expose $\mathcal{M}_{|K}$.

Theorem 3.4: (Estimability) *Let P render $\mathcal{M}_{|K}$ conditionally visible $(\Psi, \theta, \delta, m, n)$. Then, $\forall y \in \mathcal{X}^K$:*

$$P[[ch(\mathcal{M}_{|K}(y))]^{\theta+\epsilon} \supset \hat{\mathcal{M}}^{\theta,\Psi,y}_{|K} \supset \mathcal{M}_{|K}(y)] \geq 1 - \delta - \tau_n$$

where $\tau_n = 2\xi||\Psi|| \exp(\frac{-\epsilon^2 m^2}{2(n-K)})$ and $ch(\mathcal{M})$ is the convex hull of $\mathcal{M}$.

Proof: Follows immediately from Theorem 3 of Fierens and Fine 2003 [4] and Theorem 3.2 above, considering each fixed $y \in \mathcal{X}^K$. ∎

3.3 Conditional Homogeneity

There are some families of causal subsequence selection rules that are too simple to expose the structure underlying the conditional chaotic probability model, such families have the following property:

Definition 3.5: (Conditional Homogeneity) $P \in \mathcal{M}^*_{|K}$ is **conditionally homogeneous** $(\Psi, \theta, \delta, m, n)$ if $\forall y \in \mathcal{X}^K$:

$$P(\max_{\varphi_1,\varphi_2 \in \Psi} \{d(\overline{\mu}_{\varphi_1,n,y}, \overline{\mu}_{\varphi_2,n,y}) : \\ \lambda_{\varphi_1,n,y}, \lambda_{\varphi_2,n,y} \geq m\} \leq \theta) \geq 1 - \delta$$

3.4 Consistency Between Conditional Visibility and Conditional Homogeneity

Theorem 3.6: (Consistency) *Let $\epsilon > 1/m$. Assume that $\forall y \in \mathcal{X}^K$, there is an ϵ-cover of $\mathcal{M}_{|K}(y)$ by $N_\epsilon(y)$ open balls with centers in a set $\mathcal{M}_\epsilon(y) \doteq \{\mu^y_1, \mu^y_2, ..., \mu^y_{N_\epsilon(y)}\}$ such that, for each μ^y_i, there is a recursive probability measure $\nu \in B(\epsilon, \mu^y_i) \cap \mathcal{M}_{|K}(y)$. Let Ψ_0 be a set of causal subsequence selection rules. Assume also:*

$$\underline{p} \doteq \inf_{\nu \in \cup_{y \in \mathcal{X}^K} \mathcal{M}_{|K}(y)} \min_{z \in \mathcal{X}} \nu(z) > 0$$

Then, there are a process measure P and a family Ψ_1 such that, for large enough n, P will both render $\mathcal{M}_{|K}$ conditionally visible $(\Psi_1, 3\epsilon, \delta, m, n)$ and ensure conditional homogeneity $(\Psi_0, 6\epsilon, \delta, m, n)$ with

$$\delta = 2(\xi t_n + 1) \exp(\frac{-\epsilon^2 m^2}{2(n-K)})$$

where $t_n = \max\{||\Psi_0||, ||\Psi_1||\}$

Proof: It follows closely proofs contained in the Appendix C and D of [2]; we omit details here. ∎

The importance of this result is that there are conditional chaotic sources for which analysis by simple selection rules would give us the impression that the phenomena can be modeled by a standard probability model (indeed, it will look like a Markov chain where the set of states is $\mathcal{X}^K$). But if we further analyze the source with a set of more complex selection functions we can expose the underlying structure of the model. In this way, as pointed out by Fierens and Fine 2003, *the family of causal subsequence selection rules determines the power of the resolution of the model we see.*

3.5 Fierens and Fine's Approach to Conditioning

Fierens and Fine 2003 [2] also provided a model for Conditional Chaotic Probabilities, where the conditioning events are the previous K outcomes in the sequence. In their approach, they define

$$\mathbf{P}_{|K} = \{\nu : (\forall A \subseteq \mathcal{X}) \\ \nu(A, X^K) = E_\mu(I_A(X_{K+1})|X^K), \mu \in \mathcal{P}^{K+1}\}.$$

For them, a conditional chaotic probability model $\mathbf{M}_{|K}$ is any subset of $\mathbf{P}_{|K}$. They also provide an instrumental understanding of the model, by defining a selection function $\mathbf{F} : \mathcal{X}^* \to \mathbf{M}_{|K}$. It is easy to see that there is a one-to-one correspondence between their model and the one presented here. Given $\mathbf{M}_{|K}$, a conditional chaotic probability model according to our definition is given by:

$$\mathcal{M}_{|K}(y) = \{\mu \in \mathcal{P} : \forall z \in \mathcal{X}, \mu(z) = \nu(z, X^K = y), \nu \in \mathbf{M}_{|K}\}, \forall y \in \mathcal{X}^K.$$

For the converse, given $\mathcal{M}_{|K}$, a conditional chaotic probability model according to Fierens and Fine's definition is given by:

$$\mathbf{M}_{|K}(y) = \{\nu \in \mathbf{P}_{|K} : \exists y \in \mathcal{X}^K, \forall z \in \mathcal{X}, \nu(z, X^K = y) = \mu(z), \mu \in \mathcal{M}_{|K}(y)\}.$$

The major difference between both approaches is the estimation procedure; the set of subsequence selection rules Fierens and Fine allow for estimating the conditional chaotic probability model is a subset of the set we allow. Unlike us, for each fixed sequence of K outcomes y, Fierens and Fine analyze the subsequence $x^{y,n}$ of x^n, that is formed by all terms in x^n whose previous K outcomes are equal to y, using causal subsequence selection rules that depend only on past terms that appear in $x^{y,n}$, not on all past terms of the whole original sequence x^n, as we do in our approach. As the chaotic selection function both in their approach and in ours is allowed to depend on all past symbols of the sequence x^n, we believe that it is more appropriate to allow the more general set of selection rules we allow.

Although Fierens and Fine were able to prove results analogous to Theorems 3.2, 3.4, and 3.6 using their restricted set of selection rules, they did not provide a procedure for finding a family of selection rules Ψ that renders $\mathbf{M}_{|K}$ conditionally visible. We will now extend the result of Rêgo and Fine 2005 [12] providing a procedure for finding a family of selection rules Ψ that renders $\mathcal{M}_{|K}$ conditionally visible. In the next section, we provide a methodology for finding such a family of rules Ψ that works for any conditional chaotic probability source, and we call it a *universal family of selection rules*. As we see in the next section, for finding such a universal family it is crucial that we allow the more general set of subsequence selection rules that depend on the whole past terms in the sequence x^n. Unfortunately, as in the univariate case, such a family may "extract" more than $\cup_{y \in \mathcal{X}^K} \mathcal{M}_{|K}(y)$. We return to this point in Section 5.

4 Universal Family of Selection Rules

In this section we prove that there exists a universal family, which depends basically on the precision we want our estimator to have, that is able to conditionally expose all measures of any set of probabilities $\mathcal{M}_{|K}$.

Let $\lambda_{y,n} \doteq \sum_{i=K+1}^{n} I_{\{y\}}(x^{i-K:i-1})$.

Define for each family of causal selection rules, Ψ, and each $y \in \mathcal{X}^K$ the estimator based on this family as:

$$\hat{\mathcal{M}}^{\Psi,y}_{|K} \doteq \{\overline{\mu}_{\varphi,n,y} : \varphi \in \Psi, \lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq m\}$$

Approximate $F(x^{j-1})$ by $F_\epsilon(x^{j-1}) = \mu_j$ if μ_j is the closest measure to $F(x^{j-1})$ among all μ_i's that belongs to $\mathcal{M}_\epsilon(x^{j-K:j-1})$. Let $F_{\epsilon,n}$ be the restriction of F_ϵ to $\mathcal{X}^{1:n}$ (all sequences of length not greater than n). The following theorem provides the desired method of finding a universal family of selection rules for conditional chaotic probability sources.

Intuitively, Theorem 4.1 states that as long as the Kolmogorov Complexity [10] of the conditional chaotic measure selection function is not too high, and we have a long enough data sequence, then for every given sequence $y \in \mathcal{X}^K$ of past K symbols that appeared frequently enough, we are able to make visible with high probability all measures in $\mathcal{M}_{|K}(y)$ that were selected frequently enough in the sequence.

Theorem 4.1: *Choose $f, f_0 \geq 1$, $\alpha_0 = (f_0 \xi^K)^{-1}$, $\alpha = (fN_\epsilon)^{-1}$ and let $m_0 = \alpha_0(n-K)$ and $m = \alpha\lambda_{y,n}$. Define $\mathcal{M}^f_{|K}(y) \doteq \{\nu : \nu \in \mathcal{M}_{|K}(y)$ and $\exists \mu_i \in \mathcal{M}_\epsilon(y)$ such that $d(\nu, \mu_i) < \epsilon$ and μ_i is selected at least m times by $F_{\epsilon,n}$ when the previous K outcomes were equal to y and $\lambda_{y,n} \geq m_0\}$. Given β smaller than $\frac{\alpha_0^2 \alpha^2 \epsilon^2}{2}$, choose $\epsilon' \in (0, \beta \log_2 e)$ and assume the Kolmogorov complexity, $K(F_{\epsilon,n})$, of $F_{\epsilon,n}$ satisfies the following condition:*

$$\exists \kappa \geq 0,\ \exists L_{\epsilon',\kappa} \text{ such that } \forall n \geq L_{\epsilon',\kappa}, \quad \frac{K(F_{\epsilon,n})}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \epsilon' \qquad (3)$$

Define $\mathcal{M}^*_{|K,R} \doteq \{P : P \in \mathcal{M}^*_{|K}$ and the corresponding F satisfies condition (3)$\}$. Then, for $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil \log_2 Q_\epsilon \rceil}{\epsilon'}\}$, there exists a family of causal subsequence selection rules Ψ_U, depending only on α_0, α, κ and ϵ, such that $\forall \mathcal{M}_{|K}$, and $\forall P \in \mathcal{M}^*_{|K,R}$:

$$P(\bigcap_{y \in \mathcal{X}^K} \{X^n : [ch(\mathcal{M}_{|K}(y))]^{4\epsilon} \supset [\hat{\mathcal{M}}^{\Psi_U,y}_{|K}]^{3\epsilon} \supset \mathcal{M}^f_{|K}(y)\}) \geq 1 - \delta,$$

where $\gamma = \frac{\alpha_0^2\alpha^2\epsilon^2}{2} - \beta$ *and* $\delta = 2\xi^{K+1}n^{\kappa}e^{\alpha_0^2\alpha^2\epsilon^2 K}e^{-\gamma n}$.

Remark 4.2: Note that if $\lambda_{y,n} < m_0$, then by definition we have $\hat{\mathcal{M}}_{|K}^{\Psi_U,y} = \mathcal{M}_{|K}^{f}(y) = \emptyset$. Thus, we fail to estimate $\mathcal{M}_{|K}(y)$ in this case. But the fraction of times a string of outcomes $y \in \mathcal{X}^K$ such that $\lambda_{y,n} < m_0$ appears in a sequence X^n is bounded from above by $(1/f_0)$. Therefore, for f_0 sufficiently large it is reasonable to expect that such measures may not be estimated.

Remark 4.3: Note also that if $\lambda_{y,n} \geq m_0$, then the fraction of times a measure in $\mathcal{M}_{|K}(y) \setminus \mathcal{M}_{|K}^{f}(y)$ is used to generate an outcome in a sequence X^n is bounded from above by $(1/f)$. Therefore, for f sufficiently large it is reasonable to expect that such measures may not be estimated.

Proof: Define a family of selection functions, Ψ_G, that corresponds to $F_{\epsilon,n}$ as follows: $\Psi_G = \{\varphi_i^G$, for $1 \leq i \leq N_\epsilon\}$, where, for $0 \leq j \leq n-1$:

$$\varphi_i^G(x^j) \doteq \begin{cases} 1 & \text{if } F_{\epsilon,n}(x^j) = \mu_i \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

As each φ_i^G is a function of $F_{\epsilon,n}$ and μ_i, and $\lceil \log_2 Q_\epsilon \rceil$ is an upper bound on the number of bits necessary to specify the index i of the particular measure μ_i, the Kolmogorov complexity, $K(\varphi_i^G)$, of φ_i^G satisfies:

$$\max_i K(\varphi_i^G) \leq K(F_{\epsilon,n}) + \lceil \log_2 Q_\epsilon \rceil \tag{5}$$

It then follows, from our hypothesis, that for $1 \leq i \leq N_\epsilon$ and $\forall n \geq L_{\epsilon',\kappa}$, $K(\varphi_i^G)$ satisfies the following condition:

$$\frac{K(\varphi_i^G)}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \epsilon' + \frac{\lceil \log_2 Q_\epsilon \rceil}{n}$$

Therefore, for $n > \max(L_{\epsilon',\kappa}, \frac{2\lceil \log_2 Q_\epsilon \rceil}{\epsilon'})$:

$$\frac{K(\varphi_i^G)}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \frac{\epsilon'}{2}$$

Let Ψ_U consist of all rules of Kolmogorov complexity less than or equal to $\beta n \log_2 e + \kappa \log_2 n - 1$. Note that since for $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil \log_2 Q_\epsilon \rceil}{\epsilon'}\}$, $\frac{n\epsilon'}{2} > 1$, so Ψ_U includes Ψ_G for n large enough.

As $||\Psi_U|| \leq 2^{n\beta \log_2 e + \kappa \log_2 n} = n^\kappa e^{\beta n}$, $m_0 = \alpha_0(n-K)$, $m = \alpha\lambda_{y,n}$ and $\gamma = \frac{\alpha_0^2\alpha^2\epsilon^2}{2} - \beta > 0$, by the causal faithfulness theorem, for any $P \in \mathcal{M}_{|K}^*$,

$$\begin{aligned}
&P(X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq m\} \geq \epsilon) = \\
&= P(X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq \alpha\lambda_{y,n}\} \geq \epsilon) \leq \\
&\leq P(X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq \alpha m_0\} \geq \epsilon) \leq \\
&\leq P(X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{\varphi,n,y} \geq \alpha_0\alpha(n-K)\} \geq \epsilon) \leq \\
&\leq 2\xi^{K+1}n^\kappa e^{\frac{\alpha_0^2\alpha^2\epsilon^2 K}{2}} e^{-\gamma n}
\end{aligned}$$

Note that, since for $\alpha_0 = (f_0\xi^K)^{-1}$, for all X^n, there exists y such that $\lambda_{y,n} \geq m_0$. And for $\alpha = (fN_\epsilon^{-1})$, we know that for all X^n and for all y, there exists i such that $\lambda_{\varphi_i^G,n,y} \geq m$, as $\Psi_G \subset \Psi_U$, we have that for all X^n the maximum above is taken over a non-empty set.

To prove the theorem, let φ_i^G be as defined in Equation (4), then for a fixed X^n, by definition of $\mathcal{M}_{|K}^{f}(y)$, $\forall \nu \in \mathcal{M}_{|K}^{f}(y)$, $\exists \mu_i \in \mathcal{M}_\epsilon(y)$ such that $d(\nu,\mu_i) < \epsilon$, $\lambda_{\varphi_i^G,n,y} \geq m$ and $\lambda_{y,n} \geq m_0$ (Note the index i depends on X^n). Then, using the triangle inequality property:

$$\begin{aligned}
&\max_{y\in\mathcal{X}^K} \sup_{\nu\in\mathcal{M}_{|K}^{f}(y)} d(\nu, \overline{\mu}_{\varphi_i^G,n,y}) \leq \\
&\max_{y\in\mathcal{X}^K} \sup_{\nu\in\mathcal{M}_{|K}^{f}(y)} d(\overline{\mu}_{\varphi_i^G,n,y}, \overline{\nu}_{\varphi_i^G,n,y}) + \\
&\max_{y\in\mathcal{X}^K} \sup_{\nu\in\mathcal{M}_{|K}^{f}(y)} d(\overline{\nu}_{\varphi_i^G,n,y}, \mu_i) + \\
&\max_{y\in\mathcal{X}^K} \sup_{\nu\in\mathcal{M}_{|K}^{f}(y)} d(\mu_i, \nu)
\end{aligned}$$

and since $\overline{\nu}_{\varphi_i^G,n,y}$ is the time average of the actual measures selected by F in the ball $B(\epsilon,\mu_i)$, $d(\overline{\nu}_{\varphi_i^G,n,y}, \mu_i) < \epsilon$, and as $\Psi_G \subset \Psi_U$, the following holds,

$$\begin{aligned}
&\{X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq m\} < \epsilon\} \subset \\
&\{X^n : \max_{y\in\mathcal{X}^K} \sup_{\nu\in\mathcal{M}_{|K}^{f}(y)} \min_{\varphi\in\Psi_U} \{d(\nu, \overline{\mu}_{\varphi,n,y}) : \\
&\lambda_{y,n} \geq m_0, \lambda_{\varphi,n,y} \geq m\} < 3\epsilon\}.
\end{aligned} \tag{6}$$

Equation 6 implies,

$$\begin{aligned}
&\{X^n : \max_{y\in\mathcal{X}^K} \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi,n,y}, \overline{\nu}_{\varphi,n,y}) : \\
&\lambda_{\varphi,n,y} \geq m\} < \epsilon\} \subset
\end{aligned}$$

$$\{\bigcap_{y\in\mathcal{X}^K}\{X^n:$$
$$[ch(\mathcal{M}_{|K}(y))]^{4\epsilon}\supset[\hat{\mathcal{M}}_{|K}^{\Psi_U,y}]^{3\epsilon}\supset\mathcal{M}_{|K}^{f}(y)\}\} \quad (7)$$

Theorem 4.1 follows from the causal faithfulness Theorem 3.2. ∎

The problem with the sort of estimator provided by the above theorem is that on one hand it is able to conditionally expose all measures in $\mathcal{M}_{|K}(y)$ that appeared frequently enough in the process, if y also appeared frequently enough in the outcomes. On the other hand, we have that for each $y\in\mathcal{X}^K$ that appeared frequently enough, the estimator is only guaranteed to be included in an enlarged neighborhood of $\mathcal{M}_{|K}(y)$'s convex hull and in some cases this can be rather larger than $\mathcal{M}_{|K}(y)$.

The following section proves a theorem that given x^n provides a methodology for finding a universal family of subsequences, $\Psi(x^n)$, that is both able to conditionally expose all measures in $\mathcal{M}_{|K}(y)$ that appeared frequently enough in the process, if y appeared frequently enough, and contains only these subsequences whose empirical time averages are close enough to $\mathcal{M}_{|K}(y)$ with high probability. We will call this family to be **conditionally strictly faithful**.

5 Conditionally Strictly Faithful Family of Subsequences

In this section, we propose a methodology for finding a conditionally strictly faithful family of subsequences that can both conditionally expose all measures in $\mathcal{M}_{|K}(y)$ that appear frequently enough in the process, if y appears frequently enough; and contains only these subsequences whose empirical time averages are close enough to $\mathcal{M}_{|K}(y)$ with high probability.

The problem with the set of rules Ψ_U is that it may contain rules that are not conditionally homogeneous, i.e., rules that given that the previous outcomes are equal to y select subsequences generated by mixtures of measures μ_i's. In our proposed methodology in this section, we will analyze each rule $\varphi\in\Psi_U$ with a universal family Ψ_U^{φ} (see definition below) and include φ in $\Psi(x^n)$ only if it is conditionally homogeneous. As Ψ_U^{φ} is universal for the subsequence selected by φ, it will be able to identify if it is or not conditionally homogeneous with high probability. Thus, our family of sequences $\Psi(x^n)$ is constructed in a two-stage process: first we consider the family of selection rules Ψ_U which consists of all rules of at most a certain complexity value which is able to make $\mathcal{M}_{|K}$ conditionally visible; then we filter the rules contained in Ψ_U so that it contains only conditionally homogenous subsequences whose relative frequencies are close enough to a measure in $\cup_{y\in\mathcal{X}^K}\mathcal{M}_{|K}(y)$.

The following theorem proves the desired result, i.e., if the Kolmogorov complexity of the conditional chaotic measure selection function is not too high and we have a long enough data sequence, with high probability we can conditionally make visible all and only measures that were used frequently enough in the sequence.

Theorem 5.1: *Choose $f_0, f\geq 1$, $\alpha_0=(f_0\xi^K)^{-1}$, $\alpha_1=(fN_\epsilon)^{-1}$, $\alpha_2=N_\epsilon^{-1}$ and let $m_0=\alpha_0(n-K)$, $m=\alpha_1\lambda_{y,n}$. Define $\mathcal{M}_{|K}^{f}(y)\doteq\{\nu:\nu\in\mathcal{M}_{|K}(y)$ and $\exists\mu_i\in\mathcal{M}_\epsilon(y)$ such that $d(\nu,\mu_i)<\epsilon$ and μ_i is selected at least m times by $F_{\epsilon,n}$ when the previous K outcomes were equal to y and $\lambda_{y,n}\geq m_0\}$ and define $\mathcal{M}_{\epsilon}^{f}(y)\doteq\{\mu_i:\mu_i\in\mathcal{M}_\epsilon(y)$ and μ_i is selected at least $\alpha_2 m$ times by $F_{\epsilon,n}$ when the previous K outcomes were equal to y and $\lambda_{y,n}\geq m_0\}$. Given β smaller than $\frac{\alpha_0^2\alpha_1^2\alpha_2^2\epsilon^2}{2}$, choose $\epsilon'\in(0,\beta\log_2 e)$ and assume the Kolmogorov complexity, $K(F_{\epsilon,n})$, of $F_{\epsilon,n}$ satisfies the same condition (3), i.e.,:*

$$\exists\kappa\geq 0,\ \exists L_{\epsilon',\kappa}\ \textit{such that}\ \forall n\geq L_{\epsilon',\kappa},$$
$$\frac{K(F_{\epsilon,n})}{n}<\beta\log_2 e+\frac{\kappa\log_2 n}{n}-\epsilon'$$

Define $\mathcal{M}_{|K,R}^{*}\doteq\{P:P\in\mathcal{M}_{|K}^{*}$ and the corresponding F satisfies condition (3)$\}$. Then, for $n>\max\{L_{\epsilon',\kappa},\frac{2\lceil\log_2 Q_\epsilon\rceil}{\epsilon'}\}$, for each x^n, there exists a family of subsequences $\Psi(x^n)$, depending only on α_0, α_1, α_2, κ and ϵ, such that $\forall\mathcal{M}_{|K}$ and $\forall P\in\mathcal{M}_{|K,R}^{*}$.[3]

$$P(\{X^n:\max_{y\in\mathcal{X}^K}\sup_{\mu\in\mathcal{M}_{|K}^{f}(y)}\min_{\nu\in\hat{\mathcal{M}}_{|K}^{\Psi(X^n),y}}d(\mu,\nu)<3\epsilon\}$$
$$\cap\{X^n:\max_{y\in\mathcal{X}^K}\max_{\nu\in\hat{\mathcal{M}}_{|K}^{\Psi(X^n),y}}\min_{\mu\in\mathcal{M}_{\epsilon}^{f}(y)}d(\mu,\nu)<6\epsilon\})$$
$$\geq 1-\delta_1$$

where $\gamma_1=\frac{\alpha_0^2\alpha_1^2\alpha_2^2\epsilon^2}{2}-\beta$, $S_\epsilon\doteq\min\{Q_\epsilon,n^\kappa e^{\beta n}\}$ and $\delta_1=4\xi^{K+1}S_\epsilon n^\kappa e^{\frac{\alpha_0^2\alpha_1^2\alpha_2^2\epsilon^2 K}{2}}e^{-\gamma_1 n}$.

Proof: It follows closely the proof of Theorem 3 contained in the appendix of [12]; we omit details here. ∎

[3] If $\Psi(X^n)=\emptyset$, we adopt the following convention:

$$\max_{y\in\mathcal{X}^K}\sup_{\mu\in\mathcal{M}_{|K}^{f}(y)}\min_{\nu\in\hat{\mathcal{M}}_{|K}^{\emptyset,y}}d(\mu,\nu)=\infty$$

and

$$\max_{y\in\mathcal{X}^K}\max_{\nu\in\hat{\mathcal{M}}_{|K}^{\emptyset,y}}\min_{\mu\in\mathcal{M}_{\epsilon}^{f}(y)}d(\mu,\nu)=0.$$

6 Interpretation as Generalized Markov Chain

The conditional chaotic probability model studied in this paper can be given the interpretation of a generalized Markov chain (GMC). The difference from the standard Markov chain is that the transition probabilities are given by sets of probability measures instead of single probabilities. Therefore, consider the following definitions of the parameters of the GMC:

- **States:** There are ξ^K states, one state for each $y \in \mathcal{X}^K$.
- **Initial Probabilities:** They are given by the initial probability of the first K symbols of the sequence, $\mu_F \in \mathcal{P}^K$.
- **Transition Set of Probabilities:**

$$\mathcal{M}_{|K}(y_{i+1}|y_i) \doteq \begin{cases} \{\nu(y_{i+1}(K)) : \nu \in \mathcal{M}_{|K}(y_i)\}, \\ \qquad \text{if } y_i(l+1) = y_{i+1}(l), \\ \qquad \text{for } 1 \leq l \leq K-1 \\ \{0\}, \text{ otherwise.} \end{cases}$$

  where $y_i(l)$ is the l-th position of the i-th state of the GMC.

Although this GMC looks like a partially specified Markov chain, they differ in the fact that in the GMC there is no single underlying "true" transition probability as in the partially specified Markov chain.

As pointed out by an anonymous referee, we must take care with the interpretation of the conditional chaotic probability model as a GMC. On one hand, usually a Markov chain describes a random phenomenon without memory. On the other hand, the instrumental interpretation of the conditional chaotic probability model proposed is sensible to the initial conditions of the realization of the random experiment; each initial condition determines a unique process P as defined in (2). As argued in Section 2.1, the issue is that P does not reflect the reality of the underlying phenomena. In a chaotic probability model, all that can be learnt and used to predict the next outcome in the sequence is $\mathcal{M}_{|K}(y)$ for each $y \in \mathcal{X}^K$, i.e., the transition set of probabilities of the GMC. Thus, a GMC is memoryless in the sense that once one knows the transition set of probabilities of the GMC all that we can learn and use to predict about the distribution of the next outcome in the sequence is given by the present state $y \in \mathcal{X}^K$ of the GMC, and it is chaotic in the sense that given that the present state is $y \in \mathcal{X}^K$ which measure will actually produce the next outcome varies unpredictably while remaining in $\mathcal{M}_{|K}(y)$.

7 Conclusions and Future Work

For ease of exposition, in this paper we focused on the case of conditioning on the previous K outcomes. It is easy to see that the results presented can be easily generalized to conditioning on a family of selection rules Φ such that $\exists L < n-1$ such that the following two conditions hold:

1. $\forall \phi_1, \phi_2 \in \Phi,\ \phi_1 \neq \phi_2$ implies $\phi_1(x^i) \cdot \phi_2(x^i) = 0,\ L < i \leq n-1$
2. $\sum_{\phi \in \Phi} \phi(x^i) = 1,\ L < i < n$

The development of chaotic probability theory is an important conceptual achievement, since it will provide us with a more powerful and general tool for analyzing time series. With the increasing size and number of data sets available nowadays, a different way of looking at them, provided by this theory, can have a huge impact in our world.

Although we do not have analyzed any practical real world data supporting the model, the main mathematical tools that enhance our capability of recognizing such phenomena (since we believe that we are only likely to find what we expect to see) have been presented. Therefore, new concepts of probability are likely to open our perception and understanding of chance phenomena.

To further develop the chaotic probability theory, a method to evaluate self-consistency of simulation and estimation needs to be studied (for details, see Fierens and Fine 2003 [2] [4]). Also, implications of this theory for inference and decision making problems have to be investigated.

In a broader perspective, the possibility of modeling physical chance phenomena with a set of measures, raises the question about the existence of other physical quantities that have properties that cannot be quantified by a single real number, but only as a set of them.

Acknowledgements

We want to specially thanks Terry Fine and Pablo Fierens for useful talks about Chaotic Probabilities Models. We also would like to thank Terry Fine for important suggestions and comments in early drafts of this work. Last but not least, we thank anonymous referees that made useful comments about this work.

References

[1] F. Cozman and L. Chrisman, *Learning convex sets of probability from data*, Tech. Report CMU-RI-TR-97-25, Robotics Institute, Carnegie Mellon University, 1997.

[2] P. I. Fierens, *Towards a Chaotic Probability Model for Frequentist Probability*, PhD thesis, Cornell University, 2003.

[3] P. I. Fierens, *An extension of chaotic probability models to real-valued variables*, in ISIPTA'07 Proceedings, July 2007.

[4] P. I. Fierens and T. L. Fine, *Toward a Chaotic Probability Model for Frequentist Probability: The Univariate Case.*, July 2003, pp. 245–259.

[5] M. Gell-Mann, *The Quark and The Jaguar*, W. H. Freeman and Company, 1994.

[6] Y.-L. Grize and T. L. Fine, *Continuous lower probability-based models for stationary processes with bounded and divergent time averages*, Annals of Probability, 15 (1987), pp. 783–803.

[7] A. N. Kolmogorov, *On logical foundations of probability theory*, vol. 1021 of Lecture Notes in Mathematics, Springer-Verlag, 1983.

[8] A. N. Kolmogorv, *On tables of random numbers*, Sankhya: The Indian Journal of Statistics, (1963), p. 369.

[9] A. Kumar and T. L. Fine, *Stationary lower probabilities and unstable averages*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 69 (1985), pp. 1–17.

[10] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Graduate Texts in Computer Science, Springer, second ed., 1997.

[11] A. Papamarcou and T. L. Fine, *A note on undominated lower probabilities*, Annals of Probability, 14 (1986), pp. 710–723.

[12] L. C. Rêgo and T. L. Fine, *Estimation of chaotic probabilities*, in ISIPTA'05 Proceedings, 2005, pp. 297–305.

[13] A. Sadrolhefazi and T. L. Fine, *Finite-dimensional distribution and tail behavior in stationary interval-valued probability models*, Annals of Statistics, 22 (1994), pp. 1840–1870.

[14] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall Pubs., 1991.

Qualitative and Quantitative Reasoning in Hybrid Probabilistic Logic Programs

Emad Saad
College of Computer Science and Information Technology
Abu Dhabi University
Abu Dhabi, UAE
emad.saad@adu.ac.ae

Abstract

Reasoning with qualitative and quantitative uncertainty is required in some real-world applications [6]. However, current extensions to logic programming with uncertainty support representing and reasoning with either qualitative or quantitative uncertainty. In this paper we extend the language of Hybrid Probabilistic Logic programs [28, 25], originally introduced for reasoning with quantitative uncertainty, to support both qualitative and quantitative uncertainty. We propose to combine disjunctive logic programs [10, 17] with Extended and Normal Hybrid Probabilistic Logic Programs (EHPP [25] and NHPP [28]) in a unified logic programming framework, to allow directly and intuitively to represent and reason in the presence of both qualitative and quantitative uncertainty. The semantics of the proposed languages are based on the answer set semantics and stable model semantics of extended and normal disjunctive logic programs [10, 17]. In addition, they also rely on the probabilistic answer set semantics and the stable probabilistic model semantics of EHPP [25] and NHPP [28].

Keywords. Probabilistic reasoning, probabilistic logic programming, knowledge representation.

1 Introduction

Reasoning under uncertainty is crucial in most real-world applications such as planning with uncertain domains and reasoning about actions with uncertain effects—such as the actions that arise from robotics in real-world environments. The literature is rich with different forms of uncertainty in logic programming. These forms of uncertainty can be classified into qualitative and quantitative models of uncertainty. Qualitative uncertainty is represented in logic programming using disjunctive logic programs [17, 10, 2]. It often happens that $a \vee b \vee c$ occurs because we are uncertain which of these propositions is true [2]. There might be states of the world where a is true or b is true or c is true or any combinations of them might also be true [2]. Quantitative uncertainty is represented in logic programming by means of different formalisms including probability theory (see [27] for a survey). Probabilistic logic programming is motivated by the need to provide the ability to represent both logical as well as probabilistic knowledge by logic programs (see [28] for a survey). The semantics of such frameworks provide ways to systematically derive logical conclusions along with their associated probabilistic properties. Although, representing and reasoning with both forms of uncertainty is needed in some real-world applications [6], this issue has not been addressed by the current work in qualitative or quantitative uncertainty in logic programming.

We propose to combine disjunctive logic programs [10, 2] with Extended and Normal Hybrid Probabilistic Logic Programs (EHPP [25] and NHPP [28]) in a unified logic programming framework, to allow directly and intuitively to represent and reason in the presence of both qualitative and quantitative uncertainty. This is achieved by introducing the notions of *Extended and Normal Disjunctive Hybrid Probabilistic Logic Programs* (EDHPP and NDHPP). EDHPP and NDHPP generalize extended and normal disjunctive logic programs of classical logic programming [10, 2], respectively, as well as, generalizing EHPP and NHPP [25, 28]. The semantics of EDHPP and NDHPP are based on the answer set semantics and stable model semantics of extended and normal disjunctive logic programs [10, 2], as well as the probabilistic answer set semantics and the stable probabilistic model semantics of EHPP and NHPP [25, 28]. The semantics of EDHPP employs the *Open World Assumption*, whereas, the semantics of NDHPP employs the *Closed World Assumption*. Therefore, any event represented by a program in NDHPP is associated with a probability interval. Any event that cannot be derived from a program in NDHPP is assigned the probability $[0, 0]$, by default. But, an event

that can be derived from the program is assigned a probability $[a, b] \neq [0, 0]$. However, in EDHPP events may not be assigned probability intervals to represent information incompleteness. If this is the case we say that the probabilities associated with these events are *unknown* or *undecidable*. We show that EDHPP naturally subsumes extended disjunctive logic programs[10] and EHPP [25], and NDHPP naturally subsumes normal disjunctive logic programs [2] and NHPP [28]. Moreover, we show that the probabilistic answer set semantics of EDHPP is reduced to the stable probabilistic model semantics of NDHPP. The importance of that is, computational methods developed for NDHPP can be applied to the language of EDHPP. Moreover, we show that EDHPP subsumes Baral et al.'s answer set programming approach for probabilistic reasoning with causal Bayes networks [1]. We show that some commonsense probabilistic knowledge can be easily represented in EDHPP and NDHPP.

Another reason why the proposed languages are interesting is that, in addition to allowing representing and reasoning with both qualitative and quantitative uncertainty, they can also be used in some real-world applications in which quantitative uncertainty needs to be defined over qualitative uncertainty, where probabilistic measures are assigned over the possible outcomes of qualitative uncertainty. For example, flipping a fair coin leads to a head or tail with 0.5 probability each. This fact can be implicitly represented as a disjunctive logic program (since both events are equally likely) as $head(coin)$ *or* $tail(coin)$ with $\{head(coin)\}$ and $\{tail(coin)\}$ as the possible answer sets, according to the answer set semantics [10]. However, the explicit representation of probabilities and the explicit assignment of probabilities to the possible outcome of flipping the coin cannot be presented by disjunctive logic programs syntax and semantics. Moreover, consider if the coin is biased to the head, where flipping the coin produced a head with 0.58 probability or a tail with 0.42 probability. In this case a disjunctive logic program cannot represent it neither implicitly nor explicitly. On the other hand, the coin-flipping example cannot be represented intuitively and directly in NHPP or EHPP either, since a corresponding notion of disjunctions is not allowed in NHPP or EHPP.

1.1 Probabilistic Logic Programming Approaches

The current work in the literature supports either qualitative uncertainty [17, 10, 2] or quantitative uncertainty [12, 18, 23, 24, 29, 19, 20, 21, 4, 15, 16, 3, 27, 28, 25]. The closest to our work are the frameworks presented in [27, 28, 25, 29, 16, 1].

Hybrid Probabilistic Logic Programs (HPP) [27] is probabilistic logic programming frameworks that modifies the original Hybrid Probabilistic Logic Programming framework of [4], and generalizes and modifies the *probabilistic annotated logic programming framework*, originally proposed in [19] and further extended in [20]. Probabilities in [27] are presented as intervals where a probability interval represents the bounds on the degree of belief a rational agent has about the truth of an event. The semantics of HPP [27], intuitively, captures the probabilistic reasoning about how likely are the various events to occur. It was shown that the HPP [27] framework is more suitable than [4] for reasoning and decision making tasks. In addition, it subsumes Lakshmanan and Sadri's [14] probabilistic implication-based framework as well as being a natural extension of classical logic programming. As a step towards enhancing its reasoning capabilities, the framework of HPP was extended to cope with non-monotonic negation [28] by introducing the notion of Normal Hybrid Probabilistic Logic Programs (NHPP) and to provide two different semantics, namely stable probabilistic model semantics and well-founded probabilistic model semantics. Furthermore, NHPP was extended to Extended Hybrid Probabilistic Logic Programs (EHPP) [25] to cope directly with classical negation as well as non-monotonic negation to allow reasoning in the presence of incomplete knowledge.

In [28], it was shown that the relationship between the stable probabilistic model semantics and the well-founded probabilistic model semantics of NHPP *preserves* the relationship between the stable model semantics and the well-founded semantics for normal logic programs [8]. More importantly, the stable probabilistic models semantics *naturally extends* the stable model semantics [9] of normal logic programs and the well-founded probabilistic model semantics *naturally extends* the well-founded semantics [8] of normal logic programs. A consequence of this is that efficient algorithms and implementations for computing those semantics can be developed by extending the existing efficient algorithms and implementations for computing the stable model semantics and the well-found semantics for normal logic programs, e.g., SMODELS [22]. However, NHPP is developed to represent and reason in the presence of quantitative uncertainty.

However, in [25], it was shown that EHPP explicitly encodes negative information, which is important to provide the capability to reason with incomplete knowledge. The semantics of EHPP relies on a probabilistic generalization of the answer set semantics, originally developed for extended logic programs [10].

The probabilistic answer set semantics of EHPP naturally extends the answer set semantics for classical extended logic programs [10]. Moreover, it was shown that Baral et al.'s probabilistic logic programming approach for reasoning with causal Bayes networks (P-log) [1] is naturally subsumed by EHPP. Furthermore, it was shown that the probabilistic answer set semantics of EHPP is reduced to the stable probabilistic model semantics of NHPP proposed in [28]. The importance of that is computational methods developed for NHPP can be applied to the language of EHPP. Moreover, it was described in [25] that some commonsense probabilistic knowledge can be easily represented in EHPP. Similar to NHPP, EHPP is used to represent and reason in the presence of quantitative uncertainty.

Although [29] allows disjunctions in the head of rules, the probabilistic logic programming framework in [29] is used to represent and reason with quantitative uncertainty to reason with Bayes networks. In addition, EDHPP (NDHPP) is more expressive than [29], since, for example EDHPP, unlike [29], allows classical negation, non-monotonic negation, different modes of probabilistic combinations (since [29] considers independence of probabilities which is a fixed mode of probabilistic combination), and compound events to appear in the body of rules, as well as, Bayes reasoning and representation.

Similar to [29], another approach for probabilistic logic programming has been provided in [16] for quantitative uncertainty reasoning. In [16], a possible world semantics for reasoning about probabilities has been introduced by assigning probabilistic measures over the possible worlds using normal disjunctive logic programs. A probabilistic logic program in [16] consists of a set of normal disjunctive logic program clauses with associated probabilities. A normal disjunctive clause in [16] is treated as a classical formula with an associated probability, where the implication in such a clause is treated as material implication. In addition, an approximate semantics for probabilistic logic programming in [16] has been presented, where probabilities are treated as a lattice of truth values. In this case, the probability of a conjunction $Prob(A \wedge B) = min(Prob(A), Prob(B))$ and the probability of a disjunction $Prob(A \vee B) = max(Prob(A), Prob(B))$. This is considered a fixed mode of combination. Whereas, in our framework conjunctions and disjunctions are treated differently according to the type of dependency between events. In addition, unlike [16], we allow classical negation and compound events to appear in the body of rules.

A logical approach has been presented in [1] to reason with causal Bayes networks by considering a body of logical knowledge, by using the answer set semantics of classical answer set programming [10]. Although, full answer set programming (logic programs with classical negation, non-monotonic negation, and disjunctions) is used, the probabilistic logic programming framework in [1] is used to reason in the presence of quantitative uncertainty. Answer set semantics [10] has been used to emulate the possible world semantics. Probabilistic logic programs of [1] is expressive and straightforward and relaxed some restrictions on the logical knowledge representation part existed in similar approaches to Bayesian reasoning, e.g., [12, 18, 23, 24, 29]. Since [19, 20, 21, 4] provided a different semantical characterization to probabilistic logic programming, it was not clear that how these proposals relate to [1]. However, the work presented in this paper and [28, 25], which are modification and generalization of the work presented in [19, 20, 21, 4], are closely related to [1]. The framework presented in this paper, as well as the framework of [25], is strictly syntactically and semantically subsumes probabilistic logic programs of [1]. This can be argued by the fact that EDHPP naturally extends classical extended disjunctive logic programs with answer set semantics [10], and probabilistic logic programs of [1] mainly rely on extended disjunctive logic programs with answer set semantics [10] as a knowledge representation and inference mechanism for reasoning with causal Bayes networks. In this sense, the comparisons established between [1] and the existing probabilistic logic programming approaches such as [12, 18, 23, 24, 29, 19, 20, 21, 4, 15, 16, 3] also carry over to EDHPP and these approaches. In addition, unlike [1], EDHPP does not put any restriction on the type of dependency existing among events.

1.2 Paper Organization

This paper is organized as follows. Sections 2 and 3 describe the syntax, stable probabilistic model semantics of NDHPP, and the probabilistic answer set semantics of EDHPP. Finally, conclusions with some perspectives are presented in section 4.

2 Syntax

In this section we introduce the basic notions associated to the languages of EDHPP and NDHPP described throughout the rest of the paper [4, 28, 25]. EDHPP (NDHPP) is EHPP (NHPP) with disjunctions of annotated literals (atoms) in the head of rules.

2.1 Probabilistic Strategies

Let $C[0,1]$ denotes the set of all closed intervals in $[0,1]$. In the context of EDHPP, probabilities are assigned to primitive events (literals) and compound events (conjunctions or disjunctions of literals) as intervals in $C[0,1]$. Let $[\alpha_1,\beta_1],[\alpha_2,\beta_2] \in C[0,1]$. Then the *truth order* asserts that $[\alpha_1,\beta_1] \leq_t [\alpha_2,\beta_2]$ iff $\alpha_1 \leq \alpha_2$ and $\beta_1 \leq \beta_2$. The set $C[0,1]$ and the relation $\leq_t$ form a complete lattice. The type of dependency among the primitive events within a compound event is described by *probabilistic strategies*, which are explicitly selected by the user. We call ρ, a pair of functions $\langle c, md\rangle$, a probabilistic strategy (p-strategy), where $c : C[0,1] \times C[0,1] \rightarrow C[0,1]$, the *probabilistic composition function*, which is *commutative*, *associative*, *monotonic* w.r.t. $\leq_t$, and meets the following *separation* criteria: there are two functions c_1, c_2 such that $c([\alpha_1,\beta_1],[\alpha_2,\beta_2]) = [c_1(\alpha_1,\alpha_2), c_2(\beta_1,\beta_2)]$. Whereas, $md : C[0,1] \rightarrow C[0,1]$ is the *maximal interval function*. The maximal interval function md of a certain p-strategy returns an estimate of the probability range of a primitive event, e, from the probability range of a compound event that contains e. The composition function c returns the probability range of a conjunction (disjunction) of two events given the ranges of its constituents. For convenience, given a multiset of probability intervals $M = \{\!\{[\alpha_1,\beta_1],\ldots,[\alpha_n,\beta_n]\}\!\}$, we use cM to denote $c([\alpha_1,\beta_1], c([\alpha_2,\beta_2],\ldots,c([\alpha_{n-1},\beta_{n-1}],[\alpha_n,\beta_n]))\ldots)$. According to the type of combination among events, p-strategies are classified into *conjunctive* p-strategies and *disjunctive* p-strategies. Conjunctive (disjunctive) p-strategies are employed to compose events belonging to a conjunctive (disjunctive) formula (please see [4, 27] for the formal definitions).

2.2 The Languages of EDHPP and NDHPP

Let $\mathcal{L}$ be an arbitrary first-order language with finitely many predicate symbols, function symbols, constants, and infinitely many variables. In addition, let $S = S_{conj} \cup S_{disj}$ be an arbitrary set of p-strategies, where S_{conj} (S_{disj}) is the set of all conjunctive (disjunctive) p-strategies in S. The Herbrand base of $\mathcal{L}$ is denoted by $\mathcal{B}_\mathcal{L}$. A literal is either an atom a or the negation of an atom $\neg a$, where $\neg$ is the classical negation. We denote the set of all literals in $\mathcal{L}$ by Lit. More formally, $Lit = \{a | a \in \mathcal{B}_\mathcal{L}\} \cup \{\neg a | a \in \mathcal{B}_\mathcal{L}\}$. An *annotation* denotes a probability interval and it is represented by $[\alpha_1,\alpha_2]$, where α_1,α_2 are called annotation items. An *annotation item* is either a constant in $[0,1]$, a variable (*annotation variable*) ranging over $[0,1]$, or $f(\alpha_1,\ldots,\alpha_n)$ (called *annotation function*) where f is a representation of a monotonic total function $f : ([0,1])^n \rightarrow [0,1]$ and $\alpha_1,\ldots,\alpha_n$ are annotation items.

The building blocks of the language of EDHPP are *hybrid literals*. Let us consider a set of literals $l_1,\ldots,l_n$ and the p-strategies ρ and ρ'. Then $l_1 \wedge_\rho \ldots \wedge_\rho l_n$ and $l_1 \vee_{\rho'} \ldots \vee_{\rho'} l_n$ are called *hybrid literals*. A hybrid literal L is ground if each literal l_i in L is ground. $bf_S(Lit)$ is the set of all ground hybrid literals formed using distinct literals from Lit and p-strategies from S, such that for any collection of equivalent hybrid literals, $\mathsf{Y} = \{l_1 *_\rho l_2 *_\rho \ldots *_\rho l_n,\ l_2 *_\rho l_1 *_\rho \ldots *_\rho l_n, \ldots\}$, where $* \in \{\wedge,\vee\}$, only one $l_{i_1} *_\rho l_{i_2} *_\rho \ldots *_\rho l_{i_n} \in \mathsf{Y}$ is in $bf_S(Lit)$. An *annotated hybrid literal* is an expression of the form $L : \mu$, where L is a hybrid literal and μ is an annotation. Note that any hybrid literal L can be represented in terms of another hybrid literal L' such that $L = \neg L'$, since $\neg\neg a = a$, $(a_1 \wedge_\rho a_2) = \neg(\neg a_1 \vee_\rho \neg a_2)$ and $(a_1 \vee_{\rho'} a_2) = \neg(\neg a_1 \wedge_{\rho'} \neg a_2)$ and $\wedge_\rho, \vee_\rho, \vee_{\rho'}$, and $\wedge_{\rho'}$ are associative and commutative.

However, the building blocks of the language of NDHPP are *hybrid basic formulae*. Let us consider a collection of atoms $a_1,\ldots,a_n$ and the p-strategies ρ and ρ'. Then $a_1 \wedge_\rho \ldots \wedge_\rho a_n$ and $a_1 \vee_{\rho'} \ldots \vee_{\rho'} a_n$ are called *hybrid basic formulae*. A hybrid basic formula F is ground if each atom A_i in F is ground. $bf_S(\mathcal{B}_\mathcal{L})$ is the set of all ground hybrid basic formulae formed using distinct atoms from $\mathcal{B}_\mathcal{L}$ and p-strategies from S, such that for any collection of equivalent hybrid basic formulae, $\mathsf{X} = \{a_1 *_\rho a_2 *_\rho \ldots *_\rho a_n,\ a_2 *_\rho a_1 *_\rho \ldots *_\rho a_n, \ldots\}$, where $* \in \{\wedge,\vee\}$, only one $a_{i_1} *_\rho a_{i_2} *_\rho \ldots *_\rho a_{i_n} \in \mathsf{X}$ is in $bf_S(\mathcal{B}_\mathcal{L})$. An annotated hybrid basic formula is an expression of the form $F : \mu$ where F is a hybrid basic formula and μ is an annotation.

3 Extended and Normal Disjunctive Hybrid Probabilistic Logic Programs

In this section we define the syntax, declarative semantics, the probabilistic answer set semantics of *Extended Disjunctive Hybrid Probabilistic Logic Programs (EDHPP)*, and the stable probabilistic model semantics of *Normal Disjunctive Hybrid Probabilistic Logic Programs (NDHPP)*.

Definition 1 (Rules) *An extended disjunctive hybrid probabilistic rule (ed-rule) is an expression of the form*

$$l_1 : \nu_1 \ or \ldots or\ l_k : \nu_k \leftarrow L_1 : \mu_1,\ldots,L_m : \mu_m,\ not\ (L_{m+1} : \mu_{m+1}),\ldots,not\ (L_n : \mu_n),$$

whereas a normal disjunctive hybrid probabilistic rule (nd-rule) is an expression of the form

$$A_1 : \nu_1 \ or \ldots or\ A_k : \nu_k \leftarrow F_1 : \mu_1,\ldots,F_m : \mu_m,\ not\ (F_{m+1} : \mu_{m+1}),\ldots,not\ (F_n : \mu_n),$$

where $l_1, \ldots, l_k$ are literals, $A_1, \ldots, A_k$ are atoms, L_i $(1 \leq i \leq n)$ are hybrid literals, F_i $(1 \leq i \leq n)$ are hybrid basic formulae, and $\nu_i (1 \leq i \leq k), \mu_i$ $(1 \leq i \leq n)$ are annotations.

An ed-rule^{not} is an ed-rule without non-monotonic negation—i.e., $n = m$, and a d-rule is an nd-rule without non-monotonic negation—i.e., $n = m$.

The intuitive meaning of an ed-rule, in Definition 1, is that, if for each $L_i : \mu_i$, where $1 \leq i \leq m$, the probability interval of L_i is at least μ_i and for each *not* $(L_j : \mu_j)$, where $m+1 \leq j \leq n$, it is *not known (undecidable)* that the probability interval of L_j is at least μ_j, then there exist at least l_i, where $1 \leq i \leq k$, such that the probability interval of l_i is at least ν_i. However, the meaning of an nd-rule, is that, if for each $F_i : \mu_i$, where $1 \leq i \leq m$, the probability interval of F_i is at least μ_i and for each *not* $(F_j : \mu_j)$, where $m+1 \leq j \leq n$, it is *not provable* that the probability interval of F_j is at least μ_j, then there exist at least A_i, where $1 \leq i \leq k$, such that the probability interval of A_i is at least ν_i.

Definition 2 (Programs) *An extended (normal) disjunctive hybrid probabilistic logic program over S,* ed-program *(*nd-program*), is a pair $P = \langle R, \tau \rangle$, where R is a finite set of ed-rules (nd-rules) with p-strategies from S, and τ is a mapping $\tau : Lit \rightarrow S_{disj}$ $(\tau : \mathcal{B}_{\mathcal{L}} \rightarrow S_{disj})$. An extended (normal) disjunctive hybrid probabilistic logic program without non-monotonic negation is an ed-program (nd-program) where each rule in the program is an ed-rule^{not} (d-rule).*

The mapping τ in the above definition associates to each literal l_i (similarly for atoms in nd-programs) a disjunctive p-strategy that will be employed to combine the probability intervals obtained from different rules having l_i in their heads. An ed-program (nd-program) is ground if no variables appear in any of its rules.

3.1 Satisfaction and Models

In this subsection, we define the declarative semantics of EDHPP and NDHPP. We define the notions of interpretations, models, and satisfaction of ed-programs and nd-programs.

Definition 3 *A probabilistic interpretation (p-interpretation) of an ed-program is a partial or total mapping $h : bf_S(Lit) \rightarrow C[0,1]$. A probabilistic interpretation (p-interpretation) for an nd-program is a total mapping $h : bf_S(\mathcal{B}_{\mathcal{L}}) \rightarrow C[0,1]$.*

Since we allow both an event and its negation to be defined in p-interpretations for ed-programs, more conditions need to be imposed on p-interpretations to ensure their consistency. This can be characterized by the following definitions.

Definition 4 *A total (partial) p-interpretation h for an ed-program is inconsistent if there exists $L, \neg L \in bf_S(Lit)$ $(L, \neg L \in dom(h))$ such that $h(\neg L) \neq [1,1] - h(L)$.*

Definition 5 *We say a set C, a subset of Lit, is a set of consistent literals if there is no pair of complementary literals a and $\neg a$ belonging to C. Similarly, a consistent set of hybrid literals C^* is a subset of $bf_S(Lit)$ such that there is no pair of complementary hybrid literals F and $\neg F$ belonging to C^*.*

Definition 6 *A consistent p-interpretation h of an ed-program is either not inconsistent or maps a consistent set of hybrid literals C^* to $C[0,1]$.*

The notion of truth order can be extended to p-interpretations of nd-programs. Given p-interpretations h_1 and h_2 of an nd-program P, we say $(h_1 \leq_t h_2) \Leftrightarrow (\forall F \in bf_S(\mathcal{B}_{\mathcal{L}}) : h_1(F) \leq_t h_2(F))$. The set of all p-interpretations of P and the truth order $\leq_t$ form a complete lattice. In addition, given the p-interpretations h_1 and h_2 for an ed-program P', we say $(h_1 \leq_o h_2) \Leftrightarrow (dom(h_1) \subseteq dom(h_2)$ *and* $\forall L \in dom(h_1),\ h_1(L) \leq_t h_2(L))$. The set of all p-interpretations of P' and the partial order $\leq_o$ form a complete lattice.

Definition 7 (Probabilistic Satisfaction) *Let $P = \langle R, \tau \rangle$ be a ground ed-program, h be a p-interpretation, and*

$$r \equiv l_1 : \nu_1 \text{ or } \ldots \text{ or } l_k : \nu_k \leftarrow L_1 : \mu_1, \ldots, L_m : \mu_m, \text{not } (L_{m+1} : \mu_{m+1}), \ldots, \text{not } (L_n : \mu_n).$$

Then

- *h satisfies $L_i : \mu_i (l_i : \nu_i)$ (denoted by $h \models L_i : \mu_i (h \models l_i : \nu_i)$) iff $L_i \in dom(h) (l_j \in dom(h))$ and $\mu_i \leq_t h(L_i) (\nu_i \leq_t h(l_i))$.*
- *h satisfies not $(L_j : \mu_j)$ (denoted by $h \models$ not $(L_j : \mu_j)$) iff $L_j \in dom(h)$ and $h(L_j) <_t \mu_j$ or $L_j \notin dom(h)$.*
- *h satisfies*

  $$Body \equiv L_1 : \mu_1, \ldots, L_m : \mu_m, \text{not } (L_{m+1} : \mu_{m+1}), \ldots, \text{not } (L_n : \mu_n)$$

  (denoted by $h \models Body$) iff $\forall (1 \leq i \leq m), h \models L_i : \mu_i$ and $\forall (m+1 \leq j \leq n), h \models$ not $(L_j : \mu_j)$.
- *h satisfies $Head \equiv l_1 : \nu_1$ or $\ldots$ or $l_k : \nu_k$ (denoted by $h \models Head$) iff there exists at least i $(1 \leq i \leq k)$ such that $h \models l_i : \nu_i$.*
- *h satisfies $Head \leftarrow Body$ iff $h \models Head$ whenever $h \models Body$ or h does not satisfy $Body$.*
- *h satisfies P iff h satisfies every ed-rule in R and for every literal $l_i \in dom(h)$,*

$c_{\tau(l_i)}\{\!\{\nu_i \ (1 \leq i \leq k) \mid l_1 : \nu_1 \ or \ldots or \ l_k : \nu_k \leftarrow Body \in R, h \models Body, \ and \ h \models l_i : \nu_i\}\!\} \leq_t h(l_i)$.

Observe that the definition of probabilistic satisfaction for nd-programs is the same as the definition of probabilistic satisfaction for ed-programs described in Definition 7. The only difference is that classical negation is not allowed in nd-programs, in addition, p-interpretations of nd-programs are total mappings from $bf_S(\mathcal{B}_\mathcal{L})$ to $C[0,1]$.

Definition 8 (Models) *A probabilistic model (*p-model*) of an ed-program (nd-program), with or without non-monotonic negation, P is a p-interpretation of P that satisfies P.*

Definition 9 (Minimal Models) *Let P be an ed-program (nd-program). A p-model h of P is* minimal *w.r.t. $\leq_o$ ($\leq_t$) iff there does not exist a p-model h' of P such that $h' <_o h$ ($h' <_t h$).*

We call a minimal p-model of an ed-program a *probabilistic answer set*. It is possible to get a probabilistic answer set of an ed-program, P, and this probabilistic answer set is inconsistent. If this is the case, we say P is inconsistent. If P is inconsistent, LIT, where $LIT : bf_S(Lit) \rightarrow [1,1]$, is the probabilistic answer set of P. We adopt this view from the answer set semantics of classical logic programming [10].

Example 1 *Consider the following ed-program $P = \langle R, \tau\rangle$, without non-monotonic negation, where R contains*

$a : [0.1, 0.2]$	*or*	$\neg b : [0.15, 0.3]$
$\neg c : [0, 0.21]$	$\leftarrow$	$a : [0.1, 0.13]$
$d : [0.12, 0.18]$	$\leftarrow$	$\neg b : [0.1, 0.21]$
$\neg d : [0.45, 0.55]$	$\leftarrow$	$a : [0, 0.15], \neg b : [0.02, 0.22],$ $\neg c : [0.1, 0.1]$

and τ is any arbitrary assignment of disjunctive p-strategies. It is easy to verify that P has two probabilistic answer sets h_1 and h_2, where $h_1(a) = [0.1, 0.2]$ $h_1(\neg\ c) = [0, 0.21]$ *and* $h_2(\neg\, b) = [0.15, 0.3]$ $h_2(d) = [0.12, 0.18]$.

3.2 Probabilistic Answer Set and Stable Probabilistic Model Semantics

In this subsection we define the *probabilistic answer set* and the *stable probabilistic model* semantics of ed-programs and nd-programs respectively. The semantics are defined in two steps. First, we guess a probabilistic answer set (stable probabilistic model) h for a certain ed-program (nd-program) P, then we define the notion of the probabilistic reduct of P with respect to h. The probabilistic reduct is an ed-program (nd-program) without non-monotonic negation. Second, we determine whether h is a probabilistic answer set (stable probabilistic model) for P. This is verified by determining whether h is a probabilistic answer set (minimal p-model) of the probabilistic reduct of P w.r.t. h.

Definition 10 (Probabilistic Reduct) *Let $P = \langle R, \tau\rangle$ be a ground ed-program (nd-program) and h be a p-interpretation. The probabilistic reduct P^h of P w.r.t. h is $P^h = \langle R^h, \tau\rangle$ where:*

$$R^h = \left\{ \begin{array}{l} l_1 : \nu_1 \ or \ldots or \ l_k : \nu_k \leftarrow L_1 : \mu_1, \ldots, L_m : \mu_m \mid \\ \quad l_1 : \nu_1 \ or \ldots or \ l_k : \nu_k \leftarrow L_1 : \mu_1, \ldots, L_m : \mu_m, \\ \quad not \ (L_{m+1} : \mu_{m+1}), \ldots, not \ (L_n : \mu_n) \in R \ and \\ \quad \forall (m+1 \leq j \leq n), \ h(L_j) <_t \mu_j \ or \ L_j \notin dom(h) \end{array} \right.$$

Note that the definitions of the probabilistic reduct for ed-programs and nd-programs are similar. Except that classical negation is not allowed in nd-programs. In addition, p-interpretations in nd-programs are total mappings from $bf_S(\mathcal{B}_\mathcal{L})$ to $C[0,1]$, therefore, for nd-programs, the condition $L_j \notin dom(h)$ is not applicable.

The probabilistic reduct P^h is an ed-program (nd-program) without non-monotonic negation. For any *not* $(L_j : \mu_j)$ in the body of $r \in R$ with $h(L_j) <_t \mu_j$ means that it is *not known (not provable for nd-program)* that the probability interval of L_j is at least μ_j given the available knowledge, and *not* $(L_j : \mu_j)$ is removed from the body of r. In addition, for ed-program, if $L_j \notin dom(h)$, i.e., L_j is undefined in h, then it is completely *not known (undecidable)* that the probability interval of L_j is at least μ_j. In this case, *not* $(L_j : \mu_j)$ is also removed from the body of r. If $\mu_j \leq_t h(L_j)$ (similarly for nd-programs), then we know that the probability interval of L_j is at least μ_j and the body of r is not satisfied and r is trivially ignored.

Definition 11 *A p-interpretation h of an ed-program (nd-program) P is a probabilistic answer set (stable probabilistic model) of P if h is a minimal p-model of P^h.*

The domain of a probabilistic answer set of an ed-program or a stable probabilistic model of an nd-program represents an agent set of beliefs. However, the probability intervals associated to these beliefs bound the agents belief degrees on these beliefs. ed-programs without classical negation (nd-programs), i.e., ed-programs that contain no negative literals neither in head nor in the body of ed-rules, have probabilistic answer sets with hybrid literals consist of only atoms (hybrid basic formulae). Moreover, the definition of probabilistic answer sets coincides with the definition of stable probabilistic models for nd-programs. This means that the application of the probabilistic answer set semantics to nd-programs is reduced to the

stable probabilistic model semantics for nd-programs. However, there are a couple of main differences between the two semantics. A probabilistic answer set may be a partial p-interpretation, however, a stable probabilistic model is a total p-interpretation. In addition, each hybrid basic formula F with probability interval [0,0] in a stable probabilistic model of an nd-program corresponds to the fact that the probability interval of F is unknown, and hence undefined, in its equivalent probabilistic answer set.

Proposition 1 *Let P be an ed-program without classical negation. Then h is a probabilistic answer set for P iff h' is a stable probabilistic model of P, where $h(F) = h'(F)$ for each $h'(F) \neq [0,0]$ and $h(F)$ is undefined for each $h'(F) = [0,0]$.*

Proposition 1 suggests that there is a simple reduction from ed-programs to nd-programs. The importance of that is, under the consistency condition, computational methods developed for nd-programs can be applied to ed-programs.

Example 2 *Consider the following example adapted from [11]. Tom and Fred are two policemen who are challenging their firing gun skills, by shooting a bottle at a quite long distance. In one of the shoots, at the same time, both Tom and Fred shoot a bottle and the bottle shattered. In fact, we cannot determine whether Tom or Fred is the one who shattered the bottle. However, from Tom's shooting experience on similar targets at similar distances, Tom is capable of hitting targets with probability interval from 75% to 80%. Similarly, Fred can hit similar targets with probability interval from 72% to 87%. Normally, a shooter shoots a target. If a shooter sneezes while shooting, it is an exception. Hence, a shooter's shoot is abnormal with probability interval from 30% to 65% if a shooter sneezes while shooting. It was heard that somebody sneezed, however, we do not know whether Tom or Fred is the one who sneezed. A shooter shatters a bottle with probability interval from 82% to 90% if a shooter is capable of hitting similar targets with probability interval from 70% to 79%, and it is not known that a shooter's shoot is abnormal with probability interval from 30% to 60%. This can be represented by the following ed-program $P = \langle R, \tau \rangle$, where R contains:*

$sneeze(tom) : [1,1] \quad or \quad sneeze(fred) : [1,1] \leftarrow$
$ab(shoot, X) : [0.3, 0.65] \leftarrow shoot(X) : [1,1],$
$\qquad sneeze(X) : [1,1]$
$shatter(X) : [0.82, 0.9] \leftarrow hit(X) : [0.7, 0.79],$
$\qquad not\ (ab(shoot, X) : [0.3, 0.6])$
$shoot(tom) : [1,1] \leftarrow$
$shoot(fred) : [1,1] \leftarrow$
$hit(tom) : [0.75, 0.8] \leftarrow$
$hit(fred) : [0.72, 0.87] \leftarrow$

and τ is any arbitrary assignment of disjunctive p-strategies. The ed-rules in Example 2 encode two forms of uncertainty. Qualitative uncertainty represented by the first ed-rule that arises from the fact that we do not know whether Tom or Fred is the one who sneezed. And quantitative uncertainty represented by the probability intervals associated to the various events presented in R. The probability interval $[1,1]$ represents the truth value *true*. Therefore, the rule $sneeze(tom) : [1,1]$ *or* $sneeze(fred) : [1,1] \leftarrow$ is intuitively interpreted as a disjunctive rule in classical disjunctive logic programming. The above ed-program P has two probabilistic answer sets h_1 and h_2, where

$h_1(sneeze(fred)) = [1,1]$
$h_1(ab(shoot, fred)) = [0.3,0.65]$
$h_1(shatter(tom)) = [0.82, 0.9]$
$h_1(shoot(tom)) = [1,1]$
$h_1(shoot(fred)) = [1,1]$
$h_1(hit(tom)) = [0.75,0.8]$
$h_1(hit(fred)) = [0.72,0.87]$

$h_2(sneeze(tom)) = [1,1]$
$h_2(ab(shoot, tom)) = [0.3,0.65]$
$h_2(shatter(fred)) = [0.82, 0.9]$
$h_2(shoot(tom)) = [1,1]$
$h_2(shoot(fred)) = [1,1]$
$h_2(hit(fred)) = [0.72,0.87]$
$h_2(hit(tom)) = [0.75,0.8]$

For example, h_1 can be verified as a probabilistic answer set of P by computing the probabilistic reduct, $P^{h_1} = \langle R^{h_1}, \tau \rangle$, of P w.r.t. h_1, where R^{h_1} contains

$sneeze(tom) : [1,1] \quad or \quad sneeze(fred) : [1,1] \leftarrow$
$ab(shoot, tom) : [0.3, 0.65] \leftarrow shoot(tom) : [1,1],$
$\qquad sneeze(tom) : [1,1]$
$ab(shoot, fred) : [0.3, 0.65] \leftarrow shoot(fred) : [1,1],$
$\qquad sneeze(fred) : [1,1]$
$shatter(tom) : [0.82, 0.9] \leftarrow hit(tom) : [0.7, 0.79]$
$shoot(tom) : [1,1] \leftarrow$
$shoot(fred) : [1,1] \leftarrow$
$hit(tom) : [0.75, 0.8] \leftarrow$
$hit(fred) : [0.72, 0.87] \leftarrow$

It can be easily seen that h_1 is a probabilistic answer set for P^{h_1}.

Example 3 *Assume that either we believe that Tom is the one who hit the bottle or we believe that Fred is the one who hit the bottle. However, if Tom is the one who hit the bottle he can only hit it with probability interval from 75% to 80%. Similarly, if Fred is the one who hit the bottle he can only hit it with probability interval from 72% to 87%. This means that either Tom hit the bottle with probability interval from 75% to 80% or Fred hit the bottle with probability interval from 72% to 87%. This leads to the following encoding of the ed-program $P = \langle R, \tau \rangle$ presented in Example 2, where R now contains:*

$$
\begin{array}{l}
hit(tom) : [0.75, 0.8] \quad or \quad hit(fred) : [0.72, 0.87] \leftarrow \\
sneeze(tom) : [1, 1] \quad or \quad sneeze(fred) : [1, 1] \leftarrow \\
ab(shoot, X) : [0.3, 0.65] \leftarrow shoot(X) : [1, 1], \\
\qquad\qquad\qquad\qquad\qquad sneeze(X) : [1, 1] \\
shatter(X) : [0.82, 0.9] \leftarrow hit(X) : [0.7, 0.79], \\
\qquad\qquad\qquad\qquad\qquad not\ (ab(shoot, X) : [0.3, 0.6]) \\
shoot(tom) : [1, 1] \leftarrow \\
shoot(fred) : [1, 1] \leftarrow
\end{array}
$$

and τ is any arbitrary assignment of disjunctive p-strategies. The first ed-rule in R presents that quantitative uncertainty (the probability intervals $[0.75, 0.8]$ and $[0.72, 0.87]$) can be defined over qualitative uncertainty, where probabilistic measures are assigned over the possible outcomes ($hit(tom)$ and $hit(fred)$) of qualitative uncertainty. The above ed-program P has four probabilistic answer sets h_1, h_2, h_3, and h_4, where

$$
\begin{array}{lcl}
h_1(hit(tom)) & = & [0.75,0.8] \\
h_1(sneeze(tom)) & = & [1,1] \\
h_1(ab(shoot, tom)) & = & [0.3,0.65] \\
h_1(shoot(tom)) & = & [1,1] \\
h_1(shoot(fred)) & = & [1,1]
\end{array}
$$

$$
\begin{array}{lcl}
h_2(hit(fred)) & = & [0.72,0.87] \\
h_2(sneeze(fred)) & = & [1,1] \\
h_2(ab(shoot, fred)) & = & [0.3,0.65] \\
h_2(shoot(tom)) & = & [1,1] \\
h_2(shoot(fred)) & = & [1,1]
\end{array}
$$

$$
\begin{array}{lcl}
h_3(hit(tom)) & = & [0.75,0.8] \\
h_3(sneeze(fred)) & = & [1,1] \\
h_3(ab(shoot, fred)) & = & [0.3,0.65] \\
h_3(shatter(tom)) & = & [0.82,\ 0.9] \\
h_3(shoot(tom)) & = & [1,1] \\
h_3(shoot(fred)) & = & [1,1]
\end{array}
$$

$$
\begin{array}{lcl}
h_4(hit(fred)) & = & [0.72,0.87] \\
h_4(sneeze(tom)) & = & [1,1] \\
h_4(ab(shoot, tom)) & = & [0.3,0.65] \\
h_4(shatter(fred)) & = & [0.82,\ 0.9] \\
h_4(shoot(tom)) & = & [1,1] \\
h_4(shoot(fred)) & = & [1,1]
\end{array}
$$

For example, h_3 can be verified as a probabilistic answer set of P by computing the probabilistic reduct, $P^{h_3} = \langle R^{h_3}, \tau \rangle$, of P w.r.t. h_3, where R^{h_3} contains

$$
\begin{array}{l}
hit(tom) : [0.75, 0.8] \quad or \quad hit(fred) : [0.72, 0.87] \leftarrow \\
sneeze(tom) : [1, 1] or \qquad sneeze(fred) : [1, 1] \leftarrow \\
ab(shoot, tom) : [0.3, 0.65] \leftarrow shoot(tom) : [1, 1], \\
\qquad\qquad\qquad\qquad\qquad sneeze(tom) : [1, 1] \\
ab(shoot, fred) : [0.3, 0.65] \leftarrow shoot(fred) : [1, 1], \\
\qquad\qquad\qquad\qquad\qquad sneeze(fred) : [1, 1] \\
shatter(tom) : [0.82, 0.9] \leftarrow hit(tom) : [0.7, 0.79] \\
shoot(tom) : [1, 1] \leftarrow \\
shoot(fred) : [1, 1] \leftarrow
\end{array}
$$

It can be easily seen that h_3 is a probabilistic answer set for P^{h_3}.

Now we show that EDHPP and NDHPP naturally extend EHPP and NHPP respectively.

Proposition 2 *The probabilistic answer set semantics of EDHPP is equivalent to the probabilistic answer set semantics of EHPP [25] for all ed-programs $P = \langle R, \tau \rangle$ such that $\forall\ r \in R, k = 1$. In addition, the stable probabilistic model semantics of NDHPP is equivalent to the stable probabilistic model semantics of NHPP [28] for all nd-programs $P = \langle R, \tau \rangle$ such that $\forall\ r \in R, k = 1$.*

Let us show that the probabilistic answer set semantics of EDHPP and the stable probabilistic model semantics of NDHPP generalize the answer set semantics and the stable model semantics of extended and normal disjunctive logic programs [2, 10] respectively. An extended disjunctive logic program P can be represented as an ed-program $P' = \langle R, \tau \rangle$ where each extended disjunctive rule

$$l_1\ or \ldots or\ l_k \leftarrow l'_1, \ldots, l'_m, not\ l'_{m+1}, \ldots, not\ l'_n \in P$$

can be represented, in R, as an ed-rule of the form

$$
\begin{array}{l}
l_1 : [1, 1]\ or \ldots or\ l_k : [1, 1] \leftarrow l'_1 : [1, 1], \ldots, l'_m : [1, 1], \\
\qquad\qquad not\ (l'_{m+1} : [1, 1]), \ldots, not\ (l'_n : [1, 1]) \in R
\end{array}
$$

where $l_1, \ldots, l_k, l'_1, \ldots, l'_n$ are literals and $[1, 1]$ represents the truth value *true*. τ is any arbitrary assignment of disjunctive p-strategies. We call the class of ed-programs that consists of only ed-rules of the above form as $EDHPP_1$. Recall that nd-programs are ed-programs without classical negation. $NDHPP_1$ is the same as $EDHPP_1$, except that, only atoms (positive literals) are allowed to appear in rules of the above form. The following result shows that $EDHPP_1$ and $NDHPP_1$ subsume classical extended and normal disjunctive logic programs [2, 10].

Proposition 3 *Let P_1 be an extended disjunctive logic program. Then $S_1{}'$ is an answer set of P_1 iff h_1 is a probabilistic answer of $P_1{}' \in EDHPP_1$ that corre-*

sponds to P_1 *where* $h_1(l) = [1, 1]$ *iff* $l \in S_1{}'$ *and* $h_1(l')$ *is undefined iff* $l' \notin S_1{}'$. *Let* P_2 *be a normal disjunctive logic program. Then* $S_2{}'$ *is a stable model of* P_2 *iff* h_2 *is a stable probabilistic model of* $P_2{}' \in NDHPP_1$ *that corresponds to* P_2 *where* $h_2(a) = [1, 1]$ *iff* $a \in S_2{}'$ *and* $h_2(b) = [0, 0]$ *iff* $b \in \mathcal{B}_\mathcal{L} \setminus S_2{}'$.

In the following result, we show that EDHPP naturally subsumes the probabilistic logic programming framework (P-log) of [1]. This means that any P-log program can be represented as an ed-program. In [1], a logical approach has been presented to reason with causal Bayes networks, by considering a body of logical knowledge, using the answer set semantics of classical logic programming [1]. Answer set semantics has been used to emulate the possible world semantics in [1].

Proposition 4 *The language of EDHPP subsumes P-log, a probabilistic logic programming framework for reasoning with causal Bayes networks [1].*

4 Conclusions and Future Work

We extended Extended and Normal Hybrid Probabilistic Logic Programs [25, 28] to Extended and Normal Disjunctive Hybrid Probabilistic Logic Programs, to allow classical negation, non-monotonic negation, and disjunctions in the head of rules. The extension is necessary to provide the capability of reasoning in the presence of both qualitative and quantitative uncertainty in a unified logic programming framework. In addition to the ability to assign quantitative uncertainly over qualitative uncertainty, where probabilistic measures are assigned over the possible outcomes of qualitative uncertainty. We developed semantical characterizations of the extended languages, which rely on generalizations of the answer set semantics and the stable model semantics, originally developed for extended and normal disjunctive logic programs [10, 2], and the probabilistic answer set semantics and the stable probabilistic model semantics for Extended and Normal Hybrid Probabilistic Logic Programs [25, 28]. We showed that the probabilistic answer set semantics of EDHPP naturally generalizes the answer set semantics of extended disjunctive logic programs [10] and the probabilistic answer set semantics of EHPP [25]. In addition, the stable probabilistic model semantics of NDHPP generalizes the stable model semantics of normal disjunctive logic programs [2] and the stable probabilistic model semantics of NHPP [28]. Furthermore, we showed that the probabilistic answer set semantics of EDHPP is reduced to stable probabilistic model semantics of NDHPP. The importance of that is computational methods developed for NDHPP can be applied to the language of EDHPP. Moreover, we showed that some commonsense probabilistic knowledge can be easily represented in EDHPP and NDHPP. In addition, we showed that EDHPP naturally subsumes the probabilistic logic programming framework of [1].

The main topic of future research is to investigate the computational aspects of the probabilistic answer set semantics of EDHPP and stable probabilistic model semantics of NDHPP—by developing algorithms and implementations for computing these semantics. The algorithms and implementations we will develop will be based on appropriate extensions of the existing techniques for computing the answer set (stable model) semantics for extended (normal) disjunctive logic programs, e.g., DLV [7].

References

[1] C. Baral, M. Gelfond, and N. Rushton. Probabilistic reasoning with answer sets. *In Proc. 7th International Conference on Logic Programming and Nonmonotonic Reasoning*, Springer Verlag, 2004.

[2] G. Brewka and J. Dix. Knowledge representation with logic programs. *Third International Workshop on Logic Programming and Knowledge Representation*, 1997.

[3] A. Dekhtyar and I. Dekhtyar. Possible worlds semantics for probabilistic logic programs. *International Conference of Logic Programming*, 137-148, 2004.

[4] A. Dekhtyar and V.S. Subrahmanian. Hybrid probabilistic program. *Journal of Logic Programming*, 43(3): 187-250, 2000.

[5] M. Dekhtyar, A. Dekhtyar, and V. S. Subrahmanian. Hybrid Probabilistic Programs: Algorithms and Complexity. *In Proc. of Uncertainty in Artificial Intelliegence*, pages 160-169, 1999.

[6] T. Eiter and T. Lukasiewicz. Probabilistic reasoning about actions in nonmonotonic causal theories. *In Proc. of Uncertainty in Artificial Intelliegence*, pp. 192-199, 2003.

[7] T. Eiter et al. Declarative problem solving in dlv. *In Logic Based Artificial Intelligence*, 2000.

[8] A. Van Gelder, K.A. Ross, and J.S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620-650, 1991.

[9] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. *ICSLP*, 1988, MIT Press.

[10] M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9(3-4):363-385, 1991.

[11] J. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843-887, 2005.

[12] K. Kersting and L. De Raedt. Bayesian Logic Programs. *In Proc. Inductive Logic Programming*, 2000.

[13] M. Kifer and V.S. Subrahmanian. Theory of generalized annotated logic programming and its applications. *Journal of Logic Programming*, 12:335–367, 1992.

[14] L.V.S. Lakshmanan and F. Sadri. On a theory of probabilistic deductive databases. *Journal of Theory and Practice of Logic Programming*, 1(1):5-42, January 2001.

[15] T. Lukasiewicz. Probabilistic logic programming. *In Proc. 13th European Conference on Artificial Intelligence*, 388–392, 1998.

[16] T. Lukasiewicz. Many-valued disjunctive logic programs with probabilistic semantics. *In Proc. International Conference on Logic Programming and Nonmonotonic Reasoning*, 1999.

[17] J. Fernandez and J. Minker. Disjunctive deductive databases. *In Proc. International Conference on Logic Programming and Automated Reasoning*, 1992.

[18] S. Muggleton. Stochastic logic programming. *In Proc. 5th International Workshop on Inductive Logic Programming*, 1995.

[19] R.T. Ng and V.S. Subrahmanian. Probabilistic logic programming. *Information & Computation*, 101(2), 1992.

[20] R.T. Ng and V.S. Subrahmanian. A semantical framework for supporting subjective and conditional probabilities in deductive databases. *Journal of Automated Reasoning*, 10(2), 1993.

[21] R.T. Ng and V.S. Subrahmanian. Stable semantics for probabilistic deductive databases. *Information & Computation*, 110(1), 1994.

[22] I. Niemela and P. Simons. Efficient implementation of the well-founded and stable model semantics. *In Proc. Joint International Conference and Symposium on Logic Programming*, 289-303, 1996.

[23] D. Poole. The Independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, 94(1-2), 7-56, 1997.

[24] D. Poole. Abducing through negation as failure: stable models within the independent choice logic. *Journal of Logic Programming*, Vol 44, 5-35, 2000.

[25] E. Saad. Incomplete knowlege in hybrid probabilistic logic programs. *In Proc. 10th European Conference on Logics in Artificial Intelligence*, 2006.

[26] E. Saad. Classical negation in hybrid probabilistic logic programs. *Submitted*.

[27] E. Saad and E. Pontelli. Towards a more practical hybrid probabilistic logic programming framework. *In Proc. Practical Aspects of Declarative Languages*, 2005.

[28] E. Saad and E. Pontelli. Hybrid probabilistic logic programs with non-monotonic negation. *In Proc. International Conference of Logic Programming*. Springer Verlag, 2005.

[29] J. Vennekens, S. Verbaeten, and M. Bruynooghe. Logic programs with annotated disjunctions. *In Proc. International Conference of Logic Programming*, 431-445, 2004.

Coherent Choice Functions under Uncertainty

Teddy Seidenfeld
Mark J. Schervish
Joseph B. Kadane
Carnegie Mellon University
teddy@stat.cmu.edu
mark@stat.cmu.edu
kadane@stat.cmu.edu

Abstract

We discuss several features of *coherent choice functions* – where the admissible options in a decision problem are exactly those which maximize expected utility for some probability/utility pair in fixed set $\boldsymbol{S}$ of probability/utility pairs. In this paper we consider, primarily, normal form decision problems under uncertainty – where only the probability component of $\boldsymbol{S}$ is indeterminate. Coherent choice distinguishes between each pair of sets of probabilities. We axiomatize the theory of choice functions and show these axioms are necessary for coherence. The axioms are sufficient for coherence using a set of probability/almost-state-independent utility pairs. We give sufficient conditions when a choice function satisfying our axioms is represented by a set of probability/state-independent utility pairs with a common utility.

Keywords. Choice functions, coherence, *Γ-Maximin*, *Maximality*, uncertainty, state-independent utility.

1 Introduction

In this paper we continue our study of coherent choice functions, which we started in our (2004) "Rubinesque" theory of decision. Coherent choice function theory provides a more general account of Imprecise Probabilities than the theory of coherent strict preference, which we used in our (1995). Coherent choice function theory does not reduce to binary comparisons between options, as Example 1 (below) illustrates. By contrast, coherent strict preference is a binary relation that fails, in principle, to distinguish between some convex sets of probabilities that have the same convex hull.

Specifically, as we show in Section 2, with coherent choice functions, for each two different sets of probabilities it requires only a simple decision problem in order to distinguish by admissibility between them. That is, with coherent choice functions, each set of probabilities has its own footprint of admissible options. In Section 4, we illustrate this added generality with a non-convex (even a disconnected) set $\boldsymbol{S}$ of probabilities that share the common structure that, for each distribution in $\boldsymbol{S}$, two specific events are independent. Coherent choice with respect to the set $\boldsymbol{S}$ avoids making information about one event valuable in decisions that depend solely on the other event. This is in sharp contrast with theories that rely on convex sets to depict Imprecise Probabilities

Let $\boldsymbol{O}$ be a (closed) set of feasible options. A ***choice function*** $\boldsymbol{C}(\boldsymbol{O})$ identifies the (non-empty) subset of $\boldsymbol{O}$ that are the admissible options in the decision problem given by the feasible set $\boldsymbol{O}$. We say that $\boldsymbol{C}(\bullet)$ is ***coherent*** provided that there is a non-empty set $\boldsymbol{S}$ of probability/utility pairs $\boldsymbol{S} = \{(\boldsymbol{p},\boldsymbol{u})\}$ such that the admissible options under $\boldsymbol{C}$ are precisely those that are Bayes with respect to some probability/utility pair $(\boldsymbol{p},\boldsymbol{u})$ in $\boldsymbol{S}$. That is, for each admissible option, for each $o \in \boldsymbol{C}(\boldsymbol{O})$, there is a pair $(\boldsymbol{p},\boldsymbol{u}) \in \boldsymbol{S}$ such that o maximizes the $\boldsymbol{p}$-expected $\boldsymbol{u}$-utility over $\boldsymbol{O}$. For short, we will call these the *Bayes-admissible options* in $\boldsymbol{O}$ (with respect to $\boldsymbol{S}$).

Aside: If the option set $\boldsymbol{O}$ is not closed, then given a set $\boldsymbol{S}$ there may be no coherently admissible options in $\boldsymbol{O}$. For example, if utility is linear and increasing in the quantity X, then in the decision-under-certainty problem with $\boldsymbol{O} = \{0 \le x < 1\}$, each option is inadmissible with respect to $\boldsymbol{S}$.

In Section 3 we adapt Anscombe-Aumann Horse-lottery theory in order to axiomatize coherent choice functions for cases where only probability (not utility) is indeterminate. This affords a representation of choice functions in the style of our previous work (1995), where we represented coherent strict (binary) preference between options using sets of probabilities and almost-state-independent utilities. One way to understand how the new representation generalizes our previous work is to consider the partial order $\langle$ defined on pairs of ***sets*** of options $\{\boldsymbol{O}_1,\boldsymbol{O}_2\}$: where $\boldsymbol{O}_1 \langle \boldsymbol{O}_2$ obtains whenever there are no admissible options from set $\boldsymbol{O}_1$ in a choice problem given the combined set of options $\boldsymbol{O}_1 \cup \boldsymbol{O}_2$. When the two sets $\{\boldsymbol{O}_1,\boldsymbol{O}_2\}$ are singletons, this relation

reduces to the binary comparison of strict preference between options. Because our (1995) theory leads to a representation in terms of sets of probabilities and almost-state-independent utilities, that feature is inherited by our representation in Section 3.

The use of a coherent choice function coincides with Levi's (1980) principle of *E*-admissibility in cases where the set $\boldsymbol{S}$ is a cross-product of a convex set of probabilities and a convex set of utilities: $\boldsymbol{S} = \boldsymbol{P}\times\boldsymbol{U}$ for convex sets $\boldsymbol{P}$ and **U**. Also, we find that Savage [1954, pp. 123-124, particularly where he argues that option *b* is "superfluous" for the decision pictured by his Figure 1] endorses a coherent choice rule with $\boldsymbol{S}$ a convex set of probabilities and a common utility. The following example, which we repeat from our ISIPTA-03 paper, illustrates how coherent choice does not reduce to binary comparisons in a setting where only probability is indeterminate.

Example 1: Consider a binary decision problem, $\Omega = \{\omega_1, \omega_2\}$ with three feasible options $\boldsymbol{O} = \{f,g,h\}$, and where utility is determinate: $\boldsymbol{u}(f(\omega_1)) = \boldsymbol{u}(g(\omega_2)) = 0.0$, $\boldsymbol{u}(f(\omega_2)) = \boldsymbol{u}(g(\omega_1)) = 1.0$, and $\boldsymbol{u}(h(\omega_1)) = \boldsymbol{u}(h(\omega_2)) = 0.4$. Let uncertainty over the states be indeterminate, with $\boldsymbol{P} = \{\boldsymbol{p}: 0.25 \le \boldsymbol{p}(\omega_2) \le .75\}$. We rehearse three decision rules for this problem.

Γ-Maximin – Maximize minimum expected utility over the feasible options. This rule is well studied in Gilboa and Schmeidler (1989). In brief, *Γ-Maximin* induces a preference ordering over options, but fails the von Neumann-Morgenstern *Independence* postulate. Under *Γ-Maximin* only $\{h\}$ is admissible from the set $\{f,g,h\}$.

Maximality (Sen/Walley) – admissible options are those that are undominated in expectations (over $\boldsymbol{p} \in \boldsymbol{P}$) by any single alternative option. Under *Maximality* all three options are admissible from the set $\{f,g,h\}$ as none dominates the others in pairwise comparisons. *Maximality* does not induce a preference ordering over options; however, admissibility is given by pairwise comparisons. As is evident from Example 1, whether an option (e.g., option *h*) is admissible under *Maximality* depends upon whether the feasible options are closed under mixtures.

Coherent choice. Since the set of probabilities $\boldsymbol{P}$ is convex in this example, coherent choice reduces to Levi's rule of. *E*-admissibility – admissible choices have Bayes' models, i.e., they maximize expected utility for some probability in the (convex) set $\boldsymbol{P}$. Subset $\{f,g\}$ identifies the Bayes-admissible options from $\{f,g,h\}$ under *Coherent Choice*. This rule does not induce an ordering over options and does not reduce to pairwise comparisons.

Note that *h*, which is never "Bayes" with respect to $\boldsymbol{P}$, is uniformly dominated by some mixtures of *f* and *g*, e.g., the mixed option given by $.5f \oplus .5g$, with expected utility 0.5 independent of $\boldsymbol{p}$, uniformly dominates *h*. This is no coincidence, as the following result establishes.

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a finite partition of states. Let $\boldsymbol{O} = \{o_1, o_2, \ldots, o_m\}$ be a finite set of options defined on Ω, such that for $\boldsymbol{o_i} \in \boldsymbol{O}$, $\boldsymbol{u}(o_i(\omega_j)) = \boldsymbol{u_{ij}}$, a cardinal utility of the consequence of o_i when state ω_j obtains. Let $\boldsymbol{P}$ be the class of all probability distributions over Ω. Similarly, let $\boldsymbol{Q}$ be the class of all (simple) mixed acts over $\boldsymbol{O}$, with a mixed act denoted *q*.

Theorem 1 (Pearce, 1984, p.1048).
Suppose for each $\boldsymbol{p} \in \boldsymbol{P}$, act $o^* \in \boldsymbol{O}$ fails to maximize expected utility. Then there is a mixed alternative q^* that uniformly, strictly dominates o^*. That is, $\boldsymbol{u}(q^*(\omega_j)) > \boldsymbol{u}(o^*(\omega_j)) + \varepsilon$, for $j = 1, ..., n$, with $\varepsilon > 0$.

Aside: With this result we are able to apply the strict standard of de Finetti's "incoherence" (= uniform, strict dominance) to a broad class of decisions under uncertainty, analogous to traditional Complete Class Theorems for Bayes decisions (Wald, 1950). The standard of incoherence used here is notably stronger than the mere *inadmissibility* (= weak dominance) of non-Bayes decisions, as is used in those Complete Class theorems.

Let $\boldsymbol{H}(\boldsymbol{O})$ denote the result of taking the closed, convex hull of the option set $\boldsymbol{O}$. That is, $\boldsymbol{H}(\boldsymbol{O})$ is the set of all (simple) mixed acts based on $\boldsymbol{O}$. Since $\boldsymbol{O}$ is finite, q^* of Theorem 1 may be taken to be an option that also is Bayes for some $p^* \in \boldsymbol{P}$. That is, in Theorem 1 we may choose $q^* \in \boldsymbol{H}(\boldsymbol{O})$ such that $q^* \in \boldsymbol{C}(\boldsymbol{H}(\boldsymbol{O}))$ for a coherent choice function using the set $\boldsymbol{P}$ of all probability distributions on Ω.

Aside: Theorem 1 generalizes to infinite states spaces Ω and infinite, closed options sets $\boldsymbol{O}$ by using Theorem 2.1 of Kindler (1983) to replace Pearce's use of von Neumann's Minimax Theorem, which does not generalize to infinite games.

In terms of Theorem 1, in Example 1 with $o^* = h$, then $q_x^* = xf \oplus (1\text{-}x)g$ for $.4 < x < .6$ uniformly dominates o^*. But each such q_x^* is Bayes with respect to $\boldsymbol{H}(\boldsymbol{O})$ for precisely for one probability on Ω: $\boldsymbol{p}(\omega_1) = .5$. We use this fact, next, to establish that each set of probabilities has its own *unique* coherent choice function.

2. Distinguishing sets of probabilities by their coherent choice functions

Consider a finite state space $\Omega = \{\omega_1, \ldots, \omega_n\}$ with the class of all options given by horse lotteries (Anscombe and Aumann, 1963) defined on two consequences **1** and

0. In general horse-lottery theory there is a denumerable set of prizes, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots\}$. A (simple) *horse lottery* is a function from states to (simple) probability distributions over the set of prizes. In this section we use decision problems involving horse lotteries defined on only two consequences, **0** and **1**, with a strict preference for the constant act **1** over the constant act **0**, as explained below. And we consider coherent choice using a state-independent utility, $\boldsymbol{u}$ where $\boldsymbol{u}(\mathbf{1}) = 1$ and $\boldsymbol{u}(\mathbf{0}) = 0$ in each state, ω. Our goal is to show that if $\boldsymbol{P}$ and $\boldsymbol{P}'$ are different sets of probabilities, the coherent choice function based on $\boldsymbol{P}\times\{\boldsymbol{u}\}$ is different from the coherent choice function based on $\boldsymbol{P}'\times\{\boldsymbol{u}\}$.

Let $\boldsymbol{P}$ be a set of probabilities. For a (closed) set $\boldsymbol{O}$, $\boldsymbol{C}(\boldsymbol{O})$ is the non-empty set of Bayes-admissible options. Let $\boldsymbol{R}(\boldsymbol{O}) = \boldsymbol{O}\backslash\boldsymbol{C}(\boldsymbol{O})$ be the associated *Bayes-rejection function* that identifies the *inadmissible* options in $\boldsymbol{O}$. So, we assume that $\{\mathbf{0}\} = \boldsymbol{R}\{\mathbf{0}, \mathbf{1}\}$.

Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ be a probability distribution on Ω .
Let $\underline{p}$ be the smallest nonzero coordinate of $\boldsymbol{p}$.
Define the *constant* horse lottery act $\boldsymbol{a} = \underline{p}\mathbf{1} + (1-\underline{p})\mathbf{0}$.
For each $j = 1, .., n$, *define* the act h_j by

$$\begin{aligned} h_j(\omega_i) &= 1 && \text{if } i = j \text{ and } p_j = 0, \\ &= a && \text{if } i \neq j \text{ and } p_j = 0, \\ &= (\underline{p}/p_j)\mathbf{1} + (1-\underline{p}/p_j)\mathbf{0} && \text{if } i = j \text{ and } p_j > 0, \\ &= 0 && \text{if } i \neq j \text{ and } p_j > 0. \end{aligned}$$

Define the option set $\boldsymbol{O}_p = \{\boldsymbol{a}, h_1, \ldots, h_n\}$.
Theorem 2*: $\boldsymbol{p} \in \boldsymbol{P}$ if and only if $\boldsymbol{R}(\boldsymbol{O}_p) = \varnothing$.*

Proof: First, note that for all j and every utility $\boldsymbol{u}$, $\mathbf{E}_p(\boldsymbol{u}(h_j)) = \underline{p} = \mathbf{E}_p(\boldsymbol{u}(\boldsymbol{a}))$. For the "only if" direction, assume that $(\boldsymbol{p}; \boldsymbol{u}) \in \boldsymbol{S}$ for some utility $\boldsymbol{u}$. Then by this equality, every element of $\boldsymbol{O}_p$ is Bayes with respect to $(\boldsymbol{p}; \boldsymbol{u})$ and $R(\boldsymbol{O}_p) = \phi$. For the "if" direction, assume that $\boldsymbol{R}(\boldsymbol{O}_p) = \phi$. Notice that $\mathbf{E}_q(\boldsymbol{v}(\boldsymbol{a})) = \underline{p}$ for every probability /utility pair $(\boldsymbol{q}, \boldsymbol{v})$. Let $(\boldsymbol{q}, \boldsymbol{v})$ be a probability/utility pair with $\boldsymbol{q} \neq \boldsymbol{p}$. First, consider the case with $\underline{p} < 1$. Then there exists j with $\boldsymbol{q}_j > \boldsymbol{p}_j$. So,

$$\mathbf{E}_q(\boldsymbol{v}(h_j)) = \begin{cases} q_j\,\underline{p}\,/\,p_j > \underline{p} & \text{if } p_j > 0, \\ q_j + (1-q_j)\underline{p} > \underline{p} & \text{if } p_j = 0. \end{cases}$$

Hence, for each $(\boldsymbol{q}, \boldsymbol{v})$ with $\boldsymbol{q} \neq \boldsymbol{p}$, $\mathbf{E}_q(\boldsymbol{v}(h_j)) > \mathbf{E}_q(\boldsymbol{v}(\boldsymbol{a}))$. It follows that $\boldsymbol{a} \in \boldsymbol{R}(\boldsymbol{O}_p)$ unless $(\boldsymbol{p}, \boldsymbol{u}) \in \boldsymbol{S}$ for some utility $\boldsymbol{u}$. Finally, consider the case with $\underline{p} = 1$. In this case, $\boldsymbol{O}_p = \{\mathbf{1}, h_j\}$ where $p_j = 1$. So, $\mathbf{E}_q(\boldsymbol{v}(h_j)) = q_j < 1 = \mathbf{E}_q(\boldsymbol{v}(\boldsymbol{a}))$ for every probability/utility pair $(\boldsymbol{q}, \boldsymbol{v})$ with $\boldsymbol{q} \neq \boldsymbol{p}$. It follows that $h_j \in \boldsymbol{R}(\boldsymbol{O}_p)$ unless $(\boldsymbol{p}, \boldsymbol{u}) \in \boldsymbol{S}$ for some utility $\boldsymbol{u}$.◊

Corollary Let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be two distinct (nonempty) sets of probabilities with corresponding Bayes rejection functions $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. There exists a finite option set $\boldsymbol{O}_p$, as above, such that $\boldsymbol{R}_1(\boldsymbol{O}_p) \neq \boldsymbol{R}_2(\boldsymbol{O}_p)$.

Thus, each set of probabilities $\boldsymbol{P}$ has its own distinct pattern of Bayes rejection functions with respect to option sets $\boldsymbol{O}_p$ for $\boldsymbol{p} \in \boldsymbol{P}$.

Aside: This is a generalization of Theorem 1 that appears at the end of our (2004) paper. That Theorem 1 is the restriction of the corollary to pairs of convex sets of probabilities.

3. Axiomatizing coherent choice functions

We turn, next, to a system of axioms for choice functions that are necessary for coherence, and which are jointly sufficient for a representation of choice by a set $\boldsymbol{S}$ of probability/almost-state-independent utility pairs, as explained below. We provide sufficient conditions when these pairs have a common state-independent utility. In such a case the coherent choice function corresponds to choice under indeterminate uncertainty with a determinate utility.

We continue within the framework of the previous section: horse lotteries over a finite state space $\Omega = \{\omega_1, \ldots, \omega_n\}$. In that we are using choice functions over sets of options, the theory presented here extends our (1995) work, which deals solely with binary choice problems. Thus, results that follow from binary choice problems are available also within this theory. For example, it follows from Section II.6 of our (1995) theory that each agreeing cardinal utility for the choice function $\boldsymbol{C}(\bullet)$, if one exists, is a bounded utility function.

Aside: The aspects of the theory given here that compel the use of almost-state-independent utilities parallel the same issues that arise in Section IV of our (1995) representation for partially ordered preferences. In the context of this paper, that theory, which addresses binary choice only, can be taken to axiomatize choice under the Maximality rule.

In this paper, we focus on a representation for choice when utility is determinate, i.e., regarding the two distinguished prizes **1** and **0**, the constant act **1** is better than, and the constant act **0** is worse than, all other constant acts. Also, we assume that all cardinal utilities are scaled so that $\mathbf{u}(\mathbf{1}) = 1$ and $\boldsymbol{u}(\mathbf{0}) = 0$.

Given a strict preference between these two prizes, the Anscombe-Aumann (1963) theory of horse-lotteries is given by four substantive axioms, which we summarize as follows.
A-A Axiom 1: Choice over horse lotteries reduces to a pairwise comparison of options since binary preference satisfies *ordering*.
A-A Axiom 2: Preference satisfies the von Neumann-Morgenstern postulate of *Independence*.
A-A Axiom 3: An Archimedean postulate is introduced in order to assure that preference has a real-valued

representation, thus insuring also a real-valued representation for subjective probability over Ω and a real-valued cardinal utility over prizes.
A-A Axiom 4: To insure existence of a state-independent utility representation for preference over horse lotteries, a final axiom requires that the decision maker's preference for constant horse lotteries reproduces under each non-null state in the form of called-off horse lotteries.

We adapt our presentation here to match these four axioms.

Axiom 1a (Sen's property *alpha*):
If $\boldsymbol{O}_2 \subseteq \boldsymbol{R}(\boldsymbol{O}_1)$ and $\boldsymbol{O}_1 \subseteq \boldsymbol{O}_3$, then $\boldsymbol{O}_2 \subseteq \boldsymbol{R}(\boldsymbol{O}_3)$.
You cannot promote an unacceptable option into an acceptable option by adding to the choice set of options.

Axiom 1b (a variant of Aizerman's 1985 condition):
If $\boldsymbol{O}_2 \subseteq \boldsymbol{R}(\boldsymbol{O}_1)$ and $\boldsymbol{O}_3 \subseteq \boldsymbol{O}_2$, then $\boldsymbol{O}_2 \backslash \boldsymbol{O}_3 \subseteq \boldsymbol{R}(\textit{closure}[\boldsymbol{O}_1 \backslash \boldsymbol{O}_3])$.
You cannot promote an unacceptable option into an acceptable option by deleting unacceptable options from the option set. (We require $\textit{closure}[\boldsymbol{O}_1 \backslash \boldsymbol{O}_3]$ since $\boldsymbol{O}_1 \backslash \boldsymbol{O}_3$ may not be closed, despite closure of $\boldsymbol{O}_1$ and of $\boldsymbol{O}_3$.)

With Axioms 1a and 1b, define a strict partial order $\langle$ on sets of options as follows. Let $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ be two option sets.

Defn: $\boldsymbol{O}_1 \langle \boldsymbol{O}_2$ *if and only if* $\boldsymbol{O}_1 \subseteq \boldsymbol{R}[\boldsymbol{O}_1 \cup \boldsymbol{O}_2]$.

Lemma 1 of our (2004) establishes that given Axioms 1a and 1b, the binary relation $\langle$ is a strict partial order over pairs of sets of options: $\langle$ is *transitive* and *anti-symmetric*.

The role of mixtures is captured in the following pair of axioms for $\langle$. With $\boldsymbol{O}_1$ an option set and o an option, the notation $\alpha\boldsymbol{O}_1 \oplus (1-\alpha)o$ denotes the set of pointwise mixtures, $\alpha o_1 \oplus (1-\alpha)o$ for $o_1 \in \boldsymbol{O}_1$.

Axiom 2a Independence is formulated for the relation $\langle$ over sets of options. Let o be an option and $0 < \alpha \leq 1$.
$\boldsymbol{O}_1 \langle \boldsymbol{O}_2$ *if and only if* $\alpha\boldsymbol{O}_1 \oplus (1-\alpha)o \langle \alpha\boldsymbol{O}_2 \oplus (1-\alpha)o$.
Axiom 2b Mixtures If $o \in \boldsymbol{O}$ and $o \in \boldsymbol{R}[\boldsymbol{H}(\boldsymbol{O})]$, then $o \in \boldsymbol{R}[\boldsymbol{O}]$.
Axiom 2b asserts that unacceptable options from a mixed set remain so even before mixing.

Aside: *Independence* (Axiom 2a) fails in *Γ-Maximin* theory. *Mixing* (Axiom 2b) fails for the choice function determined by *Maximality*.

The Archimedean condition requires a technical adjustment, as the canonical form used by, e.g. von Neumann-Morgenstern theory or Anscombe-Aumann theory is too restrictive in this setting. (See section II.4 of our 1995.) The reformulated version of the Archimedean condition is as a continuity principle compatible with strict preference as a strict partial order. It reads as follows.

Let $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ ($n = 1, \ldots$) be sets of options converging pointwise, respectively, to the option sets $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\boldsymbol{N}$ be an option set.
Axiom 3a: If, for each n, $\boldsymbol{B}_n \langle \boldsymbol{A}_n$ and $\boldsymbol{A} \langle \boldsymbol{N}$, then $\boldsymbol{B} \langle \boldsymbol{N}$.
Axiom 3b: If, for each n, $\boldsymbol{B}_n \langle \boldsymbol{A}_n$ and $\boldsymbol{N} \langle \boldsymbol{B}$, then $\boldsymbol{N} \langle \boldsymbol{A}$.

State-neutrality / dominance is captured by the following relations. Consider horse lotteries $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, with $\boldsymbol{h}_i(\omega_j) = \beta_{ij}\mathbf{1} \oplus (1-\beta_{ij})\mathbf{0}$; $i = 1,2$ $j = 1, \ldots, n$.

Definition: $\boldsymbol{h}_2$ ***weakly dominates*** $\boldsymbol{h}_1$ if $\beta_{2j} \geq \beta_{1j}$ for $j = 1, \ldots, n$.

Axiom 4: Assume that o_2 weakly dominates o_1, and that a is an option different from each of these two.
4a: If $o_2 \in \boldsymbol{O}$ and $a \in \boldsymbol{R}(\{o_1\} \cup \boldsymbol{O})$ then $a \in \boldsymbol{R}(\boldsymbol{O})$.
4b: If $o_1 \in \boldsymbol{O}$ and $a \in \boldsymbol{R}(\boldsymbol{O})$ then $a \in \boldsymbol{R}(\{o_2\} \cup \boldsymbol{O} \backslash o_1)$.

In words, Axiom *4a* says that when a weakly dominated option is removed from the set of options, other inadmissible options remain inadmissible. So, by Axiom 1, when an option is replaced in the option set by one that it weakly dominates, other admissible options remain admissible.

Axiom *4b* says that when an option is replaced by one that weakly dominates it, (other) inadmissible options remain inadmissible. Trivially by Axiom 1, merely adding a weakly dominating option cannot promote an inadmissible option into one that is admissible.

Axiom 4 captures key aspects of what Savage's postulate **P3** asserts about state-independent utility of the prizes **1** and **0** without assuming states are not-null. That is, the intended representation for the choice function $\boldsymbol{C}(\bullet)$ is by the expected utility rule applied with a set of probability distributions $\boldsymbol{P}$. However, it may be that for each state ω_j there is a probability distribution $\boldsymbol{p}_j \in \boldsymbol{P}$ such that $\boldsymbol{P}_j(\omega_j) = 0$. In the language of our (1995) paper, each state in Ω is potentially null under $\boldsymbol{P}$. Then Savage's **P3** (or the corresponding axiom of Anscombe-Aumann theory) is vacuous. Nonetheless, Axiom 4 reports two facts about weakly dominated lotteries that obtain even when each state is potentially null.

Theorem 3: Axioms 1–4 are necessary for a coherent choice function.
That is, let $\boldsymbol{S}$ be a non-empty set of pairs of probability/state-independent utilities, and let $\boldsymbol{C}_{\boldsymbol{S}}(\bullet)$ be the coherent choice function defined by setting the admissible options in feasible set $\boldsymbol{O}$ to be exactly those that are Bayes-admissible with respect to $\boldsymbol{S}$. Then $\boldsymbol{C}_{\boldsymbol{S}}(\bullet)$ satisfies Axioms 1–4.

Proof: The argument for the necessity of Axioms 1–3 is given in our (2004). That Axiom 4 is necessary as well follows immediately by noting that whenever o_2 weakly dominates o_1 then for each $(\boldsymbol{p},\boldsymbol{u}) \in \boldsymbol{S}$, $\mathbf{E}_{\boldsymbol{p}}(\boldsymbol{u}(o_2)) \geq \mathbf{E}_{\boldsymbol{p}}(\boldsymbol{u}(o_1))$. ◊

The following result is helpful in linking our theory with Theorem 1.

Definition: $\boldsymbol{h}_2$ ***strongly dominates*** $\boldsymbol{h}_1$ if $\beta_{2j} > \beta_{1j}$ for $j = 1, \ldots, n$.

Lemma–Inadmissibility of strongly dominated options: If h_2 strongly dominates h_1 then $\{h_1\} = \boldsymbol{R}(\{h_1, h_2\})$.

Proof: The strategy of the proof is as follows: Use the Independence axiom to convert the problem with option set $\boldsymbol{O} = \{h_1, h_2\}$ into an equivalent problem $\boldsymbol{O}` = \{h`_1, h`_2\}$, where $h`_1$ is a constant horse lottery, and where $h`_2$ strongly dominates $h`_1$. Then we show that $h`_2$ weakly dominates another constant horse lottery, $h``_2$ which also strongly dominates $h`_1$. Then, by Independence $\{h`_1\} = \boldsymbol{R}(\{h`_1, h``_2\})$ and by Axiom *4b*, $\{h`_1\} = \boldsymbol{R}(\{h`_1, h`_2\})$ Last, by Independence, $\{h_1\} = \boldsymbol{R}(\{h_1, h_2\})$.

Here are the details. Let $0 \leq \beta_* = min\{\beta_{1j}\}$ and $1 > \beta^* = max\{\beta_{1j}\}$. Let $h_3(\omega_j) = \beta_{3j}\mathbf{1} \oplus (1-\beta_{3j})\mathbf{0}$, where $\beta_{3j} = \beta^* + \beta_* - \beta_{1j}$. Then the horse lottery $h`_1 = .5h_1 \oplus .5h_3$ is the constant (von Neumann-Morgenstern) lottery with $\beta`_{1j} = (\beta^*+\beta_*)/2$. Define $h`_2 = .5h_2 \oplus .5h_3$. The Independence axiom asserts that $\{h_1\} = \boldsymbol{R}(\{h_1, h_2\})$ if and only if $\{h_1`\} = \boldsymbol{R}(\{h_1`, h_2`\})$. But $h_2`$ strongly dominates $h_1`$, because h_2 strongly dominates h_1. In fact, $\beta`_{2j} - \beta`_{1j} = (\beta_{2j} - \beta_{1j})/2 > 0$. So, let $0 < \delta = min\{\beta_{2j} - \beta_{1j}\}$, and then $\delta/2 = min\{\beta`_{2j} - \beta`_{1j}\}$. Let $h``_2$ be the constant (von Neumann-Morgenstern) lottery defined with $\beta``_{2j} = \beta`_{1j} + \delta/2 = (\beta^*+\beta_* + \delta)/2 > \beta`_{1j}$. Observe, also, that $h`_2$ weakly dominates $h``_2$.

Then, as announced before, by Independence $\{h`_1\} = \boldsymbol{R}(\{h`_1, h``_2\})$ and by Axiom *4b*, $\{h`_1\} = \boldsymbol{R}(\{h`_1, h`_2\})$, and by another application of Independence, $\{h_1\} = \boldsymbol{R}(\{h_1, h_2\})$. ◊

Next we introduce two concepts central to our argument for representing coherent choice functions.

Definitions: The pair $(\boldsymbol{p},\boldsymbol{u})$ is *a <u>local Bayes model for option</u> o* provided that $\boldsymbol{o}$ maximizes $(\boldsymbol{p},\boldsymbol{u})$-expected utility with respect to the options in set $\boldsymbol{O}$.

The pair $(\boldsymbol{p},\boldsymbol{u})$ is *a <u>global Bayes model for the choice function</u>* $\boldsymbol{C}(\bullet)$ provided that, for each option set $\boldsymbol{O}$, if $\boldsymbol{o} \in \boldsymbol{O}$ maximizes $(\boldsymbol{p},\boldsymbol{u})$-expected utility with respect to the options in set $\boldsymbol{O}$ then $\boldsymbol{o} \in \boldsymbol{C}(\boldsymbol{O})$.

We adapt the concept of a set of *almost state-independent utilities*, presented in our (1995, Definition 31), as follows. Let $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$ be a set of rewards and assume that for each constant horse lottery $\boldsymbol{r} \in \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$, $\{\mathbf{0}\} \langle \{\boldsymbol{r}\} \langle \{\mathbf{1}\}$, so that the constant acts $\mathbf{0}$ and $\mathbf{1}$ strictly bound the value of the other constant acts.

The set of probability/utility pairs $\boldsymbol{S}^{\#} = \{(\boldsymbol{p}_j, \boldsymbol{u}_j): j = 1, \ldots\}$ form a set of *<u>almost state independent utilities</u>* for $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$ provided that for each $\varepsilon > 0$, there is a pair $(\boldsymbol{p}_\varepsilon, \boldsymbol{u}_\varepsilon) \in \boldsymbol{S}^{\#}$ and a set of states $\Omega_{(1-\varepsilon)} \subseteq \Omega$ with $\boldsymbol{p}_\varepsilon(\Omega_{(1-\varepsilon)}) \geq 1-\varepsilon$ such that for each $\boldsymbol{r} \in \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$

$$\boldsymbol{max}_{\omega_i,\omega_j \in \Omega_{1-\varepsilon}} | \boldsymbol{u}_{\varepsilon,\omega i}(\boldsymbol{r}) - \boldsymbol{u}_{\varepsilon,\omega j}(\boldsymbol{r}) | \leq \varepsilon.$$

The remaining theorem we seek is this one.

Theorem 4: A choice function $\boldsymbol{C}(\bullet)$ defined on the class $\mathcal{H}$ of simple Anscombe-Aumann Horse-lotteries using (at least) three prizes $\{\mathbf{0}, \boldsymbol{r}, \mathbf{1}\}$, with $\{\mathbf{0}\} \langle \{\boldsymbol{r}\} \langle \{\mathbf{1}\}$, satisfies our 4 axioms only if it is represented by a set $\boldsymbol{S}$ of probability/almost-state-independent utility pairs.

A sufficient condition is given at the end of Appendix 2 for the global Bayes models of $\boldsymbol{S}$ to be comprised solely of probability/state-independent utility pairs.

This theorem follows from three lemmas, described next.

Lemma 1: For each choice set $\boldsymbol{O}$ and admissible option $o \in \boldsymbol{C}(\boldsymbol{O})$, o has at least one local Bayes model.

Proof: By Theorem 1, an option lacking a local Bayes model is strongly dominated by a finite mixture of other options already available in the same choice problem. Then, Axiom 3 and the Lemma on inadmissibility of strongly dominated acts demonstrates Lemma 1.

Aside: Let $o \in \boldsymbol{C}(\boldsymbol{O})$. If $(\boldsymbol{p},\boldsymbol{u})$ and $(\boldsymbol{p}',\boldsymbol{u})$ both are local Bayes models for o, then so too is each pair $(\boldsymbol{q},\boldsymbol{u})$ of the form $\boldsymbol{q} = x\boldsymbol{p} + (1-x)\boldsymbol{p}'$ $(0 \leq x \leq 1)$. Likewise, if each of $(\boldsymbol{p}_j,\boldsymbol{u})$ $(j = 1, \ldots)$ is a local Bayes model for o and the sequences of distributions $\{\boldsymbol{p}_j\}$ converges to distribution $\boldsymbol{q}$, then also $(\boldsymbol{q},\boldsymbol{u})$ is a local Bayes model for o. Hence, we have the following corollary

Corollary: The set of local Bayes models for $o \in \boldsymbol{C}(\boldsymbol{O})$ with a common utility $\boldsymbol{u}$ is given by a non-empty, closed, convex set of probabilities.

Next, following the ideas presented in Section 2, given a distribution $\boldsymbol{p}$, we identify a special choice problem $\boldsymbol{O}^*$ so that precisely when all of its options are admissible, then $\boldsymbol{p}$ is a global Bayes model for the choice function. Thus, the notation for the special choice problem should be '$\boldsymbol{O}^*_{\boldsymbol{p}}$' with the subscript identifying the distribution $\boldsymbol{p}$. To make the proofs readable, that subscript is suppressed here.

Lemma 2: Suppose that $\boldsymbol{C}(\boldsymbol{O}^*) = \boldsymbol{O}^*$. Then $\boldsymbol{p}$ is a global Bayes model for the choice function $\boldsymbol{C}(\bullet)$.
Proof: See Appendix 1.

Lemma 3: For each admissible option $o \in \boldsymbol{C}(\boldsymbol{O})$ at least one of its local Bayes models is a global Bayes model or else there is a set of probability/almost-state-independent utility pairs that represent $\boldsymbol{C}$.
Proof: See Appendix 2.

4. An example of coherent choice using a non-convex set *P* reflecting "expert" opinion

In this section we illustrate how coherent choices may represent "expert" opinions while preserving independence between two events.

Example 2: Consider a decision problem among three options – three treatment plans $\{T_1, T_2, T_3\}$ defined over 4 states $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with determinate utility outcomes given in the following table. That is, the numbers in the table are the utility outcomes for the options (rows) in the respective states (columns)

	ω_1	ω_2	ω_3	ω_4
T_1	0.00	0.00	1.00	1.00
T_2	1.00	1.00	0.00	0.00
T_3	0.99	-0.01	-0.01	0.99

Let a convex set $\boldsymbol{P}$ of probabilities be generated by two extreme points, distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, defined by the following table. Distribution $\boldsymbol{p}_3$ is the .50-.50 (convex) mixture of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$.

	ω_1	ω_2	ω_3	ω_4
$\boldsymbol{p}_1$	0.08	0.32	0.12	0.48
$\boldsymbol{p}_2$	0.48	0.12	0.32	0.08
$\boldsymbol{p}_3$	0.28	0.22	0.22	0.28

Note that (for $i = 1, 2, 3$) under probability $\boldsymbol{p}_i$, only option T_i is Bayes-admissible from the option set of $\{T_1, T_2, T_3\}$.

Without convexity – that is, using only the set comprised by the two (extreme) distributions $\{\boldsymbol{p}_1, \boldsymbol{p}_2\}$ – option T_3 is the sole Bayes-inadmissible option from among the three options $\{T_1, T_2, T_3\}$.

Now, interpret these states as the cross product of two binary partitions: a medical event $\boldsymbol{A}$ (patient allergic) and its complementary event $\boldsymbol{NA}$ (patient not-allergic), with a binary meteorological partition. $\boldsymbol{S}$ (sunny) and $\boldsymbol{NS}$ (cloudy). Specifically: $\omega_1 = \boldsymbol{A\&S}$ $\omega_2 = \boldsymbol{A\&NS}$ $\omega_3 = \boldsymbol{NA\&S}$ $\omega_4 = \boldsymbol{NA\&NS}$.

Under probability distribution $\boldsymbol{p}_1$, the two partitions are independent events with $\boldsymbol{p}_1(\boldsymbol{A}) = .4$ and $\boldsymbol{p}_1(\boldsymbol{S}) = .2$. Likewise, under probability distribution $\boldsymbol{p}_2$, the two partitions are independent events with $\boldsymbol{p}_2(\boldsymbol{A}) = .6$ and $\boldsymbol{p}_2(\boldsymbol{S}) = .8$. And under distribution $\boldsymbol{p}_3$ the events $\boldsymbol{A}$ and $\boldsymbol{S}$ are positively correlated: $.56 = \boldsymbol{p}_3(\boldsymbol{A} \mid \boldsymbol{S}) > \boldsymbol{p}_3(\boldsymbol{A}) = .5$, as happens with each distribution $\boldsymbol{q}$ that is a non-trivial mixture of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$.

Continuing with the example, we see that the three options have the following interpretations: T_1 and T_2 are ordinary medical options for how to treat the patient, with outcomes that depend solely upon the patient's allergic state. T_3 is an option that makes the allocation of medical treatment a function of the meteorological state, with a "fee" of 0.01 utile assessed for that input. That is, T_3 is the option "T_1 if cloudy and T_2 if sunny, while paying a fee of 0.01."

Suppose that $\boldsymbol{p}_1$ represents the opinion of medical expert 1, and $\boldsymbol{p}_2$ represents the opinion of medical expert 2. Without convexity of the credal probabilities, T_3 is inadmissible. This captures the shared agreement between the two medical experts that T_3 is unacceptable from the choice of three $\{T_1, T_2, T_3\}$, and it captures the pre-systematic understanding that under T_3 you pay to use medically irrelevant inputs about the weather in order to determine the medical treatment. However, with convexity of the set generated by $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, then T_3 is admissible as well. Convexity of the set of indeterminate probabilities, we note, is required in each of Gilboa and Schmeidler's (1989) version of *Γ-Maximin*, in Walley's (1990) version of Maximality, and in Levi's (1980) account of *E-admissibility*..

Aside: This illustration relies on the fact that normal and extensive form decisions are generally *not* equivalent in decision theories with indeterminate probabilities. Example 2 is in the normal form, as are all the choice problems considered in this paper. In the extensive form of this decision problem, the decision maker has the opportunity to make a terminal choice between T_1 and T_2 first, or to take as a third option a sequential alternative: pay a fee of 0.01 utiles in order to learn the state of the weather before choosing between T_1 and T_2. Under decision rules for extensive form problems that we endorse, and which we believe also are endorsed by Levi, then it is *E-inadmissible* to postpone the immediate medical decision between T_1 and T_2 in order to pay an amount to acquire the irrelevant meteorological evidence. And this holds whether the indeterminate probability set is convex or not. Related results about independence with indeterminate probability are presented in Cozman and Walley (2005).

5. Concluding Remarks

We have discussed *coherent choice functions* – where the admissible options in a decision problem are exactly

those which maximize expected utility for some probability/utility pair in fixed set $\boldsymbol{S}$ of probability/utility pairs. All of the decision problems used here to characterize and axiomatize coherent choice functions are *normal form* decision problems. But, as indicated in section 4, normal and extensive form decisions generally are not equivalent when probability (or utility) is indeterminate. One of our future projects is to study coherent choice for extensive form, i.e., sequential decision problems

Also, as noted in Lemma 3, in parallel with our findings about coherent strict partial orders (1995) the axioms are sufficient for coherence using a set of probability/almost-state-independent utility pairs. Though we give sufficient conditions when a choice function satisfying our axioms is represented by a set of probability/state-independent utility pairs with a common utility, also we intend to study how to modify the axioms to avoid the use of almost-state-independent utilities.

Acknowledgements

We thank Fabio Cozman and Matthias Troffaes for helpful discussions of topics central to this paper. Seidenfeld's research efforts were supported in part by State of Sao Paulo, Brazil, grant FAPESP #05/59541-4.

Appendix 1 – Lemma 2

Lemma 2: Suppose that $\boldsymbol{C}(\boldsymbol{O}^*) = \boldsymbol{O}^*$. Then $\boldsymbol{p}$ is a global Bayes model for the choice function $\boldsymbol{C}(\bullet)$.

Proof. Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ be a probability distribution on Ω with $\underline{\boldsymbol{p}}$ its smallest nonzero coordinate. $\boldsymbol{O}^*$ is comprised by a set of acts that span all the elements of $\mathcal{H}$ with $\boldsymbol{p}$-Expected utility $\underline{\boldsymbol{p}}$.

Partition the states in Ω in two sets: $\Omega_1^p = \{\omega_1, \ldots, \omega_k\}$ those that comprise the support of $\boldsymbol{p}$ and, $\Omega_2^p = \{\omega_{k+1}, \ldots, \omega_n\}$ those states null under $\boldsymbol{p}$. Clearly, $\Omega_2^p = \phi$ if and only if $\boldsymbol{p}$ has full support. We define $\boldsymbol{O}^*$ by two cases, depending whether $\Omega_2^p = \phi$ or not.

Case 1: $\Omega_2^p = \phi$ and $\boldsymbol{p}$ has full support. $\boldsymbol{O}^*$ is comprised by n-many acts, $\{a_j : j = 1, \ldots, n\}$ For each $\boldsymbol{j} = 1, \ldots, \boldsymbol{n}$, define the act a_j by

$$a_j(\omega_i) = \frac{p}{p_j}\mathbf{1} \oplus (1-\frac{p}{p_j})\mathbf{0} \quad \text{if } i = j$$
$$= \mathbf{0} \quad \text{if } i \neq j.$$

Case 2: $\Omega_2^p \neq \phi$. $\boldsymbol{O}^*$ is defined by $\boldsymbol{k}(\boldsymbol{n}+2-\boldsymbol{k})$-many acts which can be understood to be the product of acts defined on $\Omega_1^p \times \Omega_2^p$. With respect to Ω_1^p, $\boldsymbol{O}^*$ contains k-many acts that span horse lotteries defined on Ω_1^p that have $\boldsymbol{p}$-Expected utility $\underline{\boldsymbol{p}}$, similarly to Case 1. With respect to Ω_2^p, $\boldsymbol{O}^*$ contains $(\boldsymbol{n}+2-\boldsymbol{k})$-many acts that span all horse lotteries defined on Ω_2^p, including the two constants $\mathbf{0}$ and $\mathbf{1}$.

For each $\boldsymbol{j} = 1, \ldots, \boldsymbol{k}$, and $\boldsymbol{m} = \boldsymbol{k}+1, \ldots, \boldsymbol{n}+2$ define the act a_j^m by

$$a_j^m(\omega_i) = \frac{p}{p_j}\mathbf{1} \oplus (1-\frac{p}{p_j})\mathbf{0} \quad \text{if } i = j$$
$$= \mathbf{1} \text{ if } i = m \textbf{ or } (m = n+2 \text{ and } i > k)$$
$$= \mathbf{0} \quad \text{otherwise}$$

Note that $a_j^{n+1}(\omega_i) \neq \mathbf{0}$ if and only if $i = j$. In particular, it equals $\mathbf{0}$ on Ω_2^p. And note that $a_j^{n+2}(\omega_i) \neq \mathbf{0}$ if and only if, either $i = j$ or $i > k$. It equals $\mathbf{1}$ on Ω_2^p.

Let $\boldsymbol{O}^*$ be the choice problem formed by taking the convex hull of these options:

In Case 1 $\boldsymbol{O}^* = \boldsymbol{H}\{a_j : \boldsymbol{j} = 1, \ldots, \boldsymbol{n}\}$, the convex hull of $\boldsymbol{n}$-many options. In Case 2 $\boldsymbol{O}^* = \boldsymbol{H}\{a_j^m : \boldsymbol{j} = 1, \ldots, \boldsymbol{k}; \boldsymbol{m} = \boldsymbol{k}+1, \ldots, \boldsymbol{n}+2\}$, the convex hull of $\boldsymbol{k}(\boldsymbol{n}+2-\boldsymbol{k})$-many options.

Let $\boldsymbol{a_p}$ denote the constant horse lottery that awards the identical von Neumann-Morgenstern lottery in each state, with $\boldsymbol{a_p} = \underline{\boldsymbol{p}}\mathbf{1} \oplus (1-\underline{\boldsymbol{p}})\mathbf{0}$.

Claim 1: $\boldsymbol{a_p} \in \boldsymbol{O}^*$.

Proof: In Case 1, when $\boldsymbol{p}$ has full support, $\boldsymbol{p}_1 a_1 \oplus \boldsymbol{p}_2 a_2 \oplus \ldots \oplus \boldsymbol{p}_n a_n$ is the horse lottery $\boldsymbol{a_p}$. In Case 2, when $\boldsymbol{p}$-null states exist, for each $\boldsymbol{j} = 1, \ldots, \boldsymbol{k}$, define the horse lottery $b_j = (1-\underline{\boldsymbol{p}})\, a_j^{n+1} \oplus \underline{\boldsymbol{p}}\, a_j^{n+2}$ with payoffs:

$$b_j(\omega_i) = \boldsymbol{a_p} \quad \text{if } i > k$$
$$b_j(\omega_i) = \frac{p}{p_j}\mathbf{1} \oplus (1-\frac{p}{p_j})\mathbf{0} \quad \text{if } i = j$$
$$b_j(\omega_i) = \mathbf{0} \quad \text{if } i \neq j \text{ and } i \leq k.$$

Then $\boldsymbol{p}_1 b_1 \oplus \boldsymbol{p}_2 b_2 \oplus \ldots \oplus \boldsymbol{p}_k b_k$ is the horse lottery $\boldsymbol{a_p}$. ◊–claim 1.

Note that $(\boldsymbol{p},\boldsymbol{u})$ is a local Bayes model for each element of $\boldsymbol{O}^*$ as the $\boldsymbol{p}$-Expected utility for each element of $\boldsymbol{O}^*$ is the same value, namely $\underline{\boldsymbol{p}}$.

Claim 2: If $\underline{\boldsymbol{p}} < 1$ then $(\boldsymbol{p},\boldsymbol{u})$ is the only local Bayes model for a_p

***Proof*:** Note that regardless the distribution $\boldsymbol{q}$ on Ω, a_p has $\boldsymbol{q}$-Expected utility $\underline{\boldsymbol{p}}$. We argue by cases that when $\underline{\boldsymbol{p}} < 1$, $\boldsymbol{q}$ is not a local model for $\boldsymbol{a_p}$ with respect to $\boldsymbol{O}^*$.

If $\boldsymbol{p}$ has full support ($\Omega_2^p = \phi$), the $\boldsymbol{q}$-Expected utility of $a_j = \boldsymbol{q_j}\frac{p}{p_j} > \underline{\boldsymbol{p}}$. And if $j = m > k$, so that $\boldsymbol{p_j} = 0$ and $\boldsymbol{q}(\Omega_2^p) > 0$, then the $\boldsymbol{q}$-Expected utility of

$$\boldsymbol{p}_1 a_1^{n+2} \oplus \boldsymbol{p}_2 a_2^{n+2} \oplus \ldots \oplus \boldsymbol{p}_k a_k^{n+2}$$
$$= \boldsymbol{q}(\Omega_1)\underline{\boldsymbol{p}} + \boldsymbol{q}(\Omega_2) > \underline{\boldsymbol{p}}.$$

Hence, $(\boldsymbol{q},\boldsymbol{u})$ is not a local Bayes model for $\boldsymbol{a_p}$. ◊-claim 2

Note also that for the case $\boldsymbol{p_1} = \underline{\boldsymbol{p}} = 1$, $\boldsymbol{a_p} = \boldsymbol{1}$ and then $\boldsymbol{O}^* = \boldsymbol{H}\{\boldsymbol{1}, a_1^2, \ldots a_1^{n+2}\}$. In which case if $\boldsymbol{q} \neq \boldsymbol{p}$, $\boldsymbol{q}$ is not a local Bayes model for a_1^{n+1}, which has a $\boldsymbol{q}$-expected value of $\boldsymbol{q}_1 < 1$. Thus, we have

Proposition:

$\boldsymbol{p}$ is the sole local Bayes model for all of $\boldsymbol{O}^*$.

Claim 3: $\boldsymbol{O}^*$ contains all the horse lotteries in $\mathcal{H}$ with $\boldsymbol{p}$-expected utility equal to $\underline{\boldsymbol{p}}$.

Proof: Let o be a horse lottery with $\boldsymbol{p}$-Expected utility $\underline{\boldsymbol{p}}$. Write $o(\omega_j) = \alpha_j\boldsymbol{1} \oplus (1-\alpha_j)\boldsymbol{0}$, $\boldsymbol{j} = 1, \ldots, \boldsymbol{n}$.

Case 1 ($\boldsymbol{p}$ has full support.): For $\omega_i \in \Omega = \Omega_1^p$ we have that $\sum_i \boldsymbol{p}_i\alpha_i = \underline{\boldsymbol{p}}$ and $0 \leq \alpha_i \leq 1$. The set of α-vectors satisfying these two equations is closed and convex, with extreme points given by the acts $\{\boldsymbol{a_j}\}$. That is, if $\alpha^* = \langle\alpha^*_1, \ldots., \alpha^*_n\rangle$ is an extreme point of this set of α-vectors, then $\alpha^* = a_j$ for some $1 \leq \boldsymbol{j} \leq \boldsymbol{n}$. Since a closed, convex set is identified by its extreme points, this establishes that $\boldsymbol{o} \in \boldsymbol{O}^*$.

Case 2 (There are null states under $\boldsymbol{p}$.): The reasoning is similar to Case 1, noting that $\boldsymbol{O}^*$ spans all horse lotteries defined over Ω_2^p .◊–Claim 3.

We complete the proof of Lemma 2, as follows. Let $\boldsymbol{O}$ be a choice set and let $\phi \neq \boldsymbol{O}_p \subseteq \boldsymbol{O}$ be those options for which $\boldsymbol{p}$ is a local Bayes model. So, each $a \in \boldsymbol{O}_p$ maximizes the $\boldsymbol{p}$-Expected utility of options in $\boldsymbol{O}$ at common value $\mathbf{r}$. There are two cases, depending upon whether $\mathbf{r} \geq \underline{\boldsymbol{p}}$ or $\mathbf{r} < \underline{\boldsymbol{p}}$.

In the former case, mix $\mathbf{0}$ into each act in $\boldsymbol{O}$ to form the choice set $\boldsymbol{O}` = \frac{\underline{\boldsymbol{p}}}{r}\boldsymbol{O} \oplus (1-\frac{\underline{\boldsymbol{p}}}{r})\mathbf{0}$, with the isomorphism between $\boldsymbol{O}$ and $\boldsymbol{O}`$ that associates each $o \in \boldsymbol{O}$ with $o` \in \boldsymbol{O}`$, where $o` = \frac{\underline{\boldsymbol{p}}}{r}o \oplus (1-\frac{\underline{\boldsymbol{p}}}{r})\mathbf{0}$.

In case $\mathbf{r} < \underline{\boldsymbol{p}}$ mix $\mathbf{1}$ into each act in $\boldsymbol{O}$ to form the choice set $\boldsymbol{O}` = \frac{1-\underline{\boldsymbol{p}}}{1-r}\boldsymbol{O} \oplus (\frac{\underline{\boldsymbol{p}}-r}{1-r})\mathbf{1}$, with the isomorphism between $\boldsymbol{O}$ and $\boldsymbol{O}`$ that associates each $o \in \boldsymbol{O}$ with $o` \in \boldsymbol{O}`$, where $o` = \frac{1-\underline{\boldsymbol{p}}}{1-r}o \oplus (\frac{\underline{\boldsymbol{p}}-r}{1-r})\mathbf{1}$.

The argument continues in parallel between the two cases. By the Axiom 2, $a \in \boldsymbol{C}(\boldsymbol{O})$ if and only if $a` \in \boldsymbol{C}(\boldsymbol{O}`)$. Also evident is the fact that for each $a` \in \boldsymbol{O}`_p$ the $\boldsymbol{p}$-Expected utility of $a`$ equals $\underline{\boldsymbol{p}}$. Thus, by Claim 3, for each $a` \in \boldsymbol{O}`_p$, $a` \in \boldsymbol{C}(\boldsymbol{O}^*)$.

Claim 4: Let $o` \in \boldsymbol{O}`$ and $o` \notin \boldsymbol{O}`_p$. Then each local Bayes model $\boldsymbol{q}$ for $o`$ with respect to $\boldsymbol{O}^* \cup \{o`\}$ is singular with respect to $\boldsymbol{p}$, i.e., $\Omega_1^q \cap \Omega_1^p = \phi$.

Proof: Because $o` \notin \boldsymbol{O}`_p$ then $\mathbf{E}_p(o`) < \underline{\boldsymbol{p}}$ and, trivially, $\boldsymbol{p}$ is not a local Bayes model for $o`$. Fix a distribution $\boldsymbol{q} \neq \boldsymbol{p}$ where $\Omega_1^q \cap \Omega_1^p \neq \phi$. We argue indirectly that $\boldsymbol{q}$ is not a local Bayes model for $\boldsymbol{o}$' with respect to $\boldsymbol{O}^* \cup \{o`\}$.

First consider the case where $\Omega_1^q \subseteq \Omega_1^p$, that is where $\boldsymbol{q}$ is absolutely continuous with respect to $\boldsymbol{p}$. Within the $\boldsymbol{n}$-1 dimensional simplex of distributions on Ω, let $\boldsymbol{L_{pq}}$ be the line determined by the two points $\boldsymbol{p}$ and $\boldsymbol{q}$, having endpoints denoted $\boldsymbol{q}_*$ and $\boldsymbol{q}^*$. Identify these endpoints by placing $\boldsymbol{q}$ in the closed line segment $[\boldsymbol{q}^*, \boldsymbol{p}]$, and thus $\boldsymbol{p}$ lies in the closed line segment $[\boldsymbol{q}, \boldsymbol{q}_*]$, from which we know that $\boldsymbol{p} \neq \boldsymbol{q}^*$, though it is possible that $\boldsymbol{q} = \boldsymbol{q}^*$.

Moreover, since $\Omega_2^q \supseteq \Omega_2^p$ we have that $\boldsymbol{p} \neq \boldsymbol{q}_*$, since each endpoint of $\boldsymbol{L_{pq}}$ has some null-state not shared as a null state with any other point on that line. So, $\boldsymbol{p}$ is internal to the line $\boldsymbol{L_{pq}}$. Because $\boldsymbol{q}_*$ is an endpoint of $\boldsymbol{L_{pq}}$, as just argued, $\Omega_2^{q*} \cap \Omega_1^p \neq \phi$. Assume that $\omega_k \in \Omega_2^{q*} \cap \Omega_1^p$. Since $\boldsymbol{p}$ lies on the line $[\boldsymbol{q}_*, \boldsymbol{q}^*]$, $\omega_k \in \Omega_1^{q*}$.

Consider the act a_k^{n+1} (or the act a_k if $\boldsymbol{p}$ has full support). Since $\mathbf{E}_q(\boldsymbol{o}`) \geq \mathbf{E}_q(a_k^{n+1})$ and $\mathbf{E}_p(\boldsymbol{o}`) < \mathbf{E}_p(a_k^{n+1}) = \underline{\boldsymbol{p}}$, there exists a unique distribution $\boldsymbol{r}_k$ situated on the line $\boldsymbol{L_{pq}}$ and between $\boldsymbol{p}$ and $\boldsymbol{q}$ (possibly with $\boldsymbol{r}_k = \boldsymbol{q}$), such that $E_{r_k}(o`) = E_{r_k}(a_k^{n+1})$. Because expected utility is linear in probability, for each distribution $\boldsymbol{t}$ in the half open interval $(\boldsymbol{r}_k, \boldsymbol{q}_*]$, $\mathbf{E}_t(o`) < \mathbf{E}_t(a_k^{n+1})$. But $E_{q_*}[a_k^{n+1}] = 0 > E_{q_*}[o`]$, which is a contradiction as no act has a negative expected value.

This completes the argument when $\boldsymbol{q}$ is absolutely continuous with respect to $\boldsymbol{p}$.

Next, assume that $\Omega_1^q \cap \Omega_1^p \neq \phi$ and write

$$\boldsymbol{q}(\bullet) = \boldsymbol{q}(\bullet \mid \Omega_1^p)\boldsymbol{q}(\Omega_1^p) + \boldsymbol{q}(\bullet \mid \Omega_2^p)\boldsymbol{q}(\Omega_2^p),$$

where $\boldsymbol{q}(\Omega_1^p) > 0$. So, $\mathbf{q}(\bullet|\Omega_1^p)$ is absolutely continuous with respect to $\boldsymbol{p}$.

$\mathbf{E}_q(\bullet) = \mathbf{E}_q(\bullet \mid \Omega_1^p)\boldsymbol{q}(\Omega_1^p) + \mathbf{E}_q(\bullet \mid \Omega_2^p)\boldsymbol{q}(\Omega_2^p)$. Since $a_k^{n+2} \in \boldsymbol{O}^*$ and $\mathbf{E}_{\boldsymbol{q}}(\boldsymbol{o}`) \geq \mathbf{E}_{\boldsymbol{q}}(a_k^{n+2})$, it follows that $\mathbf{E}_q(\boldsymbol{o}` \mid \Omega_1^p) \geq \mathbf{E}_q(a_k^{n+2} \mid \Omega_1^p) = \mathbf{E}_q(a_k^{n+1} \mid \Omega_1^p)$. However, as $\boldsymbol{q}(\bullet \mid \Omega_1^p)$ is absolutely continuous with respect to $\boldsymbol{p}$, we have the same situation involving $\boldsymbol{q}(\bullet \mid \Omega_1^p)$ and $\boldsymbol{p}$ as when $\boldsymbol{q}$ is absolutely continuous with respect to $\boldsymbol{p}$, completing the proof. ◊-Claim 4

Next, we show that if there is a local Bayes model for o**'** with respect to $\boldsymbol{O}^* \cup \{o`\}$, then no element of $\boldsymbol{O}^*$ becomes inadmissible by adding option $\boldsymbol{o}`$.

Claim 5: Assume that $a \in \boldsymbol{C}(\boldsymbol{O}^*)$, $o` \in \boldsymbol{O}`$ but $o` \notin \boldsymbol{O}`_p$, and let $o`$ have a local Bayes model $\boldsymbol{q}$ with respect to $\boldsymbol{O}^* \cup \{o`\}$. Then $a \in \boldsymbol{C}(\boldsymbol{O}^* \cup \{o`\})$.
Proof: Assume the premise. In the light of Axiom 4 we are done proving Claim 5 if we identify an act $a^* \in \boldsymbol{O}^*$ such that a^* weakly dominates o**'**. This we do as follows.

By Claim 4, $\boldsymbol{q}$ is singular with respect to $\boldsymbol{p}$. Consider an act a_k^{n+2} for $\omega_k \in \Omega_1^p$.

Definition: For $\boldsymbol{W} \subseteq \Omega$ and act o, define the act $o|\boldsymbol{W}$ by:
$o(\omega)|\boldsymbol{W} = o(\omega)$, for $\omega \in \boldsymbol{W}$,
and $o(\omega)|\boldsymbol{W} = \mathbf{0}$, otherwise.

Write $o`$ as an sum of three acts $o` = o`|\Omega_1^q + o`|(\Omega_2^p \cap \Omega_2^q) + o`|\Omega_1^p$, and likewise for $a_k^{n+2} = a_k^{n+2}|\Omega_1^q + a_k^{n+2}|(\Omega_2^p \cap \Omega_2^q) + a_k^{n+2}|\Omega_1^p$. Because $a_k^{n+2}(\omega) = \mathbf{1}$ for $\omega \in \Omega_2^p$, then $a_k^{n+2}|\Omega_1^q$ weakly dominates $o`|\Omega_1^q$, and likewise $a_k^{n+2}|(\Omega_2^p \cap \Omega_2^q)$ weakly dominates $o`|(\Omega_2^p \cap \Omega_2^q)$. By Claim 4, $o`|\Omega_1^p$ fails to have a local Bayes model with respect to $\boldsymbol{O}^* \cup \{o`|\Omega_1^p\}$. So, by Lemma 1, there exists an option $b \in \boldsymbol{H}(\boldsymbol{O}^*)$ that uniformly dominates $o`|\Omega_1^p$. Let $a^* = a_k^{n+2}|\Omega_1^q + a_k^{n+2}|(\Omega_2^p \cap \Omega_2^q) + b|\Omega_1^q$. Then a^* weakly dominates $o`$ and, as $\mathbf{E}_{\boldsymbol{p}}[a^*] = \mathbf{E}_{\boldsymbol{p}}[b|\Omega_1^q] = \underline{\boldsymbol{p}}$, we have $a^* \in \boldsymbol{O}^*$. ◊-Claim 5

Assume that $a` \in \boldsymbol{C}(\boldsymbol{O}^*)$. Let $\boldsymbol{N}` = \{o`: o` \in \boldsymbol{O}`$ and $o` \notin \boldsymbol{O}`_{\boldsymbol{p}}$ but $o`$ has no local Bayes model with respect to $\boldsymbol{O}^* \cup \{o`\}\}$. Then by Lemma 1, $o` \in \boldsymbol{R}(\boldsymbol{O}^* \cup \boldsymbol{N}`)$. By Axiom 1, as $a` \in \boldsymbol{C}(\boldsymbol{O}^*)$ then $a` \in \boldsymbol{C}(\boldsymbol{O}^* \cup \boldsymbol{N}`)$. If $o` \in \boldsymbol{O}`\backslash\boldsymbol{N}`$, then using Claim 5, $a` \in \boldsymbol{C}(\boldsymbol{O}^* \cup \boldsymbol{N}` \cup o`)$. By a simple induction on an arbitrary well-ordering of $\boldsymbol{O}`\backslash\boldsymbol{N}`$, then $a` \in \boldsymbol{C}(\boldsymbol{O}^* \cup \boldsymbol{N}` \cup \boldsymbol{O}`\backslash\boldsymbol{N}`) = \boldsymbol{C}(\boldsymbol{O}^* \cup \boldsymbol{O}`)$. By Axiom 1, if $a` \in \boldsymbol{O}`$ then $a` \in \boldsymbol{C}(\boldsymbol{O}`)$. Finally, by Axiom 2, $a \in \boldsymbol{C}(\boldsymbol{O})$. ◊-Lemma 2

Appendix 2 – Lemma 3

Lemma 3: For each admissible option $o \in \boldsymbol{C}(\boldsymbol{O})$ at least one of its local Bayes models is a global Bayes model or else there is a set of probability/almost-state-independent utility pairs that serve as a global Bayes-model.
Proof: The next claim, which we use to establish Lemma 3, extends the idea of Axiom 4 to the strict partial order $\langle$.
Claim 6: Suppose that for option sets $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{D}$, $\boldsymbol{B} \langle \boldsymbol{A}$ and $\boldsymbol{B} \cap \boldsymbol{C}(\boldsymbol{D}) \neq \phi$. Then $\boldsymbol{A} \cap \boldsymbol{C}(closure\{\boldsymbol{D}\backslash\boldsymbol{B} \cup \boldsymbol{A}\}) \neq \phi$.
Proof (indirect): Suppose that $\boldsymbol{A} \subseteq \boldsymbol{R}(closure\{\boldsymbol{D}\backslash\boldsymbol{B} \cup \boldsymbol{A}\})$. By Axiom 1 applied twice, $\boldsymbol{A} \subseteq \boldsymbol{R}(\boldsymbol{D} \cup \boldsymbol{A})$ and $\boldsymbol{A} \subseteq \boldsymbol{R}(\boldsymbol{D} \cup \boldsymbol{A} \cup \boldsymbol{B})$. Since $\boldsymbol{B} \langle \boldsymbol{A}$, likewise $\boldsymbol{B} \subseteq \boldsymbol{R}(\boldsymbol{D} \cup \boldsymbol{A} \cup \boldsymbol{B})$. Thus, $\boldsymbol{A} \cup \boldsymbol{B} \langle \boldsymbol{D}$. By transitivity, $\boldsymbol{B} \langle \boldsymbol{D}$ and so $\boldsymbol{B} \cap \boldsymbol{C}(\boldsymbol{D}) = \phi$. ◊-Claim 6

Given $o \in \boldsymbol{C}(\boldsymbol{O})$ and following the ideas we used in (1995, Definition 19), we introduce the notion of a ***target set*** $\boldsymbol{T}(o,\boldsymbol{O})$ of probability distributions for o with respect to choice problem $\boldsymbol{O}$. The target set for o is a subset of the local Bayes models for o which, we show, contains all of its global Bayes models. We demonstrate that whenever the target set includes a boundary point, that boundary point is a global Bayes model.

Given a probability distribution $\boldsymbol{p}$, recall the decision problem $\boldsymbol{O}_{\boldsymbol{p}} = \{\boldsymbol{a}^p, h_1^p, \ldots, h_1^p\}$ defined in Section 2. We state without proof that whenever $\boldsymbol{C}(\boldsymbol{O}_{\boldsymbol{p}}) = \boldsymbol{O}_{\boldsymbol{p}}$ then $\boldsymbol{C}(\boldsymbol{O}^*) = \boldsymbol{O}^*$ for $\boldsymbol{O}^*$ defined with respect to $\boldsymbol{p}$ as in Lemma 2, and so $\boldsymbol{p}$ is a global Bayes model**.**

Definition: $\boldsymbol{T}(o,\boldsymbol{O}) = \{\boldsymbol{p}: \boldsymbol{p}$ is local Bayes model for o in choice problem $\boldsymbol{O}$ and $\{h_1^p, \ldots, h_n^p\} \subseteq \mathrm{C}(\boldsymbol{O}_{\boldsymbol{p}})\}$

Claim 7: $\boldsymbol{T}(o,\boldsymbol{O})$ is a non-empty, convex set.
Proof: Without loss of generality, and to simplify the presentation, we give the proof for a binary state space $\Omega = \{\omega_1, \omega_2\}$. Convexity is shown as follows. Note that for $\boldsymbol{p}$ defined by $\boldsymbol{p}(\omega_2) = 0$, $h_2^p \in \mathrm{C}(\boldsymbol{O}_{\boldsymbol{p}})$, and for $\boldsymbol{p}$ defined by $\boldsymbol{p}(\omega_2) = 1$, $h_1^p \in \mathrm{C}(\boldsymbol{O}_{\boldsymbol{p}})$. And by Claim 6, if $h_2^p \in \mathrm{C}(\boldsymbol{O}_{\boldsymbol{p}})$, then for all distributions $\boldsymbol{q}$ with $\boldsymbol{q}(\omega_2) \leq \boldsymbol{p}(\omega_2)$ we have $h_2^q \in \mathrm{C}(\boldsymbol{O}_{\boldsymbol{q}})$; and if $h_1^p \in \mathrm{C}(\boldsymbol{O}_{\boldsymbol{p}})$, then for all distributions $\boldsymbol{q}$ with $\boldsymbol{q}(\omega_2) \geq \boldsymbol{p}(\omega_2)$ we have $h_1^q \in$

$C(\boldsymbol{O}_q)$. In the general case, with more than 2 states, the same result follows by noting that $\boldsymbol{T}(o,\boldsymbol{O})$ is an intersection of half-planes. We show that $\boldsymbol{T}(o,\boldsymbol{O})$ is non-empty by an indirect argument using the Archimedean axiom. So, assume that for each $\boldsymbol{p}$, $\boldsymbol{C}\{h_1^p, h_2^p\}$ is a unit set, and by the observation above, let $\boldsymbol{q}$ be the ***lub*** $\{\boldsymbol{p}(\omega_2): h_2^p \in \boldsymbol{C}\{h_1^p, h_2^p\}$. There are two cases.

Case 1: $\{h_2^q\} = \boldsymbol{C}\{h_1^q, h_2^q\}$ So $\boldsymbol{q}(\omega_2) < 1$ and then $h_1^q \prec h_2^q$ and for all $\boldsymbol{p}(\omega_2) > \boldsymbol{q}(\omega_2)$, $h_2^p \prec h_1^p$. But as $\boldsymbol{p}$ approaches $\boldsymbol{q}$, h_i^p converges to h_i^q for $i = 1, 2$. Then by Axiom 3, $h_1^q \prec h_1^q$.

Case 2: $\{h_1^q\} = \boldsymbol{C}\{h_1^q, h_2^q\}$. So $\boldsymbol{q}(\omega_2) > 0$ and then $h_2^q \prec h_1^q$ and for all $\boldsymbol{p}(\omega_2) < \boldsymbol{q}(\omega_2)$, $h_1^p \prec h_2^p$. But as $\boldsymbol{p}$ approaches $\boldsymbol{q}$, h_i^p converges to h_i^q for $i = 1, 2$. Then by Axiom 3, $h_2^q \prec h_2^q$ -◊-Claim 7

To complete the proof of Lemma 3 there are two cases to consider.

Case 1: $\boldsymbol{T}(o,\boldsymbol{O})$ contains at least one of its boundary points. Suppose, e.g., that $\boldsymbol{q}$ is the ***lub*** $\{\boldsymbol{p}(\omega_2): h_2^p \in \boldsymbol{C}\{h_1^p, h_2^p\}$ and that $\boldsymbol{R}\{h_1^q, h_2^q\} = \phi$. Then for each $0 \leq x \leq 1$, $\boldsymbol{R}\{h_1^q, h_2^q, x h_1^q \oplus (1\text{-}x) h_2^q\} = \phi$, as the following reasoning establishes.

Assume that $\boldsymbol{q}(\omega_2) < 1$, or we are done. Then for all $\boldsymbol{p}(\omega_2) > \boldsymbol{q}(\omega_2)$, $h_2^p \prec h_1^p$ as before. For $0 < x \leq 1$, by Axiom 2, $h_2^p \prec x h_1^p \oplus (1\text{-}x) h_2^p$. As $\boldsymbol{p}$ approaches $\boldsymbol{q}$, by Axiom 3, then $x h_1^q \oplus (1\text{-}x) h_2^q \in \boldsymbol{C}\{h_1^q, h_2^q, x h_1^q \oplus (1\text{-}x) h_2^q\}$, on pain of contradiction otherwise that $h_2^q \prec h_2^q$. The reasoning is similar if the target set $\boldsymbol{T}(o,\boldsymbol{O})$ is closed at the other end. Then, at each point $\boldsymbol{p}$ of closure for $\boldsymbol{T}(o,\boldsymbol{O})$, $\boldsymbol{R}(\boldsymbol{O}_p) = \phi$ and $\boldsymbol{p}$ is global Bayes model.

Case 2: If the target set is entirely open and there is no $\boldsymbol{p} \in \boldsymbol{T}(o,\boldsymbol{O})$ such that $\boldsymbol{R}(\boldsymbol{O}_p) = \phi$, we arrive at the parallel situation studied in Section IV.2 of our (1995). That situation is one where, first, a coherent choice function $\boldsymbol{C}$ is induced by a finite set $\boldsymbol{P}$ of linearly independent probabilities on Ω. The convex target sets for $\boldsymbol{C}$ include subsets of $\boldsymbol{P}$ as extreme points, i.e., $\boldsymbol{R}(\boldsymbol{O}_p) = \phi$ for each $\boldsymbol{p} \in \boldsymbol{P}$. Hence, $\boldsymbol{C}$ is represented by the set $\boldsymbol{P}$ of global Bayes models. Then, this choice function $\boldsymbol{C}$ is changed into another $\boldsymbol{C}^+$ which is formed by adding the strict preferences associated with finitely many conditions of the form $\boldsymbol{T}(o,\boldsymbol{O}) \cap \boldsymbol{R}(\boldsymbol{O}_p) \neq \phi$. The results established in Section IV.2 of our (1995) show that then $\boldsymbol{C}^+$ satisfies the axioms. Also, those results show that in a neighborhood of the extreme points of the target sets for $\boldsymbol{C}$ there are sets of probability/almost-state-independent utility pairs that are local Bayes models for $\boldsymbol{C}$, and which then represent the choice function $\boldsymbol{C}^+$. These almost-state-independent utilities result by adding at least one new prize $\{\boldsymbol{r}\}$ to the two $\{\boldsymbol{0}, \boldsymbol{1}\}$ used to create the horse lotteries studied here. ◊-Lemma 3

Corollary: If for each choice problem $\boldsymbol{O}$ and $o \in \boldsymbol{C}(\boldsymbol{O})$ and the target set $\boldsymbol{T}(o,\boldsymbol{O})$ includes at least one of its boundary points, then $\boldsymbol{C}$ is represented by a set of probability/state-independent utility pairs.

References

[1] Aizerman, M.A. (1985) "New Problems in General Choice Theory," *Soc Choice Welfare* **2**: 235-282.

[2] Anscombe, F.J. and Aumann, R.J. (1963) "A definition of subjective probability," *Ann. Math. Stat*. **34**: 199-205.

[3] Cozman, F.G. and Walley, P. (2005) "Graphoid properties of epistemic irrelevance and independence," *Ann. Math. and A.I*. **45**: 173-195.

[4] Gilboa, I. and Schmeidler, D. (1989) "Maxmin expected utility with non-unique prior," *J. Math.Econ*. **18**: 141-153.

[5] Kadane, J.B., Schervish, M.J., and Seidenfeld, T. (2004) "A Rubinesque theory of decision**,"** *IMS Lecture Notes Monograph* **45**: 1-11.

[6] Kindler, J. (1983) "A General Solution Concept for Two Person, Zero Sum Games," *J. Optimization Theory and Applications* **40**: 105-119.

[7] Levi, I. (1974) "On indeterminate probabilities" *J.Phil*. **71**: 391-418.

[8] Pearce, D. (1984) "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* **52**: 1029-1050.

[9] Savage, L.J. (1954) *The Foundations of Statistics*. Wiley, New York.

[10] Schervish, M.J., Seidenfeld, T, Kadane, J.B., and Levi, I. (2003) "Extensions of expected utility theory and some limitations of pairwise comparisons" In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications* (J-M Bernard, T.Seidenfeld, and M.Zaffalon, eds.) 496-510. Carleton Scientific.

[11] Seidenfeld, T., Schervish, M.J., and Kadane, J.B. (1995) "A representation of partially ordered preferences," *Ann Stat*. **23**: 2168-2217.

[12] Sen, A. (1977) "Social choice theory: a re-examination," *Econometrica* 45: 53-89.

[13] Wald, A. (1950) Statistical Decision Functions. John Wiley, New York.

[14] Walley, P. (1990) *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

Multilinear and Integer Programming for Markov Decision Processes with Imprecise Probabilities

Ricardo Shirota Filho
Escola Politécnica,
Universidade de São Paulo, SP, Brazil
ricardo.shirota@poli.usp.br

Fabio Gagliardi Cozman
Escola Politécnica,
Universidade de São Paulo, SP, Brazil
fgcozman@usp.br

Felipe Werndl Trevizan
Departamento de Tecnología,
Universitat Pompeu Fabra, Barcelona, Spain
felipe.trevizan@upf.edu

Cassio Polpo de Campos
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo, SP, Brazil
cassiopc@usp.br

Leliane Nunes de Barros
Instituto de Matemática e Estatística
Universidade de São Paulo, SP, Brazil
leliane@ime.usp.br

Abstract

Markov Decision Processes (MDPs) are extensively used to encode sequences of decisions with probabilistic effects. Markov Decision Processes with Imprecise Probabilities (MDPIPs) encode sequences of decisions whose effects are modeled using sets of probability distributions. In this paper we examine the computation of Γ-maximin policies for MDPIPs using multilinear and integer programming. We discuss the application of our algorithms to "factored" models and to a recent proposal, Markov Decision Processes with Set-valued Transitions (MDPSTs), that unifies the fields of probabilistic and "nondeterministic" planning in artificial intelligence research.

Keywords. Markov Decision Processes with Imprecise Probabilities, Γ-maximin criterion, multilinear and integer programming.

1 Introduction

In this paper we are concerned with the computation of *policies*, or *plans*, that aim at maximizing reward over a possibly countably infinite sequence of *stages*. At each stage, our decision maker finds herself in a *state* and she must select an *action*. As a result of this decision, she gets a *reward*, and she moves to a new state. The process is then repeated. We focus on situations where transitions between states are modeled by *credal sets*; that is, by sets of probability distributions. Thus we focus on Markov Decision Processes with Imprecise Probabilities (MDPIPs), following a sizeable literature that has steadily grown in the last few decades. We review the basic concepts on MDPIPs in Section 2; we offer a relatively long review as we attempt to capture, in a somewhat organized form, various concepts dispersed in the literature.

There are several possible criteria that we might use to evaluate policies in an MDPIP. The term *optimal policy* is used in this paper in connection with Γ-maximin expected total discounted reward; that is, highest expected total discounted reward under the worst possible selection of probabilities.

We show how to reduce the generation of optimal policies for an MDPIP to *multilinear/integer programming* in Section 3. We also discuss in that section the practical reasons to pursue such a programming solution. We comment on the relationship between multilinear programming and "factored" models in Section 4. We then move, in Section 5, to a recently proposed special type of MDPIP that has particularly pleasant properties and important applications, the Markov Decision Process with Set-valued Transitions (MDPSTs).

2 Background

In this section we review basic facts about MDPs, MDPIPs, evaluation criteria, and algorithms.

2.1 MDPs

Markov Decision Processes (MDPs) are used in many fields to encode possibly infinite sequences of decisions under uncertainty. For historical review, basic technical development, and substantial reference to related

literature, the reader may consult books by Puterman [29] and Bertsekas [5]. In this paper we consider MDPs that are described by:

- a countable set $\mathcal{T}$ of *stages*; a decision is made at each stage.
- a finite set $\mathcal{S}$ of *states*.
- a finite set of *actions* $\mathcal{A}$; the set of actions may be indexed by states, but we simplify notation here by assuming a single set of actions for all states.
- a conditional probability distribution P_t that specifies the probability of transition from state s to state r given action a at stage t. We assume that probabilities are stationary (do no depend on t) and write $P(r|s,a)$.
- a *reward* function R_t that indicates how much is gained (or lost, by using a negative value) when action a is selected in state s at stage t. We assume the reward function to be stationary and write $R(s,a)$.

We refer to the state obtained at stage t, in a particular realization of the process, as s_t; likewise, the action selected at stage t is referred to as a_t.

The history h_t of an MDP at stage t is the sequence of states and actions visited by the process, $[s_1, a_1, \ldots, a_{t-1}, s_t]$. The *Markov assumption* that is adopted for MDPs is that $P(s_t|h_{t-1}, a_t) = P(s_t|s_{t-1}, a_t)$; consequently:

$$\begin{aligned} P(h_t|s_1) &= P(s_t|s_{t-1}, a_{t-1})P(s_{t-1}|s_{t-2}, a_{t-2}) \\ &\quad \ldots \times P(s_3|s_2, a_2)P(s_2|s_1, a_1). \end{aligned} \tag{1}$$

A *decision rule* $d_t(s,t)$ indicates the action that is to be taken in state s at stage t. A *policy* π is a sequence of decision rules, one for each stage. A policy may be *deterministic* or *randomized*; that is, it may prescribe actions with certainty, or rather it may just prescribe a probability distribution over the actions. A policy may also be *history-dependent* or not; that is, it may depend on all states and actions visited in previous stages, or just on the current state. A policy that is not history-dependent is called *Markovian*. A Markovian policy induces a probability distribution over histories through Expression (1).

We also assume that an MDP with infinite horizon (that is, with infinite $\mathcal{T}$) may always stop with some probability. In fact, we assume that the process stops with geometric probability: the process stops at stage t with probability $(1-\gamma)\gamma^{t-1}$ (independently of all other aspects of the process). Then γ is called the *discount* factor of the MDP [29, p. 125].

2.2 MDPIPs

Additional realism and flexibility can be attached to MDPs by allowing imprecision and indeterminacy in the assessment of transition probabilities. A decision process with states, actions, stages and rewards as described before, but where a set of probability distributions is associated with each transition, has been called a *Markov Decision Process with Imprecise Probabilities* (MDPIP) by White III and Eldeib [44], a name we adopt in this paper. Satia and Lave Jr. use instead the name *MDP with Uncertain Transition Probabilities* [31], in what may be the first thorough analysis of this model in the literature; Harmanec uses the term *generalized MDP* to refer to MDPIPs [21].

MDPIPs can represent incomplete and ambiguous beliefs about transitions between states; conflicting assessments by a group of experts; and situations where one wishes to investigate the effect of perturbations in a "base" model. MDPIPs have also been investigated as representations for abstracted processes, where details about transition probabilities are replaced by an enveloping set of distributions [17, 20]. Similar models are encoded by the *controlled Markov set-chains* by Kurano et al [26, 24]. Slightly less related are the vector-valued MDPs by Wakuta [41]. Some of these efforts have also adopted *interval-valued rewards*; in this paper we focus on imprecision/indeterminacy only in transition probabilities.

Thus an MDPIP is composed of a set of stages $\mathcal{T}$, a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a reward function R_t and sets of probability distributions, each containing transition probabilities P_t. We assume $\mathcal{T}$ to be the non-negative integers, $\mathcal{S}$ and $\mathcal{A}$ to be finite, and $\mathcal{A}$ to be constant for all states. We assume R_t to be a stationary function $R(s,a)$. We also assume stationarity for the sets $K(r|s,a)$ of probability distributions. Note, however, that now we have to distinguish two situations. First, the sets of transition probabilities may be identical across stages, while a history of the process may be associated with different draws within these sets (that is, probabilities are selected from sets that do not depend on t, but the selection depends on t). We might refer to these MDPIPs as *set-stationary*. Alternatively, it may be that each history h_t is associated with stationary probability distributions $P(s_t|s_{t-1}, a_{t-1})$ that themselves satisfy the Markov condition (and of course $P(s_t|s_{t-1}, a_{t-1}) \in K(s_t|s_{t-1}, a_{t-1})$). We might refer to the second MDPIPs as *elementwise-stationary* or simply *stationary*. In this paper we only deal with elementwise-stationary MDPIPs; in fact it does not seem that set-stationary MDPIPs have received any attention in the literature.

In the remainder of this paper we will use the following notation and terminology regarding sets of probability distributions. A set of probability distributions is called a *credal set* [27]. The credal set $K(X)$ contains distributions for variable X, and the conditional credal set $K(X|A)$ contains conditional distributions for variable X given event A. Conditioning is elementwise: $K(X|A)$ is obtained from $K(X)$ by conditioning every distribution in $K(X)$ on the event A. The notation $K(X|Y)$ represents a *set* of credal sets: there is a credal set $K(X|Y = y)$ for each nonempty event $\{Y = y\}$. A set of credal sets $K(X|Y)$ is *separately specified* if the joint credal set $K(X, Y)$ is such that, whenever $P(X|Y = y_1) \in K(X|Y = y_1)$, $P(X|Y = y_2) \in K(X|Y = y_2)$, then $P(X|Y = y_1)$ and $P(X|Y = y_2)$ are conditional distributions obtained from a single $P(X, Y)$ in $K(X, Y)$. That is, $K(X|Y)$ is separately specified if we can select conditional distributions independently from its sets, an assumption we make throughout for our credal sets. We loosely use $K(r|s, a)$ to indicate a separately specified collection of credal sets, for a given action a, where r and s refer to states.

Given a credal set $K(X)$, we can compute *lower* and *upper* probabilities respectively as $\underline{P}(A) = \inf_{P\in K} P(A)$ and $\overline{P}(A) = \sup_{P\in K} P(A)$. We can also compute *lower* and *upper* expectations for any bounded function $f(X)$ as $\underline{E}[f] = \inf_{P\in K} E[f]$ and $\overline{E}[f] = \sup_{P\in K} E[f]$, and likewise for conditional lower/upper probabilities/expectations. We assume all credal sets to be closed, so infima and suprema can be replaced by minima and maxima.

2.3 Evaluation criteria and algorithms

Given an MDP that starts at state s, we might evaluate a policy π by its expected reward:

$$V_\pi(s) = E_{s,\pi}\left[E_T\left[\sum_{t=1}^{T} R(s_t, a_t)\right]\right]; \qquad (2)$$

that is, the expectation of the expected reward assuming the process stops at stage T. Now if the process has a geometric probability of stopping at T, with parameter γ, we have [29, p. 126]:

$$V_{\pi,\gamma}(s) = E_{s,\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t)\right]. \qquad (3)$$

We refer to $V_{\pi,\gamma}(s)$ as the expected total discounted reward. There are other criteria to evaluate policies in MDPs; for example, the expected total reward $E_{s,\pi}[\sum_{t=1}^{\infty} R(s_t, a_t)]$, and the average reward $\lim_{T\to\infty}(1/T)E_{s,\pi}\left[\sum_{t=1}^{T} R(s_t, a_t)\right]$ [5, 29]. These criteria may be useful in specific problems but they are usually less realistic than Expression (2) and the associated discounted reward (3). We focus on the latter in this paper.

When we move to MDPIPs, we find that several criteria may be used to evaluate policies, even if we adopt total discounted reward. Three possible criteria are:

- Select the policy that yields the largest value of $\min V_\pi(s)$, where the minimum applies to all transition probabilities, subject to the fact that these probabilities must belong to given credal sets [4]. That is, the optimal policy produces the highest expected total discounted reward even when probabilities are most unfavorable. This is the Γ-maximin total discounted reward, where an optimal policy starting from state s must yield

  $$\max_\pi \min_P V_{\pi,\gamma}(s),$$

  where we append a subscript P in the minimization operator, to emphasize that it applies with respect to all transition probabilities that are imprecise/indeterminate.

- Select the policy that yields, when starting from state s,

  $$\max_\pi \max_P V_{\pi,\gamma}(s).$$

  That is, both decisions and probabilities can be selected so as to maximize expected total discounted reward. This criterion is referred to as Γ-maximax total discounted reward.

- Select any policy (or perhaps select all of those policies) that maximizes $V_{\pi,\gamma}(s)$ for at least one choice of transition probabilities. This is the criterion of E-admissibility [27].

Note that Γ-maximin and Γ-maximax create a complete order over policies, while E-admissibility is content to explore the partial order of policies induced by credal sets in any convenient way. To date, most authors have adopted the Γ-maximin criterion. An exception is Harmanec's algorithm [21] which employs *interval dominance* (Harmanec presents his algorithm as providing maximal policies, however [14, 38] argue that in fact is adopts interval dominance). Several other criteria can be found in the literature [14, 37, 38].

In this paper we focus on Γ-maximin total discounted reward; we refer to it as ΓETDR (for Expected Total Discounted Reward)[1]. The work of Satia and Lave

[1]It is not our goal to discuss here the adequacy of the Γ-maximin criterion; it is investigated in this paper because of its wide application in MDP problems. Other criteria will be investigated by the authors in the future. For discussions on the different criterions see [4, 25, 34, 32, 37, 42].

Jr. has derived several important results for this situation [31]. First, there exists a deterministic stationary policy that is optimal. Second, the optimal policy induces a value function that is the unique solution of

$$V^*(s) = \sup_a \inf_P \left(R(s,a) + \gamma \sum_r P(r|s,a) V^*(r) \right). \quad (4)$$

We can take maximum and minimum in this equation whenever the set of actions $\mathcal{A}$ is finite and the credal sets $K(r|s,a)$ have finitely many vertices. We assume this to be true in the remainder of this paper.

Expression (4) can be compactly written as $V^* = \mathbf{V}V^*$, by lumping the supremum, infimum, and summation into the operator $\mathbf{V}$. Whenever the transition probabilites are fixed (or are precisely specified) at some value P, we indicate it through the operator $\mathbf{V}_P$ (where the infimum is either suppressed or unnecessary). In fact, for an MDP with transition probabilities P, the optimal policy satisfies $V^* = \mathbf{V}_P V^*$, the *Bellman equation*.

2.4 Algorithms for MDPs and MDPIPs

Consider now algorithms that solve the Bellman equation. There are three "classic" algorithms for generating optimal policies in MDPs: value iteration, policy iteration, and reduction to linear programming [5, 29]. Most of the literature focuses on value or policy iteration. However, there are at least three reasons to pay attention to linear programming solutions to MDPs. First, a linear program produces an exact solution without the need to specify any stopping criteria (as needed for value and policy iteration). This property is useful in practice and particularly important while testing other algorithms. Second, several algorithms based on approximating the value function by lower dimensional functions are based on linear programming [19, 22, 33]. Third, and perhaps more importantly, linear programs seem to offer the only organized way to deal with problems where maximization of expected total discounted reward is subject to additional constraints on expected rewards [1, 29].

The linear programming algorithm for MDPs solves the equation $V^* = \mathbf{V}_P V^*$ for the precisely specified transition probabilities as follows [16]:

$$\begin{aligned} \min_{V^*} \quad & \sum_s V^*(s) \quad (5) \\ \text{s.t.} \quad & V^*(s) \geq R(s,a) + \gamma \sum_r P(r|s,a) V^*(r), \end{aligned}$$

where each pair (s,a) corresponds to a constraint.

Policy and value iteration have known counterparts for ΓETDR. Satia and Lave Jr. presented a policy iteration algorithm for ΓETDR. The results by Satia and Lave Jr., and by Denardo [15], produce a value iteration algorithm as indicated by White III and Eldeib [44]; the same algorithm was later derived in the special case of Bounded-parameter Markov Decision Processes (BMDPs) [17]. The value iteration algorithm starts with a candidate value function $V'_0(s)$ and iterates:

$$V'_{i+1} = \mathbf{V}V'_i \quad (6)$$

until $||V'_{i+1} - V'_i||$ is sufficiently small.[2] Convergence of this procedure is based on the fact that the operator $\mathbf{V}$ is a contraction mapping.[3]

3 A multilinear/integer solution for ΓETDR

Expression (5) describes the linear program for solving MDPs with precisely specified probabilities. It does not seem possible to produce a linear programming solution for ΓETDR; however, as we show in this section, it is possible to generate solutions using well known programming problems. We do not attempt to produce algorithms that surpass value/policy iteration in execution time; rather, our reasons to pursue a programming solution mirror the reasons why others have investigated linear programming for MDPs (summarized in Section 2.4). First, the results produced by multilinear and integer programming, and in particular the latter, depend on combinatorial properties of credal sets, and can be produced exactly; this is useful, for instance, while evaluating other algorithms that only promise ϵ-optimal policies. Second, several approximate algorithms for MDPs that can possibly be extended to MDPIPs depend on linear programming; we conjecture that these potential extensions to MDPIPs will depend on the results in this section. In fact, it seems that multilinear programming is unavoidable in factored models, as we discuss in Section 4. Third, solutions based on optimization seem to be the only way to handle constraints on expected rewards, a topic we wish to pursue in connection with planning (Section 5).

Our main result is, in essence, simple. We start from Expression (4), and note that its solution can be found by solving the following optimization problem:

$$\begin{aligned} \min_{V^*} \quad & \sum_s V^*(s) \quad (7) \\ \text{s.t.} \quad & V^*(s) \geq R(s,a) + \gamma \min_P \sum_r P(r|s,a) V^*(r). \end{aligned}$$

[2] The norm $||V|| = \max_s V(s)$ is typically used in the literature.

[3] A mapping $\mathbf{V} : U \to U$, where U is a complete normed linear space, is a contraction mapping iff $||\mathbf{V}u_1 - \mathbf{V}u_2|| \leq \gamma ||u_1 - u_2||$ for some $\gamma \in [0,1)$.

This can be shown to be an instance of bilevel programming [8, 40]. Similar problems have been tackled before in connection with linear programming with uncertainty, with obvious application to ΓETDR [2, 3]. Current algorithms for bilevel programming are complex, and convergence guarantees are not as sharp as one would like. It would be interesting to reduce Program (7) to a form that were closer to existing, well studied optimization problems. We do this by reducing Program (7) to multilinear and then to integer programming.

The multilinear program we consider is:

$$\begin{aligned} \min_{V^*,P} \quad & \sum_s V^*(s) \qquad (8) \\ \text{s.t.} \quad & V^*(s) \geq R(s,a) + \gamma \sum_r P(r|s,a) V^*(r). \end{aligned}$$

Denote by (V_R^*, P_R^*) a solution of Program (7) and by (V_G^*, P_G^*) a solution of Program (8). In order to use Program (8), we must prove that V_G^* and V_R^* are identical.

Theorem 1 $V_G^* = V_R^*$

Proof. Let Ω_R and Ω_G be the solution spaces for Programs (7) and (8) respectively. We prove that Ω_R is a subset of Ω_G. Then, we show that no solution in $\Omega_G \setminus \Omega_R$ can have better performance than one in Ω_R. We have:

$$\Omega_R = \{(V,P) : V \in \mathcal{V}, P = \arg\min_{P \in \mathcal{P}} \sum_r P(r|s,a) V(r)\},$$

$$\Omega_G = \{(V,P) : V \in \mathcal{V}, P \in \mathcal{P}\}.$$

Given that the solution space in the second case is the whole space $\mathcal{V} \times \mathcal{P}$, while in the first case P can only be in a subspace $\mathcal{V} \times \mathcal{P}_R$ of $\mathcal{V} \times \mathcal{P}$ (hence restricted), Program (8) produces a value function at least as low as Program (7). So, $V_G^* \leq V_R^*$, because $\Omega_G \supset \Omega_R$. Now suppose $V_G^* < V_R^*$. For a state $s \in \mathcal{S}$ we have $V_G^*(s) = R(s,a) + \gamma \sum_r P_G^*(r|s,a) V_G^*(r)$, with $P_G^*(r|s,a) \neq \arg\min_P \sum_r P(r|s,a) V(r)$. If we take $P'(r|s,a) = \arg\min_P \sum_r P(r|s,a) V(r)$, then $V'(s) = R(s,a) + \gamma \sum_r P'(r|s,a) V_G^*(r) < V_G^*(s)$ and V_G^* is not optimal. Since V_G^* *is* optimal (given that it considers the whole state space), then $V_G^* \not< V_R^*$. This implies that $V_G^* = V_R^*$. •

Apparently we have moved from a difficult problem (bilevel programming) to another difficult problem (multilinear programming). However, the significance of this result is that multilinear programming is a widely studied field, with close connections to geometric and linear programming [18, 23, 28, 35, 39]. Implementations can deal with hundreds of variables; in our tests we resort to Sherali and Adams' algorithm [35], a branch-and-bound scheme based on linear programming relaxations. Our implementation is an optimized version of this algorithm, that has been used to solve a variety of large and challenging multilinear programs [10, 11, 13, 12]. The examples presented later in this section were solved using this implementation.

An even more interesting result obtains if we assume that the vertices of credal sets $K(r|s,a)$ are known. Consider a list of vertices (each vertex is a distribution over $\mathcal{S}$) for a credal set $K(r|s,a)$, $\{p_1, \ldots, p_M\}$. Every distribution in this credal set can be expressed as a convex combination $\sum_{i=1}^M \alpha_i p_i$ where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. We can then write our goal as:

$$\begin{aligned} \min_{V^*,\alpha_{i,s,a}} \quad & \sum_s V^*(s) \qquad (9) \\ \text{s.t.} \quad & V^*(s) \geq R(s,a) + \\ & \gamma \sum_r \sum_i \alpha_{i,s,a} p_i(r|s,a) V^*(r), \\ & \alpha_{i,s,a} \geq 0, \quad \sum_i \alpha_{i,s,a} = 1, \end{aligned}$$

where we explicitly indicate that $\alpha_{i,s,a}$ depends on (s,a).

We now use the fact that a multilinear program has a maximum at the vertices of the credal sets; thus we necessarily have $\alpha_{i,s,a} \in \{0,1\}$ at a solution. We then resort to the following transformation to produce an *integer* program out of the multilinear program (9), just assuming that we can bound V^* from above and below (such bounds can be produced quite generally using results by White III and Eldeib [44]). First, we replace $V^*(r) \in [l,u]$ by $l + (V^*(r) - l)$, and create a new variable $\beta_r = V^*(r) - l \in [0, u-l]$. Each $\alpha_{i,s,a} p_i(r|s,a) V^*(r)$ is thus replaced by $\alpha_{i,s,a} p_i(r|s,a) l + \alpha_{i,s,a} p_i(r|s,a) \beta_r$. Note that $\alpha_{i,s,a} p_i(r|s,a) l$ is easy to evaluate. As $\alpha_{i,s,a}$ can be restricted to 0 or 1, we take each term $\alpha_{i,s,a} p_i(r|s,a) \beta_r$ and replace $\alpha_{i,s,a} \beta_r$ by a new variable $\beta_{i,r,s,a}$. To ensure that this replacement does not change the original problem, we introduce linear restrictions:

$$0 \leq \beta_{i,r,s,a} \leq \beta_r,$$

$$\beta_{i,r,s,a} \leq \alpha_{i,s,a}(u-l),$$

$$\beta_r - (u-l) + \alpha_{i,s,a}(u-l) \leq \beta_{i,r,s,a}.$$

The first and second restrictions are obvious (limitations on β_r and $\alpha_{i,s,a}$. The last restriction imposes the following. When $\alpha_{i,s,a} = 1$, $\beta_r \leq \beta_{i,s,a}$. However, since from the first restriction $\beta_{i,s,a} \leq \beta_r$, then $\beta_{i,s,a} = \beta_r$, and the full $V^*(r)$ will be considered. If $\alpha_{i,s,a} = 0$, then $\beta_r - (u-l) \leq \beta_{i,r,s,a}$, but

$\beta_r - (u-l) < 0$ (since $\beta_r \leq (u-l)$), so $\beta_{i,r,s,a} = 0$, and this non-optimal pair state-action will not be considered.

We end up with the following integer program:

$$\begin{aligned}
\min_{V^*, \alpha_{i,s,a}} \quad & \sum_s V^*(s) \qquad (10)\\
\text{s.t.} \quad & V^*(s) \geq R(s,a) + \\
& \gamma \sum_r \sum_i [\alpha_{i,s,a} p_i(r|s,a) l + \\
& \qquad p_i(r|s,a)\beta_{i,r,s,a}] \\
& \alpha_{i,s,a} \geq 0, \quad \sum_i \alpha_{i,s,a} = 1 \\
& \beta_r = V^*(r) - l \\
& 0 \leq \beta_r \leq u - l \\
& 0 \leq \beta_{i,r,s,a} \leq \beta_r \\
& \beta_{i,r,s,a} \leq \alpha_{i,s,a}(u-l) \\
& \beta_r - (u-l) + \alpha_{i,s,a}(u-l) \leq \beta_{i,r,s,a}
\end{aligned}$$

We close this section with two examples of MDPIPs. We focus on multilinear programming solutions; later we will consider examples where integer programming is used.

3.1 A small MDPIP

This is a very simple, abstract example. Consider two states, s_1 and s_2. In each state, the decision maker can choose between two actions. In s_1 the transition probability for both actions are imprecisely specified, while transition probabilities in s_2 are precisely specified. Probabilities and rewards are presented in Table 1 (left). The transition probabilities are defined from the states in the first column (origin states) to the states on the first row under P (destination states). The solution given by multilinear programming leads to the optimal solution; the value function V^* is shown in Table 1 (right).

3.2 Planning airplane maintenance through MDPIPs

This example is based on a problem described by White [43, p. 171]:

> An airline classifies the condition of its planes into three categories, viz. excellent, good and poor. The annual running costs for each category are 0.25×10^6, 10^6 and 2×10^6 [monetary units] respectively. At the beginning of each year the airline has to decide whether or not to overhaul each plane individually. With no overhaul a plane in excellent condition has probabilities of 0.75 and 0.25 of its condition being excellent or good, respectively, at the beginning of the next year. A plane in good condition has probabilities of 0.67 and 0.33 of its condition being good or poor, respectively, at the beginning of the next year. A plane in poor condition will remain in a poor condition at the beginning of the next year. An overhaul costs 2×10^6 and takes no significant time to do. It restores a plane in any condition to an excellent condition with probability 0.75, and leaves it in its current condition with probability 0.25. The airline also has an option of scrapping a plane and replacing it with a new one at a cost of 5×10^6. Such a new plane will be in excellent condition initially. There is an annual discount factor of $\gamma = 0.5$.

We consider a variant of this problem where probabilities are specified as in Table 2 (left). Multilinear programming produces the value function in Table 2 (right).

4 Factored MDPs

The specification of transitions between states is particularly burdensome in large MDPs. One strategy that has been often employed is to encode transition probabilities in *factored* form; usually this means that transition probabilities are encoded by *Bayesian networks* [7]. Here the state space is defined by the configurations of variables $\{X_1, \ldots, X_n\}$. We denote by $X_{i,t}$ the ith variable at stage t. For each action a, we specify a bipartite directed acyclic graph containing $2n$ nodes denoted by X_i^+ and X_i^-; node X_i^- and X_i^+ represent respectively $X_{i,t-1}$ and $X_{i,t}$ for any $t > 0$. One layer of the graph contains nodes X_i^- for all i, and no edge between them. The other layer contains nodes X_i^+ for all i, and edges between them. Edges are allowed *from* nodes in the first layer *into* the second layer, and also between nodes in the second layer. We denote by $\text{pa}(X_i^+)$ the *parents* of X_i^+ in the graph. The graph is assumed endowed with the following Markov condition: a variable X_i^+ is conditionally independent of its nondescendants given its parents. This implies the following factorization of transition probabilities:

$$P(X_1^+, \ldots, X_n^+) = \prod_{i=1}^{n} P(X_i^+ | \text{pa}(X_i^+)). \qquad (11)$$

Now suppose that conditional probability distributions $P(X_i^+|\text{pa}(X_i^+))$, or a subset of them, are not known precisely, but rather up to inclusion in credal sets $K\big(X_i^+|\text{pa}(X_i^+)\big)$. We assume the Markov condition to operate over all combinations of distributions from these credal sets, thus producing a possibly large set of joint distributions, each one of them satisfying

$\mathcal{S}$	$\mathcal{A}$	P		$R(s,a)$
		s_1	s_2	
s_1	$a_{1,1}$	[0,0.5]	[0.5,1]	7
	$a_{1,2}$	[0,0.2]	[0.8,1]	3
s_2	$a_{2,1}$	0.3	0.7	-1
	$a_{2,2}$	0.6	0.4	9

$V^*(s_1)$	21.486474
$V^*(s_2)$	18.108099
$\sum_s V^*(s)$	39.594573

Table 1: Specification of simple MDPIP example (left), and value function V^* (right).

$\mathcal{S}$	$\mathcal{A}$	P			$R(s,a)$
		s_1	s_2	s_3	
s_1	$a_{1,1}$	[0.5,1]	[0,0.4]	[0,0.1]	-0.25×10^6
	$a_{1,2}$	1	0	0	-2×10^6
	$a_{1,3}$	1	0	0	-5×10^6
s_2	$a_{2,1}$	0	[0.67,1]	[0,0.33]	-10^6
	$a_{2,2}$	[0.75,1]	[0,0.25]	0	-2×10^6
	$a_{2,3}$	1	0	0	-5×10^6
s_3	$a_{3,1}$	0	0	1	-2×10^6
	$a_{3,2}$	[0,0.25]	[0.5,0.8]	[0,0.25]	-2×10^6
	$a_{3,3}$	1	0	0	-5×10^6

$V^*(s_1)$	-1265664.1604
$V^*(s_2)$	-2496240.6015
$V^*(s_3)$	-4000000.0
$\sum_s V^*(s)$	-7761904.7619

Table 2: Specification of MDPIP for plane maintenance (left), and value function V^* (right).

the factorization in Expression (11) — the resulting structure is a *credal network* for each action [9].

The main point of this section is to indicate that Expression (11) defines a multilinear product for the probabilities that appear in Program (8). Thus, the multilinear character of Program (8) is left unchanged: the computation of Γ-maximin policies is still a matter of multilinear programming. The development of algorithms that produce optimal policies and that exploit the factorization in Expression (11) is left for the future; this is a promising avenue of research as the most advanced algorithms for factored MDPs do use all available structure encoded in the factorization [19, 22].

5 MDPSTs

In this section we explore the properties of a class of MDPIPs that have an important application in the field of artificial intelligence planning. Roughly speaking, planning in artificial intelligence focuses on sequential decision making problems that are specified using high-level languages. There are many variants of AI planning, depending on the properties of the specification language; for example, we have *deterministic* planning, where actions have deterministic effects; *probabilistic* planning, where actions have probabilistic effects; and *nondeterministic* planning, where an action may cause a transition to a set of states without any clue at to what state will be moved into [30]. The latter name is somewhat unfortunate as "nondeterminism" is an overloaded term, but it is the usual terminology in the field. Typically deterministic and nondeterministic planning are tackled by search through state spaces, while probabilistic planning is tackled by generation of equivalent MDPs.

There has been considerable effort in the field of AI planning to develop general algorithms that can be instantiated for different types of planning problems [6]. However, until recently no model considered actions with simultaneously "probabilistic" and "nondeterministic" effects. In response to this situation, Trevizan et al. have proposed a jointly probabilistic/nondeterministic framework, based on MDPIPs [36]. Their proposal is based on a class of MDPIPs, called Markov Decision Processes with Set-valued Transitions (MDPSTs), defined as follows.

An MDPST is composed by a set of stages $\mathcal{T}$, a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a reward function R, a state transition function $F(s,a)$ mapping states s and actions $a \in \mathcal{A}$ into reachable sets of $\mathcal{S}$, i.e., into nonempty subsets of $\mathcal{S}$, and a set of mass assignments $m(k|s,a)$ for all s, $a \in \mathcal{A}$, and $k \in F(s,a)$. Here we also assume $\mathcal{T}$ to be the non-negative integers, $\mathcal{S}$ and $\mathcal{A}$ to be finite, $\mathcal{A}$ to be constant for all states, and $R(s,a)$ to be a stationary function. The state transition function $F(s,a)$ and mass assignments $m(k|s,a)$ are also stationary. MDPSTs satisfy a simplified ver-

sion of Expression (4) [36]:

$$V^*(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \sum_{k \in F(s,a)} m(k|s,a) \min_{r \in k} V^*(r) \right). \quad (12)$$

MDPSTs form a strict subset of MDPIPs [36]; thus Programs (8) or (10) can be used to solve MDPSTs. These solutions require an enumeration on mass assignments $m(k|s,a)$. However we can produce simpler programs if we study Expression (12) carefully.

Given any action $a \in \mathcal{A}$, we can collect all feasible $k \in F(s,a)$, and define a binary vector $I(s,a)$ with as many elements as sets of states in $F(s,a)$, such that $I_i(s,a) \in \{0,1\}$ for $i \in \{1,\ldots,N\}$, and $\sum_i I_i(s,a) = 1$. Because each $I_i(s,a)$ can only be equal to 0 or 1, and their sum is equal to one, only an unique $I_i(s,a)$ can be equal to one at a time. We now write Expression (12) as:

$$V^*(s) = \max_{a \in \mathcal{A}} R(s,a) + \gamma \sum_{k \in F(s,a)} m(k|s,a) \sum_{i=1}^{k} I_i(s,a) V^*(r_i). \quad (13)$$

We now transform each product $I_i(s,a)V^*(r_i)$ into a new variable, following the procedure outlined in Section 3. We first replace $V^*(r_i)$ by $l + (V^*(r_i) - l)$, where $V^*(r_i) \in [l,u]$; we then define $\beta_i = V^*(r_i) - l$, with $\beta_i \in [0, u-l]$. We define a variable $\beta_{i,s,a} = I_i(s,a)\beta_i$, and add the necessary constraints to the optimization problem. The final integer program is very similar to the Program (10):

$$\begin{aligned}
\min_{V^*, I} \quad & \sum_s V^*(s) \quad (14)\\
\text{s.t.} \quad & V^*(s) \geq R(s,a) + \\
& \gamma \sum_k \sum_i [I_i(s,a) m(k|s,a) l + \\
& \qquad m(k|s,a) \beta_{i,s,a}] \\
& I_i(s,a) \geq 0, \sum_i I_i(s,a) = \\
& \beta_i = V^*(r_i) - l \\
& 0 \leq \beta_i \leq u - l \\
& 0 \leq \beta_{i,s,a} \leq \beta_i \\
& \beta_{i,s,a} \leq I_i(s,a)(u-l) \\
& \beta_i - (u-l) + I_i(s,a)(u-l) \leq \beta_{i,s,a}.
\end{aligned}$$

This is a very useful transformation, once integer programming is much simpler than multilevel programming. There are many powerful integer program solvers that guarantee global optimal solutions, where multilevel program solvers only achieve global optimals in certain specific cases.

5.1 A small MDPST

Consider 3 states, s_1, s_2 and s_3. At state s_i, there are actions $a_{i,1}$ and $a_{i,2}$. All actions define probabilistic transitions from one state to itself or to the set composed by the other 2 states, however with different assignments of rewards and transition probabilities. The values assigned to each state and action can be found in Table 3. The optimal solution was obtained by solving an integer program.

5.2 Probabilistic/nondeterministic planning of airplane maintenance

Consider the example of airplane maintenance in Section 3. Suppose that transition probabilities follow Table 4 (left); a transition that "fills" more than a column is a nondeterministic one. The optimal solution obtained can be seen in Table 4 (right).

6 Conclusion

We have reviewed the basic theory of MDPIPs under the criterion of Γ-maximin expected total discounted reward, and we have shown how to produce policies using multilinear and integer programming. This type of solution may be useful to handle problems with further constraints on expected rewards, and to deal with factored models and factored approximations. We plan to continue the present work by exactly addressing such constraints and factorizations.

We have then looked into the recently proposed MDPSTs. We have briefly reviewed the application of these processes as a unifying language for "probabilistic" and "nondeterministic" planning, and then showed how these processes nicely lead to integer programming solutions. As indicated previously, one of the reasons to investigate a programming solution for MDPIPs is the promise it holds for treating problems with constraints on policy. For instance, it may be required that a policy, besides maximizing minimum expected total discounted reward, also guarantees the probability of some set of states to be higher than some value (in practice: maximization of profit for a company, subject to the probability that a client is left unattended being smaller than a given value). Markov decision processes subject to such constraints are called *constrained MDPs* [1, 29], and the main method of solution there is linear programming. We conjecture that constrained MDPIPs will require solutions based on multilinear/integer programming. This will be even more important in the context of MDPSTs, because "nondeterministic" planning is usually associated with contraints on policies.

$\mathcal{S}$	$\mathcal{A}$	P		$R(s,a)$
		s_i	$\mathcal{S} \setminus \{s_i\}$	
s_1	$a_{1,1}$	0.8	0.2	5
	$a_{1,2}$	0.1	0.9	-1
s_2	$a_{2,1}$	0.8	0.2	4
	$a_{2,2}$	0.3	0.7	7
s_3	$a_{3,1}$	0.7	0.3	3
	$a_{3,2}$	0.25	0.75	9

$V^*(s_1)$	17.670251
$V^*(s_2)$	19.820789
$V^*(s_3)$	22.153796
$\sum_s V^*(s)$	59.644836

Table 3: Specification of small MDPST (left), and value function V^* (right).

$\mathcal{S}$	$\mathcal{A}$	P			$R(s,a)$
		s_1	s_2	s_3	
s_1	$a_{1,1}$	0.5	0.5		-0.25×10^6
	$a_{1,2}$	1	0	0	-2×10^6
	$a_{1,3}$	1	0	0	-5×10^6
s_2	$a_{2,1}$	0	1		-10^6
	$a_{2,2}$	0.75	0.25	0	-2×10^6
	$a_{2,3}$	1	0	0	-5×10^6
s_3	$a_{3,1}$	0	0	1	-2×10^6
	$a_{3,2}$	0.8		0.2	-2×10^6
	$a_{3,3}$	1	0	0	-5×10^6

$V^*(s_1)$	-1666666.6666
$V^*(s_2)$	-3000000.0
$V^*(s_3)$	-4000000.0
$\sum_s V^*(s)$	-8666666.666

Table 4: Specification of MDPST for plane maintenance (left), and value function V^* (right).

Acknowledgements

This work has been supported by FAPESP grant 2004/09568-0; the first author is supported by FAPESP grant 05/58090-9; the second author is partially supported by CNPq grant 3000183/98-4; the third author is supported by CAPES grant BEX 4248-05-8; the fifth author is partially supported by CNPq grant 308530/03-9.

Tests were run in a multilinear programming solver that calls AMPL and CPLEX in its internal loops.

References

[1] E. Altman. *Constrained Markov decision processes*. Chapman & Hall, Boca Raton, Florida, 1999.

[2] I. Averbakh. On the complexity of a class of combinatorial optimization problems with uncertainty. *Mathematical Programming*, 90(2):263–272, 2001.

[3] A. Ben-Tal and A. Nemirovski. Robust solutions of undertain linear programs. *Operations Research Letters*, 25(1):1–13, 1993.

[4] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control (Vol. 1, 2)*. Athena Scientific, Belmont, Massachusetts, 1995.

[6] B. Bonet and H. Geffner. Learning Depth-First Search: A unified approach to heuristic search in deterministic and non-deterministic settings, and its application to MDPs. In *Proc. of the 16th ICAPS*, 2006.

[7] C. Boutilier, S. Hanks, and T. Dean. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[8] B. Colson, P. Marcotte and G. Savard. Bilevel programming: A survey. *Quaterly Journal of Operations Research*, 3(2):87–107, 2005.

[9] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

[10] F. G. Cozman, C. P. de Campos, J. S. Ide, and J. C. F. da Rocha. Propositional and relational Bayesian networks associated with imprecise and qualitative probabilistic assessments. In *Uncertainty in Artificial Intelligence*, pages 104–111. AUAI Press, 2004.

[11] C. P. de Campos and F. G. Cozman. Inference in credal networs using multilinear programming. In *Second Starting AI Researchers' Symposium (STAIRS)*, pages 50–61, Amsterdam, IOS Press, 2004.

[12] C. P. de Campos and F. G. Cozman. Belief updating and learning in semi-qualitative probabilistic networks. In *Uncertainty in Artificial Intelligence*, pages 153–160, Edinburgh, United Kingdom, 2005.

[13] C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. In *Fourth International Symposium on*

Imprecise Probabilities and Their Applications, pages 78–87, Dulles, Virginia, 2005.

[14] G. de Cooman and M. C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal Approximate Reasoning*, 39(2-3):257–278, 2005.

[15] E. V. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9(2):165–177, 1967.

[16] F. D'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10(1):98–108, 1963.

[17] R. Givan, S. M. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.

[18] W. Gochet and Y. Smeers. A branch-and-bound method for reversed geometric programming. *Operations Research*, 27(5):983–996, September/October 1979.

[19] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

[20] V. Ha and P. Haddawy. Theoretical foundations for abstraction-based probabilistic planning. In *Uncertainty in Artificial Intelligence*, pages 291–298, San Francisco, California, United States, 1996. Morgan Kaufmann.

[21] D. Harmanec. Generalizing Markov decision processes to imprecise probabilities. *Journal of Statistical Planning and Inference*, 105:199–213, 2002.

[22] M. Hauskrecht and B. Kveton. Linear program approximations for factored continuous-state Markov decision processes. In *Advances in Neural Information Processing Systems 16*, pages 895–902. 2004.

[23] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer-Verlag, 1995.

[24] M.Hosaka, J. Nakagami, and M. Kurano. Controlled Markov set-chains with set-valued rewards – the average case. *International Transactions in Operations Researach*, 9:113–123, 2002.

[25] D. Kikuti, F. G. Cozman, and C. P. de Campos. Partially ordered preferences in decision trees: computing strategies with imprecision in probabilities. In *IJCAI Workshop on Advances in Preference Handling*, pages 118–123, Edinburgh, United Kingdom, 2005.

[26] M. Kurano, J. Song, M. Hosaka, and Y. Huang. Controlled Markov set-chains with discounting. *Journal Applied Probability*, 35:293–302, 1998.

[27] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[28] C.D. Maranas and C.A. Floudas. Global optimization in generalized geometric programming. *Computers and Chemical Engineering*, 21(4):351–370, 1997.

[29] M. L. Puterman. *Markov Decision Processes*. John Wiley and Sons, New York, 1994.

[30] S. J. Russell and P. Norvig. *Artificial Intelligence: a Modern Approach*. Prentice Hall, New Jersey, 1995.

[31] J. K. Satia and R. E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21:728–740, 1970.

[32] M. Schervish, T. Seidenfeld, J. Kadane, and I. Levi. Extensions of expected utility theory and some limitations of pairwise comparisons. In *Third International Symposium on Imprecise Probabilities and their Applications*, pages 496–510. Carleton Scientific, 2003.

[33] P. J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110(2):568–582, 1985.

[34] T. Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: γ-maximin versus e-admissibility. *Synthese*, 140(1-2), 2004.

[35] H.D. Sherali and W.P. Adams. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Kluwer Academic Publishers, 1999.

[36] F. W. Trevizan, F. G. Cozman, and L. N. de Barros. Planning under risk and Knightian uncertainty. In *20th IJCAI*, pages 2023–2028, 2007.

[37] M. C. M. Troffaes. Decision making with imprecise probabilities: A short review. *SIPTA Newsletter*, pages 4–7, 2004.

[38] M. C. M. Troffaes. Learning and optimal control of imprecise Markov decision processes by dynamic programming using the imprecise Dirichlet model. pages 141–148, Berlin, 2004. Springer.

[39] H. Tuy. *Convex Analysis and Global Optimization*, volume 22 of *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, 1998.

[40] L. N. Vicente and P. H. Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global Optimization*, 5(3):291–306, 1994.

[41] K. Wakuta. Vector-valued Markov decision processes and the system of linear inequalities. *Stochastic Processes and their Applications*, 56:159–169, 1995.

[42] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[43] D. J. White. *Markov Decision Processes*. John Wiley and Sons, 1993.

[44] C. C. White III and H. K. El-Deib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, July-August 1994.

Regular finite Markov chains with interval probabilities

Damjan Škulj
Faculty of Social Sciences
University of Ljubljana, Slovenia
damjan.skulj@fdv.uni-lj.si

Abstract

In Markov chain theory a stochastic matrix P is regular if some matrix power P^n contains only strictly positive elements. Regularity of transition matrix of a Markov chain guarantees the existence of a unique invariant distribution which is also the limiting distribution. In the present paper a similar result is shown for the generalized Markov chain model that replaces classical probabilities with interval probabilities. We generalize the concept of regularity and show that for a regular interval transition matrix sets of probabilities corresponding to consecutive steps of a Markov chain converge to a unique limiting set of distributions that only depends on transition matrix and is independent of the initial distribution. A similar convergence result is also shown for approximations of the invariant set.

Keywords. Markov chains, interval probabilities

1 Introduction

Markov chains are one of the most important tools to model random phenomena evolving in time. They are enough simple to allow detailed description but also enough general to allow many possibilities for applications (see [6]). A weak point of the most widely used model is that transition probabilities have to be constant and precisely known.

An attempt to relax this restriction was proposed in [7], where classical probabilities are replaced by interval probabilities. The approach presented there extends a previous approach given in [5], where the assumption of precisely known initial and transition probabilities is relaxed so that probability intervals are used instead of precise probabilities. Their model is based on the assumption that constant classical probabilities rule the process but only approximations are known instead of precise values. Several estimates based on this model are also given in [2] and [4].

Our approach presented in [7] uses the more general model of interval probabilities based on Weichselberger's theory (see [10] or [9]) instead of simple probability intervals, and omits the assumption that transition probabilities that rule the process are constant in time. In the sequel we refer to this model as *Markov chains with interval probabilities (MCIP)*. The model allows computation of possible probability distributions at consecutive steps and estimation of invariant distributions, which are of great importance in Markov chain theory. But there is a fundamental problem of those estimations that the sets of distributions corresponding to further steps become much more complicated than sets representable by interval probabilities. A way to overcome this problem is the use of approximations.

In this paper we examine the relationship between invariant sets of distributions and long term behaviour of generalized Markov chains. In the classical theory an important class of Markov chains, so called *regular chains*, has the property that its unique invariant distribution is also the limiting distribution to which probabilities converge after long time. Here we generalize the concept of regularity to MCIP and show that generalized regular Markov chains have a similar convergence property. Moreover, we show a similar result for a class of approximations with interval probabilities.

The paper has the following structure. In Section 2 we introduce basic concepts of the theory of interval probabilities and MCIP. In Section 3 we give our main results on convergence for MCIP.

2 Markov chains with interval probabilities

2.1 Interval probabilities

First we introduce basic elements of interval probability due to Weichselberger ([10]), some of them in

a simplified form. Let Ω be a non-empty set and $\mathcal{A}$ a σ-algebra of its subsets. The term *classical probability* or *additive probability* will denote any set function $p\colon \mathcal{A} \to \mathbb{R}$ satisfying Kolmogorov's axioms. Let L and U be set functions on $\mathcal{A}$, such that $L \leq U$ and $L(\Omega) = U(\Omega) = 1$. The interval valued function $P(\,.\,) = [L(\,.\,), U(\,.\,)]$ is called an *interval probability*.

To each interval probability P we associate the set $\mathcal{M}$ of all additive probability measures on the measurable space $(\Omega, \mathcal{A})$ that lie between L and U. This set is called the *structure* of the interval probability P. The basic class of interval probabilities are those whose structure is non-empty. Such an interval probability is denoted as R-field. The most important subclass of interval probabilities, F-fields, additionally assumes that both lower bound L and upper bound U are strict according to the structure $\mathcal{M}$:

$$L(A) = \inf_{p \in \mathcal{M}} p(A) \quad \text{and} \quad U(A) = \sup_{p \in \mathcal{M}} p(A) \qquad (1)$$

for every $A \in \mathcal{A}$.

The above property is in a close relation to *coherence* in Walley's sense (see [8]). The difference is that the definition of coherence allows finitely additive probabilities while Weichselberger's model only allows σ-additive probabilities. However, in the case of finite probability spaces, both terms coincide, because finite additivity and σ-additivity then coincide. The requirement (1) implies the relation $U(A) = 1 - L(\neg A)$ for every $A \in \mathcal{A}$, and therefore, only one of the bounds L and U is needed. Usually we only take the lower one. Thus, an F-field is sufficiently determined by the triple $(\Omega, \mathcal{A}, L)$.

MCIP require several approximations involving lower expectations with respect to sets of probabilities. Let $\mathcal{C}$ be a set of probability measures on $(\Omega, \mathcal{A})$ and let a random variable $X\colon \Omega \to \mathbb{R}$ be given. The *lower* and the *upper expectation* $\underline{E}_{\mathcal{C}} X$ and $\overline{E}_{\mathcal{C}} X$ of X with respect to $\mathcal{C}$ are defined as the infimum and supremum of mathematical expectations of X with respect to members of $\mathcal{C}$:

$$\underline{E}_{\mathcal{C}} X = \inf_{p \in \mathcal{C}} E_p X$$
$$\overline{E}_{\mathcal{C}} X = \sup_{p \in \mathcal{C}} E_p X.$$

An important class of interval probabilities are those whose lower bounds L are 2-*monotone* (*convex, supermodular*), i.e. for every $A, B \subseteq \Omega$

$$L(A \cup B) + L(A \cap B) \geq L(A) + L(B). \qquad (2)$$

If equality holds in the above equation the set function L is said to be *modular*, which in the case where $L(\emptyset) = 0$ is equivalent to finite additivity.

In the finite case, 2-monotonicity implies the F-property, which is then equivalent to coherence that is always implied by 2-monotonicity. Moreover, in the case of a 2-monotone coherent lower probability L on a finite measurable space, the lower and the upper expectation operators with respect to the corresponding structure can be found in terms of Choquet integral with respect to L and the corresponding upper probability U respectively, where *Choquet integral* with respect to a set function L is defined as

$$\int_\Omega X \, dL = \int_{-\infty}^{0} (L(X > t) - L(\Omega))\, dt + \int_0^{\infty} L(X > t)\, dt.$$

The right hand side integrals are both Riemann integrals. Further, if L is an additive measure, Choquet integral coincides with Lebesgue integral.

Let Ω be a finite set. If $\mathcal{M}$ is the structure of an F-field $P = (\Omega, \mathcal{A}, L)$ with L 2-monotone, we have

$$\underline{E}_{\mathcal{M}} X = \int_\Omega X \, dL \qquad (3)$$

for every random variable X. (For the proof see e.g. [3], and note that for an infinite Ω, instead of the structure $\mathcal{M}$, the set of all finitely additive measures dominating L would be required for the above equality.) In fact, the equality in (3) for every X is equivalent to 2-monotonicity if the lower expectation is taken with respect to the set of all finitely additive measures dominating L. For a non-2-monotone L Choquet integral is in general lower than the lower expectation.

2.2 Markov chains with interval probabilities

Now we introduce the framework of MCIP model proposed in [7]. Let Ω be a finite set with elements $\{\omega_1, \ldots, \omega_m\}$ and 2^Ω the algebra of its subsets. Further let

$$X_0, X_1, \ldots, X_n, \ldots \qquad (4)$$

be a sequence of random variables such that

$$P(X_0 = \omega_i) = q^{(0)}(\omega_i) =: q_i^0,$$

where $q^{(0)}$ is a classical probability measure on $(\Omega, 2^\Omega)$ such that

$$L^{(0)} \leq q^{(0)}, \qquad (5)$$

where $Q^{(0)} = (\Omega, 2^\Omega, L^{(0)})$ is an F-probability field. Thus, $q^{(0)}$ belongs to the structure $\mathcal{M}^{(0)}$ of $Q^{(0)}$. This means that initial probability distribution is not known precisely, but only a set of possible distributions is given as a structure of an F-field.

Transition probabilities in a classical finite Markov chain can be given by a matrix whose (i,j)-th entry represents the probability that the process that is in the state ω_i at time n will be in the state ω_j at time $n+1$. Each row of a transition probability matrix is then a probability distribution on $(\Omega, 2^\Omega)$.

The idea of the generalized transition matrix is to replace classical probability distributions in rows with interval probabilities. Thus, suppose that

$$P\big(X_{n+1} = \omega_j \mid X_n = \omega_i, \\ X_{n-1} = \omega_{k_{n-1}}, \ldots, X_0 = \omega_{k_0}\big) \\ = p_i^{n+1}(\omega_j) =: p_{ij}^{n+1}, \quad (6)$$

where p_{ij}^{n+1} is independent of $X_0, \ldots, X_{n-1}$ for all $n \geq 1$, and

$$L_i \leq p_i^{n+1}, \quad (7)$$

where $P_i = (\Omega, 2^\Omega, L_i)$, for $1 \leq i \leq m$, is an F-probability field. Thus, p_{ij}^{n+1} are transition probabilities at time $n+1$, and they are not assumed to be constant in n. Instead, on each step they are only supposed to satisfy the inequality (7), where L_i are constant in time.

The above generalization of transition matrices suggests the following generalization of the concept of stochastic matrix to interval probabilities. Let $P = [P_1 \ldots P_m]^T$ where P_i are F-fields for $i = 1, \ldots, m$. We will call such P an *interval stochastic matrix*. The *lower bound* of an interval stochastic matrix is simply $P_L := [L_1 \ldots L_m]^T$, where L_i is the lower bound of P_i and the *structure* of an interval stochastic matrix is the set $\mathcal{M}(P)$ of stochastic matrices $p = (p_{ij})$ such that $p_i \geq L_i$, where p_i, for $i = 1, \ldots, m$, is the classical probability distribution on $(\Omega, 2^\Omega)$ generated by $p_i(\omega_j) = p_{ij}$ for $j = 1, \ldots, m$.

To represent an F field on a given probability space, one value has to be given for each event A; usually, this is the lower probability $L(A)$ of A. Thus, a row of an interval stochastic matrix can be represented as a row of $2^m - 2$ values, where $\emptyset$ and Ω, whose lower probabilities are always 0 and 1 respectively, are excluded. All other events correspond to each column in a given order. In general, this requires $m(2^m - 2)$ values for the transition matrix and $2^m - 2$ values for the initial distribution. The (i,j)-th entry of the transition matrix is then the lower probability of transition from the state ω_i to the set A_j.

We demonstrate this by the following example.

Example 1. Take $\Omega = \{\omega_1, \omega_2, \omega_3\}$. The algebra 2^Ω contains six non-trivial subsets, which we denote by $A_1 = \{\omega_1\}, A_2 = \{\omega_2\}, A_3 = \{\omega_3\}, A_4 = \{\omega_1, \omega_2\}, A_5 = \{\omega_1, \omega_3\}, A_6 = \{\omega_2, \omega_3\}$. Thus, besides $L(\emptyset) = 0$ and $L(\Omega) = 1$ we have to give the values $L(A_i)$ for $i = 1, \ldots, 6$. Let the lower probability L of an interval probability Q be represented through the n-tuple

$$L = (L(A_1), L(A_2), L(A_3), L(A_4), L(A_5), L(A_6)) \quad (8)$$

and take $L = (0.1, 0.3, 0.4, 0.5, 0.6, 0.7)$. Further we represent the interval transition matrix P by a matrix with three rows and six columns, each row representing an element ω_i of Ω and the values in the row representing the interval probability P_i through its lower probability L_i. Take for example the following matrix:

$$P_L = \begin{pmatrix} 0.5 & 0.1 & 0.1 & 0.7 & 0.7 & 0.4 \\ 0.1 & 0.4 & 0.3 & 0.6 & 0.5 & 0.8 \\ 0.2 & 0.2 & 0.4 & 0.5 & 0.7 & 0.7 \end{pmatrix}. \quad (9)$$

The probability of transition from ω_1 to A_2 is thus at least 0.1, and to A_5 at least 0.7. Since $A_2 = \Omega - A_5$, the corresponding upper probability of transition from ω_1 to A_2 is $1 - 0.7 = 0.3$.

Note that the case where $|\Omega| = 3$ is somewhat specific, because every non-trivial subset is either atomic or a complement of an atomic set. Therefore, lower probabilities of the non-atomic sets can be obtained from the upper probabilities corresponding to atomic sets using $L(A) = 1 - U(\neg A)$. However, in general, lower probabilities given for all non-trivial subsets carry more information than probability intervals on atomic sets alone. Another specific feature of the case with $|\Omega| \leq 3$ is that the lower probability corresponding to any F-field is 2-monotone.

2.3 Computing distributions at further steps

The main advantage of Markov chains is that knowing the probability distribution at time n we can easily compute the distribution at time $n+1$. This is done by multiplying the given distribution with the transition matrix.

In the case of MCIP, where initial distribution as well as transition matrix are interval valued, we would want the probability distribution at the next step to be of a similar form. Thus, in an ideal case, the next step probability distribution would be an interval probability or even an F-field. But this is in general not possible. According to MCIP model, the actual distribution at each step is a classical probability distribution which is assumed to be a member of some set of distributions forming a structure of an interval probability. Similarly, the transition matrix is a classical transition matrix belonging to a set of matrices, also given in terms of interval probabilities.

Let $q^{(0)}$ be an initial distribution, thus satisfying (5), and p^1 a transition probability, satisfying (7). According to the classical theory, the probability at the next step is $q^{(1)} = q^{(0)}p^1$. Thus, the corresponding set of possible probability distributions at the next step must contain all the probability distributions of this form. Consequently, in the most general form, the set of probability distributions corresponding to X_k would be

$$\mathcal{C}_k := \{q^{(0)}p^1 \dots p^k \mid q^{(0)} \in \mathcal{M}(Q^{(0)}), \\ p^i \in \mathcal{M}(P) \text{ for } i = 1, \dots, k\}. \quad (10)$$

But these sets in general cannot be represented as structures of interval probabilities. Thus, they cannot be observed in terms of interval probabilities, or even in terms of convex sets. However, a possible approach using interval probabilities is to calculate the lower and the upper envelope of the set of probabilities obtained at each step and do further calculations with this interval probability and its structure. The resulting set of possible distributions at n-th step is then in general larger than $\mathcal{C}_k$, and could only be regarded as an approximate to the true set of distributions. In a similar way also more general convex envelopes of sets $\mathcal{C}_k$ can be constructed.

Approximation with interval probabilities

Here we describe how to compute approximations of the sets $\mathcal{C}_n$ with interval probabilities. We define a sequence $(Q^{(n)})_{n\geq 0}$ of F-fields, where $Q^{(0)}$ denotes the initial interval probability distribution, such that the structure $\mathcal{M}^{(n)}$ of each member of the sequence contains the set $\mathcal{C}_n$.

For every n let $Q^{(n+1)}$ be the F-field generated by the set of all products of the form $q^{(n)}p^{n+1}$ where $q^{(n)}$ belongs to the structure $\mathcal{M}(Q^{(n)})$ and p^{n+1} is a member of $\mathcal{M}(P)$. Such $Q^{(n+1)}$ is thus the narrowest F-field whose structure contains all the products $q^{(n)}p^{n+1}$. The products $q^{(n)}p^{n+1}$ would be the possible distributions at time $n+1$ if every $q^{(n)} \in \mathcal{M}(Q^{(n)})$ was a possible distribution at time n. Clearly, the inclusions $\mathcal{C}_n \subseteq \mathcal{M}(Q^{(n)}) = \mathcal{M}^{(n)}$ hold, but the intervals are in general wider than necessary to bound the sets $\mathcal{C}_n$. However, finding exact intervals is a computationally difficult problem.

Let $L^{(n)}$ be the lower probability corresponding to $Q^{(n)}$ and $L^{(n+1)}$ the one corresponding to $Q^{(n+1)}$. Further, let $q^{(n)}$ be any member of the structure $\mathcal{M}(Q^{(n)})$ and $q^{(n+1)}$ the corresponding distribution at time $n+1$. For every $A \subseteq \Omega$ we have

$$\begin{aligned} q^{(n+1)}(A) &= \sum_{\omega_j \in A} \sum_{i=1}^{m} q_i^{(n)} p_{ij}^{n+1} \\ &= \sum_{i=1}^{m} q_i^{(n)} \sum_{\omega_j \in A} p_{ij}^{n+1} \\ &= \sum_{i=1}^{m} q_i^{(n)} p_i^{n+1}(A) \\ &\geq \sum_{i=1}^{m} q_i^{(n)} L_i(A). \end{aligned} \quad (11)$$

Since p_i^{n+1} can be chosen independently of each other and of $q^{(n)}$ and because L_i have the F-property, they can be chosen so that

$$p_i^{n+1}(A) = L_i(A) \quad \text{for every } 1 \leq i \leq m.$$

Therefore, equality can be achieved in (11). Consequently, we obtain:

$$L^{(n+1)}(A) = \inf_{q^{(n)} \geq L^{(n)}} \sum_{i=1}^{m} q_i^{(n)} L_i(A). \quad (12)$$

The above infimum can be viewed as a lower expectation with respect to $\mathcal{M}^{(n)}$ of the function $X_A(\omega_i) := L_i(A)$.

If the lower probability $L^{(n)}$ is 2-monotone, (12) can (because of finiteness) equivalently be expressed in terms of Choquet integral (see e.g. [3])

$$L^{(n+1)}(A) = \int L_i(A)\, dL^{(n)} = \int X_A\, dL^{(n)}. \quad (13)$$

The above expression is linear in $L^{(n)}$ and thus requires significantly less computation to evaluate than (12). But even if both $L^{(n)}$ and L_i, for $1 \leq i \leq m$, are 2-monotone, the resulting lower probability $L^{(n+1)}$ need not be 2-monotone. Therefore, the use of (13) would in general produce less accurate results.

2.4 Invariant distributions

The invariant set of distributions

One of the main concepts in the theory of Markov chains is the existence of an invariant distribution. In the classical theory, an invariant distribution of a Markov chain with transition probability matrix P is any distribution q such that $q\,P = q$. In the case of regular Markov chain an invariant distribution is also the limiting distribution.

In MCIP model, a single transition probability matrix as well as initial distributions are replaced by sets of distributions given by structures of interval probabilities. Consequently, an invariant distribution is replaced by a set of distributions, which is invariant for the interval transition probability matrix P. An

invariant set of distributions is thus a set $\mathcal{C}$ satisfying the condition

$$\mathcal{C} = \{qp \mid q \in \mathcal{C}, p \in \mathcal{M}(P)\}. \tag{14}$$

Thus, the invariant set of probabilities is closed for multiplication with the set of possible transition matrices. Of course, this does not mean that all its members are invariant distributions corresponding to some matrices from $\mathcal{M}(P)$, but it will follow from the construction that the largest such set must contain all those invariant distributions.

Given an interval transition matrix P it is in principle easy to find its largest invariant set of distributions. We start with the set $\mathcal{C}_0$ of all probability distributions on $(\Omega, 2^\Omega)$ and construct the sequence of sets of probability measures:

$$\mathcal{C}_{i+1} := \{q\,p \mid q \in \mathcal{C}_i, p \in \mathcal{M}(P)\}, \tag{15}$$

starting with $\mathcal{C}_0$. The above sequence corresponds to sequence (10), where the initial set of distributions is equal to the set of all probability distributions. In this case the sequence is monotone and the limiting set of distributions

$$\mathcal{C}_\infty := \bigcap_{i=1}^{\infty} \mathcal{C}_i. \tag{16}$$

is the largest invariant set of distributions.

The set $\mathcal{C}_\infty$ is non-empty because it obviously contains all invariant distributions of the matrices in $\mathcal{M}(P)$, and in the finite case invariant distributions always exist, although are not necessarily unique. Even though the invariant set of distributions is easy to find in principle, its shape can be very complicated and therefore approximations may be useful for practical purposes.

We have defined the invariant set of distributions as the limiting set of the sequence (10) starting with the set of all probability distributions. But this does not say anything about limiting set if the initial set is different. In Section 3 we show that the limiting set is unique and independent of the initial set $\mathcal{C}_0$ if a regularity condition is satisfied, which is the main result of this paper.

Approximating invariant distributions with interval probabilities

To approximate the invariant set of distributions with interval probabilities we try to find the F-field $Q = (\Omega, 2^\Omega, L)$ such that

$$L(A) = \inf_{q \in \mathcal{M}} \sum_{i=1}^{m} q_i L_i(A) \tag{17}$$

or in terms of lower expectations

$$L(A) = \underline{E}_{\mathcal{M}} X_A$$

where $X_A(\omega_i) = L_i(A)$. If the approximation with Choquet integral is used instead, the conditions become

$$L(A) = \int X_A \, dL \tag{18}$$

which is a system of linear equations with unknowns $L(A)$.

The minimal solution L of either of the sets of equations (17) or (18) approximates the largest invariant set of distributions $\mathcal{C}_\infty$ in the sense that all its members dominate L, or in other words, the set $\mathcal{C}_\infty$ is contained in the structure of the interval probability $(\Omega, 2^\Omega, L)$. This can be seen on the following way. Let $L^{(0)}$ be the lower probability with $L(A) = 0$ for every $A \subset \Omega$ and $L(\Omega) = 1$. It can be shown that both sequences of lower probabilities obtained through (12) and (13) starting with $L^{(0)}$ are monotone and therefore convergent. Clearly, their suprema are the minimal solutions of the equations (17) and (18) respectively. The inclusions $\mathcal{C}_n \subseteq \mathcal{M}^{(n)}$ for every $n \geq 0$ imply the required inclusion.

Example 2. We approximate the invariant set of distributions of the Markov chain with interval transition probability matrix given by the lower bound (9). We obtain the following solution to the system of equations (18):

$$L^{(\infty)} = (0.232, 0.2, 0.244, 0.581, 0.625, 0.6),$$

where $L^{(\infty)}$ is of the form (8). The intervals corresponding to the above solutions are then

$$P^{(\infty)} = ([0.232, 0.4], [0.2, 0.375], [0.244, 0.419], \\ [0.581, 0.756], [0.625, 0.8], [0.6, 0.768]).$$

The above lower bound is of course only an approximation (from below) of the true lower bound for the invariant set of distributions. For comparison we include the lower bound of the set of invariant distributions corresponding to 100,000 randomly generated matrices dominating P_L:

$$(0.236, 0.223, 0.275, 0.587, 0.628, 0.608).$$

Since all invariant distributions of the members of the structure $\mathcal{M}(P)$ must belong to the set $\mathcal{C}_\infty$, the above lower bound is an approximation from above of the true lower bound and yields the intervals:

$$([0.236, 0.392], [0.223, 0.372], [0.275, 0.413], \\ [0.587, 0.725], [0.628, 0.777], [0.608, 0.764]).$$

Thus, the lower bound of the true invariant set of distributions lies somewhere between the two approximations.

3 Convergence to equilibrium

3.1 Regular interval stochastic matrices

One of the main results of classical Markov chain theory is that chains with *irreducible* and *aperiodic* transition matrices always converge to a unique invariant distribution. Such transition matrices are sometimes called *regular*. In short, a transition matrix is regular if $p_{ij}^{(n)} > 0$ holds for all sufficiently large n, where $p_{ij}^{(n)}$ is the (i,j)-th entry of the matrix power P^n. Note that if all entries of P^r are strictly positive then also P^k, where $k > r$, has the same property. This follows from the properties of matrix multiplication and the fact that P has no zero rows. Therefore, a stochastic matrix is regular if all entries of P^r are strictly positive for at least one integer r.

If λ is any initial distribution and $(X_n)_{n\geq 0}$ is Markov (λ, P) with P regular then

$$P(X_n = j) \to \pi_j \quad \text{as } n \to \infty \text{ for all } j,$$

where π is the unique invariant distribution.

Regularity can similarly be defined for the case of Markov chains with interval probabilities. Let us first define the n-th power of an interval stochastic matrix P.

Definition 1. Let P be an interval stochastic matrix. We will call the set $\mathcal{P}^n = \{p_1 p_2 \dots p_n \mid p_i \in \mathcal{M}(P) \text{ for } i = 1, \dots, n\}$ the *n-th power of P*.

Note that the n-th power of an interval stochastic matrix is in general not an interval stochastic matrix, but a more general set of stochastic matrices, which are not easily tractable. Therefore, approximations in terms of interval probabilities will be useful. We also note that powers of interval stochastic matrices are associative in the sense that $\mathcal{P}^m\mathcal{P}^n = \mathcal{P}^{m+n}$, where the product of sets of matrices on the left hand side denotes the set of all products of matrices from corresponding sets.

Now we generalize the concept of regularity to interval stochastic matrices.

Definition 2. An interval stochastic matrix P is *regular* if there exists $n > 0$ such that $p_{ij} > 0$ for every $p \in \mathcal{P}^n$.

Clearly, this condition of regularity implies that every matrix in $\mathcal{M}(P)$ is regular, but inverse does not necessarily hold. In a similar way as in the classical case, it can be seen that if all matrices from $\mathcal{P}^n$ have strictly positive entries, then $\mathcal{P}^k$, where $k > n$, has the same property.

As we have pointed out before, powers of stochastic matrices as defined here are not easily tractable, thus checking regularity could be difficult in general. However, some simpler to check sufficient conditions easily follow from approximations presented before.

First we define two pseudo-powers for stochastic matrices that approximate powers from Definition 1. Both pseudo-powers are based on two operations similar to matrix multiplication, using approximations (12) and (13).

Definition 3. Let P be an interval stochastic matrix with lower probability matrix $P_L = [L_1 \dots L_m]^T$. Define $P_L^n = [L_1^n \dots L_m^n]^T$ where $L_i^1 = L_i$ and $L_i^n = \inf_{q \geq L_i^{n-1}} \sum_{i=1}^m q_j L_j(A)$ for $i = 1, \dots, m$ and $n \geq 2$.

Corollary 1. *If P is an interval stochastic matrix with lower probability matrix P_L such that $P_L^n = [L_1^n \dots L_m^n]^T$ and $L_i^n(A) > 0$ for every $i = 1, \dots, m$ and $A \subseteq \Omega,\ A \neq \emptyset$ then P is regular.*

Definition 4. Let P be an interval stochastic matrix with lower probability matrix $P_L = [L_1 \dots L_m]^T$. Define $\underline{P}_L^n = [\underline{L}_1^n \dots \underline{L}_m^n]^T$ where $\underline{L}_i^1 = L_i$ and $\underline{L}_i^n = \int L_j(A) d\underline{L}_i^{n-1}$ for $i = 1, \dots, m$ and $n \geq 2$, and the integral used is Choquet integral (as in (13)) .

Corollary 2. *If P is an interval stochastic matrix with lower probability matrix P_L such that $\underline{P}_L^n = [\underline{L}_1^n \dots \underline{L}_m^n]^T$ and $\underline{L}_i^n(A) > 0$ for every $i = 1, \dots, m$ and $A \subseteq \Omega,\ A \neq \emptyset$ then P is regular.*

The above corollaries present sufficient conditions for regularity because each power $\mathcal{P}^n$ as a set of stochastic matrices is contained within the structure of the corresponding pseudo-power, which is representable in terms of interval probabilities. Since powers from Definition 1 have no such representation, the sufficient conditions should be easier to check for pseudo-powers. Clearly, the sufficient condition in Corollary 2 implies the one in Corollary 1, but is much easier to check.

Even though the operations used in Definitions 3 and 4 resemble matrix multiplication, such a multiplication has an important weakness that it is not associative. But associativity is crucial in most methods concerning Markov chains and there is no obvious way to define an associative matrix multiplication for interval stochastic matrices, which is one of the main problems of the model.

3.2 Convergence to equilibrium

The main result of this section states that there is a unique compact set corresponding to a MCIP with a regular interval transition matrix to which its sets of distributions converge. To prove the theorem we use Banach fixed point theorem on the multivalued mapping between compact sets of probabilities corresponding to the transition matrix in Hausdorff metric.

Let (M, d) be a metric space. A mapping $T\colon M \to M$ is a *contraction* if there exists a constant $0 \leq k < 1$ such that $d(Tx, Ty) \leq k\, d(x, y)$ for all $x, y \in M$. If $k = 1$ is allowed in the above condition then the mapping T is said to be *non-expansive*.

An element $x \in M$ is a *fixed point* of an operator T if $T(x) = x$.

Theorem 1 (Banach fixed point theorem)**.** *Let (M, d) be a non-empty complete metric space and $T\colon M \to M$ a contraction. Then there exists a unique fixed point $x \in M$ of T. Furthermore, this fixed point is the limit of the sequence $\{x_n\}_{n\in\mathbb{N}}$ where $x_{i+1} = Tx_i$ and x_0 is an arbitrary element of M.*

Given a metric space M and non-empty compact subsets $X, Y \subset M$, *Hausdorff distance* is defined as

$$d_H(X, Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x, y), \sup_{y\in Y}\inf_{x\in X} d(x, y)\right\}.$$

This distance makes the set of non-empty compact sets a metric space $F(M)$. Moreover, if M is a compact space, so is $F(M)$ (see e.g. [1], p. 87). Note also that every compact metric space is complete.

To justify the use of Hausdorff metric, we show that all sets used are indeed compact. As the set of all probability distributions on a finite space is compact, we only have to note that the sets are closed. We start with a set of probabilities forming structure of an interval probability Q with lower probability L. Such a set is of the form $\mathcal{M}(Q) = \{q \mid q \text{ is a probability measure on } (\Omega, 2^\Omega), q \geq L\}$ and thus clearly closed and consequently compact. To see that $\mathcal{M}(P)$ is compact too, note that in topological sense it is a direct product of m structures corresponding to each row of P.

All sets of distributions corresponding to further steps are of the form $\mathcal{C}P = \{qp \mid q \in \mathcal{C}, p \in \mathcal{M}(P)\}$. Those sets are images of the compact sets $\mathcal{C} \times \mathcal{M}(P)$ with the continuous mapping $(q, p) \mapsto qp$, and are therefore compact too.

Proposition 1. *Let p be a stochastic matrix. Then the mapping from the set of all probability distributions $q \mapsto qp$ is non-expansive in the metric*

$$d(q, q') = \max_{A\subseteq\Omega} |q(A) - q'(A)| = \frac{1}{2}\sum_{\omega\in\Omega} |q(\omega) - q'(\omega)|.$$

Moreover, if $p_{ij} > 0$ for every $1 \leq i, j \leq m$ and $k = 1 - \inf_{1\leq i,j\leq m} p_{ij}$ then the mapping $q \mapsto qp$ is a contraction and

$$d(qp, q'p) \leq k\, d(q, q').$$

Proof. Take arbitrary $A \subseteq \Omega$ and let $p_i(A) = \sum_{\omega_j\in A} p_{ij}$. Further let q and q' be probability distributions on Ω with $q \neq q'$, and denote $B = \{\omega \in \Omega \mid q(\omega) \geq q'(\omega)\} \subsetneq \Omega$. Clearly, $k = \sup_{A\subsetneq\Omega} p_i(A)$ where $1 \leq i \leq m$.

We have

$$\begin{aligned}
&|qp(A) - q'p(A)| \\
&\quad= \left|\sum_{i=1}^m q_i p_i(A) - \sum_{i=1}^m q'_i p_i(A)\right| \\
&\quad= \left|\sum_{i=1}^m p_i(A)(q_i - q'_i)\right| \\
&\quad= \Bigg|\sum_{\omega_i\in B} p_i(A)|q_i - q'_i| \\
&\qquad\qquad - \sum_{\omega_i\notin B} p_i(A)|q_i - q'_i|\Bigg| \\
&\quad\leq \max\left\{\sum_{\omega_i\in B} p_i(A)|q_i - q'_i|,\right. \\
&\qquad\qquad \left.\sum_{\omega_i\notin B} p_i(A)|q_i - q'_i|\right\} \\
&\quad\leq \max\left\{\sum_{\omega_i\in B} k|q_i - q'_i|, \sum_{\omega_i\notin B} k|q_i - q'_i|\right\} \\
&\quad= k\max\left\{\sum_{\omega_i\in B} |q_i - q'_i|, \sum_{\omega_i\notin B} |q_i - q'_i|\right\} \\
&\quad\leq k\, d(q, q')
\end{aligned}$$

Since $k \leq 1$, the mapping is non-expansive. Furthermore, if $p_{ij} > 0$ for every $1 \leq i, j \leq m$ then $k < 1$ and thus the mapping is a contraction. □

The next proposition shows that the mapping $\mathcal{C} \mapsto \mathcal{C}\mathcal{P}^n$ is a contraction if P is a regular interval stochastic matrix and n is large enough.

Proposition 2. *Let P be a regular interval stochastic matrix and $n > 0$ an integer such that $p_{ij} > 0$ for every $p \in \mathcal{P}^n$ where $1 \leq i, j \leq m$. Let $k = 1 - \inf_{\substack{1\leq i,j\leq m\\ p\in\mathcal{P}^n}} p_{ij}$. The mapping $\mathcal{C} \mapsto \mathcal{C}\mathcal{P}^n = \{qp_1 \dots p_n \mid q \in \mathcal{C}, p_i \in \mathcal{M}(P) \text{ for } i = 1, \dots, n\}$ is then a contraction and*

$$d_H(\mathcal{C}\mathcal{P}^n, \mathcal{C}'\mathcal{P}^n) \leq k\, d_H(\mathcal{C}, \mathcal{C}').$$

Proof. By the assumption, $p_{ij} > 0$ for every $p \in \mathcal{P}^n$

and $1 \le i, j \le m$. We have

$$d_H(\mathcal{C}\mathcal{P}^n, \mathcal{C}'\mathcal{P}^n) = \max\left\{ \sup_{\substack{q\in\mathcal{C}\\ p\in\mathcal{P}^n}} \inf_{\substack{q'\in\mathcal{C}'\\ p'\in\mathcal{P}^n}} d(qp, q'p'), \sup_{\substack{q'\in\mathcal{C}'\\ p'\in\mathcal{P}^n}} \inf_{\substack{q\in\mathcal{C}\\ p\in\mathcal{P}^n}} d(qp, q'p') \right\}.$$

Take for instance

$$\begin{aligned} &\sup_{\substack{q\in\mathcal{C}\\ p\in\mathcal{P}^n}} \inf_{\substack{q'\in\mathcal{C}'\\ p'\in\mathcal{P}^n}} d(qp, q'p') \\ &\quad\le \sup_{p\in\mathcal{P}^n} \sup_{q\in\mathcal{C}} \inf_{q'\in\mathcal{C}'} d(qp, q'p) \\ &\quad\le \sup_{q\in\mathcal{C}} \inf_{q'\in\mathcal{C}'} k\, d(q, q') \\ &\quad\le k\, d_H(\mathcal{C}, \mathcal{C}'), \end{aligned}$$

where the second inequality follows from Proposition 1. Finally, this clearly implies $d_H(\mathcal{C}\mathcal{P}^n, \mathcal{C}'\mathcal{P}^n) \le k\, d_H(\mathcal{C}, \mathcal{C}')$. □

Finally we prove the main convergence theorem.

Theorem 2. *Let P be a regular interval stochastic matrix and $\mathcal{C}$ a compact set of probability distributions on $(\Omega, 2^\Omega)$ where Ω is a finite set. Then the sequence $\{\mathcal{C}\mathcal{P}^n\}_{n\in\mathbb{N}}$ converges in Hausdorff metric to a unique compact invariant set $\mathcal{C}_\infty$ that only depends on P and coincides with* (16).

Proof. Let $n > 0$ be an integer such that every $p \in \mathcal{P}^n$ satisfies $p_{ij} > 0$ for every $1 \le i, j, \le m$. By Proposition 2, the mapping $\mathcal{C} \mapsto \mathcal{C}\mathcal{P}^n$ is a contraction, and so, by Banach fixed point theorem, the sequence $\{\mathcal{C}(\mathcal{P}^n)^k\}_{k\in\mathbb{N}}$ converges to $\mathcal{C}_\infty$.

To see that the sequence $\{\mathcal{C}\mathcal{P}^k\}_{k\in\mathbb{N}}$ converges to the same set $\mathcal{C}_\infty$, we use associativity of powers of P. Thus, we have $\mathcal{C}\mathcal{P}^k = \mathcal{C}\mathcal{P}^r(\mathcal{P}^n)^s$, where $r < n$ and s goes to infinity as k goes to infinity. Since $\mathcal{C}\mathcal{P}^r$ is a compact set, the sequence converges to $\mathcal{C}_\infty$. □

3.3 Convergence of approximations

The limiting set of probabilities is computationally very difficult to find directly; therefore, approximations would be very useful in practice. Now we show that also a family of approximations converges independently from the initial distribution.

Proposition 3. *Let $P_L = [L_1 \dots L_m]^T$ be a lower transition probability matrix such that $L_i(A) < 1$ for every $1 \le i \le m$ and $A \subsetneq \Omega$. Let the mapping*

$$L \mapsto LP_L = L^*$$

be given, where L^* is the lower probability such that $L^*(A) = \int L_i(A)\, dL$. This mapping is then a contraction in the maximum distance metric

$$d(L, L') = \max_{A\subseteq\Omega} |L(A) - L'(A)|.$$

Further, if $k = \sup_{\substack{1\le i\le m\\ A\subsetneq\Omega}} L_i(A)$ then

$$d(LP_L, L'P_L) \le k\, d(L, L').$$

Proof. Take an arbitrary set $A \subsetneq \Omega$. We have:

$$\begin{aligned} &|LP_L(A) - L'P_L(A)| \\ &\qquad = \left| \int L_i(A)\, dL - \int L_i(A)\, dL' \right|. \end{aligned}$$

Let π be a permutation such that $L_{\pi(i)}(A) \ge L_{\pi(i+1)}(A)$ for every $1 \le i \le m$ and denote $S_i = \{\pi(1), \dots, \pi(i)\}$ and $x_i = L_{\pi(i)}(A)$ where $x_{m+1} = 0$. The above Choquet integrals can then be transformed into (see [3])

$$\begin{aligned} &\left| \int L_i(A)\, dL - \int L_i(A)\, dL' \right| \\ &= \left| \sum_{i=1}^m (x_i - x_{i+1}) L(S_i) - \sum_{i=1}^m (x_i - x_{i+1}) L'(S_i) \right| \\ &= \left| \sum_{i=1}^m (x_i - x_{i+1})(L(S_i) - L'(S_i)) \right| \\ &\le \sum_{i=1}^m (x_i - x_{i+1}) d(L, L') \\ &= x_1\, d(L, L') \\ &\le k\, d(L, L'). \end{aligned}$$

Thus, $d(LP_L, L'P_L) = \max_{A\subseteq\Omega} |LP_L(A) - L'P_L(A)| \le k\, d(L, L')$, which completes the proof. □

In previous sections we approximated sets of distributions corresponding to consecutive steps by lower probabilities $L^{(n)}$ where $L^{(0)} = L$ is the initial lower probability and $L^{(n)}(A) = \int L_i(A)\, dL^{(n-1)}$. A problem with this approximation is its non-associativity, but associativity was crucial in the proof of Theorem 2.

Because of this inconvenience we can only prove a convergence theorem for a slightly different approximation. The construction easily implies that for every $n > 0$ the pseudo-power $\underline{P}_L^n$ approximates the n-th power of P in the sense that every $p \in \mathcal{P}^n$

satisfies $p \geq \underline{P}_L^n$. Therefore, the sequence of lower probabilities $\{L^{(kn)}\}_{k\in\mathbb{N}}$ defined by $L^{(0)} = L$ and $L^{(kn)} = L^{((k-1)n)}\underline{P}_L^n$, where L is the initial lower probability, approximates the sets of distributions at kn-th steps in the sense that $p \geq L^{(kn)}$ for every $p \in \mathcal{C}_{kn}$.

Now we give a convergence theorem for those approximations.

Theorem 3. *Let P be an interval stochastic matrix with the lower bound P_L such that $\underline{P}_L^n = [\underline{L}_1^n \dots \underline{L}_m^n]^T$ satisfies $\underline{L}_i^n(A) < 1$ for every $1 \leq i \leq m$ and $A \subsetneq \Omega$. Further let L be any lower probability on $(\Omega, 2^\Omega)$. Then the sequence $\{L^{(kn)}\}_{k\in\mathbb{N}}$ converges to a unique lower probability $L^{(\infty)}$ that only depends on $\underline{P}_L^n$.*

Proof. By Proposition 3 the mapping $L \mapsto L\underline{P}_L^n$ is a contraction in the maximum distance metric. By Banach fixed point theorem the sequence $\{L^{(kn)}\}_{k\in\mathbb{N}}$ converges to a unique lower probability $L^{(\infty)}$ which is the fixed point for the mapping $L \mapsto L\underline{P}_L^n$. □

4 Conclusion

Results in the paper show that even if the assumptions of Markov chain model are substantially relaxed, the behaviour remains similar as in the most widely used model with constant precisely known initial and transition probabilities. However, several interesting questions still remain open. Especially those related to approximations of the intractable true sets of distributions with convex sets representable with interval probabilities.

Acknowledgements

I wish to thank the referees for their helpful comments and suggestions.

References

[1] G. Beer. *Topologies on closed and closed convex sets.* Kluwer Academic Publishers, Dordrecht, 1993.

[2] M. A. Campos, G. P. Dimuro, A. C. da Rocha Costa, and V. Kreinovich. Computing 2-step predictions for interval-valued finite stationary Markov chains. Technical Report UTEP-CS-03-20a, University of Texas at El Paso, 2003.

[3] D. Denneberg. *Non-additive measure and integral.* Kluwer Academic Publishers, Dordrecht, 1997.

[4] O. Kosheleva, M. Shpak, M. A. Campos, G. P. Dimuro, and A. C. da Rocha Costa. Computing linear and nonlinear normal modes under interval (and fuzzy) uncertainty. *Proceedings of the 25th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2006*, 2006.

[5] I. Kozine and L. V. Utkin. Interval-valued finite Markov chains. *Reliable Computing*, 8(2):97–113, 2002.

[6] J. Norris. *Markov Chains.* Cambridge University Press, Cambridge, 1997.

[7] D. Škulj. Finite discrete time Markov chains with interval probabilities. *Soft methods for integrated uncertainty modelling, (Advances in soft computing)*, pages 299–306, 2006.

[8] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, New York, 1991.

[9] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000.

[10] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I – Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica-Verlag, Heidelberg, 2001.

Minimax Regret Treatment Choice with Finite Samples and Missing Outcome Data

Jörg Stoye
New York University
j.stoye@nyu.edu

Abstract

This paper uses the minimax regret criterion to analyze choice between two treatments when one has observed a finite sample that is plagued by missing data. The analysis is entirely in terms of exact finite sample regret, as opposed to asymptotic approximations or finite sample bounds. It thus extends Manski (2007), who largely abstracts from finite sample problems, as well as Stoye (2006a), who provides finite sample results but abstracts from missing data. Core findings are: (i) Minimax regret is achieved by randomizing over two rules that were identified in the aforecited papers. (ii) For every sample size, there exists a sufficiently small (but positive) proportion of missing data such that if less data are missing, the missing data problem is ignored altogether and Stoye's (2006a) results apply. (iii) For every positive fraction of missing data, the value of additional observations drops to zero at a finite sample size. I also provide the decision problem's value function and briefly touch on optimal sample design as well as unknown propensity scores.

Keywords. Minimax regret, missing data, statistical decision theory, partial identification, treatment evaluation.

1 Introduction

Consider a planner who has to decide whether to assign a binary treatment – e.g., a medical treatment or a labor market intervention – to members of some target population. She can base her choice on observations of outcomes experienced by a sample of subjects, some of whom received the treatment and some of whom served as control group. The signal generated by these observations has two limitations: First, it is generated by a finite sample. Second, it is assumed that some data are missing, that is, a subset of the target population is not represented in the sample, and members of this subset may react to treatment differently from the observable subjects. Thus, the choice scenario simultaneously generates finite sample problems (a standard issue in statistics and econometrics) and problems of incomplete identification (a less standard issue; see Manski 2003 for a survey).

I analyze this situation using minimax regret with respect to expected outcomes as optimality criterion. Importantly, the analysis is entirely in terms of exact finite sample regret, as opposed to asymptotic approximations (as in Hirano and Porter 2005) or bounds on finite sample quantities (as in Manski 2004). It thus extends, and connects, Manski (2007), who analyzes a somewhat more general case but largely abstracts from finite sample problems, and Stoye (2006a), who provides finite sample results but abstracts from missing data.

In fact, both of the aforementioned papers analyze special cases of the present scenario, and their results are linked in a specific way here. Stoye (2006a), by ignoring missing data, characterizes finite sample minimax regret rules for the boundary case where the proportion of missing data vanishes. Manski (2007) provides a finite sample minimax regret rule if at least half the data are missing. The respective solutions are quite different from each other. The perhaps surprising upshot of this paper is that minimax regret can generally be achieved by randomizing over them, where the mixture is degenerate on the previous two findings' domains but creates a smooth transition in between. For the case where the proportion of observable data p is known a priori, two intriguing aspects of the result are the following:

- For every sample size N, there exists a critical p_N^* such that if $p \geq p_N^*$, then the presence of missing data is ignored altogether and the treatment rules from Stoye (2006a) apply. Those decision rules therefore have a certain degree of robustness to missing data.
- The minimax regret value of the decision problem

exhibits nonstandard asymptotic behavior: For every p, the limiting value of regret is exactly achieved beyond some finite N. Thus the value of additional observations drops to zero at some finite sample size.

If p is not known a priori, minimax regret is achieved by presuming that p equals the lowest value that the decision maker considers possible. In particular, if p cannot be bounded away from zero a priori, then minimax regret is achieved by a "no-data rule."

The remainder of this paper is structured as follows. I first set up the decision problem, introduce notation, and provide a brief motivation for minimax regret. The heart of this paper is section 2.2, which provides relevant results from the aforecited papers and then their joint generalization. In section 2.3, I show how to compute the decision problem's value function, section 2.4 briefly discusses optimal sample design, and section 2.5 considers unknown p. Section 3 concludes, and the appendix collects all proofs.

2 Analysis of the Treatment Choice Problem

2.1 Setup and Notation

There is a binary treatment, $T \in \{0,1\}$, that must be assigned, possibly at random, to members of a target population. Two classic examples are clinical trials, where the target population would be all people who suffer from a certain condition, $T = 1$ would denote a medical innovation, and $T = 0$ would be the status quo treatment, and job training for the unemployed, where $T = 1$ would denote training and $T = 0$ no training. To model treatment effects, I use the standard "potential outcomes" notation (Rubin 1974): For every member of the target population, the random variable $Y_1 \in [0,1]$ denotes the outcome that she would experience if assigned to treatment, whereas $Y_0 \in [0,1]$ is the outcome she would experience if assigned to the control group.[1] Of course, only one of the two random variables will be actualized; the other realization remains counterfactual.

The decision maker observed outcomes experienced in a size N simple random sample from a sample population. Members of the sample were assigned treatment according to some design that will initially be taken as given; the question of optimal sample design is considered later. The sample generates an imperfect signal for two reasons: First, it is finite, and random variation in observed outcomes will be fully considered. Second, only a subset of the target population is observable. I model this by presuming that the sample population is a subset of relative probability mass p of the target population. Importantly, it is assumed that while (Y_0, Y_1) is distributed identically across the sample population, its distribution in the unobservable part of the target population can be different. A leading example is if a study was performed on volunteers who might not be fully representative of the target population. As a consequence, the distribution of (Y_0, Y_1) would only be partially revealed even by an infinitely large sample. This is why the problem is inherently a decision problem under ambiguity, and very similar in structure to interval probability problems as well as robust Bayesian inference. See, in particular, Manski (2002, 2005). Previous analyses of the same problem either largely abstracted from the finite sample problem (Manski 2007) or from the ambiguity caused by missing data (Stoye 2006a).

To model the problem, let the random variable $Z \in \{0,1\}$ indicate whether a member of the target population is in the sample population ($Z = 1$) or not ($Z = 0$). Define the random variables $Y_{tz} \equiv (Y_t | Z = z)$ and write $p \equiv \Pr(Z = 1)$, the proportion of observable subjects in the population. I initially assume that p is known. Then a *state of nature* s can be identified with a true distribution of $(Y_{01}, Y_{00}, Y_{11}, Y_{10})$. Assume that a priori bounds on Y_0 and Y_1 are finite, coincide, and that there are no restrictions on their joint distribution, then it is without further loss of generality to set the state space $\mathcal{S}$ equal to $\Delta([0,1]^4)$, the set of distributions over $[0,1]^4$. Most of the discussion will actually restrict outcomes to be binary, i.e. set $\mathcal{S} = \Delta\left(\{0,1\}^4\right)$, more on which below. It is worth noting that $(Y_{01}, Y_{00}, Y_{11}, Y_{10})$ are not restricted to be independent. I will use the following notational conventions: If Y_i is a random variable, then μ_i denotes its expectation and $\overline{y}_i$ a sample mean.

The sample is a simple random sample from the sample population. For any sample point, one treatment is assigned according to the sample design and the according outcome is observed, thus the decision maker sees realizations (t, y_{t1}). Let $\mathcal{S}_N = (\{0,1\} \times [0,1])^N$, with typical element s_N, denote the sample space induced by a sample of size N, i.e. the collection of possible sample realizations. The decision maker has to choose a treatment rule $\delta : \mathcal{S}_N \to [0,1]$ that maps possible sample outcomes into probabilities of assigning treatment 1. In particular, she is allowed to randomize. The set of decision rules δ is labelled $\mathcal{D}$.

Any combination of state and decision rule induces an

[1] The restriction to $[0,1]$ is w.l.o.g. if, and only if, some bounds on outcomes are known a priori.

expected outcome

$$\begin{aligned} u(\delta, s) &\equiv \mu_1 \mathbb{E}\delta(s_N) + \mu_0 \left(1 - \mathbb{E}\delta(s_N)\right) \\ &= \left(p\mu_{11} + (1-p)\mu_{10}\right) \mathbb{E}\delta(s_N) \\ &\quad + \left(p\mu_{01} + (1-p)\mu_{00}\right) \left(1 - \mathbb{E}\delta(s_N)\right) \end{aligned}$$

Here, $\mathbb{E}\delta(s_N)$ is evaluated given s; although suppressed in the notation, it will also depend on the sample design. Strictly speaking, u is already a risk function with respect to an underlying loss function $L(y_t) = -y_t$. I will take for granted that if s were known, treatments rules would be evaluated according to u. With unknown s, the efficacy of δ will be measured in terms of *minimax regret* relative to u, thus a treatment rule δ^* is optimal if

$$\begin{aligned} \delta^* &\in \arg\min_{\delta \in \mathcal{D}} \left\{ \max_{s \in \mathcal{S}} R(\delta, s) \right\}, \\ R(\delta, s) &\equiv \max_{\delta' \in \mathcal{D}} \left\{ u(\delta', s) \right\} - u(\delta, s). \end{aligned}$$

The minimax regret criterion minimizes worst-case performance relative to the ex-post optimal expected outcome or, equivalently, relative to the performance of an infeasible "oracle" treatment rule that utilizes full knowledge of $\mathbb{E}(Y_{01}, Y_{00}, Y_{11}, Y_{10})$. Minimax regret was originally suggested by Savage (1951). In the present formulation – which is not the only possible one – it was recently reconsidered in statistics and related fields (Droge 1998, 2006; Eldar et al. 2003; Hirano and Porter 2005; Manski 2004, 2005, 2007; Schlag 2003, 2006; Stoye 2006a, 2007).[2] A motivation for it is that it avoids the imposition of priors and optimizes against states of the world in which the decision maker's action has a large effect. This sets it apart from its main competitors: The Bayesian decision rule requires specification of a subjective prior over states; maximin utility also avoids priors but optimizes against states in which outcomes are very bad, irrespective of whether they are affected by actions. For a historical overview and further heuristic as well as axiomatic discussion, see Stoye (2006b).

Of course, there are many possible sample designs. I will focus on those considered in Stoye (2006a); they may serve as stylized models of real-world sampling schemes and will turn out to be minimax regret optimal. By *stratified assignment*, I henceforth mean that N is even and that exactly half of the sample is allocated to treatment 1. By *randomized assignment*, I mean that sample points are assigned to treatments by independent tosses of a fair coin.

Some comments on this setup are in order.

- In this paper, p cannot depend on t, and N is not a random variable. The story behind this setting is that missing data occur before treatments are assigned, an example being selection of subjects into experimental pools. Manski (2007) considers the more general case where attrition from an experimental pool can be selective by, and potentially in reaction to, treatment assignment. Unfortunately, finite sample analysis of this case is extremely involved, because sample composition becomes a random variable whose exact distribution must be taken into account. Although one specific such case is analyzed below, a general treatment is left to future research.

- The below results presume binary outcomes, i.e. $Y_0, Y_1 \in \{0, 1\}$. For lemma 2, it will be pointed out that this is not necessary. For the other cases, minimax regret treatment rules for $Y_0, Y_1 \in [0, 1]$ can – under regularity conditions on the state space – be generated by a technique due to Schlag (2003, 2006). The trick, which will be called *binary randomization*, is to replace every sample realization y_i by the outcome of one independent toss of a coin with parameter y_i and then apply the below treatment rules to the resulting, binary samples.

- Covariates are not introduced into this paper's notation, but the results immediately extend to the case of finite-valued covariates by means of proposition 3 in Stoye (2006a). Specifically, let there be a covariate X and let the sample be stratified by covariate, then minimax regret is achieved by applying the below treatment rules separately across covariates. For treatment assignment conditional on $X = x$, the treatment rule therefore utilizes only the subsample with covariate value x. The surprising aspect of this is that there is no inference across covariates. See Stoye (2006a) for an in-depth discussion.

- The decision problem can clearly be interpreted as an imprecise probability problem. The present specification respresents a very special case, however, because complete ignorance about true probabilities is presumed. Prior information can be introduced by restricting the state space $\mathcal{S}$, as is done in Stoye (2006a). This poses no conceptual difficulties but may, of course, change computations.

The remainder of the paper is concerned with finding δ^* for different decision scenarios. The proofs exploit the fact that δ^* can be represented as the decision maker's equilibrium strategy in a fictitious zero-sum

[2] Minimax regret is also closely related to the competititve ratio; indeed, it could as well be called *competititve difference*.

game against Nature. This allows one to infer existence of a minimax regret treatment rule from known game theoretic results (Glicksberg 1952). Other than that, it just restates that the minimax regret decision rule can be characterized as a saddle point, but the game theoretic interpretation facilitates the import of heuristics and solution strategies developed by economists.

2.2 Treatment Rules

The first step is to analyze the aforementioned boundary cases, namely $p = 1$ and $p \leq 1/2$.

Lemma 1 *If $p = 1$, minimax regret is achieved by*

$$\delta_1^* \equiv \begin{cases} 0, & I_N < 0 \\ 1/2, & I_N = 0 \\ 1, & I_N > 0 \end{cases},$$

where

$$\begin{aligned} I_N \equiv & \#\left(\textit{observed successes of treatment } 1\right) \\ & +\#\left(\textit{observed failures of treatment } 0\right) \\ & -\#\left(\textit{observed failures of treatment } 1\right) \\ & -\#\left(\textit{observed successes of treatment } 0\right) \\ \propto & N_1(\overline{y}_{11} - 1/2) - N_0(\overline{y}_{01} - 1/2), \end{aligned}$$

where N_t is the number of sample subjects assigned to treatment t. For a stratified sample design, this is equivalent to

$$\delta_1^* \equiv \begin{cases} 0, & \overline{y}_{11} < \overline{y}_{01} \\ 1/2, & \overline{y}_{11} = \overline{y}_{01} \\ 1, & \overline{y}_{11} > \overline{y}_{01} \end{cases}.$$

Lemma 2 *If $p \leq 1/2$, minimax regret is achieved by*

$$\delta_2^* \equiv \frac{1}{2} + \frac{p}{2(1-p)} \frac{I_N}{N}.$$

This applies for either stratified or randomized sample design; in the former case, it can be rewritten as

$$\delta_2^* \equiv \frac{1}{2} + \frac{p\left(\overline{y}_{11} - \overline{y}_{01}\right)}{2(1-p)} = \frac{(p\overline{y}_{11} + 1 - p) - p\overline{y}_{01}}{2(1-p)}$$

as in Manski's (in press) proposition 2.

Lemma 1 is from Stoye (2006a, proposition 1). Lemma 2 follows from Manski (2007, proposition 2) for stratified samples but not for randomized ones. The latter design allows for empty sample cells, a case that Manski has to exclude. The generalization presented here is new.

Also, while lemma 2 is here stated for $Y_0, Y_1 \in \{0, 1\}$, inspection of the proof reveals that δ_2^* can be extended to $Y_0, Y_1 \in [0, 1]$ by using the above definition of I_N in terms of $(N_0, N_1, \overline{y}_{01}, \overline{y}_{11})$. This is not the rule that would emerge from applying the binary randomization technique and then operating δ_2^* on the resulting, binary sample; in particular, it randomizes with probability 0 if (Y_{01}, Y_{11}) has a continuous distribution. This illustrates that minimax regret treatment rules need not be unique.

The next lemma and definition set the stage for the general problem, i.e. $p \in [0, 1]$. From Manski (2007, see also Stoye 2007), we know a minimax regret decision rule for the limiting case where the expectations (μ_{01}, μ_{11}) of (Y_{01}, Y_{11}) are known.

Lemma 3 *Let (μ_{01}, μ_{11}) be known, then minimax regret is achieved by*

$$\begin{aligned} \delta_3^* \equiv & \begin{cases} 0, & \delta < 0 \\ \delta, & 0 \leq \delta \leq 1 \\ 1, & 1 < \delta \end{cases} \\ \delta \equiv & \frac{1}{2} + \frac{p}{2(1-p)}(\mu_{11} - \mu_{01}). \end{aligned}$$

This rule is essentially the population analog of δ_2^*; it just adds a truncation to insure that $\delta_3^* \in [0, 1]$. As final preliminary step, I define its sample analog:

Definition 1 *The sample analog of δ_3^* is*

$$\begin{aligned} \delta_4^* \equiv & \begin{cases} 0, & \delta < 0 \\ \delta, & 0 \leq \delta \leq 1 \\ 1, & 1 < \delta \end{cases} \\ \delta \equiv & \frac{1}{2} + \frac{p}{2(1-p)} \frac{I_N}{N}. \end{aligned}$$

To accommodate both sample designs, δ_4^* is based on I_N/N rather than $(\overline{y}_{11} - \overline{y}_{01})$. Of course, these expressions coincide under the stratified design.

I am now ready to state this paper's main result.

Theorem 4 *Consider any fixed $N < \infty$ as well as $p \in (0, 1]$. Then minimax regret is achieved by the following randomization over δ_1^* and δ_4^*:*

$$\delta^* \equiv \begin{cases} \delta_1^* & \textit{with probability } \alpha^* \\ \delta_4^* & \textit{with probability } (1 - \alpha^*) \end{cases},$$

where

$$\begin{aligned}
\alpha^* &\equiv \min\left\{\frac{\frac{p}{2(1-p)}-A}{B-A},1\right\}\\
A &\equiv 2^{-N^*}\sum_{n\geq N^*(1-\frac{1}{2p})}\left(\begin{array}{c}N^*\\ n\end{array}\right)(2n-N^*)\\
&\quad\times\min\left\{\frac{1}{2}+\frac{p}{2(1-p)}\frac{2n-N^*}{N^*},1\right\}\\
B &\equiv 2^{-N^*}\sum_{n>N^*/2}\left(\begin{array}{c}N^*\\ n\end{array}\right)(2n-N)
\end{aligned}$$

and

$$N^*=\left\{\begin{array}{ll}N, & N \text{ is odd}\\ N-1, & N \text{ is even}\end{array}\right. .$$

In particular, α^* equals 1, and the decision rule therefore collapses to δ_1^*, iff $p\geq p_N^*\equiv\frac{2B}{2B+1}$. This threshold value converges to 1 as $N\to\infty$. On the other hand, α^* equals 0, and the decision rule therefore collapses to δ_4^*, iff $p\leq 1/2$.

If $p=0$, then $\delta^*=1/2$.

Substantively, it turns out that minimax regret is achieved by randomizing over δ_1^* and δ_4^*. In words, the decision maker should toss a (biased) coin and then use rule δ_1^* if the coin came up head. The randomization parameter α^* changes with p and N in interesting ways:[3]

- For any given N, δ_4^* applies for $p\leq 1/2$ and its weight then decreases, with δ_1^* being attained for $p\geq p_N^*$, a value that is strictly below 1. Thus for every N, a sufficiently small but nonzero mass of missing data can be ignored. Although p_N^* converges to 1 at rate $N^{-1/2}$, it significantly differs from 1 for rather large N, so that δ_1^*, which was developed for fully observable data, exhibits quite some robustness to missing data.
- For any given $p\in(1/2,1)$, the randomization changes with N as follows: For N small enough, the presence of missing data is ignored, i.e. $\alpha^*=1$, but α^* converges to zero as N grows, so that the limit rule is approximated (but not attained) for large N. Again, the convergence of α^* to 0 is perhaps surprisingly slow; it is also nonuniform in p. It should be pointed out, however, that for any N, δ_4^* becomes similar to δ_1^* as $p\to 1$; thus, it does not follow that convergence of the treatment rule to its limit is "slow" (or nonuniform in p) in every interesting metric.

[3]MATLAB code that evaluates α^* is available on the author's webpage at `http://homepages.nyu.edu/~js3909`.

2.3 Value Function

By evaluating regret on the fictitious game's equilibrium path, one can find the minimax regret achievable under either treatment assignment rule.

Proposition 5 *For either treatment assignment scheme, the decision problem has minimax regret value $(1-p)/2$ if $p<p_N^*$ and*

$$\max_{a\in[1/2,1]}\left\{\begin{array}{c}(2p(a-1)+1)\\ \times\sum_{n<\frac{N^*}{2}}\left(\begin{array}{c}N^*\\ n\end{array}\right)a^n(1-a)^{N^*-n}\end{array}\right\}$$

otherwise, where N^* is as in theorem 4. In particular, if $p\leq 1/2$, then the minimax regret value equals $(1-p)/2$ for any N.

As in Stoye (2006a) and Schlag (2006), learning only occurs with every other sample point. Unlike in those papers, learning is incomplete: As $N\to\infty$, regret does not converge to zero but to $(1-p)/2$. This reflects the fact that in the presence of missing data, even the asymptotic decision problem will generate positive regret. An especially unusual feature is that learning occurs only as long as $p\geq p_N^*\Leftrightarrow\alpha^*=1$. When this region of parameter space is left, regret "locks in" at its limiting value.[4] Recall that this occurs for some finite sample size for any $p<1$; what's more, it is the case for *any* sample size if $p\leq 1/2$. This insight generalizes Manski's (2007) finding that when at least half of the data are missing, minimax regret is independent of sample size. More generally, the presence of any missing data whatsoever means that the limiting decision quality is exactly attained for some finite N, and additional observations are useless beyond that threshold. At least from an econometrician's perspective, this finding is unexpected.

2.4 Optimal Sample Design

The preceding analysis took two different sample designs as given. They turn out to generate the same maximal expected regret whenever both are feasible, i.e. when N is even. An obvious question is whether this regret is optimal when sample design is itself a choice variable. The answer is in the affirmative.

[4]The intuition is that in the fictitious game, the switch to $\alpha^*<1$ marks the transition to a *pooling equilibrium* in which the signal generated by sample data is noninformative about the true state of the world. Hence, the decision maker ceases to learn from the signal.

To formalize this idea, let $h_n \equiv (t^i, y^i_{t1})^n_{i=1}$ denote the sample history up to realization n (with the understanding that $h_0 = \varnothing$). If assignment of treatments to sample points is a choice variable, this can be modelled by letting the decision maker choose a vector of mappings $\tau = (\tau_n)^N_{n=1}$, where $\tau_n(N, h_{n-1}) \in [0,1]$ specifies the probability of assigning treatment 1 to sample point n conditional on history h_{n-1} in a sample of overall size N. The randomized assignment scheme corresponds to $\tau^{\text{rand}}_n = 1/2, \forall n$, whereas one way to realize the stratified sample design is to set $\tau^{strat}_n \equiv \mathbb{I}\{n \text{ is even}\}$. Observe that the sampling scheme τ may depend on sample history, but that the decision maker has to specify it before seeing any sample points. This is a common formalization in econometrics because it corresponds to the concept of risk functions in statistical decision theory, but also because it is often realistic for the problems that economists consider. For example, assignment to job training is typically planned by the researcher but executed by caseworkers or other third parties. However, if one wanted to model an online problem, one might also want to allow the decision maker to re-optimize τ along the sample path, which does not in general lead to the same problem.[5]

Proposition 6 *Both τ^{rand} and τ^{strat} (when applicable) are minimax regret optimal assignment schemes.*

This result and proposition 3 in Stoye (2006a) jointly imply that if one faces a random sample from a population with a finite-valued covariate and can choose the sample design, then one can achieve minimax regret by using the randomized assignment scheme and applying theorem 4 separately across covariates. Compared to Manski (2007), this result applies to a narrower range of missing data scenarios but is more general on two other dimensions: p may exceed $1/2$, and the treatment scheme is defined even when some sample cells are empty.

2.5 Treatment Choice with Unknown Propensity Score

I now turn to the case where p is not known a priori but has to be learned from the data. Thus, assume that p can merely be restricted to lie in an interval $\left[\underline{p}, \overline{p}\right] \subseteq [0,1]$. The sample size N continues to refer to the number of observed units. This leaves open the question of how exactly learning about p happens as samples are realized. It turns out that this question is irrelevant: Minimax regret can be achieved by presuming that p equals $\underline{p}$.

Proposition 7 *Consider the setting of theorem 4, except that $p \in \left[\underline{p}, \overline{p}\right]$. Then minimax regret is achieved by setting $p = \underline{p}$ and applying δ^*.*

The intuition for this result is straightforward: As can be seen from proposition 6, minimax regret decreases in p. This is also intuitive since a higher p means that signals are representative of a larger part of the target population, and should thus be more informative. It implies that the worst case scenario is given by $p = \underline{p}$.

Proposition 4 has an unsettling implication: If $\underline{p} = 0$, minimax regret is achieved by setting $\delta^* = 1/2$ irrespective of the observed sample, i.e. by a "no-data rule." As observed by Savage (1954) and recently by Manski (2004), Schlag (2003), and Stoye (2006a), no-data rules are a frequent problem with the maximin utility criterion. Only Stoye (2006a) previously found similar problems to arise with minimax regret. Proposition 7 provides another, realistic problem that leads to a no-data minimax regret treatment rule. It suggests that the issue of no-data rules may not constitute the most compelling argument for minimax regret over maximin utility.

3 Summary and Outlook

This paper added to the recently growing literature on minimax regret and specifically to research by Manski (2007) and Stoye (2006a). It provided a joint generalization of much of those papers' analyses by considering a treatment choice problem where information is incomplete in two ways, firstly because of finite sample variation but also, and more fundamentally, because of missing data and hence incomplete identification of population distributions. The core finding is that results by Manski (2007) and Stoye (2006a) can be linked in a particular way: Each of them identifies a minimax regret treatment rule for a boundary case of the present problem, and a smooth transition between these solutions is generated by randomizing over them. This insight also strengthens the general finding that minimax regret tends to prescribe randomization, a point stressed by Schlag (2003, 2006). The result was extended by presenting the decision problem's value function, by allowing for unknown or partially known propensity scores, and by showing optimality of certain sample designs.

Many questions remain open on minimax regret treat-

[5] One might think that the difference cannot matter, because the definition of τ allows the decision maker to prescribe reactions to sample observations. In fact, this depends on how the decision maker reacts to the arrival of information, that is, on her updating rule. Under the most intuitive such rule, namely pointwise Bayesian updating of the state space, both minimax regret and maximin utility are dynamically inconsistent, meaning that the conjecture is false. What's more, the present choice of suppressing updating then appears not only realistic but sensible (e.g., Augustin 2003). Hanany and Klibanoff (2005) provide an updating rule that renders the conjecture true.

ment choice. For example, Stoye (2007) generalizes Manski (2007) in a different direction, namely by allowing for a multi-valued treatment. This generalization could be further extended by considering finite samples. The same remark holds for additional results that Stoye (2006a) presents with respect to covariates and the effect of restricting $\mathcal{S}$, as well as Manski's (2007) consideration of sample attrition that may vary by assigned treatment. When designing sample designs, one could also consider the possibility that outcomes experienced by the sample subjects are taken into account when evaluating the design. This possibility was ignored here for simplicity; it leads to intricate "bandit" problems as in Schlag (2003).

Acknowledgements

I thank the referees, the program chairs, Gary Chamberlain, and seminar audiences at Harvard/MIT and Pittsburgh for helpful comments.

A Proofs

Preliminaries Most proofs proceed by analyzing the following zero-sum game: (i) The decision maker (DM) chooses a statistical treatment rule $\delta : \theta \to [0,1]$, Nature chooses a mixed strategy $\sigma \in \Delta(\mathcal{S})$ over states. (ii) A neutral meta-player draws s according to σ, then θ according to s. (iii) DM's payoff is $\int R(\delta, s) d\sigma$. This game is useful because of the following fact (e.g., Berger 1985).

Lemma 8 *Assume that $\sigma^* \in \Delta(\mathcal{S})$ and δ^* are such that (δ^*, σ^*) is a Nash equilibrium of the above game, that is, $\delta^* \in \arg\min_{\delta \in \mathcal{D}} \int R(\delta, s) d\sigma^*$ and $\sigma^* \in \arg\max_{\sigma \in \Delta(\mathcal{S})} \int R(\delta^*, s) d\sigma$. Then δ^* is a minimax regret treatment rule.*

Proofs will, therefore, proceed by conjecturing and then verifying Nash equilibria of the fictitious game (as also in Schlag 2003, 2006, and Stoye 2006a, 2007).

Lemma 1 See Stoye (2006a, propositions 1 and 2).

Lemma 2 Consider first the randomized treatment assignment scheme. Assume that DM plays δ^*, then any $P(Y_{01}, Y_{00}, Y_{11}, Y_{10})$ in the support of σ^* must maximize $R(\delta^*, s)$. Expansion of $R(\delta^*, s)$ yields

$$\begin{aligned}\max\{&(p\mu_{11} + (1-p)\mu_{10} - p\mu_{01} - (1-p)\mu_{00}) \\ &\times \mathbb{E}\left(\frac{1}{2} + \frac{p}{2(1-p)}\frac{I_N}{N}\right), \\ &(p\mu_{01} + (1-p)\mu_{00} - p\mu_{11} - (1-p)\mu_{10}) \\ &\times \mathbb{E}\left(\frac{1}{2} - \frac{p}{2(1-p)}\frac{I_N}{N}\right)\}.\end{aligned}$$

Simple calculations show that distribution of I_N depends on $P(Y_{01}, Y_{00}, Y_{11}, Y_{10})$ only through (μ_{01}, μ_{11}). Thus $R(\delta^*, s)$ depends on $P(Y_{01}, Y_{00}, Y_{11}, Y_{10})$ only through $(\mu_{01}, \mu_{00}, \mu_{11}, \mu_{10})$. Furthermore, symmetry of the two components of the max-operator means that $(\mu_{01}, \mu_{00}, \mu_{11}, \mu_{10}) = (a, b, c, d)$ maximizes $R(\delta^*, s)$ iff $(\mu'_{01}, \mu'_{00}, \mu'_{11}, \mu'_{10}) \equiv (c, d, a, b)$ does. One can thus construct a best response to δ^* by finding some $(\mu^*_{01}, \mu^*_{00}, \mu^*_{11}, \mu^*_{10})$ that maximizes

$$\begin{aligned}&(p\mu_{11} + (1-p)\mu_{10} - p\mu_{01} - (1-p)\mu_{00}) \\ &\qquad \times \mathbb{E}\left(\frac{1}{2} + \frac{p}{2(1-p)}\frac{I_N}{N}\right)\end{aligned}$$

and presuming that Nature randomizes evenly between $(\mu^*_{01}, \mu^*_{00}, \mu^*_{11}, \mu^*_{10})$ and its symmetric counterpart. I will now find $(\mu^*_{01}, \mu^*_{00}, \mu^*_{11}, \mu^*_{10})$ and then verify that δ^* is a best response to Nature's strategy.

In the proof of proposition 1(ii) in Stoye (2006a), it is established that the distribution of I_N depends on (μ_{01}, μ_{11}) only through $\mu_{11} - \mu_{01}$. Without loss of generality, I therefore presume that $(\mu_{01}, \mu_{11}) = \left(\frac{1-\Delta}{2}, \frac{1+\Delta}{2}\right)$ for some $\Delta \in [-1, 1]$. Observe furthermore that since I_N is a sum of N realizations of an i.i.d. random variable, $\mathbb{E}(I_N/N) = \mathbb{E}I_1 = \frac{1}{2}(\mu_{11} - (1-\mu_{11})) - \frac{1}{2}(\mu_{01} - (1-\mu_{01})) = \Delta$. Thus, we can define $(\mu^*_{01}, \mu^*_{00}, \mu^*_{11}, \mu^*_{10})$ as maximizer of

$$\begin{aligned}&(p\mu_{11} + (1-p)\mu_{10} - p\mu_{01} - (1-p)\mu_{00}) \\ &\qquad \times \left(\frac{1}{2} + \frac{p(\mu_{01} - \mu_{11})}{2(1-p)}\right).\end{aligned}$$

Clearly this requires that $\mu^*_{10} = 1$, $\mu^*_{00} = 0$, and that $\Delta^* \equiv \mu^*_{11} - \mu^*_{01}$ maximize

$$(p\Delta + 1 - p) \times \left(\frac{1}{2} + \frac{p\Delta}{2(1-p)}\right) = \frac{1-p}{2} - \frac{p^2\Delta^2}{2(1-p)},$$

which obtains whenever $\Delta = 0 \Leftrightarrow \mu_{01} = \mu_{11}$. It follows that under Nature's best response, observations of Y_{01} and Y_{11} are uninformative, and the decision maker is indifferent between all treatment rules. In particular, δ^* is a best response.

The proof is essentially the same for stratified sampling. In that case, $\mathbb{E}I_N$ can be directly written as linear function of (μ_{01}, μ_{11}), so that proposition 1(ii) from Stoye (2006a) need not be invoked.

Lemma 3 Follows from Manski (2007, proposition 1); see also Stoye (2007, corollary 1).

Theorem 4 I restrict attention to the randomized treatment assignment scheme and also assume N to be odd; the extension to stratified sampling as well as even N follows along the lines of proposition 1

in Stoye (2006a). The core idea is to restrict DM's (pure) strategy space to $\{\delta_1^*, \delta_4^*\}$, rendering the game more tractable. Of course, it must be shown that equilibria of the simplified game are also equilibria of the original one. Thus, identify DM's strategy with $\alpha \in [0,1]$, the probability with which δ_1^* is played. As in lemma 2, the distribution of δ^* depends on s only through $\Delta \equiv \mu_{11} - \mu_{01}$. Nature will therefore pick $(\mu_{01}^*, \mu_{00}^*, \mu_{11}^*, \mu_{10}^*) \in [0,1]^4$ to maximize $R(\alpha, s)$, which can be expanded to

$$\begin{aligned}\max\{&(p\mu_{11} + (1-p)\mu_{10} - p\mu_{01} - (1-p)\mu_{00}) \\ &\times (1 - \alpha f_1(\mu_{11} - \mu_{01}) - (1-\alpha) f_0(\mu_{11} - \mu_{01})), \\ &(p\mu_{01} + (1-p)\mu_{00} - p\mu_{11} - (1-p)\mu_{10}) \\ &\times (\alpha f_1(\mu_{11} - \mu_{01}) + (1-\alpha) f_4(\mu_{11} - \mu_{01}))\},\end{aligned}$$

where

$$f_i(d) \equiv \mathbb{E}(\delta_i^* | \mu_{11} - \mu_{01} = d).$$

Some observations from the proof of lemma 2 apply: The objective function is symmetric, so that to find best responses, one can restrict attention to maximizers of the first element. Such maximizers must have $(\mu_{10}, \mu_{00}) = (1,0)$, and the optimization problem can be reduced to maximization over $\Delta \in [-1,1]$ of

$$\phi(\Delta; p, \alpha) \equiv (p\Delta + 1 - p)(1 - \alpha f_1(\Delta) - (1-\alpha) f_4(\Delta)).$$

Notice that f_i and ϕ are differentiable in their arguments; this will be used as first-order conditions will be evaluated. To construct Nash equilibria, it will be assumed that Nature randomizes evenly over the maximizer such found and its symmetric counterpart. The new arguments relative to lemma 2 are as follows.

Step 1: By the same arguments that apply to the original game, the simplified game possesses Nash equilibria. These must fall into one of three classes:

(i) Separating equilibria: Assume $\Delta > 0$, then the better treatment is the one that has higher expected success in observable units. The sampling distribution is binomial and thus possesses a monotone likelihood ratio property. It follows that δ_1^* (respectively $\alpha = 1$) is a best response.

(ii) Pooling equilibria: Assume $\Delta = 0$, then the signal generated by the sample is uninformative. Any decision rule constitutes a best response to this. The equilibrium from lemma 2 is an example of this case.

(iii) Negatively separating equilibria: Assume $\Delta < 1$, then the sample generates an informative signal, but the decision maker wants to act *against* this signal. In the simplified game, her best response would therefore be δ_4^*, which is less sensitive to the signal than δ_1^*.

The first two cases have in common that DM's equilibrium strategy remains a best response in her unrestricted strategy space. Whenever the simplified game's equilibrium falls into one of these cases, it therefore is an equilibrium of the original game as well. This does not hold if the equilibrium is negatively separating, in which case the decision maker's unrestricted best response would be $\delta_5^* \equiv 1 - \delta_1^*$.

Step 2: I will now show that a negatively separating equilibrium cannot obtain. It follows that equilibria of the simplified game are either separating or pooling, and thus coincide with equilibria of the original game.

To show the claim, suppose that DM plays δ_4^*. This leads to a negatively separating equilibrium iff Nature's best response is some $\Delta^* < 0$. The accordingly constrained value of her response problem is

$$\begin{aligned}&\sup_{\Delta \in [-1,0)} \phi(\Delta; p, 0) \\ = &\sup_{\Delta \in [-1,0)} (p\Delta + 1 - p)(1 - f_4(\Delta)).\end{aligned}$$

For comparison, the problem of maximizing

$$\rho(\Delta; p) \equiv (p\Delta + 1 - p)(1 - f_3(\Delta))$$

was considered in lemma 2; recall it is solved by $\Delta = 0$ and has value $(1-p)/2$. Substitute the definitions of δ_3^* and δ_4^* into the definition of f_i to find

$$\begin{aligned}f_3(\Delta) &= \mathbb{E}_{B(\Delta,N)} d^* \\ f_4(\Delta) &= \mathbb{E}_{B(\Delta,N)} d,\end{aligned}$$

where

$$\begin{aligned}d^* &= \frac{1}{2} + \frac{p(2n - N)}{2N(1-p)} \\ d &= \begin{cases} 0, & d^* < 0 \\ d^*, & 0 \le d^* \le 1 \\ 1, & d^* > 1 \end{cases}\end{aligned}$$

and where $\mathbb{E}_{B(\Delta,N)}$ denotes expectation with respect to the distribution of n, which is binomial with parameters (Δ, N). From inspection of these, it is elementary that $f_4(\Delta)$ lies between $f_3(\Delta)$ and $1/2$ for any (Δ, p); specifically, $f_4(\Delta) \ge f_3(\Delta)$ whenever $\Delta < 0$. It follows that $\Delta \le 0 \Rightarrow \phi(\Delta; p, 0) \le \rho(\Delta; p)$. Hence,

$$\sup_{\Delta \in [-1,0)} \phi(\Delta; p, 0) \le \sup_{\Delta \in [-1,0)} \rho(\Delta; p) = (1-p)/2,$$

and this supremum is furthermore not attained on $[-1, 0)$. But $\phi(\Delta; p, 0) = (1-p)/2$, so $\Delta = 0$ is a strictly better response to δ_4^* than any $\Delta < 0$.

Step 3: It remains to characterize separating respectively pooling equilibria. The main tool for this will be evaluation of first-order conditions. For a separating equilibrium, one must have $0 \le \arg\max_\Delta \phi(\Delta; p, 1)$. Consider the partial derivatives

$$\begin{aligned}\phi_\Delta(\Delta; p, 1) &= -f_1'(\Delta)\,(p\Delta + 1 - p) + p(1 - f_1(\Delta)) \\ \phi_{\Delta p}(\Delta; p, 1) &= (1-\Delta) f_1'(\Delta) + 1 - f_1(\Delta) > 0.\end{aligned}$$

Since the cross-derivative is positive, $\arg\max_\Delta \phi(\Delta; p, 1)$ increases in p in strong set order (that is, its smallest and largest element increase) by standard supermodularity arguments. Hence, the separating equilibrium can be maintained for $p > p_N^*$, where p_N^* is implicitly defined by

$$0 \in \arg\max_{\Delta \in [-1,1]} \phi(\Delta; p_N^*, 1).$$

An expression for p_N^* can be derived by inspecting the first-order condition:

$$\phi_\Delta(0, p_N^*; 1) = 0.$$

The previous expression for $\phi_\Delta(\Delta, p; 1)$ can be simplified at $\Delta = 0$. Write

$$\begin{aligned} f_1(\Delta) &= \Pr(I_N > N/2) \\ &= \sum_{n > N/2} \left(\begin{array}{c} N \\ n \end{array} \right) \left(\frac{1+\Delta}{2} \right)^n \left(\frac{1-\Delta}{2} \right)^{N-n}, \end{aligned}$$

which implies that (after some simplification)

$$f_1'(0) = 2^{-N} \sum_{n > N/2} \left(\begin{array}{c} N \\ n \end{array} \right) (2n - N) \equiv B.$$

Also observing that $f_1(0) = 1/2$, the first-order condition becomes

$$-(1-p)B + \frac{p_N^*}{2} = 0 \Longrightarrow p_N^* = \frac{2B}{2B+1}.$$

To see convergence of $p^* = \frac{2B}{2B+1}$ to 1, notice that B can be rewritten as

$$B = \mathbb{E}(|n| - N/2),$$

where n is the number of successes recorded in N independent coin tosses. The convergence rate of binomial distributions to the Normal immediately implies that $B = O\left(N^{1/2}\right)$ and hence that $1 - p_N^* = O(N^{-1/2})$.

Consider now the pooling equilibrium. This equilibrium requires that $\Delta = 0$ maximizes $\phi(\Delta, p; \alpha)$. A necessary condition for this is

$$\begin{aligned} 0 &= \phi_\Delta(0; p, \alpha) \\ &= (p \cdot 0 + 1 - p)\left(-\alpha f_1'(0) - (1-\alpha) f_4'(0)\right) \\ &\quad + p\left(1 - \alpha f_1(0) - (1-\alpha) f_4(0)\right). \end{aligned}$$

Some of the previous simplifications apply again; in particular, $f_4(0) = 1/2$. Substituting in for $f_4(\Delta) = \mathbb{E}_{B(\Delta,N)} d$, one finds that (after simplification)

$$f_4'(0) = 2^{-N} \sum_{n=0}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) (2n - N)\, d \equiv A.$$

The first-order condition thus simplifies to

$$\begin{aligned} 0 &= -(1-p)(\alpha B + (1-\alpha)A) + \frac{p}{2} \\ \Longrightarrow \alpha^* &= \frac{\frac{p}{2(1-p)} - A}{B - A}. \end{aligned}$$

This yields an equilibrium iff $\alpha^* \in [0, 1]$. I will show below that $B > A$ and that $A \leq \frac{p}{2(1-p)}$, with equality iff $p \leq 1/2$. Hence, $\alpha^* \geq 0$ as required, and $\alpha^* = 0$ iff $p \leq 1/2$, yielding the equilibrium from lemma 2. Furthermore, α^* equals 1 if $\frac{p}{2(1-p)} = B \Leftrightarrow p = \frac{2B}{2B+1}$, the condition identified for a separating equilibrium.

I conclude by filling the gaps in the preceding paragraph. To see that $A \leq \frac{p}{2(1-p)}$, with equality iff $p \leq 1/2$, observe that $f_4'(0) \leq f_3'(0)$ because $f_4'(\Delta) \leq f_3'(\Delta)$ was shown for $\Delta < 0$ in step 2, yet these derivatives are continuous. Hence $A \leq f_3'(0)$, but

$$f_3'(0) = \frac{d}{d\Delta} \mathbb{E}_{B(\Delta,N)} \left(\frac{1}{2} + \frac{p(2n-N)}{2N(1-p)} \right) = \frac{p}{2(1-p)}$$

because $\mathbb{E}_{B(\Delta,N)} \left(\frac{2n-N}{N} \right) = \Delta$. For $p > 1/2$, one can minimally expand on arguments from step 2 to show $f_4'(0) < f_3'(0)$, hence $A < \frac{p}{2(1-p)}$.

To see $B > A$, take explicit derivatives of binomial expectations to find (after some simplification)

$$B = 2^{-N} \sum_{n=0}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) (2n - N)\, \mathbb{I}\{d^* > 1/2\},$$

thus

$$B - A = 2^{-N} \sum_{n=0}^{N} \left(\begin{array}{c} N \\ n \end{array} \right) (2n - N) \left(\mathbb{I}\{d^* > 1/2\} - d \right),$$

but $(2n - N)\left(\mathbb{I}\{d^* > 1/2\} - d\right)$ is easily seen to be over nonnegative for any (n, N). Furthermore, the above sum is strictly positive whenever there exists n for which $\mathbb{I}\{d^* > 1/2\} - d \neq 0$, that is, whenever δ_1^* and δ_4^* do not agree. (They agree iff $p \geq N/(N+1)$, a number that is well above p^* for all N.)

Proposition 5 Follows by algebraically evaluating $\max_s R(\delta^*, s)$ using the above simplifications.

Proposition 6 Assume the decision maker can pick (τ, δ) and that the worst-case prior σ^* is as in proposition 1, then due to that prior's symmetry, the distribution of $(I_{n+1} | I_n)$ does not depend on τ_n. Hence, the decision maker is indifferent between all possible τ; in particular, the randomized design, in conjunction with δ^*, constitutes a best response. That σ^* remains a best response follows immediately from the proof of proposition 1. The conclusion extends to the stratified design because that design generates the same

maximal regret as the randomized one, yet a zero-sum game cannot have two Nash equilibria with different values.

Proposition 7 The proof is just as in theorem 4, with the following adjustment: Extend the decision problem, and hence the fictitious game, by identifying the state space with $\mathcal{S} \times [0,1]$ with typical element (s,p). Assume DM sets $p = \underline{p}$ and then uses δ^* from proposition 4. Then by following steps from theorem 4, Nature's best-response problem can be reduced to

$$\max_{p \in [\underline{p}, \overline{p}], \Delta \in [-1,1]} ! \; \{(p\Delta + 1 - p)(1 - \alpha f_1(\Delta) - (1-\alpha) f_4(\Delta))\}.$$

The objective decreases in p – notice especially that since DM uses $p = \underline{p}$, $f_4(\Delta)$ is not a function of Nature's choice of p. Hence, Nature will choose $p = \underline{p}$. The remainder of the proof is unchanged.

References

[1] T. Augustin. On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View — a Cautionary Note on Updating Imprecise Priors. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon (Eds.), *ISIPTA 03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*. Carleton Scientific, 2003.

[2] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, 1985.

[3] B. Droge. Minimax Regret Analysis of Orthogonal Series Regression Estimation: Selection Versus Shrinkage. *Biometrika* 85:631-643, 1998.

[4] —. Minimax Regret Comparison of Hard and Soft Thresholding for Estimating a Bounded Normal Mean. *Statistics and Probability Letters* 76:83–92, 2006.

[5] Y.C. Eldar, A. Ben-Tal, and A. Nemirovski. Linear Minimax Regret Estimation of Deterministic Parameters with Bounded Data Uncertainties. *IEEE Transactions on Signal Processing* 52:2177-2188, 2004.

[6] I.L. Glicksberg. A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points. *Proceedings of the American Mathematical Society* 3:170-174, 1952.

[7] E. Hanany and P. Klibanoff. Dynamically Consistent Updating of MaxMin EU and MaxMax EU Preferences. In F.G. Cozman, R. Nau, T. Seidenfeld (Eds.), *ISIPTA 05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*. Carnegie Mellon University, 2005.

[8] K. Hirano and J.R. Porter. Asymptotics for Statistical Treatment Rules. Technical Report, University of Arizona and University of Wisconsin.

[9] C.F. Manski. Treatment Choice under Ambiguity Induced by Inferential Problems. *Journal of Statistical Planning and Inference* 105:67-82, 2002.

[10] —. *Partial Identification of Probability Distributions*. Springer Verlag, 2003.

[11] —. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica* 72:1221-1246, 2004.

[12] —. *Social Choice with Partial Knowledge of Treatment Response*. Princeton University Press, 2005.

[13] —. Minimax-Regret Treatment Choice with Missing Outcome Data. *Journal of Econometrics* 139:105-115, 2007.

[14] D.B. Rubin. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66:688-701, 1974.

[15] L.J. Savage. The Theory of Statistical Decision. *Journal of the American Statistical Association* 46:55-67, 1951.

[16] K.H. Schlag. How to Minimize Maximum Regret in Repeated Decision-Making. Technical Report, European University Institute, 2003.

[17] —. Eleven. Technical Report, European University Institute, 2006.

[18] J. Stoye. Minimax Regret Treatment Choice with Finite Samples. Technical Report, New York University, 2006a.

[19] —. Statistical Decisions Under Ambiguity. Technical Report, New York University, 2006b.

[20] —. Minimax Regret Treatment Choice with Missing Data and Many Treatments. *Econometric Theory* 23:190–199, 2007.

Finite Approximations To Coherent Choice

Matthias C. M. Troffaes
Department of Mathematical Sciences, Durham University
matthias.troffaes@gmail.com

Abstract

This paper studies and bounds the effects of approximating loss functions and credal sets, under very weak assumptions, on choice functions. In particular, the credal set is assumed to be neither convex nor closed. The main result is that the effects of approximation can be bounded, although in general, approximation of the credal set may not always be practically possible. In case of pairwise choice, I demonstrate how the situation can be improved by showing that only approximations of the extreme points of the closure of the convex hull of the credal set need to be taken into account, as expected.

Keywords. decision making, E-admissibility, maximality, numerical analysis, lower prevision, sensitivity analysis

1 Introduction

Classical decision theory tells a decision maker to choose that option which maximises his expected utility. A generalisation of this principle is compelling when the probabilities and utilities relevant to the problem are not well known. Choice functions are one such generalisation, and select a set of optimal options: instead of pointing to a single solution based on possibly wrong assumptions, choice functions provide a set of optimal options. The decision maker can then investigate further if the set is too large, or not, if for instance the optimal set is a singleton, or if a single option from the set stands out from the rest by other arguments.

However, in modelling decision problems, we often afford ourselves the luxury of infinite spaces and infinite sets, making those problems sometimes hard to solve analytically. In such cases we must resort to computers, and these cannot handle gambles on infinite spaces, let alone arbitrary infinite sets of probabilities. Hence, in that case we must approximate our infinite sets by finite ones. By taking the finite sets sufficiently large, hopefully the approximation reflects the true result accurately. This paper confirms this intuition when modelling choice functions induced by arbitrary (not necessarily convex) sets of probabilities and a single cardinal utility, extending similar results known in classical decision theory [5, 10].

The paper is organised as follows. Section 2 introduces notation, and briefly reviews the theory of coherent choice functions and their role in decision theory. In Section 3 the building blocks for a theory of approximation are introduced, along with some useful results on what they imply for loss functions, sets of probabilities, and expected utility. The main part of the paper begins in Section 4, studying and bounding the effects of approximation on coherent choice functions. Section 5 improves the results of the previous section for pairwise choice. Section 6 concludes the paper. Some essential but technical results on approximating the standard simplex in $\mathbb{R}^n$ are deferred to an appendix.

2 Choice Functions

Let Ω denote an arbitrary set of states. Bounded random quantities on Ω, i.e. bounded maps from Ω to $\mathbb{R}$, are also called *gambles* [17], and will be denoted by f, g, ... $\mathcal{L}(\Omega)$ denotes the set of all gambles on Ω. Finitely additive probability measures, or briefly *probability charges* [2], are denoted by P, Q, ... and $\mathcal{P}(\Omega)$ denotes the set of all probability charges on the power set $\wp(\Omega)$ of Ω.

In a decision problem, we desire to choose an optimal option d from a set D of options. Choosing d induces an uncertain reward r from a set R of rewards, with probability charge $\mu_d(\cdot|w)$ over $\wp(R)$, depending on the outcome of the uncertain state $w \in \Omega$. For each $w \in \Omega$, $\mu_d(\cdot|w)$ is a *lottery* over R, and as a function of w, $\mu_d(\cdot|\cdot)\colon w \mapsto \mu_d(\cdot|w)$ is a *horse lottery* or *act*.

If we model our belief about states and rewards by a probability charge P on $\wp(\Omega)$ and a state dependent utility function $U(\cdot|w)$ on R, then utility theory [16, 1, 4] tells us to choose a decision d which maximises the expected utility, or prevision:

$$\begin{aligned} E(d) &= \int_\Omega \left(\int_R U(r|w)\,\mathrm{d}\mu_d(r|w) \right) \mathrm{d}P(w) \\ &= \int_\Omega f_d(w)\,\mathrm{d}P(w) \end{aligned}$$

where $f_d(w) = \int_R U(r|w)\,\mathrm{d}\mu_d(r|w)$ is the gamble associated with decision d, and the integrals are Dunford integrals [2]. For simplicity, in this paper, we assume $U(r|w)$ to be bounded, i.e.

$$\sup_{r,w} U(r|w) - \inf_{r,w} U(r|w) < +\infty$$

Among other things, this ensures that relative approximation can be defined, as in Section 3, without technical complications.

A decision which maximises expected utility is called a *Bayes decision* for the decision problem (Ω, D, P, U).

However, if we are not sure about the probability of all events and the utility of all rewards, a more reliable design is to use a family $(P_\alpha, U_\alpha)_{\alpha\in\aleph}$ of probability-utility pairs (where $\aleph$ is an index arbitrary set), and to elicit from D those options which maximise expected utility with respect to at least one of the pairs (P_α, U_α). First, for each $\alpha \in \aleph$, let

$$E_\alpha(d) = \int_\Omega f_d^\alpha(w)\,\mathrm{d}P_\alpha(w)$$

where $f_d^\alpha(w) = \int_R U_\alpha(r|w)\,\mathrm{d}\mu_d(r|w)$ is the gamble associated with decision d and model $\alpha \in \aleph$. Then we define:

Definition 1. *A decision $d \in D$ is called an* optimal decision *for the decision problem $(\Omega, D, (P_\alpha, U_\alpha)_{\alpha\in\aleph})$ if d belongs to the set*

$$\begin{aligned} &\mathrm{opt}(\Omega, D, (P_\alpha, U_\alpha)_{\alpha\in\aleph}) \\ &= \{d \in D\colon (\exists \alpha \in \aleph)(\forall e \in D)(E_\alpha(d) \geq E_\alpha(e))\} \\ &= \left\{ d \in D\colon (\exists \alpha \in \aleph) \left(E_\alpha(d) = \sup_{e\in D} E_\alpha(e) \right) \right\} \end{aligned}$$

As such, the operator opt selects a *set* of optimal decisions, namely all decisions which are Bayes with respect to $(\Omega, D, P_\alpha, U_\alpha)$ for at least one $\alpha \in \aleph$. Such an operator is called a *choice function* or *optimality operator* [3, 15].

In case $(P_\alpha, U_\alpha)_{\alpha\in\aleph} = \mathcal{M} \times \mathcal{U}$ for some convex sets $\mathcal{M}$ and $\mathcal{U}$, optimality as defined above is also called *E-admissibility* [9, Sec. 4.8].

There are many ways to define a choice function starting from a set $(P_\alpha, U_\alpha)_{\alpha\in\aleph}$ (see [9, 12, 17, 8, 15]). The one in Definition 1 satisfies an interesting set of axioms [8, 13], and is subject of a representation theorem in case utility is precise and state independent (i.e. if $U_\alpha(r|w)$ depends on neither on α nor on w) and Ω is finite (for infinite Ω the representation theorem is subject to additional constraints, which preclude merely finitely additive probabilities over Ω) [13].

For the sake of simplicity, we shall only be concerned about decision problems with precise and state independent utility functions, i.e. when $(P_\alpha, U_\alpha)_{\alpha\in\aleph} = \mathcal{M} \times \{U\}$ with $U\colon R \to \mathbb{R}$ a bounded state independent utility over R and

$$\mathcal{M} = \{P_\alpha\colon \alpha \in \aleph\}$$

The set $\mathcal{M}$ is called a *credal set* as it represents our belief about $w \in \Omega$. We can identify $\mathcal{M}$ itself as index set, and write

$$E_P(d) = \int_\Omega f_d(w)\,\mathrm{d}P(w)$$

with $f_d(w) = \int_R U(r)\,\mathrm{d}\mu_d(r|w)$, for any $P \in \mathcal{M}$.

Finally, defining the loss function $L\colon D \times \Omega \to \mathbb{R}$ as $L(d, w) = -f_d(w)$, the expected value $E_P(d)$ is uniquely determined by P and L alone: we need not be concerned explicitly with R, $\mu_d(r|w)$, and $U(r)$.

3 Approximate Gambles, Probabilities, and Previsions

Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ denote a finite partition of Ω. As we approximate Ω by the finite set $\mathcal{A}$, we also need to approximate decisions, gambles, and probability charges on Ω.

Let $\epsilon \geq 0$. For a gamble f in $\mathcal{L}(\Omega)$ and a gamble $\hat{f}$ in $\mathcal{L}(\mathcal{A})$, we shall write $f \sim_\epsilon \hat{f}$ if

$$\max_{A\in\mathcal{A}} \sup_{w\in A} \left| f(w) - \hat{f}(A) \right| \leq [\sup f - \inf f]\epsilon$$

Note that $f \sim_\epsilon \hat{f}$ implies $af + b \sim_\epsilon a\hat{f} + b$, for any real numbers a and b, $a > 0$. Therefore, the relation $\sim_\epsilon$ is invariant with respect to positive linear transformations of utility: it only depends on our preferences over lotteries, and not on our particular choice of utility scale.

For a probability charge P in $\mathcal{P}(\Omega)$, and a probability charge $\hat{P}$ in $\mathcal{P}(\mathcal{A})$, we shall write $P \sim_\epsilon \hat{P}$ if

$$\sum_{A\in\mathcal{A}} \left| P(A) - \hat{P}(A) \right| \leq \epsilon$$

Note that this implies $|P(A) - \hat{P}(A)| \leq \epsilon$ for any $A \in \wp(\mathcal{A})$. Also note the differences between the definitions of $\sim_\epsilon$ for gambles and bounded charges.

For a loss function L on $D \times \Omega$ and a loss function $\hat{L}$ on $D \times \mathcal{A}$ we write $L \sim_\epsilon \hat{L}$ if for all $d \in D$

$$f_d \sim_\epsilon \hat{f}_d$$

(with $f_d(w) = -L(d, w)$ and $\hat{f}_d(A) = -\hat{L}(d, A)$).

For a subset $\mathcal{M}$ of $\mathcal{P}(\Omega)$ and a subset $\hat{\mathcal{M}}$ of $\mathcal{P}(\mathcal{A})$, we write $\mathcal{M} \sim_\epsilon \hat{\mathcal{M}}$ if for every P in $\mathcal{M}$ there is a $\hat{P}$ in $\hat{\mathcal{M}}$ such that $P \sim_\epsilon \hat{P}$, and for every $\hat{P}$ in $\hat{\mathcal{M}}$ there is a P in $\mathcal{M}$ such that $P \sim_\epsilon \hat{P}$.

A few useful results about approximations are stated in the next lemmas.

Lemma 2. *Assume that D is finite. Then, for every loss function L on $D \times \Omega$ and every $\epsilon > 0$, there is a finite partition $\mathcal{A}$ of Ω and a loss function $\hat{L}$ on $D \times \mathcal{A}$ such that $L \sim_\epsilon \hat{L}$ and $|\mathcal{A}| \leq (1 + 1/\epsilon)^{|D|}$.*

Proof. Consider any d in D, and let $R_d = \sup f_d - \inf f_d$. Because f_d is bounded, we can embed the range of f_d in k intervals $I_1, \ldots, I_k$ of length $R_d\epsilon$, say

$$[\inf f_d, \inf f_d + R_d\epsilon), [\inf f_d + R_d\epsilon, \inf f_d + 2R_d\epsilon), \ldots, [\inf f_d + (k-1)R_d\epsilon, \inf f_d + kR_d\epsilon)$$

with k such that $\sup f_d \in I_k$. Therefore, $\inf f_d + (k-1)R_d\epsilon \leq \sup f_d < \inf f_d + kR_d\epsilon$ and hence $k - 1 \leq 1/\epsilon < k$. Observe that k is independent of $d \in D$.

The sets $A_1, \ldots, A_k$ defined by

$$A_j = f_d^{-1}(I_j)$$

form a finite partition $\mathcal{A}_d = \{A_j \colon A_j \neq \emptyset\}$ of cardinality $|\mathcal{A}_d| \leq k \leq 1 + 1/\epsilon$ and the gamble $\hat{f}_d \in \mathcal{L}(\mathcal{A}_d)$ defined by

$$\hat{f}_d(A_i) = \inf_{w \in A_i} f_d(w)$$

satisfies

$$\begin{aligned}\sup_{w \in A_j} \left| f_d(w) - \hat{f}_d(A_j) \right| &= \sup_{f_d(w) \in I_j} \left| f_d(w) - \inf_{f_d(w) \in I_j} f_d(w) \right| \\ &\leq \sup I_j - \inf I_j = R_d\epsilon\end{aligned}$$

for all $A_j \in \mathcal{A}_d$; hence $f_d \sim_\epsilon \hat{f}_d$. Defining $\hat{L}(d, A) = -\hat{f}_d(A)$ for all $d \in D$, we have $L \sim_\epsilon \hat{L}$.

The finite collection of partitions $\{\mathcal{A}_d \colon d \in D\}$ has a smallest common refinement $\mathcal{A}$. Since each $\mathcal{A}_d$ has no more than $1 + 1/\epsilon$ elements, $\mathcal{A}$ has no more than $(1 + 1/\epsilon)^{|D|}$ elements. Indeed, two partitions of cardinalities k_1 and k_2 respectively have a smallest common refinement of cardinality no more than $k_1 k_2$.By induction, n partitions of cardinalities $k_1, \ldots, k_n$ have a smallest common refinement of cardinality no more than $\prod_{j=1}^n k_j$ and hence,

$$|\mathcal{A}| \leq (1 + 1/\epsilon)^{|D|}$$

□

	ϵ: 0.2	0.1	0.05	0.02	0.01
$\lvert D\rvert$: 2	1.6	2.1	2.6	3.4	4.0
4	3.1	4.2	5.3	6.8	8.0
8	6.2	8.3	10.6	13.7	16.0
16	12.5	16.7	21.2	27.3	32.1
32	24.9	33.3	42.3	54.6	64.1

Table 1: Upper bound on $\log_{10}(|\mathcal{A}|)$, i.e. the logarithm of the cardinality of the finite partition $\mathcal{A}$ for various values of precision $\epsilon > 0$ and number of decisions (see Lemma 2).

Table 1 lists upper bounds on the size of the partition, to ensure $L \sim_\epsilon \hat{L}$, for various values of ϵ and $|D|$, according to Lemma 2.

Let $\binom{a}{b}$ be the binomial coefficient, defined for all real numbers $a \geq b \geq 0$ by

$$\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}$$

with Γ the Gamma function.

Lemma 3. *For every subset $\mathcal{M}$ of $\mathcal{P}(\Omega)$, every $\delta > 0$, and every finite partition $\mathcal{A}$ of Ω, there is a finite subset $\hat{\mathcal{M}}$ of $\mathcal{P}(\mathcal{A})$ such that $\mathcal{M} \sim_\delta \hat{\mathcal{M}}$ and $|\hat{\mathcal{M}}| \leq \binom{|\mathcal{A}|(1+1/\delta)}{|\mathcal{A}|-1}$.*

Proof. Consider any P in $\mathcal{M}$. Let $n = |\mathcal{A}|$ and let the elements of $\mathcal{A}$ be $A_1, \ldots, A_n$. Consider the vector $\underline{x} = (P(A_1), \ldots, P(A_n))$ in Δ^n. Let N be the smallest natural number such that $N \geq n/\delta$.

By Lemma 13 there is a vector $\underline{y}$ in Δ_N^n such that

$$|\underline{x} - \underline{y}|_1 < n/N \leq \delta$$

Define $\hat{P}$ in $\mathcal{P}(\mathcal{A})$ by

$$\hat{P}(A_i) = y_i$$

for all $i \in \{1, \ldots, n\}$—by finite additivity, $\hat{P}$ is well defined on $\wp(\mathcal{A})$. By construction, $P \sim_\delta \hat{P}$ because

$$\sum_{i=1}^n \left| P(A_i) - \hat{P}(A_i) \right| = |\underline{x} - \underline{y}|_1 < \delta$$

	δ:		
	0.2	0.1	0.05
$\lvert\mathcal{A}\rvert$: 4	3.3	4.1	5.0
8	7.9	9.8	11.8
12	12.5	15.5	18.7
16	17.1	21.3	25.6
20	21.8	27.1	32.6
24	26.4	32.9	39.5
28	31.1	38.6	46.5
32	35.8	44.4	53.4
$\log_{10}(\lvert\mathcal{A}\rvert)$: 0.7	4.4	5.5	6.7
1.4	27.6	34.3	41.3
2.1	144.6	179.5	215.5
2.8	731.3	906.8	1088.2
3.5	3666.1	4544.7	5452.8
4.2	18341.5	22735.9	27277.5
4.9	91719.7	113693.0	136402.5

Table 2: Upper bound on $\log_{10}(|\hat{\mathcal{M}}|)$, i.e. the logarithm of the cardinality of the finite set of probability charges $\hat{\mathcal{M}}$, for various values of precision $\delta > 0$ and cardinality of the partition $|\mathcal{A}|$ (see Lemma 3).

Approximating each P in $\mathcal{M}$ in this manner, the set

$$\hat{\mathcal{M}} = \{\hat{P}\colon P \in \mathcal{M}\}$$

is finite as each of its elements corresponds to an element of the finite set Δ_N^n, and therefore $|\hat{\mathcal{M}}| \leq |\Delta_N^n|$. By Lemma 12,

$$\begin{aligned} |\hat{\mathcal{M}}| &\leq \binom{N+n-1}{n-1} \\ &\leq \binom{n/\delta+1+n-1}{n-1} = \binom{|\mathcal{A}|(1+1/\delta)}{|\mathcal{A}|-1} \end{aligned}$$

The second inequality follows from the fact that $\binom{a}{b}$ is strictly increasing in a, for fixed b (for integer a and b this follows immediately from Pascal's triangle; the general case follows from the properties of the Gamma function). □

Table 2 lists upper bounds on the cardinality of $\hat{\mathcal{M}}$ on a logarithmic scale, for some values of $|\mathcal{A}|$ and δ. The cardinality follows an exponential trend in $|\mathcal{A}|$ and in $1/\delta$. The table shows that the influence of $|\mathcal{A}|$ is much larger than the influence of δ: more precisely, doubling $|\mathcal{A}|$ increases $|\hat{\mathcal{M}}|$ by far more than halving δ.

Next, we study the effect on the expectation if both gambles and probabilities are approximated. Let us use the notation $E_P(f) = \int_\Omega f(w)\,\mathrm{d}P(w)$. In the lemma below, assume $0 < \epsilon < 1/2$.

Lemma 4. *For every finite partition $\mathcal{A}$ of Ω, every $f \in \mathcal{L}(\Omega)$, $\hat{f} \in \mathcal{L}(\mathcal{A})$, $P \in \mathcal{P}(\Omega)$, and $\hat{P} \in \mathcal{P}(\mathcal{A})$, the following implications hold. If $f \sim_\epsilon \hat{f}$ and $P \sim_\delta \hat{P}$ then*

$$\left|E_P(f) - E_{\hat{P}}(\hat{f})\right| \leq [\sup f - \inf f](\epsilon + \delta(1+2\epsilon))$$

and

$$\left|E_P(f) - E_{\hat{P}}(\hat{f})\right| \leq [\sup \hat{f} - \inf \hat{f}]\left(\frac{\epsilon}{1-2\epsilon} + \delta\right)$$

Proof. Let $R = \sup f - \inf f$, $\hat{R} = \sup \hat{f} - \inf \hat{f}$, and write $\inf_A f$ for $\inf_{w\in A} f(w)$ and $\sup_A f$ for $\sup_{w\in A} f(w)$. Then

$$\left|E_P(f) - E_{\hat{P}}(\hat{f})\right| = \left|\sum_{A\in\mathcal{A}}\left(\int_A f\,\mathrm{d}P - \hat{f}(A)\hat{P}(A)\right)\right|$$

and since $P(A)\inf_A f \leq \int_A f\,\mathrm{d}P \leq P(A)\sup_A f$, there is an $r_A \in [\inf_A f, \sup_A f]$ such that $P(A)r_A = \int_A f\,\mathrm{d}P$, and hence

$$= \left|\sum_{A\in\mathcal{A}}\left(r_A P(A) - \hat{f}(A)\hat{P}(A)\right)\right|$$

but, because $|f(w) - \hat{f}(A)| \leq R\epsilon$ for all $w \in A$, and $\inf_A f \leq r_A \leq \sup_A f$, it must also hold that $|r_A - \hat{f}(A)| \leq R\epsilon$, and therefore $\left|\sum_{A\in\mathcal{A}}\left(r_A P(A) - \hat{f}(A)P(A)\right)\right| \leq \sum_{A\in\mathcal{A}}\left|r_A - \hat{f}(A)\right|P(A) \leq \sum_{A\in\mathcal{A}} R\epsilon P(A) = R\epsilon$, whence

$$\begin{aligned} &\leq \left|\sum_{A\in\mathcal{A}}\left(\hat{f}(A)P(A) - \hat{f}(A)\hat{P}(A)\right)\right| + R\epsilon \\ &= \left|\sum_{A\in\mathcal{A}}\hat{f}(A)\left(P(A) - \hat{P}(A)\right)\right| + R\epsilon \end{aligned}$$

and because $\sum_{A\in\mathcal{A}}(P(A) - \hat{P}(A)) = 0$,

$$\begin{aligned} &= \left|\sum_{A\in\mathcal{A}}(\hat{f}(A) - \inf\hat{f})\left(P(A) - \hat{P}(A)\right)\right| + R\epsilon \\ &\leq \sum_{A\in\mathcal{A}}(\hat{f}(A) - \inf\hat{f})\left|P(A) - \hat{P}(A)\right| + R\epsilon \\ &\leq (\sup\hat{f} - \inf\hat{f})\sum_{A\in\mathcal{A}}\left|P(A) - \hat{P}(A)\right| + R\epsilon \\ &\leq \hat{R}\delta + R\epsilon \end{aligned}$$

and since $R(1+2\epsilon) \geq \hat{R} \geq R(1-2\epsilon)$

$$\leq \begin{cases} R(1+2\epsilon)\delta + R\epsilon = R(\epsilon + \delta(1+2\epsilon)) \\ \hat{R}\delta + \hat{R}\epsilon/(1-2\epsilon) = \hat{R}\,(\epsilon/(1-2\epsilon) + \delta) \end{cases}$$

□

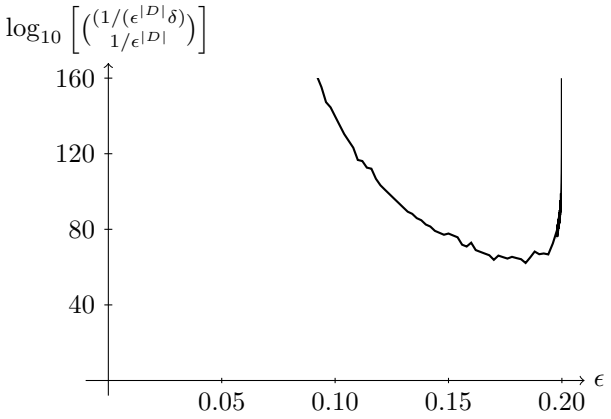


Figure 1: Upper bound on $\log_{10}|\hat{\mathcal{M}}|$ for various values of ϵ, with $\epsilon+\delta=0.2$ and $|D|=2$.

Let us now investigate what is the most optimal choice for $\epsilon>0$ and $\delta>0$. The cardinality of $\hat{\mathcal{M}}$ is of largest concern as it follows an exponential trend in the cardinality of the finite partition $\mathcal{A}$ and in the inverse of the required precision $\delta>0$ (see Table 2). Therefore, as a first step, let us see how we can minimise $|\hat{\mathcal{M}}|$, assuming a fixed relative error $\epsilon+\delta$ on the expectation (see Lemma 4)—omitting higher order terms in ϵ and δ to simplify the analysis.

We wish to minimise the upper bound (neglecting lower order terms)

$$\binom{(1/(\epsilon^{|D|}\delta))}{1/\epsilon^{|D|}}$$

on $|\hat{\mathcal{M}}|$ along the ϵ–δ-curve $\gamma(\epsilon,\delta)=\epsilon+\delta=\gamma_*$. Figure 1 demonstrates a typical case: the ϵ–δ-ratio has a large impact on the upper bound of $|\hat{\mathcal{M}}|$. In particular, the curve grows extremely large for small ϵ, because a small ϵ corresponds to a large partition $\mathcal{A}$, and the cardinality of the partition has a huge impact on the cardinality of $\mathcal{M}$ as shown in Table 2.

4 Approximate Choice

Let us now consider again the decision problem $(\Omega,D,\mathcal{M},L)$ with state space Ω, decision space D, credal set $\mathcal{M}$, and loss function L, and reflect upon how the results in the previous section could be of use in finding the optimal decisions $\mathrm{opt}(\Omega,D,\mathcal{M},L)$. Can we still find the optimal decisions after approximating the loss function L and the set of probabilities $\mathcal{M}$?

As we admit a relative error on gambles and probabilities, and therefore also on previsions, we should admit a relative error on the choice function as well. Let R_D be defined by (recall that $f_d(w)=-L(d,w)$)

$$R_D=\sup_{d\in D}[\sup f_d-\inf f_d]$$

Definition 5. *Let $\epsilon\geq 0$. A decision d in D is called an* ϵ-optimal decision *for the decision problem $(\Omega,D,\mathcal{M},L)$ if it belongs to the set*

$$\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},L)=$$
$$\left\{d\in D\colon(\exists P\in\mathcal{M})\left(\sup_{e\in D}E_P(e)-E_P(d)\leq\epsilon R_D\right)\right\}$$

Note that

$$\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},aL+b)=\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},L)$$

for any real numbers a and b, $a>0$. In other words, $\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},L)$ is invariant with respect to positive linear transformations of utility: ϵ-optimality does not depend on our choice of utility scale.

Clearly,

$$\mathrm{opt}(\Omega,D,\mathcal{M},L)\subseteq\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},L)$$

because

$$\mathrm{opt}^\epsilon(\Omega,D,\mathcal{M},L)\subseteq\mathrm{opt}^\delta(\Omega,D,\mathcal{M},L)$$

whenever $\epsilon\leq\delta$, and

$$\mathrm{opt}^0(\Omega,D,\mathcal{M},L)=\mathrm{opt}(\Omega,D,\mathcal{M},L)$$

In approximating a decision problem $(\Omega,D,\mathcal{M},L)$, we start with a finite partition $\mathcal{A}$, consider a (possibly finite) set $\hat{\mathcal{M}}$ such that $\mathcal{M}\sim_\delta\hat{\mathcal{M}}$, and approximate the loss $L(d,w)$ by a loss $\hat{L}(d,A)$ such that $L\sim_\epsilon\hat{L}$.

Theorem 6. *Consider two decision problems $(\Omega,D,\mathcal{M},L)$ and $(\mathcal{A},D,\hat{\mathcal{M}},\hat{L})$. If $L\sim_\epsilon\hat{L}$ and $\mathcal{M}\sim_\delta\hat{\mathcal{M}}$ then, for any $\gamma\geq 0$,*

$$\mathrm{opt}^\gamma(\Omega,D,\mathcal{M},L)\subseteq\mathrm{opt}^{\frac{\gamma}{1-2\epsilon}+2(\frac{\epsilon}{1-2\epsilon}+\delta)}(\mathcal{A},D,\hat{\mathcal{M}},\hat{L})\quad(1)$$

and

$$\mathrm{opt}^\gamma(\mathcal{A},D,\hat{\mathcal{M}},\hat{L})\subseteq\mathrm{opt}^{\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega,D,\mathcal{M},L)\quad(2)$$

Proof. We prove Eq. (1). Let $d\in\mathrm{opt}^\gamma(\Omega,D,\mathcal{M},L)$. Then

$$\sup_{e\in D}E_P(f_e)-E_P(f_d)\leq\gamma R_D\quad(3)$$

for some $P \in \mathcal{M}$. Let $\hat{P}$ be such that $P \sim_\delta \hat{P}$. Because, by Lemma 4,

$$\begin{aligned}
&\left|\sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - \sup_{e'\in D} E_P(f_{e'})\right| \\
&\leq \sup_{e\in D}\left|E_{\hat{P}}(\hat{f}_e) - E_P(f_e)\right| \\
&\leq \sup_{e\in D}[\sup \hat{f}_e - \inf \hat{f}_e](\epsilon/(1-2\epsilon)+\delta) \\
&= (\epsilon/(1-2\epsilon)+\delta)\hat{R}_D \qquad (4)
\end{aligned}$$

it follows that

$$\begin{aligned}
&\sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - E_{\hat{P}}(\hat{f}_d) \\
&\leq \sup_{e\in D} E_P(f_e) - E_{\hat{P}}(\hat{f}_d) + (\epsilon/(1-2\epsilon)+\delta)\hat{R}_D
\end{aligned}$$

and again by Lemma 4,

$$\leq \sup_{e\in D} E_P(f_e) - E_P(f_d) + 2(\epsilon/(1-2\epsilon)+\delta)\hat{R}_D$$

and by Eq. (3),

$$\begin{aligned}
&\leq \gamma R_D + 2(\epsilon/(1-2\epsilon)+\delta)\hat{R}_D \\
&\leq [\gamma/(1-2\epsilon) + 2(\epsilon/(1-2\epsilon)+\delta)]\hat{R}_D
\end{aligned}$$

hence, $d \in \mathrm{opt}^{\gamma/(1-2\epsilon)+2(\epsilon/(1-2\epsilon)+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L})$.

Next, we prove Eq. (2). Let $d \in \mathrm{opt}^{\gamma}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L})$. Then

$$\sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - E_{\hat{P}}(\hat{f}_d) \leq \gamma \hat{R}_D \qquad (5)$$

Because, by Lemma 4,

$$\begin{aligned}
&\left|\sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - \sup_{e'\in D} E_P(f_{e'})\right| \\
&\leq \sup_{e\in D}\left|E_{\hat{P}}(\hat{f}_e) - E_P(f_e)\right| \\
&\leq \sup_{e\in D}[\sup f_e - \inf f_e](\epsilon + \delta(1+2\epsilon)) \\
&= (\epsilon+\delta(1+2\epsilon))R_D \qquad (6)
\end{aligned}$$

we have that

$$\begin{aligned}
&\sup_{e\in D} E_P(f_e) - E_P(f) \\
&\leq \sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - E_P(f) + (\epsilon+\delta(1+2\epsilon))R_D
\end{aligned}$$

and again by Lemma 4,

$$\leq \sup_{e\in D} E_{\hat{P}}(\hat{f}_e) - E_{\hat{P}}(\hat{f}_e) + 2(\epsilon+\delta(1+2\epsilon))R_D$$

and by Eq. (5)

$$\begin{aligned}
&\leq \gamma\hat{R}_D + 2(\epsilon+\delta(1+2\epsilon))R_D \\
&\leq [\gamma(1+2\epsilon) + 2(\epsilon+\delta(1+2\epsilon))]R_D
\end{aligned}$$

so $d \in \mathrm{opt}^{\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega, D, \mathcal{M}, L)$. □

If we ignore higher order terms in γ, ϵ, and δ, then the above theorem says that when moving from an original decision problem to an approximate decision problem, or the other way around, with relative error ϵ in gambles and relative error δ in probabilities, the relative error in optimality increases by $2(\epsilon+\delta)$. For example, for small ϵ and δ the following holds, up to a small error: if $L \sim_\epsilon \hat{L}$ and $\mathcal{M} \sim_\delta \hat{\mathcal{M}}$, then

$$\begin{aligned}
&\mathrm{opt}(\Omega, D, \mathcal{M}, L) \\
&\qquad \subseteq \mathrm{opt}^{2(\epsilon+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L}) \subseteq \\
&\qquad\qquad \mathrm{opt}^{4(\epsilon+\delta)}(\Omega, D, \mathcal{M}, L)
\end{aligned}$$

So, the approximate problem with relative error $2(\epsilon+\delta)$ will contain all solutions to the original problem with no relative error, and will, so to say, not contain any solutions to the original problem with relative error over $4(\epsilon+\delta)$. Because of this property, $\mathrm{opt}^{2(\epsilon+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L})$ seems a logical choice when solving decision problems in practice.

5 Pairwise Choice

Table 2 reveals that the credal set is a serious computational bottleneck. Therefore, it is worth to investigate how the size of $\hat{\mathcal{M}}$ can be reduced, without compromising the accuracy $\delta > 0$. One way to this end is to restrict to pairwise comparisons, i.e. using maximality [17, Sec. 3.7–3.9].

5.1 Maximality

Definition 7. *A decision $d \in D$ is called a* maximal decision *for the decision problem $(\Omega, D, \mathcal{M}, L)$ if d belongs to the set*

$$\begin{aligned}
&\max(\Omega, D, \mathcal{M}, L) \\
&= \{d \in D \colon (\forall e \in D)(\exists P \in \mathcal{M})\,(E_P(d) \geq E_P(e))\}
\end{aligned}$$

Denote by $\mathrm{co}(\mathcal{M})$ the convex hull of $\mathcal{M}$. Obviously $\max(\Omega, D, \mathcal{M}, L) = \max(\Omega, D, \mathrm{co}(\mathcal{M}), L)$, because for any $\lambda \in [0,1]$ and any two P and Q in $\mathcal{M}$, the inequalities $E_P(d) \geq E_P(e)$ and $E_Q(d) \geq E_Q(e)$ imply the inequality

$$E_{\lambda P+(1-\lambda)Q}(d) \geq E_{\lambda P+(1-\lambda)Q}(e)$$

This does not hold for optimality as defined in Definition 1: assuming Ω finite, for any two distinct subsets $\mathcal{M}$ and $\mathcal{M}'$ of $\mathcal{P}(\Omega)$, we can always find a set D and a loss function L such that $\mathrm{opt}(\Omega, D, \mathcal{M}, L) \neq \mathrm{opt}(\Omega, D, \mathcal{M}', L)$ (see Kadane, Schervish, and Seidenfeld [8, Thm. 1, p. 53]).

To understand why the above notion of optimality is called maximality, consider the strict partial ordering

$>$ on D defined by

$$(e > d) \iff (\forall P \in \mathcal{M})\,(E_P(e) > E_P(d))$$

for any d and e in D, that is, e is strictly preferred to d if e is strictly preferred to d with respect to every $P \in \mathcal{M}$. Then,

$$\max(\Omega, D, \mathcal{M}, L) \;=\; \{d \in D\colon (\forall e \in D)(e \not> d)\}$$

so $\max(\Omega, D, \mathcal{M}, L)$ elects those decisions d which are maximal with respect to $>$. Therefore, maximality can be expressed through pairwise preferences only—again in contrast to $\mathrm{opt}(\Omega, D, \mathcal{M}, L)$ as demonstrated by Kadane, Schervish, and Seidenfeld [8, Sec. 4, p. 51].

However, because

$$\mathrm{opt}(\Omega, D, \mathcal{M}, L) \subseteq \max(\Omega, D, \mathcal{M}, L)$$

we may interpret $\max(\Omega, D, \mathcal{M}, L)$ as an approximation to $\mathrm{opt}(\Omega, D, \mathcal{M}, L)$, an approximation which discards all preferences but the pairwise ones.

Let us admit a relative error on the choice function max as well. Recall, $R_D = \sup_{d \in D}[\sup f_d - \inf f_d]$.

Definition 8. *Let $\epsilon \geq 0$. A decision d in D is called an* ϵ-maximal decision *for the decision problem $(\Omega, D, \mathcal{M}, L)$ if it belongs to the set*

$$\begin{aligned}\max^{\epsilon}(\Omega, D, \mathcal{M}, L) = \\ \{d \in D\colon (\forall e \in D)(\exists P \in \mathcal{M}) \\ (E_P(e) - E_P(d) \leq \epsilon R_D)\}\end{aligned}$$

5.2 Approximating Extreme Points

It turns out that we can restrict our attention to the extreme points of the closure of the convex hull of $\mathcal{M}$, with respect to the topology of pointwise convergence on members of $\mathcal{L}(\Omega)$. This topology is characterised by the following notion of convergence: $\lim_n P_n = P$ if

$$\lim_n E_{P_n}(f) = E_P(f) \text{ for all } f \in \mathcal{L}(\Omega)$$

Without further mentioning, I will assume this topology on $\mathcal{P}(\Omega)$.

There is a nice connection between the closure of $\mathcal{M}$, denoted by $\mathrm{cl}(\mathcal{M})$, and ϵ-maximality.

Lemma 9. *Assume that $R_D > 0$. Let $\epsilon \geq 0$. For any decision problem $(\Omega, D, \mathcal{M}, L)$, the following equality holds:*

$$\max^{\epsilon}(\Omega, D, \mathrm{cl}(\mathcal{M}), L) = \lim_{\delta \xrightarrow{\geq} 0} \max^{\epsilon+\delta}(\Omega, D, \mathcal{M}, L)$$

Proof. Assume $d \in \max^{\epsilon}(\Omega, D, \mathrm{cl}(\mathcal{M}), L)$. Consider any $e \in D$. By assumption, there is a $P \in \mathrm{cl}(\mathcal{M})$ such that $E_P(e) - E_P(d) \leq R_D\epsilon$. Hence, there is a sequence $(P_n \in \mathcal{M})_{n\in\mathbb{N}}$ such that $\lim_n E_{P_n}(f) = E_P(f)$ for all gambles f, and therefore also $\lim_n E_{P_n}(e) - \lim_n E_{P_n}(d) \leq R_D\epsilon$. This implies that for every $\delta > 0$, there is an $n \in \mathbb{N}$ such that $E_{P_n}(e) - E_{P_n}(f) \leq (\epsilon + \delta)R_D$. So, for every $\delta > 0$, there is a $P \in \mathcal{M}$ such that $E_P(e) - E_P(f) \leq (\epsilon + \delta)R_D$. Whence, because this holds for any $e \in D$, $d \in \max^{\epsilon+\delta}(\Omega, D, \mathcal{M}, L)$ for all $\delta > 0$, and therefore, $d \in \lim_{\delta \xrightarrow{\geq} 0} \max^{\epsilon+\delta}(\Omega, D, \mathcal{M}, L)$.

Conversely, assume $d \in \lim_{\delta \xrightarrow{\geq} 0} \max^{\epsilon+\delta}(\Omega, D, \mathcal{M}, L)$. Consider any $e \in D$. Then, for all $\delta > 0$, there is a $P_\delta \in \mathcal{M}$ such that $E_{P_\delta}(e) - E_{P_\delta}(f) \leq (\epsilon + \delta)R_D$. Hence, for all $n \in \mathbb{N}$, there is a $P_n \in \mathcal{M}$ such that

$$E_{P_n}(e) - E_{P_n}(d) \leq 1/n + \epsilon R_D \tag{7}$$

For any $m \in \mathbb{N}$, consider the following closed subset of $\mathcal{P}(\Omega)$:

$$\mathcal{R}_m = \mathrm{cl}(\{P_n\colon n \geq m\})$$

The collection $\{\mathcal{R}_m\colon m \in \mathbb{N}\}$ satisfies the finite intersection property. By the Banach-Alaoglu-Bourbaki theorem [11, §28.29(UF26)] $\mathcal{P}(\Omega)$ is compact, and hence

$$\mathcal{R} = \cap_{m\in\mathbb{N}}\mathcal{R}_m$$

is non-empty as well [11, §17.2].

Take any $R \in \mathcal{R}$. Since each $P_n \in \mathcal{M}$, it follows that each $\mathcal{R}_m \subseteq \mathrm{cl}(\mathcal{M})$, and hence $R \in \mathrm{cl}(\mathcal{M})$. If we can show that $E_R(e) - E_R(d) \leq \epsilon R_D$, then $d \in \max^{\epsilon}(\Omega, D, \mathrm{cl}(\mathcal{M}), L)$ is established.

Indeed, because $R \in \mathcal{R}_m$, there is a sequence $(P_{n_k})_{k\in\mathbb{N}}$ in $\{P_n\colon n \geq m\}$—so $n_k \geq m$, but n_k is not necessarily an increasing function of k—such that $\lim_k P_{n_k}(f_e - f_d) = R(f_e - f_d)$. Hence, by Eq. (7), for each $\gamma > 0$, there is a $k \in \mathbb{N}$ such that $E_R(e) - E_R(d) \leq P_{n_k}(e) - P_{n_k}(d) + \gamma$, and hence $E_R(e) - E_R(d) \leq 1/n_k + \epsilon R_D + \gamma$. Because this inequality holds for every m and every $\gamma > 0$, and $n_k \geq m$, it follows that $E_R(e) - E_R(d) \leq \epsilon R_D$. □

In particular, assuming $R_D > 0$, if for any $\delta > \epsilon > 0$

$$\max^{\epsilon}(\Omega, D, \mathcal{M}, L) = \max^{\delta}(\Omega, D, \mathcal{M}, L)$$

then

$$\max^{\epsilon}(\Omega, D, \mathcal{M}, L) = \max^{\epsilon}(\Omega, D, \mathrm{cl}(\mathcal{M}), L)$$

As a special case, Lemma 9 implies an interesting connection between maximality and ϵ-maximality:

Corollary 10. *Assume that $R_D > 0$. For any decision problem $(\Omega, D, \mathcal{M}, L)$, the following equality holds:*

$$\max(\Omega, D, \mathrm{cl}(\mathcal{M}), L) = \lim_{\epsilon \xrightarrow{\geq} 0} \max^{\epsilon}(\Omega, D, \mathcal{M}, L)$$

In the following theorem, assume that $0 < \epsilon < 1/2$.

Theorem 11. *Consider two decision problems* $(\Omega, D, \mathcal{M}, L)$ *and* $(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L})$. *Assume that* $R_D > 0$. *If* $L \sim_\epsilon \hat{L}$ *and* $\mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M}))) \sim_\delta \hat{\mathcal{M}}$ *then, for any* $\gamma \geq 0$,

$$\max^\gamma(\Omega, D, \mathcal{M}, L) \subseteq \max^{\frac{\gamma}{1-2\epsilon}+2(\frac{\epsilon}{1-2\epsilon}+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L}) \quad (8)$$

and

$$\max^\gamma(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L}) \subseteq \lim_{\eta \gtrdot 0} \max^{\eta+\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega, D, \mathcal{M}, L) \quad (9)$$

Proof. First, note that

$$\begin{aligned}&\max^\gamma(\Omega, D, \mathcal{M}, L)\\ &= \max^\gamma(\Omega, D, \mathrm{co}(\mathcal{M}), L)\\ &\subseteq \max^\gamma(\Omega, D, \mathrm{cl}(\mathrm{co}(\mathcal{M})), L)\\ &= \max^\gamma(\Omega, D, \mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M}))), L)\end{aligned}$$

because, by convexity of $\mathrm{cl}(\mathrm{co}(\mathcal{M}))$ [11, §26.23] and the Krein-Milman theorem [6, p. 74], the convex hull of $\mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M})))$ is $\mathrm{cl}(\mathrm{co}(\mathcal{M}))$. Now apply the same argument as in the proof of Theorem 6 to recover Eq. (8).

To establish Eq. (9), observe that $\mathcal{M}' \sim_\delta \hat{\mathcal{M}}'$ implies $\mathrm{co}(\mathcal{M}') \sim_\delta \mathrm{co}(\hat{\mathcal{M}}')$ because if $P \sim_\delta \hat{P}$ and $Q \sim_\delta \hat{Q}$ then, for any $\lambda \in [0, 1]$,

$$\lambda P + (1-\lambda)Q \sim_\delta \lambda\hat{P} + (1-\lambda)\hat{Q}$$

In particular, because $\mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M}))) \sim_\delta \hat{\mathcal{M}}$, and because the convex hull of $\mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M})))$ is $\mathrm{cl}(\mathrm{co}(\mathcal{M}))$ (again see [11, §26.23] and [6, p. 74]), it follows that

$$\mathrm{cl}(\mathrm{co}(\mathcal{M})) \sim_\delta \mathrm{co}(\hat{\mathcal{M}})$$

So,

$$\max^\gamma(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L}) = \max^\gamma(\mathcal{A}, D, \mathrm{co}(\hat{\mathcal{M}}), \hat{L})$$

and again from the same argument as in the proof of Theorem 6

$$\begin{aligned}&\subseteq \max^{\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega, D, \mathrm{cl}(\mathrm{co}(\mathcal{M})), L)\\ &= \lim_{\eta \gtrdot 0} \max^{\eta+\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega, D, \mathrm{co}(\mathcal{M}), L)\\ &= \lim_{\eta \gtrdot 0} \max^{\eta+\gamma(1+2\epsilon)+2(\epsilon+\delta(1+2\epsilon))}(\Omega, D, \mathcal{M}, L)\end{aligned}$$

□

Again, if we ignore higher order terms in γ, ϵ, and δ, then the above theorem says that when moving from the original decision problem to the approximate decision problem, with relative error ϵ in gambles and relative error δ in probabilities, the relative error in maximality increases by $2(\epsilon + \delta)$. Hence, for small ϵ and δ the following holds, up to a small error: if $L \sim_\epsilon \hat{L}$ and $\mathrm{ext}(\mathrm{cl}(\mathrm{co}(\mathcal{M}))) \sim_\delta \hat{\mathcal{M}}$, then

$$\max(\Omega, D, \mathcal{M}, L) \subseteq \max^{2(\epsilon+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L}) \subseteq \max^{4(\epsilon+\delta)}(\Omega, D, \mathcal{M}, L)$$

Again, $\max^{2(\epsilon+\delta)}(\mathcal{A}, D, \hat{\mathcal{M}}, \hat{L})$ seems a logical choice when calculating maximal decisions in practice.

6 Conclusion and Remarks

With this paper, I hope to have consolidated at least part of our every day intuition when approximating decision problems involving sets of probabilities, for instance when those problems have to be solved by computer.

One result is quite depressing: Lemma 2 and Lemma 3 seem to tell us that except in the simplest cases, any approximation will need too many resources to be of any practical value, as demonstrated by Table 1 and Table 2.

Fortunately, not all is lost. If we resort to pairwise comparison, we may restrict ourselves to the extreme points of the closure of the convex hull of the credal set, which can be *much* smaller than the original credal set. Closing the credal set only has an arbitrary small effect on maximality, and in part for this reason, it turns out that approximating extreme points suffices when restricting to pairwise preference.

I wish to emphasise that the bounds on the cardinalities of the approximating partition and the approximating credal set are only upper bounds under very weak assumptions. These bounds are only attained in extreme situations. In many cases the credal set and the loss function have additional structure which may allow for much lower upper bounds.

In case the problem has sufficient structure, an alternative approach is to develop algorithms which do not need to traverse the complete credal set (or an approximation thereof) to compute the optimal solution. The imprecise Dirichlet model has already been given considerable attention in this direction [7].

Finally, one could also simply sample elements from the credal set, for instance through Monte-Carlo techniques, and solve a classical decision problem for

each of these samples. If the sample s from $\hat{\mathcal{M}}$ is large enough, then—since $\bigcup_{P\in s}\operatorname{opt}(\mathcal{A},D,P,L)=\operatorname{opt}(\mathcal{A},D,s,L)$—hopefully

$$\operatorname{opt}(\mathcal{A},D,\mathcal{M},L)=\bigcup_{P\in s}\operatorname{opt}(\mathcal{A},D,P,L)$$

The question how large a sample we need to ensure convergence is definitely worth further investigation.

Acknowledgements

I am grateful to Teddy Seidenfeld for the many helpful discussions on issues related to this paper, and also for encouraging me to extend my view on approximations to choice functions. I thank Max Jensen for his help in characterising the discretisation of the simplex in $\mathbb{R}^n$, presented in the appendix. I also thank all three referees for their constructive comments and useful suggestions which have improved the presentation of this paper. The research reported in this paper has been supported in part by the Belgian American Educational Foundation.

A Discretisation Of The Standard Simplex In $\mathbb{R}^n$

In this appendix a simple discretisation of Δ^n, the standard simplex in $\mathbb{R}^n$, is studied—these results are not new and are in fact related to well known notions from combinatorics, in particular multisets [14]. The standard simplex Δ^n is defined as

$$\Delta^n=\{\underline{x}\in\mathbb{R}^n\colon \underline{x}\geq 0,\ |\underline{x}|_1=1\}$$

where $|\cdot|_1$ denotes the 1-norm, i.e. $|\underline{x}|_1=\sum_{i=1}^n|x_i|$.

For any non-zero natural number N, let Δ^n_N denote the following finite subset of Δ^n:

$$\Delta^n_N=\{\underline{m}/N\colon \underline{m}\in\mathbb{N}^n,\ |\underline{m}|_1=N\}$$

(above, $\mathbb{N}$ is the set of natural numbers including 0).

Lemma 12. *The cardinality of Δ^n_N is $\binom{N+n-1}{N}$.*

Proof. There is an obvious one-to-one and onto correspondence between Δ^n_N and all multisets of cardinality N with elements taken from $\{1,\ldots,n\}$—for any $\underline{m}/N\in\Delta^n_N$, interpret m_i as the multiplicity of i. The number of all such multisets is precisely $\binom{N+n-1}{N}$ (see Stanley [14]). □

Lemma 13. *For every $\underline{x}$ in Δ^n there is a $\underline{y}$ in Δ^n_N such that*

$$|\underline{x}-\underline{y}|_1<n/N$$

Proof. For each $i\in\{1,\ldots,n\}$, let m_i be the unique natural number such that $x_i\in[m_i/N,(m_i+1)/N)$, or equivalently, let m_i be the largest natural number such that $m_i/N\leq x_i$. Define $M=\sum_{i=1}^n m_i$. Then, $M\leq N<M+n$ since $M/N=|\underline{m}/N|_1\leq|\underline{x}|_1=1$ and $(M+n)/N=|(\underline{m}+1)/N|_1>|\underline{x}|_1=1$. Define

$$e_i=\begin{cases}1 & \text{if } i\in\{1,\ldots,N-M\}\\ 0 & \text{if } i\in\{N-M+1,\ldots,n\}\end{cases}$$

and let $\underline{y}=(\underline{m}+\underline{e})/N$. Note that $\underline{y}\in\Delta^n_N$ because $|\underline{y}|_1=|\underline{m}+\underline{e}|_1/N=(M+(N-M))/N=1$. Finally,

$$|\underline{x}-\underline{y}|_1=\sum_{i=1}^{N-M}|x_i-\tfrac{m_i+1}{N}|+\sum_{i=N-M+1}^{n}|x_i-\tfrac{m_i}{N}|<n/N$$

as $|x_i-\frac{m_i+1}{N}|\leq 1/N$ and $|x_i-\frac{m_i}{N}|<1/N$. □

References

[1] F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34(1):199–205, March 1963.

[2] K.P.S. Bhaskara Rao and M. Bhaskara Rao. *Theory of Charges, a Study of Finitely Additive Measures*. Academic Press, London, 1983.

[3] Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal of Approximate Reasoning*, 39(2–3):257–278, Jun 2004.

[4] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. Wiley, New York, 1974–5. Two volumes.

[5] Peter C. Fishburn, Allan H. Murphy, and Herbert H. Isaacs. Sensitivity of decisions to probability estimation errors: A reexamination. *Operations Research*, 16(2):254–267, 1968.

[6] Richard B. Holmes. *Geometric Functional Analysis and Its Applications*. Springer, New York, 1975.

[7] Marcus Hutter. Robust estimators under the imprecise dirichlet model. In Jean-Marc Bernard, Teddy Seidenfeld, and Marco Zaffalon, editors, *ISIPTA '03 – Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, pages 274–289. Carleton Scientific, July 2003.

[8] J. B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. A Rubinesque theory of decision. In *A*

festschrift for Herman Rubin, volume 45 of *IMS Lecture Notes – Monograph Series*, pages 45–55. Inst. Math. Statist., Beachwood, Ohio, 2004.

[9] Isaac Levi. *The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability, and Chance*. MIT Press, Cambridge, 1983.

[10] Donald A. Pierce and J. Leroy Folks. Sensitivity of Bayes procedures to the prior distribution. *Operations Research*, 17(2):344–350, 1969.

[11] Eric Schechter. *Handbook of Analysis and Its Foundations*. Academic Press, San Diego, 1997.

[12] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. A representation of partially ordered preferences. *The Annals of Statistics*, 23:2168–2217, 1995.

[13] Teddy Seidenfeld, Mark Schervish, and Jay Kadane. Coherent choice functions under uncertainty. Submitted to ISIPTA'07.

[14] Richard P. Stanley. *Enumerative Combinatorics*. Cambridge University Press, 1997.

[15] Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45:17–29, 2007.

[16] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[17] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Computing expectations with p-boxes: two views of the same problem

L. Utkin
Department of computer science
State Forest Technical Academy
St.Petersburg, Russia
lev.utkin@mail.ru

S. Destercke
Institut de Radioprotection
et de sûreté nucléaire (IRSN),
Cadarache, France
sebastien.destercke@irsn.fr

Abstract

Given an imprecise probabilistic model over a continuous space, computing lower (upper) expectations is often computationally hard to achieve, even in simple cases. Building tractable methods to do so is thus a crucial point in applications. In this paper, we concentrate on p-boxes (a simple and popular model), and on lower expectations computed over non-monotone functions. For various particular cases, we propose tractable methods to compute approximations or exact values of these lower expectations. We found interesting to compare two approaches: the first using general linear programming, and the second using the fact that p-boxes are special cases of random sets. We underline the complementarity of both approaches, as well as the differences.

Keywords. P-boxes, Random sets, Linear programming, Lower/upper expectation, Optimization

1 Introduction

When dealing with scarce information or with indeterminate beliefs, imprecise probability theory [11], together with lower previsions (expectations), offer a very appealing framework, for its mathematical soundness as well as for its well-defined behavioral interpretation. Nevertheless, computing lower previsions by means of the so-called natural extension when beliefs are modeled by a set of precise probability distributions on a continuous space (here, the reals) is often a very hard problem. Thus, building tractable methods to compute good approximations or exact values of such lower previsions is essential in applications.

First, let us note that the methods proposed here are "interpretation"-independent, and are valid both in Walley's behavioral theory (where the existence of an "ideal" precise distribution is not generally assumed) as well as in a more classical Bayesian sensitivity analysis framework (where not enough information is available to precisely know the "true" probability distribution). Thus, although interpretation issue is very important, we won't deal with it in the sequel, where an "interpretation-free" vocabulary is adopted.

In this paper, we concentrate on the case when probabilistic models are p-boxes and when the function (i.e. gamble in Walley's theory) over which is computed the lower expectation is non-monotone and whose behavior is (partially) known . In other words, we propose efficient algorithms for computing lower and upper expectations of non-monotone functions of various types under the condition that the given uncertainty model is p-box.

P-boxes are one of the simplest and most popular model of sets of probability distributions, directly extending cumulative distributions used in the precise case. Although we admit that the poor expressive power of p-boxes (a price to pay for the simplicity of the model) is a limitation, we believe that they can be a good first approximation that allows for more efficient computations, and that if a decision can be taken using them, there is no reason to use a more complex model. Moreover, we should be able to efficiently compute with simple models before thinking of stepping towards more complex ones.

Although we will briefly deal with the trivial case of monotone functions, they are, as well as functions whose behavior is completely unknown, two extreme cases that will seldom be encountered in real applications (at least in "human sized" models). In most real applications, the function of interest is non-monotone but some of its characteristics are known.

Methods developed in the paper are based on two different approaches, and we found interesting to emphasize similarities and differences between these approaches, as well as how one approach can help the other: the first is based on the fact that natural extension can be viewed as a linear programming problem,

while the second use the fact that a p-box is a particular case of random set.

The next section states the problem we're going to deal with. Section 3 then explores how to compute both the unconditional and conditional interval-valued expectations of a function of one variable having one maximum. The multivariate case when the function of a set of variables has one maximum is then explored in section 4 . Finally, section 5 illustrates how results could be extended to more complicated functions.

2 Problem statement

We assume that the information about a (real) variable X is (or can be) represented by some (continuous) lower $\underline{F}$ and upper $\overline{F}$ probability distributions defining the p-box $[\underline{F}, \overline{F}]$ [5]. Lower $\underline{F}$ and upper $\overline{F}$ thus define a set of precise distributions s.t.

$$\underline{F}(x) \leq F(x) \leq \overline{F}(x), \ \forall x \in \mathbb{R}. \quad (1)$$

Given a function $h(X)$, lower ($\underline{\mathbb{E}}$) and upper ($\overline{\mathbb{E}}$) expectations over $[\underline{F}, \overline{F}]$ of $h(X)$ can be computed by means of a procedure called natural extension [11, 12], which corresponds to the following equations:

$$\underline{\mathbb{E}}(h) = \inf_{\underline{F} \leq F \leq \overline{F}} \int_{\mathbb{R}} h(x) \mathrm{d}F, \overline{\mathbb{E}}(h) = \sup_{\underline{F} \leq F \leq \overline{F}} \int_{\mathbb{R}} h(x) \mathrm{d}F \quad (2)$$

and computing the lower (upper) expectation can be seen as finding the "optimal" distribution F reaching the infimum (supremum) in equations (2). We now detail the two generic approaches used throughout the paper.

2.1 Linear programming view

Numerically solving the above problem can be done by approximating the probability distribution function F by a set of N points $F(x_i)$, $i = 1, ..., N$, and by translating equations (2) into the corresponding linear programming problem with N optimization variables and where constraints correspond to equation (1). Those linear programming problems are of the form

$$\underline{\mathbb{E}}^*(h) = \inf \sum_{k=1}^{N} h(x_k) z_k \ \text{ or } \ \overline{\mathbb{E}}^*(h) = \sup \sum_{k=1}^{N} h(x_k) z_k$$

subject to

$$z_i \geq 0, \ \ i = 1, ..., N, \ \sum_{k=1}^{N} z_k = 1,$$

$$\sum_{k=1}^{i} z_k \leq \overline{F}(x_i), \ \sum_{k=1}^{i} z_k \geq \underline{F}(x_i), \ i = 1, ..., N.$$

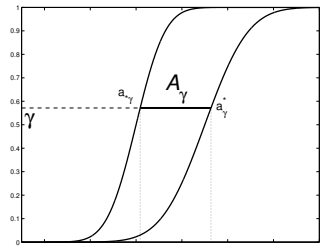


Figure 1: P-box as random set, illustration

where the z_k are the optimization variables, and objective function $\underline{\mathbb{E}}^*(h)$ ($\overline{\mathbb{E}}^*(h)$) is an approximation of the lower (upper) expectation. This way of determining the lower and upper expectations meets some computation difficulties when the value of N is rather large. Indeed, the optimization problems have N variables and $3N + 1$ constraints. On the other hand, by taking a small value of N, we take the risk of obtaining bad approximations of the exact value.

2.2 Random set view

Now that we have given a global sketch of the linear programming approach, we can detail the one using random sets. Formally, a random set is a mapping Γ from a probability space to the power set $\wp(X)$ of another space X, also called a multi-valued mapping. This mapping induces lower and upper probabilities on X [3]. Here, we shall consider the probability space $[0, 1]$ equipped with Lebesgue measure, and space $\wp(X)$ will be the measurable subsets of the real line $\mathbb{R}$.

Given the (uniformly continuous) p-box $[\underline{F}, \overline{F}]$, we will note $A_\gamma = [a_{*\gamma}, a^*_\gamma]$ the set s.t.

$$a_{*\gamma} := \sup\{x \in [a_{inf}, a_{sup}] : \overline{F}(x) < \gamma\} = \overline{F}^{-1}(\gamma),$$
$$a^*_\gamma := \inf\{x \in [a_{inf}, a_{sup}] : \underline{F}(x) > \gamma\} = \underline{F}^{-1}(\gamma),$$

By extending results from [7, 5, 4] to the continuous real line, we can conclude that the p-box $[\underline{F}, \overline{F}]$ is equivalent to the continuous random set with a uniform mass density on $[0, 1]$ and a mapping (see figure 1) s.t.

$$\Gamma(\gamma) = A_\gamma = [a_{*\gamma}, a^*_\gamma] \ \gamma \in [0, 1].$$

The interest of this mapping is that it allows us to rewrite equations (2) and the Choquet integral in a

"Lebesgue" type integral, namely

$$\underline{\mathbb{E}}(h) = \int_0^1 \inf_{x\in A_\gamma} h(x)\, d\gamma, \tag{3}$$

$$\overline{\mathbb{E}}(h) = \int_0^1 \sup_{x\in A_\gamma} h(x)\, d\gamma. \tag{4}$$

Finding analytical solutions of such integrals is not easy in the general case, but approximations (either inner or outer) can be more or less easy to compute by discretizing the p-box on a finite number of levels γ_i, the main difficulty in the general case being to find the infimum or supremum of $h(X)$ for each discretized level. As in the case of linear programming, choosing too few levels γ_i or using poor heuristics can lead to bad approximations.

In both cases, it is obvious that the optimal probability distribution F providing the minimum (maximum) expectation of h depends on the form of the function h. If this form follows some typical cases, efficient solutions can be found to compute lower (upper) expectations. The simplest examples (for which solutions are well known) of such typical cases are monotone functions.

2.3 The simple case of monotone functions

Let h be a monotone function that is non-decreasing (non-increasing) in $\mathbb{R}$, then we have [12]:

$$\underline{\mathbb{E}}(h) = \int_{\mathbb{R}} h(x)\mathrm{d}\overline{F} \quad \underline{\mathbb{E}}(h) = \int_{\mathbb{R}} h(x)\mathrm{d}\underline{F} \quad , \tag{5}$$

$$\overline{\mathbb{E}}(h) = \int_{\mathbb{R}} h(x)\mathrm{d}\underline{F} \quad \overline{\mathbb{E}}(h) = \int_{\mathbb{R}} h(x)\mathrm{d}\overline{F} \quad , \tag{6}$$

and we see from (5)-(6) that lower and upper expectations are completely determined by bounding distributions $\underline{F}$ and $\overline{F}$. Using equations (3)-(4), we get the following formulas

$$\underline{\mathbb{E}}(h) = \int_0^1 h(a_{*\gamma})d\gamma \quad \underline{\mathbb{E}}(h) = \int_0^1 h(a^*_\gamma)d\gamma \quad , \tag{7}$$

$$\overline{\mathbb{E}}(h) = \int_0^1 h(a^*_\gamma)d\gamma \quad \overline{\mathbb{E}}(h) = \int_0^1 h(a_{*\gamma})d\gamma \quad , \tag{8}$$

which are the counterparts of equations (5)-(6). Here, expectations are totally determined by extreme values of the mappings. When h is non-monotone, equations (5)-(8) provide inner approximations of $\underline{\mathbb{E}}(h)$,$\overline{\mathbb{E}}(h)$.

We then explore the cases where our knowledge of h can greatly improve those approximations (and even make them become exact values) without too much extra computational cost.

3 Function with one maximum - univariate case

In this section, we study the case where the function h has one maximum at point a, i.e. h is increasing (decreasing) in $(-\infty, a]$ ($[a, \infty)$). The case of h having one minimum easily follows.

3.1 Unconditional expectations

Proposition 1. *If the function h has one maximum at point $a \in \mathbb{R}$, then the upper and lower expectations of $h(X)$ on $[\underline{F}, \overline{F}]$ are*

$$\overline{\mathbb{E}}(h) = \int_{-\infty}^{a} h(x)\mathrm{d}\underline{F} + h(a)\left[\overline{F}(a) - \underline{F}(a)\right] + \int_{a}^{\infty} h(x)\mathrm{d}\overline{F}, \tag{9}$$

$$\underline{\mathbb{E}}(h) = \left[\int_{-\infty}^{\overline{F}^{-1}(\alpha)} h(x)\mathrm{d}\overline{F} + \int_{\underline{F}^{-1}(\alpha)}^{\infty} h(x)\mathrm{d}\underline{F}\right], \tag{10}$$

or, equivalently

$$\overline{\mathbb{E}}(h) = \int_0^{\underline{F}(a)} h(a^*_\gamma)d\gamma + [\overline{F}(a) - \underline{F}(a)]h(a) + \int_{\overline{F}(a)}^{1} h(a_{*\gamma})d\gamma \tag{11}$$

$$\underline{\mathbb{E}}(h) = \int_0^{\alpha} h(a_{*\gamma})d\gamma + \int_{\alpha}^{1} h(a^*_\gamma)d\gamma, \tag{12}$$

where α is one of the solution of the equation

$$h\ \ \overline{F}^{-1}(\alpha)\ \ = h\ \ \underline{F}^{-1}(\alpha)\ \ . \tag{13}$$

Proof using linear programming. We assume that the function $h(x)$ is differentiable in $\mathbb{R}$ and has a finite value by $x \to \infty$. The lower and upper cumulative probability functions $\underline{F}$ and $\overline{F}$ are also differentiable. Then the following primal and dual optimization problems can be written for computing the lower expectation of the function h:

Primal problem:
Minimize $v = \int_{-\infty}^{\infty} h(x)\,\rho(x)\,\mathrm{d}x$
subject to
$\rho(x) \geq 0, \int_{-\infty}^{\infty} \rho(x)\,\mathrm{d}x = 1,$
$-\int_{-\infty}^{x} \rho(x)\,\mathrm{d}x \geq -\overline{F}(x),$
$\int_{-\infty}^{x} \rho(x)\,\mathrm{d}x \geq \underline{F}(x).$
Dual problem:
Max. $w = c_0 + \int_{-\infty}^{\infty} -c(t)\,\overline{F}(t) + d(t)\,\underline{F}(t)\ \ \mathrm{d}t$
subject to
$c_0 + \int_x^{\infty} (-c(t) + d(t))\,\mathrm{d}t \leq h(x), c_0 \in \mathbb{R},$
$c(x) \geq 0,\ d(x) \geq 0.$

The proof that equations (9)-(10) and (13) are right then follows in three main steps:

1. We propose a feasible solution of the primal problem.
2. We then consider the feasible solution of the dual problem corresponding to the one proposed for the primal problem.
3. We show that the two solutions coincide and, therefore, according to the basic duality theorem of linear programming, these solutions are optimal ones.

First, we consider the primal problem. Let a' and a'' be real values. The function

$$\rho(x) = \begin{cases} \mathrm{d}\overline{F}(x)/\mathrm{d}x, & x < a' \\ 0, & a' \leq x \leq a'' \\ \mathrm{d}\underline{F}(x)/\mathrm{d}x, & a'' < x \end{cases}$$

is a feasible solution to the primal problem if the following conditions are respected:

$$\int_{-\infty}^{\infty} \rho(x)\,\mathrm{d}x = 1,$$

which, given the above solution, can be rewritten

$$\int_{-\infty}^{a'} \mathrm{d}\overline{F} + \int_{a''}^{\infty} \mathrm{d}\underline{F} = 1,$$

which is equivalent to the equality

$$\overline{F}(a') = \underline{F}(a''). \tag{14}$$

We now interest ourselves in the dual problem. Let us first consider the sole constraint

$$c_0 + \int_x^{\infty} (-c(t) + d(t))\,\mathrm{d}t \leq h(x), \tag{15}$$

which is the equivalent of the primal constraint $\rho(x) \geq 0$. We then consider the following feasible solution to the dual problem as $c_0 = h(\infty)$,

$$c(x) = \begin{array}{ll} h'(x), & x < a' \\ 0, & x \geq a' \end{array} \quad d(x) = \begin{array}{ll} 0, & x < a'' \\ -h'(x), & x \geq a'' \end{array}.$$

The inequalities $c(x) \geq 0$ and $d(x) \geq 0$ are valid provided we have the inequalities $a' \leq a \leq a''$ (i.e. interval $[a', a'']$ encompasses maximum of h). By integrating $c(x)$ and $d(x)$, we get the increasing function

$$C(x) = -\int_x^{\infty} c(t)\,\mathrm{d}t = \begin{array}{ll} h(x) - h(a'), & x < a' \\ 0, & x \geq a' \end{array}$$

and the decreasing function

$$D(x) = \int_x^{\infty} d(t)\,\mathrm{d}t = \begin{array}{ll} h(a'') - h(\infty), & x < a'' \\ h(x) - h(\infty), & x \geq a'' \end{array}.$$

Let us rewrite condition (15) as follows:

$$c_0 + C(x) + D(x) \leq h(x). \tag{16}$$

If $x < a'$, equation (16) reads

$$c_0 + h(x) - h(a') + h(a'') - h(\infty) = h(x).$$

Hence

$$h(a'') = h(a'). \tag{17}$$

If $a' < x < a''$, we have $c_0 + h(a'') - h(\infty) \leq h(x)$ which means that for all $x \in (a', a'')$ we have $h(a'')\,(= h(a')) \leq h(x)$ (i.e. $h(a'')$ and a' are the minimal values of the function $h(x)$ in interval $x \in (a', a'')$.) If $x \geq a''$, then we get the trivial equality $c_0 + h(x) - h(\infty) = h(x)$. The two proposed solutions are valid iff equation (14) is valid for the primal problem and equation (17) is valid for the dual problem. To show that they are actually valid, let us consider the function

$$\varphi(\alpha) = h\ \overline{F}^{-1}(\alpha)\ - h\ \underline{F}^{-1}(\alpha)\ ,$$

which, being a substraction of two continuous functions (by supposition), is continuous. Since the function h has its maximum at point $x = a$, then, by taking $\alpha = \underline{F}(a)$, we get the inequality

$$\varphi(\underline{F}(a)) = h\ \overline{F}^{-1}(\underline{F}(a))\ - h(a) \leq 0$$

and, by taking $\alpha = \overline{F}(a)$, we get the inequality

$$\varphi\ \overline{F}(a)\ = h(a) - h\ \underline{F}^{-1}\ \overline{F}(a)\ \geq 0.$$

Consequently, there exists α in the interval $\underline{F}(a), \overline{F}(a)$ such that $\varphi(\alpha) = 0$ (since φ is continuous). Therefore, there exist $a' = \overline{F}^{-1}(\alpha)$ and $a'' = \underline{F}^{-1}(\alpha)$ (hence, equality (14) holds) such that equality $h(a') = h(a'')$ in (17) is valid. We find the values of the objective functions

$$v_{\min} = \int_0^{a'} h(x)\,\mathrm{d}\overline{F} + \int_{a''}^{\infty} h(x)\,\mathrm{d}\underline{F},$$

$$w_{\max} = c_0 + \int_0^{\infty} -c(t)\,\overline{F}(t) + d(t)\,\underline{F}(t)\ \mathrm{d}t.$$

and, by using integration by parts together with equations (14)-(17), we can show that equality $w_{\max} = v_{\min}$ holds, with α the particular solution of equation (13) for which optimum is reached, as was to be proved. □

Proof using random sets. Let us now consider equations (4)-(3). Looking first at equation (4), we see that before $\gamma = \underline{F}(a)$, the supremum of h on A_γ is

$h(a^*_\gamma)$, since h is increasing between $[\infty, a]$. Between $\gamma = \underline{F}(a)$ and $\gamma = \overline{F}(a)$, the supremum of h on A_γ is $f(a)$. After $\gamma = \overline{F}(a)$, we can make the same reasoning as for the increasing part of h (except that it is now decreasing). Finally, this gives us the following formula:

$$\overline{\mathbb{E}}(h) = \int_0^{\underline{F}(a)} h(a^*_\gamma)d\gamma + \int_{\underline{F}(a)}^{\overline{F}(a)} h(a)d\gamma + \int_{\overline{F}(a)}^{1} h(a_{*\gamma})d\gamma \quad (18)$$

which is equivalent to (11). Let us now turn to the lower expectation. Before $\gamma = \underline{F}(a)$ and after $\gamma = \overline{F}(a)$, finding the infinimum is again not a problem (it is respectively $h(a_{*\gamma})$ and $h(a^*_\gamma)$). Between $\gamma = \underline{F}(a)$ and $\gamma = \overline{F}(a)$, since we know that h is increasing before $x = a$ and decreasing after, infinimum is either $h(a_{*\gamma})$ or $h(a^*_\gamma)$. This gives us equation

$$\underline{\mathbb{E}}h = \int_0^{\underline{F}(a)} h(a_{*\gamma})d\gamma + \int_{\underline{F}(a)}^{\overline{F}(a)} \min(h(a_{*\gamma}), h(a^*_\gamma))d\gamma + \int_{\overline{F}(a)}^{1} h(a^*_\gamma)d\gamma \quad (19)$$

and if we use equations (14),(17) as in the first proof (reasoning used in the first proof to show that they have a solution is general, and thus applicable here), we know that there is a level α s.t. $h(\overline{F}^{-1}(\alpha)) = h(\underline{F}^{-1}(\alpha))$, and for which equation (19) reduce to equation (13). □

Solutions for a function h having a minimum directly follow, due to the duality between lower and upper expectations [12] (i.e. $\underline{\mathbb{E}}(-h) = -\overline{\mathbb{E}}(h)$ and $\overline{\mathbb{E}}(-h) = -\underline{\mathbb{E}}(h)$). Of course, both proofs lead to similar formulas and, in applications, would lead to the same lower and upper expectations. Nevertheless, each view suggests a different way to solve the problem or to approximate the solution.

The proof using linear programming and the associated formulas suggest a more analytical and explicit solution, where we have to find the level α satisfying equation (14). If an analytical solution is not available, then the solution is generally approximated by scanning a larger or smaller range of possible values for α(see [10] for an example). On the other side, the proof is shorter in the case of random set, but the presence of a level α is hardly visible at first sight, and analytical results are more difficult to derive. Compared to the linear programming view, equations (11),(12),(19) suggest numerical methods based on a discretization of the levels γ rather than a heuristic search of the level α satisfying equation (14). Let us note that in the worst case, two evaluations are needed at each of the discretized levels (using equation (19)).

If the function h is symmetric about a, i.e., the equality $h(a - x) = h(a + x)$ is valid for all $x \in \mathbb{R}$, then the value of α in (13) does not depend on h and is determined as

$$a - \overline{F}^{-1}(\alpha) = \underline{F}^{-1}(\alpha) - a.$$

Note that expressions (5),(6) can be obtained from (9),(10) by taking $a \to \infty$.

3.2 Conditional expectations

Suppose that we observe an event $B = [b_0, b_1]$. Then the lower and upper conditional expectations under condition of B can be determined as follows:

$$\underline{\mathbb{E}}(h|B) = \inf_{\underline{F} \le F \le \overline{F}} \frac{\int_{\mathbb{R}} h(x) I_B(x) \mathrm{d}F}{\int_{\mathbb{R}} I_B(x) \mathrm{d}F},$$

$$\overline{\mathbb{E}}(h|B) = \sup_{\underline{F} \le F \le \overline{F}} \frac{\int_{\mathbb{R}} h(x) I_B(x) \mathrm{d}F}{\int_{\mathbb{R}} I_B(x) \mathrm{d}F}.$$

Generally speaking, the above problems can numerically be solved by approximating the probability distribution function F by a set of N points $F(x_i)$, $i = 1, ..., N$, and by writing linear-fractional optimization problems and then linear programming problems. Problems mentioned for the unconditional case can again occur. Figure 2 illustrates a potential optimal distribution F for which upper conditional expectation is reached (under the condition $B = [1, 8]$) when h has one maximum (which value is 5 in figure 2).

Proposition 2. *If the function h has one maximum at point $a \in \mathbb{R}$, then the upper and lower conditional expectations of $h(X)$ on $[\underline{F}, \overline{F}]$ after observing the event B are*

$$\overline{\mathbb{E}}(h|B) = \sup_{\substack{\underline{F}(b_0) \le \alpha \le \overline{F}(b_0) \\ \underline{F}(b_1) \le \beta \le \overline{F}(b_1)}} \frac{1}{\beta - \alpha} \Psi(\alpha, \beta),$$

$$\underline{\mathbb{E}}(h|B) = \inf_{\substack{\underline{F}(b_0) \le \alpha \le \overline{F}(b_0) \\ \underline{F}(b_1) \le \beta \le \overline{F}(b_1)}} \frac{1}{\beta - \alpha} \Phi(\alpha, \beta),$$

$$\begin{aligned}
\Psi(\alpha, \beta) &= I(\alpha < \underline{F}^{-1}(a)) \int_{\underline{F}^{-1}(\alpha)}^{a} h(x) \mathrm{d}\underline{F} \\
&+ I(\beta > \overline{F}^{-1}(a)) \int_{a}^{\overline{F}^{-1}(\beta)} h(x) \mathrm{d}\overline{F} \\
&+ h(a) \ \min(\overline{F}(a), \beta) - \max(\underline{F}(a), \alpha) \\
&= \int_\alpha^\beta \sup_{x \in A_\gamma} h(x) \mathrm{d}\gamma.
\end{aligned}$$

$$\begin{aligned}
\Phi(\alpha, \beta) &= h(b_0) \ \overline{F}(b_0) - \alpha \ + \int_{b_0}^{\overline{F}^{-1}(\varepsilon)} h(x) \mathrm{d}\overline{F} \\
&+ h(b_1) \left(\beta - \underline{F}(b_1)\right) + \int_{\underline{F}^{-1}(\varepsilon)}^{b_1} h(x) \mathrm{d}\underline{F} \\
&= \int_\alpha^\beta \inf_{x \in A_\gamma} h(x) \mathrm{d}\gamma.
\end{aligned}$$

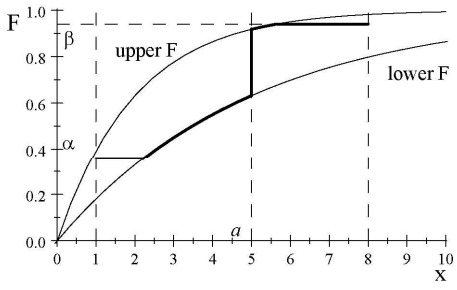


Figure 2: Optimal distribution (thick) for computing upper conditional expectation on $B = [1, 8]$

Here $I(a < b)$ is the indicator function taking 1 if $a < b$ and 0 if $a \geq b$; ε is one of the roots of the following equation:

$$h\ \ \overline{F}^{-1}(\varepsilon)\ \ = h\ \ \underline{F}^{-1}(\varepsilon)\ \ . \tag{20}$$

General proof. We consider only upper expectation. We do not know how the optimal distribution function behaves outside the interval B. Therefore, we suppose that the value of the optimal distribution function at point b_0 is $F(b_0) = \alpha \in [\underline{F}(b_0), \overline{F}(b_0)]$ and its value at point b_1 is $F(b_1) = \beta \in [\underline{F}(b_1), \overline{F}(b_1)]$ (see Fig. 2). Then there holds

$$\int_{\mathbb{R}} I_B(x)\mathrm{d}F(x) = \beta - \alpha.$$

Hence, we can write

$$\overline{\mathbb{E}}(h|B) = \sup_{\substack{\underline{F}(b_0)\leq\alpha\leq\overline{F}(b_0)\\ \underline{F}(b_1)\leq\beta\leq\overline{F}(b_1)\\ \underline{F}\leq F\leq\overline{F}}} \frac{1}{\beta-\alpha}\int_{\mathbb{R}} h(x)I_B(x)\mathrm{d}F(x)$$

$$= \sup_{\substack{\underline{F}(b_0)\leq\alpha\leq\overline{F}(b_0)\\ \underline{F}(b_1)\leq\beta\leq\overline{F}(b_1)}} \frac{1}{\beta-\alpha}\left(\sup_{\substack{\underline{F}\leq F\leq\overline{F}\\ F(b_0)=\alpha\\ F(b_1)=\beta}}\int_{\mathbb{R}} h(x)I_B(x)\mathrm{d}F(x)\right)$$

$$= \sup_{\substack{\underline{F}(b_0)\leq\alpha\leq\overline{F}(b_0)\\ \underline{F}(b_1)\leq\beta\leq\overline{F}(b_1)}} \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta}\sup_{x\in A_\gamma} h(x)\mathrm{d}\gamma. \tag{21}$$

Here $A_\gamma = B\cap[\underline{F}^{-1}(\gamma), \overline{F}^{-1}(\gamma)]$. By using the results obtained for the unconditional upper expectation, we can see that the integrand is equal to $\Psi(\alpha, \beta)$. The lower expectation is similarly proved, and conditional expectations when h has one minimum immediately follow. □

Equation (21) shows that, as value $\beta - \alpha$ increases, so do the numerator and denominator, thus playing opposite role in the evolution of the objective function. Hence, computing the upper (lower) conditional expectation consists in finding the values β and α s.t. any increase (decrease) in the value $\beta - \alpha$ is greater (lower) than the corresponding increase (decrease) in $\Psi(\alpha, \beta)$.

A crude algorithm to approximate the solution would be to start from the largest (tightest) interval $[\alpha, \beta]$ and then to gradually shrink (enlarge) it, evaluating each time equation (21) and retaining the highest obtained value (let us note that we can have $\overline{F}(b_0) \geq \underline{F}(b_1)$, thus the tightest interval can be void).

Another interesting point to note is that the proof takes advantage of both views, since the idea to use levels α and β comes from fractional linear programming, while the final equation (21) can be elegantly formulated by using the random set view.

4 Function with one maximum - multivariate case

Now, let h be a function from $\mathbb{R}^2 \to \mathbb{R}$ which depends on two variables X and Y. The uncertainty model becomes the following bivariate p-box:

$$\underline{F}(x, y) \leq F(x, y) \leq \overline{F}(x, y),\ \forall (x, y) \in \mathbb{R}^2.$$

Again, we assume that h has one global maximum at point (x_0, y_0) and that $\forall z,\ \ \partial h(x, z)/\partial x = 0$ and $\partial h(z, y)/\partial y = 0$ respectively have solutions at points $x = x_0$ and $y = y_0$, making the task to find infinima and suprema easier in further equations . In the next sections, we explore how we would solve the problem, under some common independence hypothesis existing in the framework of imprecise probabilities [2]. In this paper, we only provide an outline, giving general ideas and underlining the most interesting points.

In the sequel, we will consider, for the marginal random set of variable Y, the sets $B_\kappa = [b_{*\kappa}, b^*_\kappa]$ s.t.

$$b_{*\kappa} := \sup\{y \in [b_{inf}, b_{sup}] : \overline{F}(y) < \kappa\} = \overline{F}^{-1}(\kappa),$$
$$b^*_\kappa := \inf\{y \in [b_{inf}, b_{sup}] : \underline{F}(y) > \kappa\} = \underline{F}^{-1}(\kappa).$$

Moreover, following Smets [9], we will note $f^{\mathcal{M}}_X$ and $f^{\mathcal{M}}_Y$ the basic belief densities corresponding to the continuous random sets of $[\underline{F}, \overline{F}]_X$,$[\underline{F}, \overline{F}]_Y$ when needed.

4.1 Strong Independence

In the case of strong independence, we can write

$$\underline{\mathbb{E}}(h) = \inf_{\underline{F}_1\leq F_1\leq\overline{F}_1}\ \inf_{\underline{F}_2\leq F_2\leq\overline{F}_2}\int_{\mathbb{R}}\int_{\mathbb{R}} h(x, y)\mathrm{d}F_1\mathrm{d}F_2,$$

$$\overline{\mathbb{E}}(h) = \sup_{\underline{F}_1 \le F_1 \le \overline{F}_1} \sup_{\underline{F}_2 \le F_2 \le \overline{F}_2} \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y) \mathrm{d}F_1 \mathrm{d}F_2.$$

The simplest case is when the function h can be represented as $h(X,Y) = h_1(X)h_2(Y)$. Then $\underline{\mathbb{E}}(h) = \underline{\mathbb{E}}(h_1) \cdot \underline{\mathbb{E}}(h_2)$ and $\overline{\mathbb{E}}(h) = \overline{\mathbb{E}}(h_1) \cdot \overline{\mathbb{E}}(h_2)$. However, we consider a more complex case. Let us fix the second variable Y at point z. Denote

$$\xi(z) = \int_{\mathbb{R}} h(x,z) \mathrm{d}F_1(x).$$

Then we have

$$\mathbb{E}(h(X,Y)) = \int_{\mathbb{R}} \xi(z) \mathrm{d}F_2(z).$$

Let us fix variable Y to value z. Given our particular $h(X,Y)$ and Proposition 1, we have

$$\begin{aligned}\overline{\mathbb{E}}(h(X,z)) &= \sup_{\underline{F}_1 \le F_1 \le \overline{F}_1} \xi(z) \\ &= h(x_0,z)\left[\overline{F}_1(x_0) - \underline{F}_1(x_0)\right] + \\ &\int_{-\infty}^{x_0} h(x,z) \mathrm{d}\underline{F}_1 + \int_{x_0}^{\infty} h(x,z) \mathrm{d}\overline{F}_1. \end{aligned} \quad (22)$$

Given the assumption we've made on $h(X,Y)$ behavior, function $\xi(z)$ has a maximum at point $z = y_0$ and is monotone in intervals $(-\infty, z_0)$ and (z_0, ∞), whatever the value of x. This implies that the optimal distribution F_2 is of the form considered in Proposition 1. Moreover, the following inequality

$$\sup_{\underline{F}_2 \le F_2 \le \overline{F}_2} \int_{\mathbb{R}} \xi(z) \mathrm{d}F_2(z) \ge \sup_{\underline{F}_2 \le F_2 \le \overline{F}_2} \int_{\mathbb{R}} \widehat{\xi}(z) \mathrm{d}F_2(z)$$

holds if $\xi(z) \ge \widehat{\xi}(z)$. Then it follows from the above and from the form of the optimal distribution F_2 determined in Proposition 1 that

$$\begin{aligned}\overline{\mathbb{E}}(h(X,Y)) &= \sup_{\underline{F}_2 \le F_2 \le \overline{F}_2} \int_{\mathbb{R}} \overline{\mathbb{E}}(h(X,z)) \mathrm{d}F_2(z) \\ &= \sup \xi(y_0)\left[\overline{F}_2(y_0) - \underline{F}_2(y_0)\right] + \\ &\int_{-\infty}^{y_0} \sup \xi(z) \mathrm{d}\underline{F}_2(z) + \int_{y_0}^{\infty} \sup \xi(z) \mathrm{d}\overline{F}_2(z)\end{aligned}$$

and $\sup \xi(z)$ is given by equation (22). Upper expectation under strong independence can then be found in an almost explicit form. The same is not true for the lower expectation, since, relying on the first proof of Proposition 1, $\inf \xi(z)$ is obtained in this case by solving the equation

$$h(\overline{F}_1^{-1}(\alpha), z) = h(\underline{F}_1^{-1}(\alpha), z).$$

where the root α obviously depends on z. By denoting this dependency as α_z, we can nevertheless derive the following formula

$$\begin{aligned}\underline{\mathbb{E}}(h(X,Y)) &= \inf_{\underline{F}_2 \le F_2 \le \overline{F}_2} \int_{\mathbb{R}} \underline{\mathbb{E}}(h(X,z)) \mathrm{d}F_2(z) \\ &= \int_{-\infty}^{\overline{F}_2^{-1}(\beta)} \int_{-\infty}^{\overline{F}_1^{-1}(\alpha_z)} h(x,z) \mathrm{d}\overline{F}_1 \mathrm{d}\overline{F}_2 \\ &+ \int_{-\infty}^{\overline{F}_2^{-1}(\beta)} \int_{\underline{F}_1^{-1}(\alpha_z)}^{\infty} h(x,z) \mathrm{d}\underline{F}_1 \mathrm{d}\overline{F}_2 \\ &+ \int_{\underline{F}_2^{-1}(\beta)}^{\infty} \int_{-\infty}^{\overline{F}_1^{-1}(\alpha_z)} h(x,z) \mathrm{d}\overline{F}_1 \mathrm{d}\underline{F}_2 \\ &+ \int_{\underline{F}_2^{-1}(\beta)}^{\infty} \int_{\underline{F}_1^{-1}(\alpha_z)}^{\infty} h(x,z) \mathrm{d}\underline{F}_1 \mathrm{d}\underline{F}_2.\end{aligned}$$

where β is a root of the equation

$$\underline{\mathbb{E}}(h(X, \underline{F}_2^{-1}(\beta))) = \underline{\mathbb{E}}(h(X, \overline{F}_2^{-1}(\beta)).$$

and only an approximation of such a lower bound can be found.

For the strong independence case, results rely heavily on the linear programming view and allow us to derive nice analytical formulas. Although we could set the problem in a random set framework, it would lead to numerical solutions less efficient than the one presented here (difficult problems already arise when computing lower and upper probabilities [6]).

Next cases emphasize the random set view, since this view makes solutions easier to state (especially, as could be expected, in the random set independence case).

4.2 Random set Independence

In the case of random set independence, lower and upper expectations can be computed by the following formulas:

$$\underline{\mathbb{E}}(h) = \int_0^1 \int_0^1 \inf_{(x,y) \in [B_\kappa \times A_\gamma]} h(x,y) \mathrm{d}\kappa \mathrm{d}\gamma,$$

$$\overline{\mathbb{E}}(h) = \int_0^1 \int_0^1 \sup_{(x,y) \in [B_\kappa \times A_\gamma]} h(x,y) \mathrm{d}\kappa \mathrm{d}\gamma,$$

for which we can get a numerical approximation as close as we want to the exact value, by discretizing each integral. Moreover, in our particular case, evaluating $\inf h(x,y)$ or $\sup h(x,y)$ is easy.

Indeed, if h is as stated above, finding the supremum or infimum of h on $[B_\kappa \times A_\gamma]$ will often require only one computation: when $b_\kappa^* \le y_0$ and $a_\kappa^* \le x_0$, the supremum and infimum values are respectively on the vertices (b_κ^*, a_κ^*) and $(b_{*\kappa}, a_{*\gamma})$ of the square. when $b_{*\kappa} \le y_0 \le b_\kappa^*$ and $a_\gamma^* \le x_0$, the supremum is at point

(a^*_γ, y_0) and the infimum is either at point $(a_{*\gamma}, b_{*\kappa})$ or $(a_{*\gamma}, b^*_\kappa)$. In the case where the square contains point (x_0, y_0), this point is the supremum and the infimum is on one of the four vertices of the square. All other situations easily follow.

From a numerical standpoint, we can note that assuming random set independence is equivalent to assuming independence in a Monte-Carlo sampling scheme where each sample consists of two randomly chosen intervals A_γ and B_κ.

4.3 Unknown Interaction

Since p-boxes are special case of random sets, we can follow Fetz and Oberguggenberger [6], who show that considering unknown interaction when marginals are random sets is equivalent to consider the set of all possible joint random sets having the latter for marginals, and using results from [9] (where the extension of continuous belief functions to n-dimensional case is briefly sketched), computing lower (upper) expectation can be expressed as follows:

$$\underline{\mathbb{E}}(h) = \inf_{f^{\mathcal{M}}_{XY} \in \mathcal{J}_{XY}} \iiiint_{x_1 x_2 y_1 y_2} \inf_{\substack{x \in [x_1, x_2] \\ y \in [y_1, y_2]}} h(x,y) \mathrm{D} f^{\mathcal{M}}_{XY},$$

$$\overline{\mathbb{E}}(h) = \sup_{f^{\mathcal{M}}_{XY} \in \mathcal{J}_{XY}} \iiiint_{x_1 x_2 y_1 y_2} \sup_{\substack{x \in [x_1, x_2] \\ y \in [y_1, y_2]}} h(x,y) \mathrm{D} f^{\mathcal{M}}_{XY},$$

with

$$\mathrm{D} f^{\mathcal{M}}_{XY} = f^{\mathcal{M}}_{XY}(x_1, x_2, y_1, y_2) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}y_1 \mathrm{d}y_2,$$

where $\mathcal{J}_{XY}$ is the set of all possible joint basic belief densities $f^{\mathcal{M}}_{XY}$ over $\mathbb{R}^4$ which have $f^{\mathcal{M}}_X$ and $f^{\mathcal{M}}_Y$ as their marginals.

Although the above equations are nice ways to formulate the problem, solving them analytically will be impossible in most cases. Again, the result can be approximated by approximating each p-box by a finite random set.

For instance, let us consider the two random sets $\Gamma_\gamma, \Gamma_\kappa$ approximating the p-boxes $[\underline{F}, \overline{F}]_X$,$[\underline{F}, \overline{F}]_Y$ with sets $A_{\gamma_i}, B_{\kappa_j}$, where $i, j = 1, \ldots, n$ and where all sets have equal weights (i.e. $\gamma_i - \gamma_{i-1} = \kappa_j - \kappa_{j-1} = 1/n \ \forall\ i, j$). The problem of approximating lower expectation then comes down to finding

$$\underline{\mathbb{E}}^*(h) = \inf_{\Gamma_{\gamma,\kappa} \in \Gamma^*_{\gamma,\kappa}} \sum \inf_{\substack{x \in A_{\gamma_i} \\ y \in B_{\kappa_j}}} h(x,y) m_{\Gamma_{\gamma,\kappa}}(A_{\gamma_i} \times B_{\kappa_j})$$

subject to

$$\sum_{j=1}^{n} m_{\Gamma_{\gamma,\kappa}}(A_{\gamma_i} \times B_{\kappa_j}) = m_{\Gamma_\gamma}(A_{\gamma_i}),$$

$$\sum_{i=1}^{n} m_{\Gamma_{\gamma,\kappa}}(A_{\gamma_i} \times B_{\kappa_j}) = m_{\Gamma_\gamma}(B_{\kappa_j}),$$

where $\Gamma^*_{\gamma,\kappa}$ is the set of joint random sets having $\Gamma_\gamma, \Gamma_\kappa$ for marginals, and $m_{\Gamma_{\gamma,\kappa}}(A_{\gamma_i} \times B_{\kappa_j})$ the mass attached to the focal element $A_{\gamma_i} \times B_{\kappa_j}$. Approximation of upper expectation can be derived in a similar way (i.e. replacing the inf by sup).

Although solving the above equations is not easy, we can hope to find efficient solutions, provided we can easily evaluate $\inf h(x,y)$ on elements of the Cartesian product (we have seen that it is the case here) . Also, this method can be seen as an extension of some existing methods (see [13, 8]) to functions $h(x)$ more general than indicator functions of events. Hence, we could extend some previous results concerning indicators functions to integrate some information about dependencies [1]. Another interesting thing to point out is that approximating the result in the case of unknown interaction naturally leads to a linear programming problem.

Methods given for unknown interaction and random set independence are applicable to all random sets (and only to random sets, which is a limitation compared to general linear programming), and considering special cases such as p-boxes or possibility distributions often allow the derivation of more efficient algorithms for solving the problems.

5 Function with local maxima/minima - univariate case

Now we consider a general form of the function h, i.e., the function $h(x)$ has alternate local maximum at point a_i and minimum at point b_{i-1}, $i = 1, 2, ...$, such that

$$b_0 < a_1 < b_1 < a_2 < b_2 < ... \tag{23}$$

Proposition 3. *If local maxima (a_i) and minima (b_i) of the function h satisfy condition (23), then the optimal distribution F for computing the lower unconditional expectation $\underline{\mathbb{E}}(h)$ has (vertical) jumps at points b_i, $i = 1, ...$ of the size*

$$\min\ \overline{F}(b_i), \alpha_{i+1}\ \ - \max\left(\underline{F}(b_i), \alpha_i\right).$$

Between (vertical) jumps with numbers $i-1$ and i, the optimal probability distribution function F is of

the form:

$$F(x) = \begin{cases} \overline{F}(x), & x < a' \\ \alpha_i, & a' \le x \le a'' \\ \underline{F}(x), & a'' < x \end{cases},$$

where α_i is the root of the equation

$$h \ \max\ \overline{F}^{-1}(\alpha_i), b_{i-1} \quad = h \ \min\ \underline{F}^{-1}(\alpha_i), b_i$$

in interval $\left[\underline{F}(a_i), \overline{F}(a_i)\right]$,

$$a' = \max\ \overline{F}^{-1}(\alpha_i), b_{i-1} \ , \ a'' = \min\ \underline{F}^{-1}(\alpha_i), b_i \ .$$

The upper expectation $\overline{\mathbb{E}}(h)$ can be found from the condition $\overline{\mathbb{E}}(h) = -\underline{\mathbb{E}}(-h)$.

Proof using linear programming (brief sketch). The first proof is based on the investigation of the following local primal and dual optimization problems for computing the lower expectation of h in finite interval $[b_0, b_1]$ where h has one maximum at point a_1:

Primal problem:
$v = \int_{b_0}^{b_1} h(x) f(x) \mathrm{d}x \rightarrow \min$ subject to
$f(x) \ge 0,\ F_0 \ge 0,\ F_1 \ge 0,$
$-\int_{b_0}^{x} f(t)\,\mathrm{d}t - F_0 \ge -\overline{F}(x),$
$\int_{b_0}^{x} f(t)\,\mathrm{d}t + F_0 \ge \underline{F}(x),$
$-F_0 \ge -\overline{F}(b_0), F_0 \ge \underline{F}(b_0),$
$-F_1 \ge -\overline{F}(b_1), F_1 \ge \underline{F}(b_1),$
$\int_{b_0}^{b_1} f(t)\,\mathrm{d}t + F_0 - F_1 = 0.$
Dual problem:
$w = -c_0\overline{F}(b_0) + d_0\underline{F}(b_0) - c_1\overline{F}(b_1) + d_1\underline{F}(b_1)$
$+\int_{b_0}^{b_1} \ -\overline{F}(x)\,c(x) + \underline{F}(x)\,d(x) \ \mathrm{d}x \rightarrow \max$
subject to
$e + \int_{x}^{b_1} (-c(t) + d(t))\,\mathrm{d}t \le h(x),$
$e - c_0 + d_0 + \int_{b_0}^{b_1} (-c(t) + d(t))\,\mathrm{d}t \le 0,$
$-e - c_1 + d_1 \le 0,$
$c(x) \ge 0, c_0 \ge 0, c_1 \ge 0,$
$d(x) \ge 0, d_0 \ge 0, d_1 \ge 0, e \in \mathbb{R}$

All inequalities in the above primal and dual problems are valid only for $x \in [b_0, b_1]$. Results similar to those of proposition 1 can then be derived, and it is interesting to note that b_0, b_1 play similar roles to those of α,β in the conditional case. Finding the optimal distribution between each b_o, b_1 leads to four cases, depending on the situation. Figures 3.A-D illustrate these situations. The optimal F for which the lower expectation is reached is then a succession of such subcases, with a vertical jump between each of them (in figures 3.A-D, α, b_0 and b_1 are respectively equivalent to α_i, b_i and b_{i+1} of proposition 3) . □

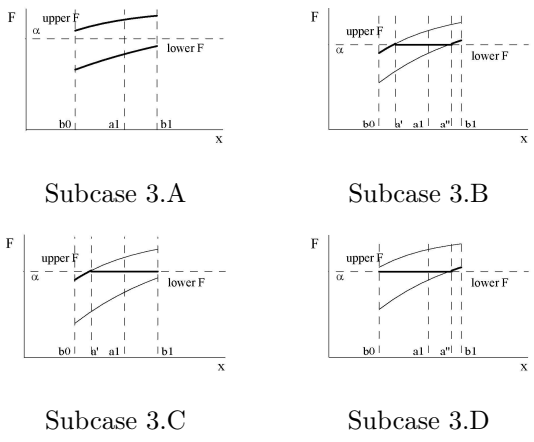


Figure 3: Subcases of piecewise optimal F

Proof using random sets (brief sketch). For convenience, we will consider that h begins with a local minimum and ends with a local maxima a_n. Formulas when h begins/ends with a local maximum (minimum) are similar. Lower/upper expectations can be computed as follows:

$$\underline{\mathbb{E}}(h) = \int_{0}^{\underline{F}(b_n)} \min_{b_i \in A_\gamma} (h(a_{*\gamma}), h(b_i), h(a^*_\gamma))d\gamma + \int_{\underline{F}(b_n)}^{1} h(a_{*\gamma})d\gamma,$$

$$\overline{\mathbb{E}}(h) = \int_{0}^{\underline{F}(a_1)} h(a^*_\gamma)d\gamma + \int_{\underline{F}(a_1)}^{\overline{F}(a_n)} \max_{a_i \in A_\gamma} (h(a_{*\gamma}), h(a_i), h(a^*_\gamma))d\gamma.$$

Let us explain a bit the equation for the lower expectation (details for upper one are similar). The most interesting part is the first integral. Let $B = [b_i, \ldots, b_j]$ $(i \le j)$ be the set of local minima included in any particular set A_γ (B can be empty). b_{i-1} and b_{j+1} are the closest local minima outside A_γ. Let us consider the situation for which the lowest local minima $h(b_k)$ s.t. $b_k \in B$ (an empty B is a degenerated case of this one) is higher than $h(b_{i-1})$,$h(b_{j+1})$. As γ increases and as set A_γ evolves, various situations can happen. Either the infinimum shifts from $h(a_{*\gamma})$ to $h(b_k)$ at some point (this is subcase 3.C) or it shifts from $h(b_k)$ to $h(a^*_\gamma)$ (subcase 3.D), or it shifts from $h(a_{*\gamma})$ to $h(a^*_\gamma)$ if $h(b_k)$ is too high (subcase 3.B). Subcase 3.A corresponds to the case of a local minimum b_i always dominating two other local minima (equivalent to b_0,b_1) in any set A_γ. The jumps in proposition 3 correspond to the situations where the infinimum of $h(x)$ has value $h(b_k)$, either until $h(b_k) = h(a^*_\gamma)$ or until b_k is on the border of A_γ as γ increases. In the first case, it corresponds to an "horizontal" jump and to one of the root α in proposition 3, while in the latter case, the vertical jump

collapses with the upper cumulative distribution. □

Similarly to figure 2, the optimal F will be a succession of vertical and horizontal jumps, sometimes following either $\overline{F}$ or $\underline{F}$ after a vertical jump has "collapsed" with $\overline{F}$ or an horizontal jump with $\underline{F}$. The proof using linear programming concentrates on "horizontal" jumps, while the proof using continuous random set emphasize vertical jumps. Again, each view suggests a different way to approximate the result.

An appealing way of formulating lower expectation is the following: let b_j $i = 1, \ldots, m$ be the local minima where we have the "vertical" jumps and $\gamma_{j_*}, \gamma_{j^*}$ the associated levels on the set $[0, 1]$. Then we have

$$\underline{\mathbb{E}}(h) = \sum_{j=1}^{m} \left(\int_{\gamma_{(j-1)^*}}^{\gamma_{j_*}} \inf_{x \in A_\gamma} (h(a^*_\gamma), h(a_{*\gamma}))d\gamma + (\gamma_{(j+1)^*} - \gamma_{j_*})h(b_j) \right). \tag{24}$$

6 Conclusions

We have considered the problem of computing lower and upper expectations on p-boxes and particular functions under two different approaches: by using linear programming and by using the fact that p-boxes are special cases of random sets. Although the two approaches try to solve identical problems, their differences suggest different ways to approximate the solutions of those problems. Moreover, some particular problems are easier to state (solve) in one approach than in the other (for example, the solutions explored in section 4). But more important than their differences is the complementarity of both approaches. Indeed, one approach can shed light on some problems shaded by the other approach (e.g. the α level of proposition 1). Another advantage of combining both approaches is the ease with which some problems are solved and the elegant formulation resulting from this combination (like in the conditional case). Let us nevertheless note that the constraint programming approach can apply to imprecise probabilities in general, while the random set approach is indeed limited to random sets.

Further works should concentrate on two directions: exploring further some ideas that were stated in the multivariate case (as well as deriving similar results for independence types not considered here), and extending the presents results to the general case of a function having many local extrema.

References

[1] D. Berleant and J. Zhang. Using pearson correlation to improve envelopes around the distributions of functions. *Reliable Computing*, 10(2):139–161, 2004. [http://ifsc.ualr.edu/jdberleant/]

[2] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[3] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[4] S. Destercke and D. Dubois. A unified view of some representations of imprecise probabilities. *Int. Conf. on Soft Methods in Probability and Statistics (SMPS)*, Advances in Soft Computing, pages 249–257, Bristol, 2006. Springer.

[5] S. Ferson, L. Ginzburg, V. Kreinovich, D. Myers, and K. Sentz. Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Laboratories, 2003. [http://www.sandia.gov/epistemic/Reports/SAND2002-4015.pdf].

[6] T. Fetz and M. Oberguggenberger. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering and System Safety*, 85:73–87, 2004.

[7] E. Kriegler and H. Held. Utilizing belief functions for the estimation of future climate change. *I. J. of Approximate Reasoning*, 39:185–209, 2005.

[8] H. Regan, S. Ferson, and D. Berleant. Equivalence of methods for uncertainty propagation of real-valued random variables. *I. J. of Approximate Reasoning*, 36:1–30, 2004.

[9] P. Smets. Belief functions on real numbers. *I. J. of Approximate Reasoning*, 40:181–223, 2005. [http://iridia.ulb.ac.be/ psmets/]

[10] L. Utkin. Risk analysis under partial prior information and non-monotone utility functions. *I. J. of Information Technology and Decision Making*, To appear. [http://www.levvu.narod.ru/select_refern.htm]

[11] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, 1991.

[12] P. Walley. Measures of uncertainty in expert systems. *Artifical Intelligence*, 83:1–58, 1996.

[13] R. Williamson and T. Downs. Probabilistic arithmetic : Numerical methods for calculating convolutions and dependency bounds. *I. J. of Approximate Reasoning*, 4:8–158, 1990.

Linear Regression Analysis under Sets of Conjugate Priors

Gero Walter
Department of Statistics,
Ludwig-Maximilians-University
Munich, Germany
gero.walter@campus.lmu.de

Thomas Augustin
Department of Statistics,
Ludwig-Maximilians-University
Munich, Germany
thomas@stat.uni-muenchen.de

Annette Peters
GSF – National Research Center
for Environment and Health,
Institute for Epidemiology
Neuherberg, Germany
peters@gsf.de

Abstract

Regression is *the* central concept in applied statistics for analyzing multivariate, heterogenous data: The influence of a group of variables on one other variable is quantified by the regression parameter β. In this paper, we extend standard Bayesian inference on β in linear regression models by considering imprecise conjugated priors. Inspired by a variation and an extension of a method for inference in i.i.d. exponential families presented at ISIPTA'05 by Quaeghebeur and de Cooman, we develop a general framework for handling linear regression models including analysis of variance models, and discuss obstacles in direct implementation of the method. Then properties of the interval-valued point estimates for a two-regressor model are derived and illustrated with simulated data. As a practical example we take a small data set from the AIRGENE study and consider the influence of age and body mass index on the concentration of an inflammation marker.

Keywords. AIRGENE study, analysis of variance, exponential family, (imprecise) conjugate priors, imprecise probability models, interval probability, prior-data conflict, regression, robust Bayesian inference

1 Introduction and Sketch of the Argument

From engineering science over econometrics to sociology, from psychology over biometrics to medicine, one of the omnipresent questions is how certain variables (called covariates/confounders, regressors, stimulus or independent variables, here denoted by x) influence a certain outcome (called response or dependent variable z). The answer is obtained from regression models, and so regression modelling is *the* central concept in applied statistics.

The most common and simple case, dating back already to Gauß, is linear regression (see Section 3.1 for more details on the model), where, possibly after some transformations, for every unit i, taken from a sample of size k, the dependent variable z_i is assumed to be of metric scale and to be linearly related to p independent variables $x_{i1}, x_{i2} \ldots, x_{ip}$:

$$z_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \,, \qquad (1)$$

where ε_i is a stochastic error term subsuming the residual variation beyond the linear relationship between the variables.

The so-called regression coefficients β_j, $j = 1 \ldots, p$ measure the extent to which the dependent variable is expected to change if the value of the j-th regressor is enlarged by one unit and all other regressors remain unchanged. Often $x_{i1} = 1$ for all i and then β_1 is called intercept, describing some baseline level. The special case where all regressors are categorical (and coded via (several) 0/1 variables) is known as ANOVA (analysis of variance). The coefficients β_j can, for example, be estimated by the classical least squares method, or, relying on the Bayesian paradigm, by assigning a prior on the β_j's and updating it in the light of the data. This update step is especially elegant and simple to perform in situations we will call LUCK-models (described in Section 2.1). In our situation there are several possibilities to construct such a LUCK-model. We will rely on that multivariate normal as the prior for β_j, $j = 1 \ldots, p$, which has become the standard for regression analysis (see, e.g., [13]).

Although imprecise probabilities and related concepts [21, 26, 18] have proven to be quite powerful and convincing in many areas of application, regression analysis has only lived in the shadows there:[1] population heterogeneity, i.e. individual variation related to different covariate values, has mainly been incorporated by means of classification models [30, 1]; generalized inference (in particular along the lines of Walley's generalized Bayes's rule [21]) and decision theory (see the

[1] Very rare exceptions can be found in the robust Bayesian context, including: [6, 12]

survey in [20]) have almost exclusively confined themselves to the case of homogenous populations (i.i.d. case) or two-sample models (like [22, Section 5] and [8]).

Sound regression models would make imprecise probabilities quite attractive for applied scientists. As a step towards this ambitious aim we show that the approach of Quaeghebeur and de Cooman [16], who developed a concept of imprecise conjugated priors that nicely generalizes the widely applied imprecise Dirichlet model (IDM) [22, 2, 3], can be extended in an appropriate way.[2] Indeed, at least two different ways are successful. We briefly comment on the first one, which directly adopts to regression analysis Quaeghebeur and de Cooman's [16] original way to proceed, and investigate in more detail the second one, which is in straight line with the standard model for regression analysis (cf., e.g., [13].) For that purpose we interpret [16]'s approach, beyond its direct application in their work, as a general method for powerfully introducing imprecision into a huge class of Bayesian models, which we will call, for sake of brevity, LUCK-models in this paper, and demonstrate that the standard model for Bayesian regression analysis indeed fits into this framework.

In more detail, the paper is organized as follows: In Section 2 we collect some basic ingredients from Bayesian analysis, identify the special case of Bayesian analysis (LUCK-models) that underlies our basic argument, and then turn to the method for introducing imprecision into conjugate priors [16]. Section 3.1 firstly recalls classical[3] Bayesian regression analysis and puts it into the framework of LUCK-models. After having utilized [16] as a powerful method to extend LUCK-models to imprecise probabilities, we arrive at a general framework for regression analysis under sets of conjugate priors. We then focus consideration on a special case with two regressors, where some complex constraints underlying the general situation can be made easily tractable. The results are illustrated in Section 4 by simulated data and in Section 5 by a small data set from the AIRGENE study [15]. We conclude with some remarks on modifications and extensions, including the possibility to incorporate modelling of prior-data conflict, which was explicitly named by Walley as one of the main arguments for imprecise probabilities [21, p. 6], but which cannot be captured by the original method along the lines of [16].

[2] A different approach to generalize regression analysis has been proposed quite recently [28, Chapter 13] in the framework of the symmetric theory based on logical probability ([28], see also [27]).

[3] We use the term 'classical' for all concepts relying on precise probabilities / linear previsions.

2 Bayes Inference and LUCK-models

2.1 Classical Bayesian Inference and LUCK-models

As a preparation, some basic notions from the *classical Bayesian approach* will be recalled: Central is the assumption that the knowledge on a (possibly multivariate) parameter ϑ can be perfectly expressed by a single precise probability distribution on ϑ. So, inference from a (possibly multidimensional) sample w, the distribution of which is described by a density or probability function $f(w \mid \vartheta)$ (called likelihood in this context), consists in updating the so-called *prior* $p(\vartheta)$ to the so-called *posterior* $p(\vartheta \mid w)$ via Bayes's rule

$$p(\vartheta \mid w) \propto f(w \mid \vartheta) \cdot p(\vartheta)\,. \tag{2}$$

For a Bayesian, the prior describes the knowledge before having seen the sample, and the posterior subsumes the complete knowledge on ϑ after having seen the sample, and therefore it underlies all inferences drawn from the data.

For the intended application presented later on, it is quite convenient to distinguish certain standard situations (called *models with 'Linearly Updated Conjugate prior Knowledge'* (LUCK) here) of Bayesian updating with classical probabilities, where prior and posterior fit nicely together in the sense that

i) they belong to the same class of parametric distributions, a case where they are called *conjugate*, and, in addition,
ii) the updating of one parameter ($y^{(0)}$ below) of the prior is linear.[4]

More precisely, we introduce the following definition:

Definition 1 *Consider classical Bayesian inference on a parameter ϑ based on a sample w as described in (2), and let the prior $p(\vartheta)$ be characterized by the (vectorial) parameter $\vartheta^{(0)}$. The pair $\big(p(\vartheta), p(\vartheta \mid w)\big)$ is said to constitute a* LUCK-model of size q in the natural parameter ψ with prior parameters $n^{(0)} \in \mathbb{R}^+$ and $y^{(0)}$ and sample statistic $\tau(w)$, *iff there exist $q \in \mathbb{N}$ as well as transformations of ϑ into ψ and $\mathbf{b}(\psi)$ and of $\vartheta^{(0)}$ into $n^{(0)}$ and $y^{(0)}$, such that $p(\vartheta)$ and $p(\vartheta \mid w)$ can be rewritten in the following way:*[5]

$$p(\vartheta) \propto \exp\big\{n^{(0)}\big[\langle \psi, y^{(0)}\rangle - \mathbf{b}(\psi)\big]\big\} \tag{3}$$

and

$$p(\vartheta \mid w) \propto \exp\big\{n^{(1)}\big[\langle \psi, y^{(1)}\rangle - \mathbf{b}(\psi)\big]\big\} \tag{4}$$

with

$$n^{(1)} = n^{(0)} + q \quad \textit{and} \quad y^{(1)} = \frac{n^{(0)} y^{(0)} + \tau(w)}{n^{(0)} + q}\,. \tag{5}$$

[4] The second parameter $n^{(0)}$ possesses a vivid interpretation as "prior strength", which will become clearer in Section 2.2.

[5] $\langle a, b\rangle$ denotes the scalar product of a and b.

2.2 Imprecise Priors for Inference in LUCK-models

Several powerful approaches have been proposed to overcome the "dogma of ideal precision" (Walley) underlying classical Bayesian inference (cf., in particular, [14, 7, 5, 16]; see also Section 6). We rely in the following on the work of Quaeghebeur and de Cooman [16], who consider, by utilizing a general result (see, e.g., [4, Proposition 5.4]), certain LUCK-models for Bayesian inference based on independently and identically distributed (i.i.d.) observations from *regular, linear canonical exponential families* [4, p. 202 and p. 272f]. The central idea of [16] is that the seemingly strange parameterization in terms of $y^{(0)}$ and $n^{(0)}$ in (3) and (4) is perfectly suitable to be generalized to credal sets of priors. The crucial point is that these parameters are updated *linearly*, thus allowing for an easily tractable imprecise calculus: When sets of priors are defined via sets of parameters, and these sets of parameters are defined by lower and upper bounds, the lower and upper bounds of the sets of posterior parameters can be obtained directly from (5). So, just as in the popular IDM, which is contained as the special case of a multinomial sampling model with conjugated Dirichlet priors, minimization and maximization problems on the set of posteriors can be reduced to minimization and maximization problems on the set of priors in the case when the parameter $y^{(1)}$ (or a linear function of it) is the quantity of interest.

It shall be noted explicitly that this line of argumentation simply uses the linearity of the updating in the parameters, not the concrete derivation of the conjugate prior. Consequently, Quaeghebeur and de Cooman's technique to construct *imprecise* conjugate priors can be applied to any LUCK-model.

In more detail, the following technique will be applied: Given the situation in Definition 1, let $y^{(0)}$ vary in some set $\mathcal{Y}^{(0)} \subset \mathcal{Y}$, where the parameter space $\mathcal{Y}$ is taken as the convex hull (without the boundary) of the range of $\tau(w_i)$, and take as the imprecise prior the credal set consisting of all convex mixtures of all $p(\vartheta)$ from (3) created by varying $y^{(0)}$ in $\mathcal{Y}^{(0)}$. After having evaluated the sample w, the posterior credal set arising from applying Bayes's rule element by element has to be determined. For its calculation it is sufficient to consider the extreme points, and so it is obtained as the set of all convex mixtures of posteriors $p(\vartheta \,|\, w)$ arising from (4) by varying $y^{(1)}$ in $\mathcal{Y}^{(1)}$, where

$$\mathcal{Y}^{(1)} = \left\{ \frac{n^{(0)} y^{(0)} + \tau(w)}{n^{(0)} + n} \;\middle|\; y^{(0)} \in \mathcal{Y}^{(0)} \right\} \subset \mathcal{Y} \,. \quad (6)$$

$\mathcal{Y}^{(1)}$ can actually be seen as a shifted and rescaled version of $\mathcal{Y}^{(0)}$:

$$\mathcal{Y}^{(1)} = \frac{n^{(0)}}{n^{(0)}+n} \cdot \mathcal{Y}^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{1}{n} \sum_{i=1}^{n} \tau(w_i) \,, \quad (7)$$

which immediately suggests a vivid interpretation of $n^{(0)}$ as "prior strength" or as "pseudocounts", as it plays the same role for the prior as n for the sample. So, $n^{(0)}$ can be interpreted as the size of an imaginary sample that corresponds to the trust on the prior information in the same way as the sample size of a real sample corresponds to the trust in conclusions based on that sample.

For posterior inference, lower and upper posterior expectations are derived as the infimum and the supremum over all classical expectations with $y^{(1)}$ varying in $\mathcal{Y}^{(1)}$. The resulting relations between $n^{(0)}$, prior and posterior bounds are in essence the same as in the IDM:[6] In particular, for $n^{(0)} = n$, the width of the posterior expectation interval is half the width of the prior interval.

The choice of $\mathcal{Y}^{(0)}$ should reflect the prior information on the parameters. When there is very little or no information at all available, $\mathcal{Y}^{(0)}$ should be chosen as large as possible, that is, as the set of all possible parameter values, $\mathcal{Y}^{(0)} = \mathcal{Y}$. However, in most cases this would lead to the posterior set $\mathcal{Y}^{(1)}$ being vacuous as well, whatever the number of observations used for updating; for any $\overline{y}_j^{(0)} = \infty$, it holds that $\overline{y}_j^{(1)} = \infty$ as well. To avoid this, $\mathcal{Y}^{(0)}$ must be bounded by (element-wise) finite lower and upper boundaries.[7] This need to bound $\mathcal{Y}^{(0)}$ is not perceived as a severe restriction in practical application; typically, as in our example in Section 5, the very rough magnitude of reasonable parameter values is known, and there exist some trivial bounds.

3 Towards Imprecise Normal Regression Models

3.1 The Linear Regression Model, and its Classical Bayesian Treatment

In handling the linear regression model it is helpful to arrange the components in column vectors, denoted without index, i.e., $z = (z_1, \ldots, z_k)^\mathsf{T}$, and to collect all regressors column by column in the so-called design matrix $\mathbf{X}$. Equation (1) then reads as

$$z = \mathbf{X}\beta + \varepsilon \,, \quad \mathbf{X} \in \mathbb{R}^{k \times p},\ \beta \in \mathbb{R}^p,\ z \in \mathbb{R}^k,\ \varepsilon \in \mathbb{R}^k \,;$$

ε is assumed to have expected value $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, i.e. $\mathbb{V}(\varepsilon_i) = \sigma^2$, the variance of ε_i does

[6] $n^{(0)}$ corresponds to the parameter s in the IDM.

[7] For the IDM, this is not necessary, as the parameter space $\mathcal{Y}$ itself is already bounded, being the unit simplex.

not vary among the units (homoscedasticity) and all units are uncorrelated.

There are several methods to construct estimators $\hat{\beta}$ for the regression parameter β. With the least squares method, $\hat{\beta}$ is chosen to minimize the squared difference between the observed response values z and the values estimated by $\mathbf{X}\hat{\beta}$, yielding the well-known *least squares (LS) estimator*

$$\hat{\beta}_{LS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} z\,. \tag{8}$$

Other estimation techniques additionally assume that the error term ε is normally distributed. Then, as the design matrix $\mathbf{X}$ is considered to be non-stochastic,[8] also z is normally distributed,

$$z \sim N_k(\mathbf{X}\beta, \sigma^2\mathbf{I})\,. \tag{9}$$

This point of view is very helpful for several types of generalizations.[9] Reinterpreting (9) as a likelihood on β and σ^2 and applying the maximum likelihood (ML) principle again leads to the estimator from (8). In the Bayesian context appropriate priors related to the parameters and the likelihood (9) have to be found. Several choices for the prior seem attractive,[10] even different LUCK-models can be produced: In the light of the intended generalization below one natural possibility would be to follow the path of Quaeghebeur and de Cooman closely, by constructing a conjugate prior along the general construction method for LUCK-models (see, e.g., [4, Proposition 5.4]), also mentioned at the beginning of Section 2.2. For the case of known (or in advance estimated) σ^2 considered throughout the paper, one obtains by this procedure

$$p(\beta) \propto \exp\left\{n^{(0)}\left[\langle \beta, y^{(0)}\rangle - \mathbf{b}(\beta)\right]\right\},$$

where $\mathbf{b}(\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^{k}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)^2$. This prior can be shown to be a normal distribution on β, with its parameters being some transformations of $n^{(0)}$ and $y^{(0)}$ depending on $\mathbf{X}$.[11] It was maybe this strange dependency of the prior on the covariates $\mathbf{X}$ that resulted in this prior rarely being used for estimating regression parameters in the Bayesian framework.

[8] If $\mathbf{X}$ is stochastic, then it is common, and legitimate, practice (cf., e.g., [9]) to perform the analysis conditional on $\mathbf{X}$, hence (9) is replaced by $z \mid \mathbf{X} \sim N_k(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

[9] It makes also clear how the heterogeneity in the data is modelled: Each response z_i is assumed to be normally distributed, but the corresponding mean value depends on the individual characteristics (regressors) $x_{i1}, \ldots, x_{ip}$ and the effect size (expressed by β).

[10] So-called 'objective Bayesian estimation' of β, using the 'non-informative' prior $p(\beta) \propto$ *const.* leads to the same results as LS and ML when the expected or the maximum value of the posterior is used as the estimate. Therefore, when the interval-valued estimations of β proposed in this work are compared with the LS estimate, they are implicity compared to the ML and the objective Bayesian estimates as well.

[11] See [25] for more details on this model.

Instead, the commonly used approach specifies

$$\beta \sim \mathrm{N}_p\left(\beta^{(0)}, \sigma^2\boldsymbol{\Sigma}^{(0)}\right) \tag{10}$$

as the conjugate prior.[12] Advocated, e.g., by [13], it has become the standard, why we call the model based on this prior *normal regression model* throughout the paper. Applying Bayes's rule (2) to (10) yields

$$\beta \,|\, z \sim \mathrm{N}_p\left(\beta^{(1)}, \sigma^2\boldsymbol{\Sigma}^{(1)}\right), \tag{11}$$

where the updated parameters $\beta^{(1)}$ and $\boldsymbol{\Sigma}^{(1)}$ are obtained as

$$\beta^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda}^{(0)}\right)^{-1}\left(\mathbf{X}^\mathsf{T} z + \boldsymbol{\Lambda}^{(0)}\beta^{(0)}\right) \tag{12}$$

$$\boldsymbol{\Sigma}^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda}^{(0)}\right)^{-1}, \tag{13}$$

$\boldsymbol{\Lambda}^{(0)} = {\boldsymbol{\Sigma}^{(0)}}^{-1}$ being the so-called precision matrix.[13]

3.2 The Normal Regression Model as a LUCK-model

Now the argument that the extension proposed in [16] is neither limited to the i.i.d. case of homogenous populations nor to the special construction of LUCK-models considered there becomes fruitful: The standard Bayesian treatment of regression models based on the prior (10) can be shown to fit into the framework of LUCK-models, a fact that luckily immediately enables an appropriate generalization to imprecise models.

Theorem 2 *Consider the normal regression model described by the prior $p(\beta)$ from (10) with prior parameters $\beta^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$, and the posterior $p(\beta \,|\, z)$ from (11) with (12) and (13).*

Fixing a value $n^{(0)}$, $\big(p(\beta), p(\beta \,|\, z)\big)$ constitutes a LUCK*-model of size* 1 *with prior parameters*

$$y^{(0)} = \frac{1}{n^{(0)}}\begin{pmatrix} \boldsymbol{\Lambda}^{(0)} \\ \boldsymbol{\Lambda}^{(0)}\beta^{(0)} \end{pmatrix} =: \begin{pmatrix} y_a^{(0)} \\ y_b^{(0)} \end{pmatrix} \tag{14}$$

and $n^{(0)}$ and sample statistic

$$\tau(z) = \tau(\mathbf{X}, z) = \begin{pmatrix} \mathbf{X}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T} z \end{pmatrix} =: \begin{pmatrix} \tau_a(\mathbf{X}, z) \\ \tau_b(\mathbf{X}, z) \end{pmatrix}. \tag{15}$$

[12] Throughout the paper we denote prior parameters by the superscript $^{(0)}$ and, if appropriate, corresponding parameters of the posterior by $^{(1)}$. Here, the mean vector $\beta^{(0)} \in \mathbb{R}^p$ and the (positive definite) covariance matrix $\boldsymbol{\Sigma}^{(0)} \in \mathbb{R}^{p\times p}$ are the prior parameters defining the concrete distribution on β.

[13] If, in addition, σ^2 is considered unknown, too, the commonly used prior distribution conjugate to the likelihood in (9) is the so-called normal-inverse gamma distribution (e.g., [13, §9.4]). Unfortunately, this model will turn out to be not generalizable in the same way as it is done here for the normal regression model (cf., [23, Appendix], and also briefly in [25]). A first way out would be to estimate σ^2 in advance, and then to apply the normal regression model with the estimated value of σ^2, a strategy that we followed in our examples in Sections 4 and 5.

Proof: The proof is given in [24].

Knowing now the form of $y^{(0)}$, we can finally start with the imprecise probability calculus: By varying $y^{(0)}$ from (14) in a set $\mathcal{Y}^{(0)} \subset \mathcal{Y}$, the set of priors is generated. Since $\mathcal{T}$, the range of the sample statistic, is the product of the set of positive semidefinite $(p \times p)$ matrices and arbitrary vectors of dimension p, $\mathcal{Y}$ is taken as the convex hull of $\mathcal{T}$ without the boundary, thus $y_a^{(0)}$ having to be a positive definite $(p \times p)$ matrix. On the one hand, $\mathcal{Y}^{(0)}$ is chosen in order to reflect prior knowledge on β; on the other hand, this set must, as mentioned above at the end of Section 2.2, be bounded in order to avoid the possibility of vacuous posterior inference. In the case of a multidimensional parameter space $\mathcal{Y}$, [16] suggest to relate the element-wise bounds to each other. Their suggestion for the multivariate normal distribution is adopted here, leading to the following constraints of positive definiteness (p.d.):

$$\frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)} \quad \text{p.d.}, \qquad \text{and} \tag{16}$$

$$\frac{1}{n^{(0)}}\left(\mathbf{\Lambda}^{(0)} - \frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\right) \quad \text{p.d.} \tag{17}$$

If the normal regression model is to be applied as an imprecise probability model, we have to proceed as follows:

1. Prior knowledge on β must be expressed as a set of values of $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$.
2. This set must be "translated" into a set of values of $y^{(0)}$ in a way such that the resulting set $\mathcal{Y}^{(0)}$ satisfies the constraints (16) and (17).
3. Then each $y^{(0)}$ in $\mathcal{Y}^{(0)}$ is linearly updated by (5) to $y^{(1)}$.
4. The obtained set $\mathcal{Y}^{(1)}$ must be "retranslated" into an interpretable set of values of $\beta^{(1)}$ and $\mathbf{\Lambda}^{(1)}$.

The sets can be defined by lower and upper bounds for each element, e.g., for $\beta^{(0)}$ by

$$\beta_j^{(0)} \in \left[\underline{\beta}_j^{(0)}, \overline{\beta}_j^{(0)}\right] \qquad j = 1, \ldots, p\,.$$

The bounds for the components $\beta_j^{(0)}$ of $\beta^{(0)}$ can be chosen independently of each other, as any vector of reals forms an admissible regression parameter. For $\mathbf{\Lambda}^{(0)}$ the situation is more complex, because all the element-wise bounds $\underline{\lambda}_{ij}^{(0)}$ and $\overline{\lambda}_{ij}^{(0)}$ have to be chosen such that for any combination of values between the bounds the resulting $\mathbf{\Lambda}^{(0)}$ is positive definite. Choosing bounds for the precision matrix $\mathbf{\Lambda}^{(0)}$ instead of bounds for $\mathbf{\Sigma}^{(0)}$ facilitates the "translation" issues in application to a great extent.[14]

[14] Defining the bounds for $\mathbf{\Lambda}^{(0)}$ is in fact not as complicated as it might seem as the elements are interpretable in a quite straightforward way (and maybe even closer to intuition than the elements of $\mathbf{\Sigma}^{(0)}$): According to [17] (who is referring to [29, p. 142ff]), it holds that $\lambda_{ii} = [\mathbb{V}(\beta_i \,|\, \beta_{\smallsetminus i})]^{-1}$, where $\beta_{\smallsetminus i}$ is vector β without element i, and $\lambda_{ij} = -(\lambda_{ii}\lambda_{jj})^{\frac{1}{2}} \cdot \rho(\beta_i, \beta_j \,|\, \beta_{\smallsetminus\{i,j\}})$, with the second factor being the correlation of β_i and β_j conditioned on the linear effect of $\beta_{\smallsetminus\{i,j\}}$.

In the "translation" step the bounds on $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$ must be turned into bounds on $y^{(0)}$ that have to satisfy conditions (16) and (17). For $y_a^{(0)}$, this is simple, as multiplying by $\frac{1}{n^{(0)}}$ does not change positive definiteness. But deriving bounds on $y_b^{(0)}$ is more difficult, as it holds that

$$\underline{y}_{bi}^{(0)} = \min_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)}$$

$$\overline{y}_{bi}^{(0)} = \max_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)}\,.$$

The minima and maxima are to be taken over a joint set of $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$ that satisfies the constraint (17). Note that for obtaining the bounds for a *single* $y_{bi}^{(0)}$ the bounds of all elements of $\beta^{(0)}$ and of the i-th row on $\mathbf{\Lambda}^{(0)}$ have to be taken into account on the one hand, but on the other hand maximization and minimization must be executed only on combinations of all values between these bounds that are admissible according to (17). The obstacle is that (16) and (17) are nonlinear constraints (polynomial of degree p when checking wether all eigenvalues are positive), so that $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$ can hardly be calculated analytically. The satisfaction of the highly complex constraint (17) can be taken into account when "translating" to $\mathcal{Y}^{(0)}$ or already when defining the sets for $\beta^{(0)}$ and $\mathbf{\Lambda}^{(0)}$.

3.3 The Case of Two Regressors

In order to be able to give vividly interpretable analytical expressions, we now focus attention on the case of two regressors. Here, (16) turns out to demand only that, for any given $\lambda_{11}^{(0)}$ and $\lambda_{22}^{(0)}$, $\lambda_{12}^{(0)}$ must be chosen such that it leads to a correct non-deterministic correlation ρ. Still, with the five parameters $\beta_1^{(0)}$, $\beta_2^{(0)}$, $\lambda_{11}^{(0)}$, $\lambda_{12}^{(0)}$, $\lambda_{22}^{(0)}$ in this model, (17) turns out be quite complex, leading to an inequality in six parameters (the above five plus $n^{(0)}$) that does not seem to produce an easily interpretable condition on their choice.

Therefore a further simplification was made by assuming $\lambda_{11}^{(0)} = \lambda_{22}^{(0)} =: a$ and $\lambda_{12}^{(0)} = 0$. Then, (16) is trivially satisfied and (17) requires only

$$a\left({\beta_1^{(0)}}^2 + {\beta_2^{(0)}}^2\right) < n^{(0)}\,. \tag{18}$$

If the bounds for $\beta_1^{(0)}$, $\beta_2^{(0)}$ and a are chosen such that all possible combinations of values satisfy this

constraint, minimization and maximization can be performed for every parameter independently. Now, most but not all parameters of the posterior can be specified analytically, and the results to be sketched here[15] will turn out to be highly plausible:

We consider the following prior:

$$\beta \sim \mathrm{N}_2\left(\beta^{(0)},\, \frac{\sigma^2}{a}\mathbf{I}\right),$$

with $a \in A := [\underline{a}, \overline{a}]$, $\underline{a} > 0$ and

$$\beta^{(0)} = \begin{pmatrix}\beta_1^{(0)}\\ \beta_2^{(0)}\end{pmatrix} \in B := \begin{pmatrix} B_1 = [\underline{b}_1, \overline{b}_1] \\ B_2 = [\underline{b}_2, \overline{b}_2]\end{pmatrix}.$$

In the description we jump directly to the "retranslated" results. Denoting the elements of the updated covariance matrix $\mathbf{\Sigma}^{(1)}$ by $\sigma_{ij}^{(1)}$, $i,j = 1,2$, we obtain for any $a \in A$ by abbreviating

$$D = \left(\sum_{l=1}^k x_{l1}^2 + a\right)\left(\sum_{l=1}^k x_{l2}^2 + a\right) - \left(\sum_{l=1}^k x_{l1}x_{l2}\right)^2 :$$

$$\begin{aligned}\sigma_{11}^{(1)} &= D^{-1}\cdot\left(\textstyle\sum_{l=1}^k x_{l2}^2 + a\right)\\ \sigma_{22}^{(1)} &= D^{-1}\cdot\left(\textstyle\sum_{l=1}^k x_{l1}^2 + a\right)\\ \sigma_{12}^{(1)} &= D^{-1}\cdot\left(-\textstyle\sum_{l=1}^k x_{l1}x_{l2}\right).\end{aligned}$$

Their basic properties are summarized in

Remark 3

i) As $\frac{\partial}{\partial a}\sigma_{11}^{(1)}$ *and* $\frac{\partial}{\partial a}\sigma_{22}^{(1)}$ *are always negative, the higher the prior precision* a*, the lower the posterior variance of* β_1 *and* β_2*. The trend of the covariance* $\sigma_{12}^{(1)}$ *depends on the sign of* $\sum_{l=1}^k x_{l1}x_{l2}$*.*

ii) $\lim_{a\to 0}\sigma^2\mathbf{\Sigma}^{(1)} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \mathbb{V}(\hat{\beta}_{LS})$*. Therefore, for* $\underline{a} > 0$ *and monotonicity, it holds that the posterior variance of the regression parameters in the imprecise normal regression model is always smaller than the one of the LS estimator.*

iii) $\lim_{a\to\infty}\sigma^2\mathbf{\Sigma}^{(1)} = \mathbf{0}$*: An infinitely high prior precision causes naturally an infinitely small posterior variance.*

Most of the results on $\beta^{(1)}$ are reported in terms of $\beta_1^{(1)}$ only; by noting the symmetry underlying β_1 and β_2, analogous results for $\beta_2^{(1)}$ are immediately achieved mutatis mutandis. We obtain

$$\beta_1^{(1)} = \frac{1}{D}\left\{\left(\sum_{l=1}^k x_{l2}^2 + a\right)\left[a\cdot b_1 + \sum_{l=1}^k x_{l1}z_l\right] - \left(\sum_{l=1}^k x_{l1}x_{l2}\right)\left[a\cdot b_2 + \sum_{l=1}^k x_{l2}z_l\right]\right\}.$$

[15] See [23, Section 4.3] for a detailed derivation.

As these expressions are linear in b_1 and b_2 and optimizations in B can be taken independently of a, it holds that

$$\beta_1^{(1)} \to \max \text{ for } b_1 \to \overline{b}_1 \text{ and } \begin{cases} b_2 \to \overline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} < 0 \\ b_2 \to \underline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} > 0\end{cases}$$

$$\beta_1^{(1)} \to \min \text{ for } b_1 \to \underline{b}_1 \text{ and } \begin{cases} b_2 \to \underline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} < 0 \\ b_2 \to \overline{b}_2 & \sum_{l=1}^k x_{l1}x_{l2} > 0\end{cases}.$$

Unfortunately, calculating $\frac{\partial}{\partial a}\beta_1^{(1)}$ yields neither monotonicity nor an easily interpretable condition so that the bounds for $\beta^{(1)}$ can not be given analytically. But still asymptotic results can be obtained, which are summarized in

Remark 4

i) For any $b_j \in B_j$, $j = 1,2 : \lim_{a\to\infty}\beta^{(1)} = (b_1, b_2)^\mathsf{T}$*, and so, for very high values of* a *implying a very high prior precision, each* $b \in B$ *is updated to a value very near to itself; very high trust in the given prior values in* B *means sticking on the prior values and results in learning from the sample only to a very small extent.*

ii) On the other hand, $\lim_{a\to 0}\beta^{(1)} = \hat{\beta}_{LS}$*: Very low trust in the prior values in* B *results in relying almost only on the information given by the sample, and so, any given* $b \in B$ *will be updated to a value close to the least squares estimate* $\hat{\beta}_{LS}$*.*

On a first view, it is disturbing that none of the above formulae for deriving posterior parameters depends on $n^{(0)}$, the second prior parameter. The reason for this is that in proving Theorem 2, the parameter $n^{(0)}$ had to be introduced 'artificially' to match Relations (3) to (5) for the LUCK-model. When 'retranslating' $y^{(1)}$ into $\beta^{(1)}$ and $\mathbf{\Sigma}^{(1)}$, the parameter $n^{(1)}$ is eliminated immediately, and, as a consequence, the dependency on $n^{(0)}$ seems to vanish. In fact, the posterior bounds do actually depend on $n^{(0)}$ via Equation (18). Through this restriction on the prior bounds, the range of posterior bounds is constrained. When using the imprecise normal regression model, the bounds for B are quite easy to derive; a possible strategy is then to set a value for $n^{(0)}$ according to the interpretation as pseudocounts or sample size equivalent, and then to determine $\overline{a}$ from (18).

4 Results Based on Simulated Data

To illustrate the performance of the two-parameter model developed in Section 3.3, three data sets were simulated, each with 20 observations, but with a different arrangement of parameters. For data set 1,

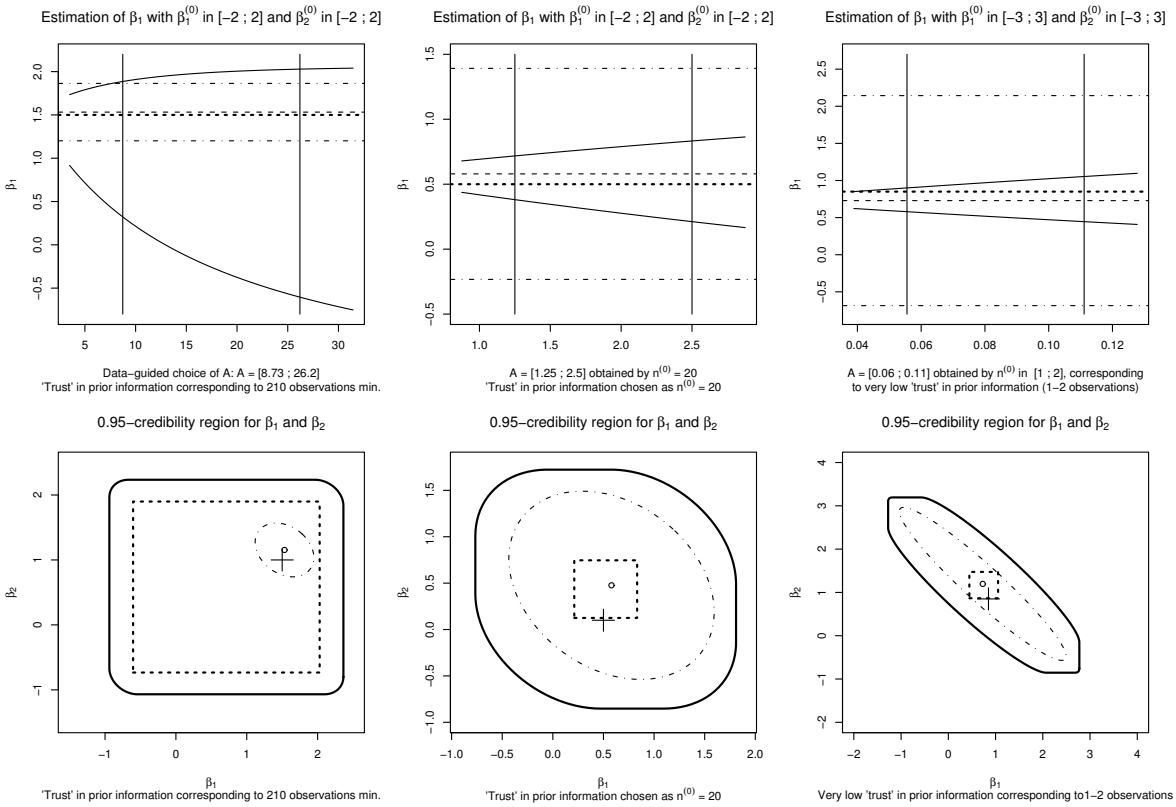


Figure 1: Exemplary results for two-regressor models based on three simulated data sets with 20 observations each.

realizations x_1 and x_2 of two independent standard normal variables were generated as regressors; the error ε was simulated with variance $\sigma^2 = 0.5$. Then the response z was calculated by choosing $\beta_1 = 1.5$, $\beta_2 = 1$. Data set 2 was generated analogously but with $\beta_1 = 0.5$, $\beta_2 = 0.1$ and $\sigma^2 = 3$. In data set 3, multi-collinearity was modeled by simulating x_1 and x_2 by a two-dimensional normal with correlation $\rho = 0.9$, taking $\beta_1 = \beta_2 = 0.85$ and $\sigma^2 = 1$ for calculating z. The regressors were standardized and z was centered with the observed moments in order to make the estimation of an additional intercept unnecessary. Exemplary results are shown in Figure 1, where the graphs from the left to the right show results for data set 1, 2 and 3, respectively.

The upper graphs show the estimation of β_1 for each data set. In each of these graphs, the thick short-dashed line represents the actual value of β_1, the thinner dashed line the LS estimate, and dot-dashed lines indicate the bounds of the 0.95% confidence interval for the LS estimate. The 'horizontal' solid lines represent the estimated lower and upper bound for $\beta_1^{(1)}$ as a function of a, and the vertical lines mark the chosen values of $\underline{a}$ and $\overline{a}$. The lower graphs compare the classical ellipsoid confidence region (dash-dotted line) for the LS estimate of β_1 and β_2 (indicated by the small circle) with the interval-valued estimate (thick short-dashed line) and a 0.95-credibility region for it (thick solid line). The actual value of (β_1, β_2) is indicated by the big cross.

For the "large β, small σ^2" data set 1, relatively high values of a were chosen 'data-guided' by taking the estimated variance of the LS estimator to calculate a central value of A. Because standardized regression parameters are to be estimated, their absolute value is interpretable, and the choice of $B_1 = B_2 = [-2\,;\,2]$ seems reasonable, as higher values are very rare in application. Note that the course of the 'horizontal' solid lines illustrates clearly the statement in Remark 4: The prior assignment results in a quite broad posterior interval for $\beta_1^{(1)}$ (lowest and highest intersection of vertical with 'horizontal' solid lines), as the induced value of $n^{(0)} = 210$ is quite high with respect to the sample size of 20. Consequently, the interval-valued estimate displayed in the lower graph covers a wide area compared to the frequentist confidence region. So does the 0.95-credibility region, which was approximated by the union of 0.95-credibility regions for all combinations of β_1 and β_2 in the interval-valued estimate. Because the maximum posterior variance is lower than the variance for the LS estimate (as mentioned in Remark 3), the distance between the bounds

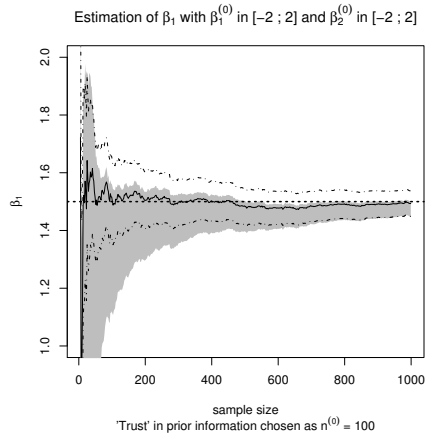


Figure 2: Illustration of asymptotic behavior of interval-valued regression parameter estimates.

of the interval-valued estimate and its credibility region is smaller than the distance between the LS estimate and its confidence region.

In the "small β, large σ^2" situation of data set 2, $\overline{a}$ was chosen such that $n^{(0)} = 20$, implying that the influence of prior and data information are evenly balanced. $\underline{a}$ was chosen ad hoc as $0.5 \cdot \overline{a}$ to illustrate the effect of values of $\underline{a} > 0$ on the variance of $\beta^{(1)}$. (Here, the choice of $\underline{a}$ has no influence on the interval-valued estimates for $\beta^{(1)}$.) Now, as smaller values of $\overline{a}$ result in shorter intervals for $\beta^{(1)}$ (which can be seen clearly from each of the top graphs), the resulting interval-valued estimate for β_1 is less wide, being now shorter than the confidence interval that is quite wide due to the high value of σ^2. This can be seen also in the lower graph, where the confidence region and the credibility region differ to a much lesser extent than in the lower left graph.

For the analysis of data set 3 with a "moderate β_1 and β_2, moderate σ^2" arrangement, A was chosen by using values for $n^{(0)}$ that are commonly suggested for s in the IDM to represent prior ignorance. So $\underline{a}$ was derived from $n^{(0)} = 1$, and $\overline{a}$ from $n^{(0)} = 2$. Together with, as a precaution, even wider prior intervals $B_1 = B_2 = [-3\,;\,3]$, this still yields a very short posterior interval for β_1, as can be seen in the top right graph. Note the exceedingly wide confidence interval for the LS estimate, as this shows the troublesome property of the LS estimate in the case of multi-collinearity: the high resulting variance of estimates can, in many cases, even cause the 'observed' estimates having the wrong sign. In the lower right graph, both the confidence as well as the credibility region show the effect of multi-collinearity through their diagonal shape: estimates for β_1 and β_2 are negatively correlated because both x_1 and x_2 contain similar information. Still, the interval-valued estimate covers a quite small area around the LS estimate, illustrating again the results achieved for the limiting case $a \to 0$.

In Figure 2, asymptotic behavior of the interval-valued estimation for β_1 is illustrated using the situation of data set 1 and choosing $n^{(0)} = 100$. With increasing sample size $k = \dim(z)$, the range of the interval, marked as a gray colored vertical line for each value of k, is becoming shorter and shorter, tightening around the LS estimate, represented by the thin solid line, which approaches the actual value of β_1, marked by the thick short-dashed horizontal line. The dot-dashed lines indicate again the bounds of the 0.95% confidence interval for the LS estimate.

5 The AIRGENE Study

In addition, the model was applied to a data set that is a part of the data collected for the AIRGENE study [15], an EU financed panel study which was conducted to assess the association between air pollutants and inflammation markers in the high-risk group of myocardial infarction survivors. As epidemiological studies show that inflammation markers are associated with the BMI (Body-Mass-Index) and the age of subjects [19], their influence on inflammation marker levels must be taken into account when estimating the effect of air pollutants. To this end, estimations for the parameters of these interfering factors (confounders) are derived in a separate regression model and then are used – in the main analysis not to be presented here – to adjust the main model that estimates the influence of air pollutant variables.

Here, the 200 cases collected by KORA [11] in Augsburg, which was one of the six study centers, are analyzed. The reduced data set consists of the variables `bmi` and `age` as regressors and `log(fib)` as the response variable, being the log of the concentration of the inflammation marker fibrinogen, averaged over the several blood samples collected for each subject during the study period.

Just as for the simulated data sets, the response was centered and the regressors standardized to make an estimation of an intercept unnecessary. Prior bounds for $\beta_{\texttt{bmi}}$ and $\beta_{\texttt{age}}$ were derived each in a straightforward way by considering the lowest and highest possible values (e.g. for `age`, these were, according to the inclusion criterion of the study, 35 and 80 respectively) that were transformed on the standardized scale and then linked to the range of the centered response.When choosing A in the same way as for data set 3 to model weak prior knowledge, the retransformed interval-valued estimates can be combined to the following regression equation:

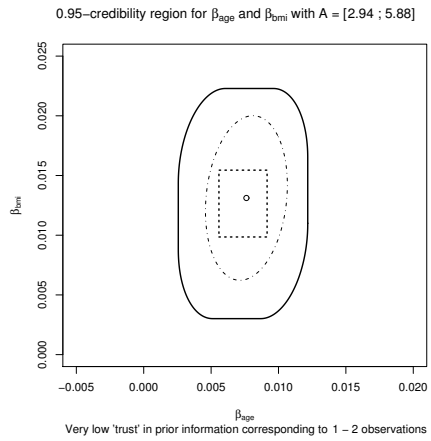


Figure 3: Exemplary results for the AIRGENE data.

$$\begin{aligned}\log(\texttt{fib})_i = \quad & \texttt{age}_i \cdot [0.00558,\, 0.00915] \\ & + \texttt{bmi}_i \cdot [0.00985,\, 0.01545] \\ & + \qquad [0.180,\, 0.562] \qquad\qquad + \varepsilon_i\end{aligned}$$

The fact that the 0.95-credibility region displayed in Figure 3 does not cover the origin is a strong hint that, also when considering complex uncertainty in the prior, `age` and `bmi` have a noteworthy effect on the fibrinogen level. So, the established evidence on this association can be confirmed and set on a more stable base with respect to the model assumptions.

6 Concluding Remarks

We have suggested a first approach to linear regression with imprecise conjugate priors. Of course, the approach needs further investigation, including a comparison to modifications and alternative ways to proceed. This applies in particular to the approach briefly described at the beginning of Section 3.1 where an (imprecise) LUCK-model is constructed directly along the lines of [16].

Further research should also clarify whether other powerful models generalizing classical Bayesian inference in the i.i.d. case (like [14, 7, 5]) can also be extended to linear regression models by similar arguments. The results should also be compared with the approach currently being developed by [28, Chapter 13], whose so-called symmetric theory based on logical probabilities ([28], see also [27]) allows the derivation of probability distributions on parameters without prior modelling.

A possible drawback of the approach introduced by [16], which consequently is shared by the models developed here, is that, in some sense, it does not entirely utilize the expressive power of imprecise probabilities: As $n^{(0)}$ is fixed (like s in the IDM), the behavior of the model – outside the situation of prior ignorance – is not optimal in the case of prior-data conflict in the sense of [21, p. 6]. To see this, note that, if in the situation of Section 2.2 $y^{(0)}$ varies between $\underline{y}^{(0)}$ and $\overline{y}^{(0)}$, then the difference between the updated bounds $\overline{y}^{(1)}$ and $\underline{y}^{(1)}$ is given by $\frac{n^{(0)}\left(\overline{y}^{(0)}-\underline{y}^{(0)}\right)}{n^{(0)}+n}$. So the imprecision decreases by the same amount for any sample of size n, irrespectively whether or not there is substantial discrepancy between prior assignments and the sample. A natural attempt to find a way out would be to vary $n^{(0)}$ in addition. This idea still has to be explored, but the model developed in [21, Ch. 5.4], where such effects are described for an IDM with two categories, may give some hint.

From the applied point of view it is quite important to extend the modelling to generalized linear models, which in particular allow regression analysis for non-metric responses. Here the adaption of auxiliary variable models, considered by [10] in a simulation-based classical Bayesian setting, seems to be very promising.

Acknowledgements

We thank the GSF, KORA and the AIRGENE study group for the help with the data set used for illustration. We are very grateful to the three referees for very helpful and stimulating remarks.

References

[1] J. Abellán and S. Moral. Upper Entropy of Credal Sets. Applications to credal classification. *Intern. J. Approx. Reasoning*, 39: 235–255, 2005.

[2] J.-M. Bernard. An Introduction to the Imprecise Dirichlet Model for Multinomial Data. *Intern. J. Approx. Reasoning*, 39: 123–150, 2005.

[3] J.-M. Bernard. Special Issue on the Imprecise Dirichlet Model. *Intern. J. Approx. Reasoning*, to appear.

[4] J. Bernardo and A. Smith. *Bayesian Theory.* Wiley & Sons, 1993.

[5] A. Boratyńska. Stability of Bayesian Inference in Exponential Families. *Stat. Probabil. Lett.*, 36: 173–178, 1997.

[6] A. Chaturvedi. Robust Bayesian analysis of the linear regression model. *J. Stat. Plan. Inf.*, 50: 175–186, 1996.

[7] F. P. A. Coolen. Imprecise Conjugate Prior Densities for the One-parameter Exponential Family of Distributions. *Stat. Probabil. Lett.*, 16: 337–342, 1993.

[8] F. P. A. Coolen and P. Coolen-Schrijner. Nonparametric Predictive Comparison of Proportions. *J. Stat. Plan. Inf.*, 137: 23–33, 2007.

[9] L. Fahrmeir and G. Tutz. *Multivariate Statistical Models Based on Generalized Linear Models.* Springer, 2001.

[10] C. C. Holmes and L. Held. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, 1: 145–168, 2006.

[11] H. Löwel, C. Meisinger, M. Heier, and A. Hormann. The Population-based Acute Myocardial Infarction (AMI) Registry of the MONICA/KORA Study Region of Augsburg. *Gesundheitswesen*, 67, Supplement 1: S31–S37, 2005.

[12] M. Lavine. An approach to evaluating sensitivity in Bayesian regression analyses. *J. Stat. Plan. Inf.*, 40: 233–244, 1994.

[13] A. O'Hagan. *Bayesian Inference*. Kendall´s Advanced Theory of Statistics Vol. 2B, Arnold, 1994.

[14] L. P. Pericchi and P. Walley. Robust Bayesian Credible Intervals and Prior Ignorance. *Int. Stat. Rev.*, 58: 1–23, 1991.

[15] A. Peters, A. Schneider, S. Greven, T. Bellander, F. Forastiere, A. Ibald-Mulli, T. Illig, B. Jacquemin, K. Katsouyanni, W. Koenig, T. Lanki, J. Pekkanen, G. Pershagen, S. Picciotto, R. Rückerl, A. Schaffrath-Rosario, C. Stefanadis, and J. Sunyer. Air Pollution and Inflammatory Response in Myocardial Infarction Survivors: Gene-environment-interactions in a High-risk Group: Study Design of the AIRGENE Study. *Inhalation Toxicology*, in press.

[16] E. Quaeghebeur and G. de Cooman. Imprecise Probability Models for Inference in Exponential Families. In: F. G. Cozman, R. Nau, and T. Seidenfeld (eds.) *Proc. of the 4th Intern. Symp. on Imprecise Probabilities and their Applications (ISIPTA'05)*, 2005.

[17] F. Reithinger. Zusammenhangsstrukturen, *Technical Report, Department of Statistics, LMU Munich*, 2006.
`http://www.statistik.lmu.de/~flo/ss06/strukturen.pdf`

[18] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[19] B. Thorand, J. Baumert, A. Döring, C. Herder, H. Kolb, W. Rathmann, G. Giani, and W. Koenig; Kora Group. National Research Center for Environment and Health, Institute of Epidemiology, Neuherberg, Germany. Sex Differences in the Relation of Body Composition to Markers of Inflammation, *Atherosclerosis*, 184: 216–224, 2006.

[20] M. C. M. Troffaes. Decision Making Under Uncertainty using Imprecise Probabilities. *Intern. J. Approx. Reasoning*, in press.

[21] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[22] P. Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J. Roy. Stat. Soc. B*, 58: 3–57, 1996.

[23] G. Walter. Robuste Bayes-Regression mit Mengen von Prioris — Ein Beitrag zur Statistik unter komplexer Unsicherheit. *Diploma thesis, Department of Statistics, LMU Munich,* 2006.
`http://www.statistik.lmu.de/~thomas/team/diplomathesis_GeroWalter.pdf`

[24] G. Walter. The Normal Regression Model as a LUCK-model. *Discussion Paper*, 2007.
`http://www.statistik.lmu.de/~thomas/team/isipta07_proof.pdf`

[25] G. Walter. Sketch of an Alternative Approach to Linear Regression Analysis under Sets of Conjugate Priors. *Discussion Paper*, 2007.
`http://www.statistik.lmu.de/~thomas/team/isipta07_conjugate_prior.pdf`

[26] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung, volume I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, 2001.

[27] K. Weichselberger. The Logical Concept of Probability and Statistical Inference. In: F. G. Cozman, R. Nau, and T. Seidenfeld (eds.) *Proc. of the 4th Intern. Symp. on Imprecise Probabilities and their Applications (ISIPTA'05)*, 396–405, 2005.

[28] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung, volume II: Symmetrische Wahrscheinlichkeitstheorie*, in cooperation with A. Wallner. In preparation, 2008.

[29] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley & Sons, 1990.

[30] M. Zaffalon, K. Wesnes, and O. Petrini. Reliable Diagnoses of Dementia by the Naive Credal Classifier Inferred from Incomplete Cognitive Data. *Artif. Intell. Med.*, 29: 61–79, 2003.

The Logical Concept of Probability: Foundation and Interpretation

Kurt Weichselberger
(In cooperation with Anton Wallner)
Ludwig-Maximilians-University Munich, Germany,
Kurt.Weichselberger@stat.uni-muenchen.de

Abstract

The Logical concept of probability, introduced to ISIPTA 2005 in a tutorial ([15]), is based on the theory of Interval probability. Since the main feature of the Logical concept is given by the evaluation of arguments consisting of premises and conclusions, it proves necessary to define exactly which kinds of propositions can be employed hereby. If this is done, the analysis allows the definition of independent arguments by examination of the contents of premises and conclusions. If Interval probability is attributed to arguments according to the relevant axioms, a frequency interpretation becomes feasible which decisively relies on the autonomous concept of independence.

Keywords. Interval probability, evaluation of arguments, concept of independence, frequency interpretation, Symmetric theory of probability.

1 Introduction

The motivation to develop the Logical concept of probability and the main features of this concept are thoroughly described in [15]. While in this paper the emphasis was laid on the merits of this concept and of the Symmetric theory of probability which is based on this concept, the present article concentrates on the understanding of the elements which constitute the Logical concept.

Since by the Logical concept probability exclusively is attributed to arguments, the question must be answered, which kind of arguments are suitable for such an evaluation: It is easy to find counter-examples. In Section 2 the concept of P-argument is introduced, meant to cover all situations of interest in statistical analysis and reasoning. This approach allows the definition of mutually independent P-arguments, based on the contents of the respective premises and conclusions. Irrelevance of one proposition with respect to another one and mutual independence of P-arguments constitute the most important aspects of this approach — prior to any kind of evaluating the arguments.

The axioms governing the introduction of interval probability via the establishment of W-fields must be based on this conceptual foundation: Axioms L III and L IV produce the result that mutual independence has the effect of multiplicativity — but not vice versa[1] (Section 3).

A generalization of the classical Binomial law produces a Weak law of large numbers based on mutual independence of arguments. Its result is described by a proposed new expression: weak invergence (Section 4).

The frequency interpretation of the Logical concept which arises out of these results (Section 5) is not liable to objections of circular reasoning. It is seen as the basis of understanding the statements of the Logical concept and of the Symmetric theory of probability.

Section 6 contains a few historical remarks and a comparison of some different approaches to the combined aspects of probability and independence. It also outlines the importance of these results with respect to the Symmetric theory.

2 P-Arguments

Let A and B be propositions describing contingent facts, which may be right or wrong. The propositions A and B are therefore neither tautologies nor antinomies. Only propositions of this kind will be considered in this article. Concerning the ordered pairs (A, B) the question arises what kind of consequences can be drawn from B concerning A.

[1] The paper of ISIPTA'05 does not cover the aspects described in Section 2 of the present article. As a consequence the axioms of W-fields in 2005 are partially different from that in Section 3 of the present article.

Four categories of such pairs $(A,\, B)$ may be distinguished. They are characterized by:

$$\begin{aligned} K(A,\, B) &= +1, \\ K(A,\, B) &= -1, \\ K(A,\, B) &= 0, \\ K(A,\, B) &= P. \end{aligned}$$

Note that here $+1$, -1, 0, P are mere symbols and don't represent numbers!

Definition 1 *$K(A,\, B) = +1$ is meant to describe pairs $(A,\, B)$, where A can be derived logically from B.* □

Definition 2 *$K(A,\, B) = -1$ describes pairs $(A,\, B)$ where $\neg A$ logically can be derived from B.* □

Definition 3 *$K(A,\, B) = 0$ distinguishes such pairs $(A,\, B)$, where absolutely no consequences about A can be drawn from B, in particular that for all potential premises B_1 (where $B \wedge B_1$ is not an antinomy)*

$$K(A,\, B_1) = K(A,\, B \wedge B_1)$$

holds and for all potential conclusions A_1 (provided that $A \vee A_1$ is not a tautology)

$$K(A_1,\, B) = K(A \vee A_1,\, B)$$

holds: B is irrelevant with respect to A. □

Corollary 1 *With respect to the meaning of "irrelevance" it can be concluded that the following implications hold:*

$$K(A,\, B) = 0 \;\Rightarrow\; K(\neg A,\, B) = 0;$$

$$K(A,\, B_1) = 0,\, K(A,\, B_2) = 0 \Rightarrow K(A,\, B_1 \vee B_2) = 0,\, K(A,\, B_1 \wedge B_2) = 0;$$

$$K(A_1,\, B) = 0,\, K(A_2,\, B) = 0 \Rightarrow K(A_1 \vee A_2,\, B) = 0,\, K(A_1 \wedge A_2,\, B) = 0.$$ □

The attachment $K(A,\, B) = 0$ is determined by the contents of A and of B, and its meaning is generally unquestioned. It is, however, possible that persons with different background disagree with respect to the eventuality of consequences which the facts described by the proposition B can have for the facts described by the proposition A.

It may be expected that a parapsychologist or a supporter of Chaos-theory refuse attachments $K(A,\, B) = 0$, which are selfunderstanding for other scientists. Concerning characteristic types of $(A,\, B)$ with religious background there will be an influence of creed.

On the other hand distinction of pairs $(A,\, B)$ with $K(A,\, B) = 0$ constitutes fundamental prerequisites in most scientific disciplines as far as empirical research is concerned. Consequently these attachments are inevitable tools of statistical modeling.

Historically the idea that the circumstances of one game of chance must have no consequence whatever for the following games, was prior and fundamental to the idea of introducing probability in analyzing the results of games of chance.

Definition 4 *All ordered pairs $(A,\, B)$ in consideration which do not belong to the categories $+1$, -1 or 0, are attached to category P. A pair $(A,\, B)$ belonging to this category is named a* partial argument *or* P-argument *$(A||B)$. B is named the* premise*, A is named the* conclusion *of the P-argument $(A||B)$.* □

It must be agreed that generally P-arguments are the most important tools of learning and therefore are the means of evidential reasoning. The class of P-arguments is huge and extremely heterogeneous.

For clearness it must be pointed out that the category of P-arguments contains pairs $(A,\, B)$, which at first sight could be expected to belong to category 0:

A pair of propositions $(A,\, B)$ where $B = R \vee M$ and $K(A,\, R) = +1$, $K(A,\, M) = -1$ does not qualify for $K(A,\, B) = 0$ since it violates a criterion of this category. This aspect may be demonstrated by

Example 1 *Let*

$$\begin{aligned} A &= \text{“It is freezing”}, \\ R_1 &= \text{“The temperature is } -3^\circ C\text{”}, \\ M_1 &= \text{“The temperature is } +2^\circ C\text{”} \end{aligned}$$

and $B_1 = R_1 \vee M_1$. Obviously $K(A,\, B_1) \in \{0,\, P\}$. Now let

$$\begin{aligned} R_2 &= \text{“The temperature is } -1^\circ C\text{”}, \\ M_2 &= \text{“The temperature is between } 0^\circ C \text{ and } +3^\circ C\text{”}. \end{aligned}$$

$B_2 = R_2 \vee M_2$ produces $K(A,\, B_2) \in \{0,\, P\}$.

$B_1 \wedge B_2 = (R_1 \vee M_1) \wedge (R_2 \vee M_2) =$

$$M_1 = \text{“The temperature is } +2^\circ C\text{”}.$$

Therefore: $K(A,\, B_1 \wedge B_2) = -1$. Comparison with Definition 3 reveals that $K(A,\, B_1) = 0$ as well as $K(A,\, B_2) = 0$ would be in contradiction with the requirements of this definition. Accordingly $K(A,\, B_1) = P$ and $K(A,\, B_2) = P$ must hold true and: $(A||B_1)$ as well as $(A||B_2)$ are P-arguments. □

This example points at the possibility of pairs $(A,\ B)$, where B is not informative directly with respect to A, but nevertheless $(A,\ B)$ does not belong to category 0, because B contains information which can be relevant with respect to A, if combined with some complementary information.

On the other hand it reveals the existence of P-arguments where the premises are not informative with respect to the conclusions — as long as both of them stand alone.

This possibility contrasts sharply to another type of P-arguments describing reliable empirical knowledge which may be classified as "practically sure".

Altogether the kinds of treatment with P-arguments are very different in different fields of application: in daily life, in court, in science or in humanities. It is, however, possible to define generally a relation between P-arguments which is of special importance for establishing concepts to evaluate P-arguments.

Definition 5 *The P-arguments* $(A_1||B_1)$ *and* $(A_2||B_2)$ *are* independent of each other*, iff* $K(A_1,\ B_2) = 0$ *and* $K(A_2,\ B_1) = 0$. □

Mutual independence of P-arguments is distinguished, therefore, solely by the reciprocal irrelevance of premises with respect to the conclusion of the other P-argument. This definition is prior to all attempts to introduce the concept of probability and in the context it is seen as a prerequisite for establishing a suitable theory.

Generalizations of Definition 5 seemingly can be established in two different ways.

Definition 6 *Let* $(A_i||B_i),\ i = 1, ..., r,$ *be P-arguments. Iff*

$$K(A_i,\ B_j) = 0,\ \forall\, i,\ j \in \{1, ..., r\},\ i \neq j,$$

holds, the P-arguments $(A_1||B_1), ..., (A_r||B_r)$ *are* pairwise independent. □

Definition 7 *Iff, under the assumptions of Definition 6,*

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2} B_j\right) = 0,$$
$$\forall \emptyset \subsetneqq I_1,\ I_2 \subsetneqq \{1, ..., r\},\ I_1 \cap I_2 = \emptyset,$$

holds, the P-arguments $(A_1||B_1), ..., (A_r||B_r)$ *are* totally independent from each other. □

Obviously P-arguments which are totally independent from each other are pairwise independent, too. But additionally the following Lemma holds:

Lemma 1 *If* $(A_i||B_i),\ i = 1, ..., r,$ *are pairwise independent P-arguments, then they are totally independent from each other.* □

The *proof* of this Lemma is based on Corollary 1 by induction on r. Obviously Definition 6 and 7 coincide for $r = 2$. It is now presupposed that the assertion of Lemma 1 holds for $r \geq 2$.

Therefore, if $(A_i||B_i),\ i = 1, ..., r+1,$ are taken as pairwise independent, for every $I_0 \subseteq \{1, ..., r+1\}$ with $|I_0| \leq r$, the relation

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2} B_j\right) = 0,$$
$$\forall \emptyset \subsetneqq I_1,\ I_2,\ I_1 \cup I_2 \subseteq I_0,\ I_1 \cap I_2 = \emptyset,$$

must be valid.

Now let $\emptyset \subsetneqq I_1,\ I_2 \subsetneqq \{1, ..., r+1\}$ and $I_1 \cap I_2 = \emptyset$. If there exists $I_0 \subseteq \{1, ..., r+1\},\ |I_0| \leq r,$ and $I_1 \cup I_2 \subseteq I_0$, according to the assumption

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2} B_j\right) = 0$$

must hold.

If, however, $I_1 \cup I_2 = \{1, ..., r+1\}$, no such I_0 exists. Now two cases have to be distinguished:

a) If $|I_1| < r,\ |I_2| \geq 2$, let

$$I_2 = I_2' \cup I_2'',\ \emptyset \subsetneqq I_2',\ I_2'',\ I_2' \cap I_2'' = \emptyset.$$

Therefore $|I_1 \cup I_2'| \leq r,\ |I_1 \cup I_2''| \leq r$.

Due to the assumption

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2'} B_j\right) = K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2''} B_j\right) = 0,$$

and Corollary 1 produces

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2} B_j\right) =$$
$$= K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2'} B_j \wedge \bigwedge_{j\in I_2''} B_j\right) = 0.$$

b) If $|I_1| = r,\ |I_2| = 1$, let

$$I_1 = I_1' \cup I_1'',\ \emptyset \subsetneqq I_1',\ I_1'',\ I_1' \cap I_1'' = \emptyset.$$

$|I_1' \cup I_2| \leq r,\ |I_1'' \cup I_2| \leq r,$

$$K\left(\bigwedge_{i\in I_1'} A_i,\ \bigwedge_{j\in I_2} B_j\right) = K\left(\bigwedge_{i\in I_1''} A_i,\ \bigwedge_{j\in I_2} B_j\right) = 0,$$

and due to Corollary 1:

$$K\left(\bigwedge_{i\in I_1} A_i,\ \bigwedge_{j\in I_2} B_j\right) = \\ = K\left(\bigwedge_{i\in I_1'} A_i \wedge \bigwedge_{i\in I_1''} A_i,\ \bigwedge_{j\in I_2} B_j\right) = 0.$$

In both cases the conditions for total independence of $(A_i||B_i)$, $i = 1, ..., r$, are satisfied. □

Consequently in the following only the concept of r mutual independent P-arguments has to be taken into consideration.

This result characterizes the difference between the concept of independence in the theory employed here and in the Classical theory: Independence is a more demanding relation in the theory of the Logical concept.

3 W-Fields

The most advanced method of evaluating P-arguments is that of attaching interval-probability. It affords the selection of two sets of propositions: $\mathcal{A}_P$, $\mathcal{B}_P$ with $\mathcal{A}_P \cap \mathcal{B}_P = \emptyset$, so that

$$K(A, B) \in \{0, P\},\ \forall\, A \in \mathcal{A}_P,\ B \in \mathcal{B}_P.$$

$\mathcal{A}_P$ as well as $\mathcal{B}_P$, in the second step, have to be completed, if necessary, to generate sets $\mathcal{A}_P^*$ and $\mathcal{B}_P^*$, which are closed under the logical operations $\vee$, $\wedge$, and $\setminus$ ("logical difference"). Additional potential conclusions A and additional potential premises B may produce additional P-arguments, but ordered pairs (A, B) of category $+1$, -1 or 0 as well. It must be secured that all assignments are in concordance with the definitions of $K(A, B)$.

According to the tradition and the actual practice of probability theory conclusions as well as premises should be described by sets. Therefore the elements of $\mathcal{A}_P^*$ and $\mathcal{B}_P^*$ must be represented by sets in a way guaranteeing that logical operations on $\mathcal{A}_P^*$ and on $\mathcal{B}_P^*$ are transformed to the corresponding set operations on the representing sets $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A} \cap \mathcal{B} = \emptyset$.

Obviously this representation is by no means uniquely determined. It always must be borne in mind that the tools of representation must not influence the decisive probabilistic reasoning.

Then any assignment of interval-probability is produced by

$$P(A||B) = [L(A||B),\ U(A||B)],\ \forall\, A \in \mathcal{A},\ B \in \mathcal{B}.$$

It may be understood as the result of evaluating P-arguments completed by the following attachments:

$L(A||B) = 1$, $U(A||B) = 1$, if for the corresponding propositions $K(A, B) = 1$ holds;

$L(A||B) = 0$, $U(A||B) = 0$, if $K(A, B) = -1$ holds;

$L(A||B) = 0$, $U(A||B) = 1$, if $K(A, B) = 0$ holds.

The probability of any P-argument $(A||B)$ determines the interval-limits $L(A||B)$ and $U(A||B)$ for the representing sets A and B. The rules governing this assessment are given in a three-level hierarchy:

1) Classical theory of probability:

Any function $p(.)$ on a measure space $(\Omega; \mathcal{A})$ which obeys Kolmogorov's three axioms is called a *K-function*.

2) Theory of interval probability (see [14])**:**

An *F-(probability-)field* $\mathcal{F} = (\Omega; \mathcal{A}; L(.))$ is given, iff the following three axioms hold[2]:

T IV: $P(A) = [L(A); U(A)] \subseteq [0; 1],\ \forall\, A \in \mathcal{A}$.

T V: The set $\mathcal{M}$ of K-functions $p(.)$ on $(\Omega; \mathcal{A})$ with $L(A) \le p(A) \le U(A)$, $\forall\, A \in \mathcal{A}$, is not empty.

T VI: $\inf_{p(.)\in\mathcal{M}} p(A) = L(A)$,
$\sup_{p(.)\in\mathcal{M}} p(A) = U(A)$, $\forall\, A \in \mathcal{A}$.

3) Logical concept of probability:

Let $(\Omega_A; \mathcal{A})$ and $(\Omega_B; \mathcal{B})$, $\Omega_A \cap \Omega_B = \emptyset$, be two measure spaces, where $\{x\} \in \mathcal{A}$, $\forall\, x \in \Omega_A$, $\{y\} \in \mathcal{B}$, $\forall\, y \in \Omega_B$.

A *W-field* $\mathcal{W} = (\Omega_A; \mathcal{A}; \Omega_B; \mathcal{B}; L(.||.))$ is given, iff the following four axioms hold:

L I: To each $B \in \mathcal{B}^+ := \mathcal{B} \setminus \{\emptyset\}$ an F-field $\mathcal{F}(B) = (\Omega_A; \mathcal{A}; L(.||B))$ is attached.

L II: Let $I \neq \emptyset$ be an index set, $B_0 \in \mathcal{B}^+$, $B_i \in \mathcal{B}^+$, $i \in I$, and

$$B_0 = \bigcup_{i\in I} B_i.$$

Then[3]:

$$\mathcal{F}(B_0) = \bigcup_{i\in I} \mathcal{F}(B_i).$$

[2] According to Axioms T IV–T VI the function $U(.)$ is conjugate to $L(.)$: $U(A) = 1 - L(\neg A)$, $\forall\, A \in \mathcal{A}$.

[3] The union $\mathcal{F} = \cup_{i\in I}\mathcal{F}_i = (\Omega_A; \mathcal{A}; L(.))$ of F-fields $\mathcal{F}_i = (\Omega_A; \mathcal{A}; L_i(.))$, $i \in I$, is defined by $L(.) := \inf_{i\in I} L_i(.)$. Hence $U(.) = \sup_{i\in I} U_i(.)$, and $\mathcal{F}$ is an F-field too. The employment of this procedure in assigning probability of arguments characterizes the Logical concept in contrast to the Bayesian approach.

L III: Let $A \in \mathcal{A}$, $B_1 \in \mathcal{B}^+$ irrelevant for A. Then:

$$L(A||B_1 \cap B_2) = L(A||B_2),\ \forall B_2 \in \mathcal{B}^+.$$

L IV: Let $A_i \in \mathcal{A}^+$, $B_i \in \mathcal{B}^+$, $i = 1, 2$, $(A_1||B_1)$ and $(A_2||B_2)$ independent from each other. Then:

$$L(A_1 \cap A_2||B_1 \cap B_2) = \\ = L(A_1||B_1 \cap B_2) \cdot L(A_2||B_1 \cap B_2)$$

$$U(A_1 \cap A_2||B_1 \cap B_2) = \\ = U(A_1||B_1 \cap B_2) \cdot U(A_2||B_1 \cap B_2).$$

The Logical concept of probability defined by Axioms L I–L IV as a general principle employs probability as a two-place-function: $P(A||B)$ is to be interpreted as probability of the argument with premise B and with conclusion A and never must be mistaken as conditional probability. According to this concept $P(A)$ and $P(B)$ do not exist and therefore $P(A|B)$ never exists either. (On the other hand $P((A_1|A_2)||B)$ is a valuable information in many situations.) Axiom L II characterizes the distinction of the Logical concept and any kind of Bayesian concept.

The fact that $\Omega_A \cap \Omega_B = \emptyset$ and therefore $\mathcal{A}$ and $\mathcal{B}$ always are disjoint, demonstrates the basic distinction between W-fields and Popper-spaces (cf. [12] and [11]). This does not prevent the idea of combining both aspects — but the success of such a program cannot be foreseen.

On the other hand there is no relationship of the Logical concept with approaches of Default reasoning (cf. [7] and [10]) or of Plausibility measures and Possibility measures. The Logical Concept does not extend the field of application beyond that of classical probability: Its main goal is to improve the methodology of statistical reasoning by introducing duality between appropriate W-fields and hereby allowing the employment of probability to describe results of statistical inference. With respect to the intention there is a relationship to approaches by R.A. Fisher ([5]), D.A.S. Fraser ([6]), A. Dempster ([4]), A. Birnbaum ([2]), and I. Hacking ([8]), but there exist fundamental differences in methodology[4].

A survey of the resulting Symmetric theory of probability was given in the ISIPTA 05 paper, a short survey of duality in statistical inference can be found in a report for the 56th Session of ISI in Lisboa, 2007 ([16]).

[4] A review of these approaches is given by T. Seidenfeld ([13]).

4 Independence and Multiplicativity

Axioms L I–L IV allow to establish a corpus of definitions and statements constituting the theory of the Logical concept of probability. With one important exception the theory of the classical concept can be regarded as a special case of this theory. The difference between the two approaches with respect to the concept of independence is characterized by the results of this section.

Corollary 2 *From Axiom L III and Corollary 1 it follows that under the conditions for L III:*

$$U(A||B_1 \cap B_2) = U(A||B_2),\ \forall B_2 \in \mathcal{B}^+,$$

holds. □

From Axioms L III and L IV together with Corollary 2, it may be concluded:

Corollary 3 *If $(A_1||B_1)$ and $(A_2||B_2)$ are mutually independent,*

$$\begin{aligned} L(A_1 \cap A_2||B_1 \cap B_2) &= L(A_1||B_1) \cdot L(A_2||B_2) \\ U(A_1 \cap A_2||B_1 \cap B_2) &= U(A_1||B_1) \cdot U(A_2||B_2) \end{aligned}$$

hold. □

This result says that, according to the Logical Concept, mutual independence of P-arguments produces total multiplicativity of probabilities. However, on the other hand, it is not possible in this theory to infer mutual independence of P-arguments from multiplicativity of probabilities. This is a decisive difference to the objectivistic view of classical theory, where independence of events is defined by means of multiplicativity of probabilities. It should be emphasized that mutual independence of P-arguments can only be understood as the fact that each premise is irrelevant to the conclusion of the other P-argument.

Now let $(A_i||B_i)$ with $P(A_i||B_i) = [L; U]$, $i = 1, 2, ...$, be a potentially infinite series of mutually independent P-arguments and $r \in \mathbb{N}$.

Due to independence, the probability for the combined P-argument $(\vec{A}||\vec{B})$ with $\vec{A} = A_1^* \times ... \times A_r^*$ where $A_i^* \in \{A_i, \neg A_i\}$, $\vec{B} = B_1 \times ... \times B_r$ is multiplicative:

$$P^{[r]}(\vec{A}||\vec{B}) = \left[\prod_{i=1}^{r} L(A_i^*||B_i); \prod_{i=1}^{r} U(A_i^*||B_i)\right].$$

Let $I \subseteq \{1, ..., r\}$, $A_i^* = A_i$, $\forall\, i \in I$, $A_i^* = \neg A_i$, $\forall\, i \notin I$. Then I describes the conclusion $\vec{A} =: \vec{A}(I)$ uniquely. Because of

$$\begin{aligned} P^{[r]}(A_i^*||B_i) &= [L; U],\ \forall\, i \in I, \\ P^{[r]}(A_i^*||B_i) &= [1-U; 1-L],\ \forall\, i \notin I, \end{aligned}$$

one arrives at

$$P^{[r]}(I||\vec{B}) := P(\vec{A}(I)||\vec{B}) = \left[L^{|I|}\cdot(1-U)^{r-|I|};\, U^{|I|}\cdot(1-L)^{r-|I|}\right].$$

Let $\rho := |I|$. In order to calculate the probability of an argument with a conclusion of very few $\neg A_i$ and almost all A_i,

$$P^{[r]}(\rho \geq \rho_0||\vec{B}) = \left[\inf_{\substack{L\leq p_i\leq U\\ i\in\{1,\ldots,r\}}} \sum_{|I|\geq\rho_0}\prod_{i\in I} p_i \prod_{i\in\{1,\ldots,r\}\setminus I}(1-p_i);\ \sup_{\substack{L\leq p_i\leq U\\ i\in\{1,\ldots,r\}}} \sum_{|I|\geq\rho_0}\prod_{i\in I} p_i \prod_{i\in\{1,\ldots,r\}\setminus I}(1-p_i)\right]$$

has to be calculated. Due to the monotonicity of the function

$$p^{[r]}(\rho \geq \rho_0||\vec{B}) = \sum_{|I|\geq\rho_0}\prod_{i\in I} p_i \prod_{i\in\{1,\ldots,r\}\setminus I}(1-p_i)$$

in each of the $p_i \in [L;\, U]$, $i = 1,\, ...,\, r$,

$$P^{[r]}(\rho \geq \rho_0||\vec{B}) = \left[\sum_{\rho\geq\rho_0}\binom{r}{\rho}L^{\rho}(1-L)^{r-\rho};\ \sum_{\rho\geq\rho_0}\binom{r}{\rho}U^{\rho}(1-U)^{r-\rho}\right]$$

holds.

On the other hand:

$$P^{[r]}(\rho \leq \rho_0||\vec{B}) = \left[\sum_{\rho\leq\rho_0}\binom{r}{\rho}U^{\rho}(1-U)^{r-\rho};\sum_{\rho\leq\rho_0}\binom{r}{\rho}L^{\rho}(1-L)^{r-\rho}\right].$$

Therefore the probabilities of arguments with conclusions of extremely many or extremely few factors A_i can be calculated employing classical Binomial law and Tschebysheff's inequality:

Let

$$\rho_0 = rU + r\delta: \quad U^{[r]}(\rho \geq \rho_0||\vec{B}) \leq \frac{U(1-U)}{r\delta^2},$$

let

$$\rho_0^* = rL - r\delta: \quad U^{[r]}(\rho \leq \rho_0^*||\vec{B}) \leq \frac{L(1-L)}{r\delta^2}.$$

As a consequence:

$$L^{[r]}\left(L-\delta < \frac{\rho}{r} < U+\delta\right) \geq 1 - \frac{L(1-L)+U(1-U)}{r\delta^2}$$

and

$$L^{[r]}(L-\delta < \frac{\rho}{r} < U+\delta) \geq 1-\varepsilon, \quad \text{if } r \geq \frac{L(1-L)+U(1-U)}{\varepsilon\delta^2}. \tag{1}$$

This result can be interpreted by means of appropriate concepts of converging sequences of W-fields:

Definition 8 *Let $\mathcal{W}^{[r]} = (\Omega_A;\, \mathcal{A};\, \Omega_B^{[r]};\, \mathcal{B}^{[r]};\, L^{[r]}(.||.))$, $r \in \mathbb{N}$, be a sequence of W-fields and $\mathcal{Z} \in \mathcal{A}$ be a non-empty conclusion. If for $\vec{B}_0^{[r]} = B_0^{[1]} \times ... \times B_0^{[r]}$, $r \in \mathbb{N}$, and for every $\mathcal{Z}^* \in \mathcal{A}$ with $\mathcal{Z}^* \supsetneqq \mathcal{Z}$ there exists a function $N(\mathcal{Z}^*,\, \varepsilon) \in \mathbb{N}$, so that*

$$L^{[r]}(\mathcal{Z}^*||B_0^{[r]}) \geq 1-\varepsilon,\ \forall\, r \geq N(\mathcal{Z}^*,\, \varepsilon), \tag{2}$$

then with respect to the arguments $(\mathcal{Z}||\vec{B}_0^{[r]})$, $r \in \mathbb{N}$, the sequence $\mathcal{W}^{[r]}$ is named stochastically convergent *to a sequence $\overline{\mathcal{W}}^{[r]}$ of W-fields, $\overline{\mathcal{W}}^{[r]} = (\Omega_A;\, \mathcal{A};\, \Omega_B^{[r]};\, \mathcal{B}^{[r]};\, \overline{L}^{[r]}(.||.))$, $r \in \mathbb{N}$, with $\overline{P}^{[r]}(\mathcal{Z}||B_0^{[r]}) = [1;\, 1] =: [1]$.* □

According to Definition 8 the result (1) can be utilized to formulate

Corollary 4 *Let the sequence of P-arguments $(A_0^{(i)}||B_0^{(i)})$, $i \in \mathbb{N}$, with $P^{(i)}(A_0^{(i)}||B_0^{(i)}) = [L;\, U]$ be mutually independent. For $r \in \mathbb{N}$ let*

$$\vec{B}_0^{[r]} = B_0^{(1)} \times ... \times B_0^{(r)},$$
$$t = \frac{\rho}{r},\ \rho \in \{0,\, ...,\, r\},$$
$$A_0^{[r]}(t) := \bigcup_{\substack{I\subseteq\{0,\,\ldots,\,r\}\\ |I|=r\cdot t}}\left[\bigcap_{i\in I} A_0^{(i)} \cap \bigcap_{i\in\{0,\,\ldots,\,r\}\setminus I}\neg A_0^{(i)}\right].$$

Let $\mathcal{W}^{[r]} = (\Omega_A^{[r]};\, \mathcal{A}^{[r]};\, \Omega_B^{[r]};\, \mathcal{B}^{[r]};\, L^{[r]}(.||.))$, $r \in \mathbb{N}$, be W-fields containing the probability of arguments with premise $\vec{B}_0^{[r]}$ and conclusions of the kind $A_0^{[r]}(J) = \bigcup_{r\cdot t\in J} A_0^{[r]}(t)$, $J \subseteq \{0,\, ...,\, r\}$, so that

$$\Omega_A^{[r]} = \{0,\, \tfrac{1}{r},\, \tfrac{2}{r},\, ...,\, 1\},$$
$$\mathcal{A}^{[r]} = \mathcal{P}ot(\Omega_A^{[r]}),$$
$$\vec{B}_0^{[r]} \in \mathcal{B}^{[r]}.$$

The sequence $\mathcal{W}^{[r]}$, $r \in \mathbb{N}$, is then with respect to the arguments $(A_0^{[r]}(t)||\vec{B}_0^{[r]})$ stochastically convergent to the sequence $\overline{\mathcal{W}}^{[r]}$, $r \in \mathbb{N}$, of W-fields

$$\overline{\mathcal{W}}^{[r]} = (\overline{\Omega}_A^{[r]};\, \overline{\mathcal{A}}^{[r]};\, \Omega_B^{[r]};\, \mathcal{B}^{[r]};\, \overline{L}^{[r]}(.||.))$$

with

$$\begin{aligned}
\overline{\Omega}_A^{[r]} &= [0;\,1], \\
\overline{\mathcal{A}}^{[r]} &= \mathcal{B}or(\overline{\Omega}_A^{[r]}), \\
\overline{L}^{[r]}(A||\vec{B}_0^{[r]}) &= \begin{cases} [1], & A \supseteq [L;\,U] \\ [0], & A \cap [L;\,U] = \emptyset \\ [0;\,1], & else. \end{cases}
\end{aligned}$$

□

Corollary 4 is the obvious consequence of (1) if applied to a sequence of mutually independent arguments with probability $[L;\,U]$.

Introducing abbreviations, the result (1) may be expressed by the statement

$$\lim_{r\to\infty} P^{[r]}(L \le t_r \le U||\vec{B}_0^{[r]}) = [1] \qquad (3)$$

and may be interpreted in a way similar to that concerning convergence of a sequence of variables in classical statistics — if the decisive difference is seen, that $[L;\,U]$ is an interval and normally not a single number. From the facts given, it is not possible to describe the result by more information than, what may be characterized by "finally: $L \le t_r \le U$". Neither convergence nor divergence of the sequence t_r, $r = 1, ...$, can be excluded as a possible conclusion. As an appropriate new expression to denote results of this type the sentence "The sequence t_r inverges the set $[L;\,U]$" is proposed.

5 Frequency Interpretation

As a consequence of these results a frequency interpretation of the Logical concept of probability is available.

If $(A||B)$ is a P-argument with $P(A||B) = [L;\,U]$ in a kind of thought experiment, then $(A||B)$ may be conceived as one out of a potentially infinite sequence of mutually independent P-arguments $(A_i||B_i)$ with exactly the same probability assessment:

$$P(A_i||B_i) = [L;\,U],\ i = 1,\,2,\,...$$

Then the P-argument $(\hat{A}^{[r]}||\hat{B}^{[r]})$ is considered, where

$$\hat{B}^{[r]} := \bigcap_{i=1}^{r} B_i$$

is the conjunction of all single premises and

$$\hat{A}^{[r]} := \bigcup_{\substack{I\subseteq\{1,\,...,\,r\} \\ rL\le|I|\le rU}} \left[\bigcap_{i\in I} A_i \cap \bigcap_{i\in\{1,\,...,\,r\}\setminus I} \neg A_i\right]$$

is the adjunction of all combined conclusions, for which the proportion of A_i lies between L and U.

Due to (2) and (3)

$$\lim_{r\to\infty} P^{[r]}(\hat{A}^{[r]}||\hat{B}^{[r]}) = [1]$$

holds.

If only r is large enough, the conclusion $\hat{A}^{[r]}$ can be derived from the premise $\hat{B}^{[r]}$ with practical surety.

Therefore the P-argument $(A||B)$ with $P(A||B) = [L;\,U]$ can be interpreted as if it was one out of a huge set of mutual independent P-arguments $(A_i||B_i)$, for which the proportion of successful arguments $(A_i||B_i)$ — producing unsuccessful arguments $(\neg A_i||B_i)$ — lies between L and U, the proportion of unsuccessful arguments $(A_i||B_i)$ — and therefore successful arguments $(\neg A_i||B_i)$ — lies between $1-U$ and $1-L$.

The conceptual basis of this kind of procedure is given by Cournot's Lemma, which was formulated with reference to the objectivistic view on probability, and can be transferred to the Logical concept in the following way:

Cournot's Lemma: *If $L(A||B) = 1-\varepsilon$ and ε is extremely small, the P-argument $(A||B)$ may be understood as if $P(A||B) = [1]$.* □

The validity of this interpretation is founded on the fact that in a set of mutually independent P-arguments, for which only $P(A_i||B_i) = [L;\,U]$ is known, obviously no subset can be identified, for which an additional information about the proportion of successful arguments $(A_i||B_i)$ would be possible.

It must be pointed to the fact, that the value of this frequency-interpretation of the Logical concept of probability in the first line depends on the concept of independent P-arguments: A set of mutual independent P-arguments is defined by contents of premises and conclusions. If additionally all of them are evaluated by the same $P(A_i||B_i)$ according to the axioms of the Logical concept, and if the set is large enough, inference about the proportion of successful ones in the set is possible.

The availability of an unassailable frequency interpretation of the Logical concept is of high importance with respect to the Symmetric theory (see [15]): In this theory probabilistic statements about arguments are employed not only to describe statistical modeling but also — by means of dual W-fields — for statistical inference. The concept of imaging any evaluated argument as one out of a potentially infinite sequence of mutually independent arguments with the same probability guarantees the uniformity in understanding the assessments employed in modeling as well as in inference.

Finally it must be mentioned that it is possible to improve the result of weak invergence: If a more general concept of W-field is introduced, the concept of strong invergence can be defined and it can be proven, that t_r inverges $[L; U]$ with probability [1] according to this concept. However, this result does not influence the understanding of a frequency interpretation of the Logical concept.

6 Conclusions and Prospect

When early in the 17th century the concept of probability arose from the study of games of chance, the possibility of repeating any game was a fundamental idea, comprising the concept of mutual independence of the repetitions. Multiplicativity of probability under this supposition was accepted as intuition resulting in the close relationship between frequency and probability ([14], pp. 42 ff.).

When the theory of probability developed in the following centuries it was obvious that mutual independence of events is a relation which cannot be defined without employment of exogenous concepts.

It was A. N. Kolmogorov who produced a solution by defining mutual independence via multiplicativity of probabilities ([9], pp. 8–12 of the English version), hereby accepting some border cases which are more or less counter-intuitive.

Employment of imprecise probabilities, as it is propagated by Peter Walley must rely on definitions as they are given for precise probabilities. Owing to a behaviouristic approach the concept of irrelevance is near at hand ([3]). As long as probability is seen as a one-place-function — attributed to events or statements — this remains conditional irrelevance of one event with respect to another one. Considerations of this kind come near to a justification of multiplicativity, but fail to explain independence without using the concept of probability ([3]). An additional aspect is provided by the question which type of conditional interval probability is to be employed in such consideration ([1]). Altogether: The difference between the concept of independence as employed in the present approach and other concepts of independence introduced in methodology of imprecise probability is fundamental: Independence of P-arguments is defined by the contents of propositions employed, while independence of events with imprecise probability is defined by relations of — total or conditional — probabilities.

This remains the situation of classical probability theory according to the objectivistic view. By means of the weak law of large numbers a frequency interpretation of classical probability can be derived, based on the concept of a sequence of mutual independent events with the same probability.

Criticism of the objective view stresses that probability being the conceptual basis defining independence of events, it should not be interpreted by a characteristic of a sequence of mutual independent events.

If probability is attributed to arguments — instead of events — the situation is different, since independence of two arguments can be identified with irrelevance of both premises with respect to the conclusion of the other argument.

The ISIPTA'05 paper ([15]) defines mutual independence of arguments by means of the probability assessment: It is in fact Axiom L III of the present paper which is employed in [15] to distinguish mutual independent arguments without denoting this procedure explicitly. Simultaneously multiplicativity of probability for independent arguments is required: Thus Axiom L III of the '05-paper combines two different aspects in one equation.

This procedure can be criticized not only because of its complexity, but also with respect to the detail that it encourages objections against a frequency interpretation of probability employing a sequence of independent arguments because of the role of probability in defining independence.

The present paper relies on definitions of irrelevance and independence through the contents of the propositions involved. Axiom L III contains the demand that irrelevance always is reflected by the probability assessment, and Axiom L IV insists on multiplicativity with respect to conclusions in case of mutual independence — in analogy to multiplicativity in classical probability.

Consequently the Logical concept of probability according to Section 5 is characterized by a frequency interpretation employing an autonomous concept of independence — a feature not to be found elsewhere.

This result characterizes the Symmetric theory of probability, which relies decisively on the Logical concept. The establishment of duality between W-fields generates a methodology of statistical inference employing the concept of probability in the same way as in statistical modeling attached to P-arguments.

Any statement of probability arising from applied Symmetric theory therefore is of the same quality and should be understood by means of the frequency interpretation: one out of a very large series of assessments with the same $[L; U]$ concerning mutually independent P-arguments. The proportion of "successful" arguments in this series lies between L and U.

Present research into Symmetric theory aims at the range of possible applications. Which types of problems in classical statistics can be solved by means of the concept of duality?! A comprehensive report [17] is in preparation.

Acknowledgements

I am grateful to Joseph Kadane, with whom I had a short but very impressive talk together with filet Mignon at Andy Warhol-Museum in Pittsburgh on July 21, 2005, I am grateful to Thomas Augustin whose remarks upon the draft were very instructive, and I am grateful to Anton Wallner who supports me through many years and corrects me — hopefully always when it is necessary!

References

[1] T. Augustin, K. Weichselberger. On the symbiosis of two concepts of conditional interval probability, in: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (eds.), *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, Carleton Scientific, 608–629, 2003.

[2] A. Birnbaum. Concepts of statistical evidence, in: S. Morgenbesser, P. Suppes, M. White (eds.), *Philosophy, Science, and Method*, St. Martin's Press, New York, 1969.

[3] I. Couso, S. Moral, P. Walley. A survey of concepts of independence for imprecise probabilities, in: *Risk, Decision and Policy*, vol. 5, 165–181, 2000.

[4] A.P. Dempster. Upper and lower probabilities generated by a random closed interval, in: *Annals of Mathematical Statistics* 39, 957–966, 1968.

[5] R.A. Fisher. *Statistical Methods and Scientific Inference*, 3rd ed., Hafner, New York, 1973.

[6] D.A.S. Fraser. *The Structure of Inference*, Wiley, New York, 1968.

[7] N. Friedman, J.Y. Halpern, D. Koller. First-order conditional logic for default reasoning revisited, in: *ACM Transactions on Computational Logic*, Vol. 1, No. 2, 175–207, 2000.

[8] I. Hacking. *Logic of Statistical Inference*, Cambridge University Press, Cambridge, 1965.

[9] A. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, Heidelberg, New York, 1933. English version: *Foundations of the Theory of Probability*, Chelsea Publishing Company, New York, 1950.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[11] K.R. Popper. *The Logic of Scientific Discovery (revised edition)*, Hutchison, London, 1968. The first version of this book appeared as *Logik der Forschung*, 1934.

[12] A. Rényi. On conditional probability spaces generated by a dimensionally ordered set of measures, in: *Theory of Probability and its Applications 1*, 61–71, 1956. Reprinted as paper 120 in *Selected Papers of Alfred Rényi, I: 1948–1956*, Akadémia Kiadó, 554–557, 1976.

[13] T. Seidenfeld. *Philosophical Problems of Statistical Inference*, Reidel, Dordrecht, 1979.

[14] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I — Intervallwahrscheinlichkeit als umfassendes Konzept*, in cooperation with T. Augustin and A. Wallner, Physica, Heidelberg, 2001.

[15] K. Weichselberger. The logical concept of probability and statistical inference, in: F.G. Cozman, R. Nau, T. Seidenfeld (eds.), *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, published by SIPTA, 396–405, 2005.

[16] K. Weichselberger. Applying symmetric probability theory, in: *Proceedings of the 56th Session of International Statistical Institute ISI 2007*, CD, Internet.

[17] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung II — Symmetrische Wahrscheinlichkeitstheorie*, in cooperation with A. Wallner, in preparation.

AUTHOR INDEX