ISIPTA '15

Proceedings of the 9th International Symposium on Imprecise Probability: Theories and Applications

20-24 July 2015, Pescara, Italy

Edited by

Thomas Augustin
Serena Doria
Enrique Miranda
Erik Quaeghebeur

ISIPTA '15 – http://www.sipta.org/isipta15

Electronic version published by SIPTA – Society for Imprecise Probability: Theories and Applications – http://www.sipta.org

Paper version published by *Aracne Editrice* – http://www.aracneeditrice.it/

This book was typeset with LaTeX using the `memoir` *class,*
supported by the `datatool` *bundle and the* `pdfpages`*,* `imakeidx`*, and* `pgffor` *packages.*

Contents

Poster abstracts

Preface

The ISIPTA meetings are the primary forum for presenting and discussing advances in imprecise probability research. They are organized once every two years by SIPTA, the *Society for Imprecise Probability: Theories and Applications*. The first meeting was held in Ghent in 1999. It was followed by meetings in Ithaca, Lugano, Pittsburgh, Prague, Durham, Innsbruck, and Compiègne. After having successfully hosted the SIPTA Summer School in 2012, we now return to the beautiful and welcoming Italian city of Pescara for

The 9th International Symposium
on Imprecise Probability: Theories and Applications

It is held from Monday 20 to Friday 24 July 2015.

As with previous ISIPTA meetings, there are only plenary sessions in the program. In total, 31 papers are presented by a short talk and a poster, which guarantees ample time for discussion. The papers are included in these proceedings and are also available on the SIPTA website (`www.sipta.org`). Each submitted paper has undergone a thorough reviewing process by multiple expert reviewers, ensuring the quality of the accepted contributions.

To provide a platform for preliminary ideas and challenging applications for which the research is not yet completed, poster-only presentations were introduced at ISIPTA '09. It has become a tradition that is continued at ISIPTA '15: during the conference 17 additional posters will be presented. Short abstracts for these poster-only presentations are included in these proceedings and are also available on the SIPTA website.

The contributions bring us a large number of new results—both theoretical and applied—within the field of imprecise probability. The broad impact of imprecise probability is shown by the wide variety in the contributions' domains: decision making, statistical inference, belief aggregation, artificial intelligence, and stochastic processes, amongst others.

We are pleased to have three eminent invited speakers: *Itzhak Gilboa*, from Tel Aviv University and HEC Paris, will propose a unified model of inductive reasoning; *Peter Williams*, from the University of Sussex and BW Mining, will review the intellectual background for the development of coherent lower previsions; and *Massimo Marinacci*, from Bocconi University, will discuss approaches to model uncertainty in decision problems.

We are also pleased to have two tutorials to highlight specific subdomains of the wide field of imprecise probability: *Barbara Vantaggi*, from Universitá "La Sapienza" di Roma, will lecture on de Finetti's work on coherence and its extensions to an imprecise context; whereas *Gregory Wheeler*, from Ludwig-Maximilians Universität in Munich, will teach us about the philosophical foundations of imprecise probabilities.

During the conference two sets of prizes are awarded: the *Best Poster Award*, sponsored by Springer and Wiley, and the *IJAR Young Researcher Award*, granted by the International Journal of Approximate Reasoning. We express our gratitude for their support.

This conference is a result of the productive cooperation between the members of the Steering Committee, formed by *Gert de Cooman*, *Teddy Seidenfeld*, and ourselves. We wish to thank all of those that have contributed to the organization of this conference: all the members of the Local Organizing Committee; the Department of Engineering and Geology of the University G. d'Annunzio for its financial support; the many members of the Program Committee and the extra reviewers for their dedicated work in evaluating the contributions. Last but not least, we would like to particularly thank *Matthias Troffaes* and *Sébastien Destercke* for their assistance with many aspects of the conference, and for sharing their previous organizational experience.

Finally, we thank all who have contributed to the success of ISIPTA '15, be it by submitting their research results, presenting them at the conference, or by attending sessions and participating in discussions. In particular, we would like to welcome the delegates from Statistics Korea, whom we thank for their effort to become part of our research network.

Thomas Augustin
Serena Doria
Enrique Miranda
Erik Quaeghebeur

June 2015

Organization

Steering Committee

Thomas Augustin
Ludwig-Maximilians-Universität, Munich, Germany

Gert de Cooman
Ghent University, Belgium

Serena Doria
University G. d'Annunzio, Chieti-Pescara, Italy

Enrique Miranda
University of Oviedo, Spain

Erik Quaeghebeur
Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Teddy Seidenfeld
Carnegie Mellon University, Pittsburgh, USA

Local Organization

Andrea Di Cencio
University G. d'Annunzio, Chieti-Pescara, Italy

Serena Doria
University G. d'Annunzio, Chieti-Pescara, Italy

Attilio Grilli
University G. d'Annunzio, Chieti-Pescara, Italy

Mariangela Scorrano
University of Trieste, Italy

Program Committee

Board

Thomas Augustin
Ludwig-Maximilians-Universität, Munich, Germany

Enrique Miranda
University of Oviedo, Spain

Erik Quaeghebeur
Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Members

Alessandro Antonucci
IDSIA, Lugano, Switzerland

Michael Beer
Univesity of Liverpool, UK

Alessio Benavoli
IDSIA, Lugano, Switzerland

Mik Bickis
University of Saskatchewan, Canada

Seamus Bradley
Munich Centre for Mathematical Philosophy, Ludwig-Maximilians-Universität, Munich, Germany

Andrey Bronevich
Research National University "Higher School of Economics", Russia

Andrea Capotorti
University of Perugia, Italy

Marco Cattaneo
University of Hull, UK

Giulanella Coletti
University of Perugia, Italy

Frank Coolen
Dept. of Mathematical Sciences, Durham University, UK

Giorgio Corani
IDSIA, Lugano, Switzerland

Inés Couso
University of Oviedo, Spain

Fabio G. Cozman
Universidade de São Paulo, Brazil

Fabio Cuzzolin
Oxford Brookes University, United Kingdom

Milan Daniel
Institute of Computer Science, Academy of Sciences of the Czech Republic

Jasper De Bock
Ghent University, Belgium

Cassio de Campos
Queen's University Belfast, UK

Gert de Cooman
Ghent University, Belgium

Thierry Denoeux
Universite de Technologie de Compiegne, France

Sebastien Destercke
CNRS, UMR Heudiasyc, France

Serena Doria
University G. d'Annunzio, Chieti-Pescara, Italy

Didier Dubois
Université Paul Sabatier, Toulouse, France

Love Ekenberg
Stockholm University and KTH, Sweden

Scott Ferson
Applied Biomathematics, Inc.

Thomas Fetz
University of Innsbruck, Austria

Itzhak Gilboa
Tel Aviv University, Israel & HEC, Paris, France

Michel Grabisch
Université Paris I, France

Brian Hill
CNRS-HEC Paris, France

Radim Jiroušek
University of Economics, Czech Republic

Jim Joyce
University of Michigan, United States

Jay Kadane
Carnegie Mellon University, Pittsburgh, USA

Alexander Karlsson
University of Skövde, Sweden

Vladik Kreinovich
University of Texas at El Paso, USA

Tomáš Kroupa
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic

Shoumei Li
Beijing University of Technology, China

Massimo Marinacci
Bocconi University, Italy

Andrés Masegosa
University of Granada, Spain

Denis Mauá
Universidade de São Paulo, Brazil

Ignacio Montes
University of Oviedo, Spain

Serafín Moral
University of Granada, Spain

Yasuo Narukawa
Tokyo Institute of Technology, Japan

Michael Oberguggenberger
University of Innsbruck, Austria

Arthur Paul Pedersen
Max Planck Institute for Human Development, Berlin, Germany

Renato Pelessoni
Dept. D.E.A.M.S. 'Bruno de Finetti', University of Trieste, Italy

David Schmeidler
Tel Aviv University, Israel

Teddy Seidenfeld
Carnegie Mellon University, Pittsburgh, USA

Glenn Shafer
Rutgers Business School, USA

Damjan Skulj
University of Ljubljana, Slovenia

Michael Smithson
The Australian National University

Joerg Stoye
Cornell University, USA

David Sundgren
Dept. of Computer and Systems Sciences, Stockholm University, Sweden

Matthias Troffaes
Dept. of Mathematical Sciences, Durham University, UK

Lev Utkin
Saint Petersburg State Forest Technical University, Russia

Linda Van der Gaag
Dept. of Information and Computing Sciences, Utrecht University, Netherlands

Barbara Vantaggi
Dept. Me.Mo.Mat., Università "La Sapienza"

Jiřina Vejnarova
Institute of Information Theory and Automation of the AS CR

Paolo Vicig
Dept. D.E.A.M.S. 'Bruno de Finetti', University of Trieste, Italy

Volodya Vovk
Royal Holloway, University of London, UK

Peter Wakker
Erasmus University, Rotterdam, Netherlands

Gero Walter
Eindhoven University of Technology, Netherlands

Gregory Wheeler
Munich Centre for Mathematical Philosophy, Ludwig-Maximilians-Universität, Munich, Germany

Andrea Wiencierz
Dept. of Mathematics, University of York, UK

Marco Zaffalon
IDSIA, Lugano, Switzerland

Additional Reviewers

Marco Baioletti
University of Perugia, Italy

Davide Petturiti
University of Perugia, Italy

IJAR Award Committee

Fabio G. Cozman
Universidade de Sao Paulo, Brazil

Thierry Denoeux
Universite de Technologie de Compiegne, France

Enrique Miranda
University of Oviedo, Spain

Teddy Seidenfeld
Carnegie Mellon University, Pittsburgh, USA

Sponsors

Abstracts of Invited Talks

A Unified Model of Inductive Reasoning

Itzhak Gilboa
Eitan Berglas School of Economics, Tel-Aviv University, Israel
HEC, Paris, France

We offer a model that can capture three types of reasoning.[1] The first, which is the most common in economic modeling, is *Bayesian*. The agent formulates the set of possible states of the world and a prior probability distribution over this state space. The agent's predictions are a relatively straightforward matter of applying Bayes' rule, as new observations allow her to rule out some states and condition her probability distribution on the surviving states.

An alternative mode of reasoning is *case-based*. The agent considers past observations and predicts the outcome that appeared more often in those past cases that are considered similar. If all past observations are considered equally similar, the case-based prediction is simply the mode, that is, the outcome that is most frequent in the database. If the agent uses a similarity function that puts all its weight on the most recent outcome, her prediction will simply be that outcome.

Finally, *rule-based* reasoning calls for the agent to base her predictions on regularities that she believes characterize the phenomenon in question.

The boundaries between the three modes of reasoning are not always sharp. Our focus is on the Bayesian approach. By "Bayesian reasoning" we refer to the common approach in economic theory, according to which *all* reasoning is Bayesian. *Any* source of uncertainty is modeled in the state space, and all reasoning about uncertainty takes the form of updating a prior probability via Bayes' rule.

We present a framework that unifies these three modes of reasoning (and potentially others), allowing us to view them as special cases of a general learning process. The agent attaches weights to conjectures. Each conjecture is a set of states of the world, capturing a way of thinking about how outcomes in the world will develop. The associated weights capture the relative influence that the agent attaches to the various conjectures. The weighted sum of these conjectures is a Belief Function as in Dempster (1967) and Shafer (1976).

Given a sequence of observations, the agents rules out the conjectures that have been refuted by them, and continues with the weighted sum of the remaining ones. This turns out to be equivalent to Dempster-Shafer rule of combination, or updating of a belief function.

To generate a prediction, the agent sums the weight of all nontrivial conjectures consistent with each possible outcome, and then ranks outcomes according to their associated total weights. In the special case where each conjecture consists of a single state of the world, our framework is the standard Bayesian model, and the learning algorithm is equivalent to Bayesian updating. Employing other conjectures, which include more than a single state each, we can capture other modes of reasoning, as illustrated by simple examples of case-based and of rule-based reasoning.

Our model could be used to address either positive or normative questions. We focus on positive ones, describing how the reasoning process of an agent evolves as observations are gathered. Within the class of such questions, our model could be used to capture a variety of psychological biases and errors, but the focus of this paper is on the reasoning of an agent who makes no obvious errors in her reasoning. Such an agent may well be surprised by circumstances that she has deemed unlikely, that is, by "black swans," but will never be surprised by a careful analysis of her own reasoning. The optimality of this reasoning process is a normative question, which we do not address here.

Our main results concern the dynamics of the weight put on Bayesian vs. non-Bayesian reasoning. We suggest conditions under which Bayesian reasoning will give way to other modes of reasoning, and alternative conditions under which the opposite conclusion holds. Importantly, if the agent does not know the type of

[1] The talk is based on joint work with (i) Larry Samuelson and David Schmeidler (2013); (ii) Gabrielle Gayer (2014); (iii) Alfredo Di Tillio and Larry Samuelson (2013).

process she is facing, and attempts to be open-minded about it, Bayesian reasoning will disappear in the limit. The very simple reason is that there are many Bayesian conjectures, whereas other families of conjectures may be small. Specifically, the weight put on the Bayesian conjectures (as a whole) has to be divided among exponentially many disjoint subset, whereas the case-based ones (as well as some families of rule-based ones) are only polynomially large.

In a similar vein, we can also ask how the relative weight of rule-based and case-based conjectures changes with evidence. If a "rule" has to provide a prediction at each and every node, and be computable, we find that (i) if reality is simple enough (say, computable), then rule-based reasoning takes over; (ii) if reality isn't simple enough, then case-based reasoning is likely to be dominant.

Finally, the model can also be used to reason about counterfactuals.

References

Dempster, Arthur P. (1967). "Upper and lower probabilities induced by a multivalued mapping." *The Annals of Mathematical Statistics* 38.2, pp. 325–339. URL: http://www.jstor.org/stable/2239146.

Di Tillio, Alfredo, Itzhak Gilboa, & Larry Samuelson (2013). "The predictive role of counterfactuals." *Theory and Decision* 74.2, pp. 167–182. DOI: 10.1007/s11238-011-9263-6. URL: https://hal-hec.archives-ouvertes.fr/hal-00712888.

Gayer, Gabrielle & Itzhak Gilboa (2014). "Analogies and theories: The role of simplicity and the emergence of norms." *Games and Economic Behavior* 83, pp. 267–283. DOI: 10.1016/j.geb.2013.11.003. URL: https://hal-hec.archives-ouvertes.fr/hal-00712917.

Gilboa, Itzhak, Larry Samuelson, & David Schmeidler (2013). "Dynamics of Inductive Inference in a Unified Model." *Journal of Economic Theory* 148.4, pp. 1399–1432. DOI: 10.1016/j.jet.2012.11.004. URL: https://hal-hec.archives-ouvertes.fr/hal-00712823.

Shafer, Glenn (1976). *A mathematical theory of evidence*. Princeton University Press.

Model Uncertainty

Massimo Marinacci
Università Bocconi, Milan, Italy

We study decision problems in which the consequences of the alternative actions depend on states determined by a generative mechanism representing some natural or social phenomenon. Model uncertainty arises as decision makers may not know such mechanism. Two types of uncertainty result, a state uncertainty within models and a model uncertainty across them. We discuss some two-stage static decision criteria proposed in the literature that address state uncertainty in the first stage and model uncertainty in the second one (by considering subjective probabilities over models). We consider two approaches to the Ellsberg-type phenomena that these decision problems feature: a Bayesian approach based on the distinction between subjective attitudes toward the two kinds of uncertainty, and a non-Bayesian one that permits multiple subjective probabilities. Several applications are used to illustrate concepts as they are introduced.

Early Approaches to Exact Imprecision

Peter M. Williams
Associate, Department of Informatics, University of Sussex, Brighton, UK
Principal, BW Mining, Brighton, UK

The 1960s and 70s were a period of widespread interest in the philosophical and mathematical foundations of probability. Bayesian ideas were recognized though not well understood, and treated with caution by mainstream statisticians. This talk surveys the intellectual climate of the period, including the impact of de Finetti's ideas, then becoming more widely known in English translation, and traces the motivation and development of non-additive measures of uncertainty, together with their impact on the then developing treatment of uncertainty in artificial intelligence.

References

De Finetti, Bruno (1970). *Teoria Delle Probabilità*. English translation: (de Finetti 1974–1975). Giulio Einaudi.

– (1974–1975). *Theory of Probability*. Two volumes; translation of (de Finetti 1970). John Wiley & Sons.

Dempster, Arthur P. (1968). "A generalization of Bayesian inference." *Journal of the Royal Statistical Society. Series B (Methodological)* 30.2, pp. 205–247. JSTOR: 2984504.

Good, Irving John (1962). "Subjective Probability as the Measure of a non-Measurable Set." *Logic, Methodology and Philosophy of Science*. International Congress on Logic, Methodology and Philosophy of Science. (1960). Ed. by Ernest Nagel, Patrick Suppes, & Alfred Tarski. Stanford University Press, pp. 319–329.

Shafer, Glenn (1976). *A mathematical theory of evidence*. Princeton University Press.

Smith, Cedric A. B. (1961). "Consistency in statistical inference and decision." *Journal of the Royal Statistical Society. Series B (Methodological)* 23.1, pp. 1–37. JSTOR: 2983842.

Walley, Peter (1991). *Statistical reasoning with imprecise probabilities*. Vol. 42. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Williams, Peter M. (1975a). "Coherence, strict coherence and zero probabilities." *Fifth International Congress of Logic, Methodology and Philosophy of Science*. Vol. VI, pp. 29–33.

– (1975b). *Notes on conditional previsions*. Tech. rep. Published as (Williams 2007). School of Mathematical & Physical Sciences, University of Sussex.

– (1976). "Indeterminate probabilities." *Formal Methods in the Methodology of Empirical Sciences*. Conference for Formal Methods in the Methodology of Empirical Sciences. (Warsaw, Poland, June 17–21, 1974). Ed. by Marian Przełęcki, Klemens Szaniawski, Ryszard Wójcicki, & Grzegorz Malinowski. Vol. 103. Synthese Library. D. Reidel Publishing Company & Ossolineum Publishing company, pp. 229–246. DOI: 10.1007/978-94-010-1135-8_16.

– (2007). "Notes on conditional previsions." *International Journal of Approximate Reasoning* 44. Published version of (Williams 1975b), pp. 366–383. DOI: 10.1016/j.ijar.2006.07.019.

Abstracts of Tutorials

De Finetti Coherence and Beyond

Barbara Vantaggi
"La Sapienza" University of Rome, Italy
barbara.vantaggi@sbai.uniroma1.it

The aim of the tutorial is to present the concept of coherence, which dates back to de Finetti, showing its role in managing incomplete (or missing) information.

We will start recalling the notion of coherence for (unconditional) probabilities and the related fundamental theorem.

Then, in order to generalize this notion to assessments on a set of conditional events, the axiomatic definition of conditional probability, essentially due to Renyi, de Finetti and Dubins, needs to be recalled together with the representation theorem of a conditional probability by means of a linearly ordered class of finitely additive measures.

Both for the unconditional and conditional case, de Finetti's coherence has a betting scheme interpretation and it can also be characterized in terms of solvability of a sequence of linear systems for each finite subset of conditional events.

One of the main peculiarities of de Finetti's coherence is that a coherent assessment can always be extended, generally not in a unique way, to any superset of (conditional) events, giving rise to a class of coherent extensions.

The relationship of coherence with the first fundamental theorem of the asset pricing will be underlined.

The role of coherence is particularly meaningful in Bayesian statistics where the extensions of a likelihood function and a prior probability need to be found. Even in this case the coherent extensions are not necessarily unique, and the whole class of coherent extensions needs to be considered. This leads to study lower and upper envelopes.

However, the coherent extensions could be required to satisfy some further properties such as disintegrability and conglomerability: this leads to distinguish different subclasses of extensions.

Models able to handle uncertainty in a more flexible way have favored the emergence of theories more general than classical probability.

The resulting uncertainty calculi, such as possibility measures, belief functions and k-monotone Choquet capacities, can be interpreted in terms of envelopes of de Finetti's coherent probabilities, also referred to as imprecise probabilities.

The main features of de Finetti's coherence are discussed in connection with its "generalizations" to imprecise probabilities, essentially given by Williams and Walley.

The coherence criteria given by Williams and Walley for imprecise probabilities differ in the way they face conditioning, so a comparison of the different notions will be presented.

Finally, the different notions of coherence for (conditional) random quantities will be reviewed by comparing Williams and Walley theories.

Some examples coming from applications will be used to illustrate key concepts.

Introduction to the Philosophical Foundations of Imprecise Probabilities

Gregory Wheeler
Munich Center for Mathematical Philosophy
Ludwig Maximilians University
Geschwister-Scholl-Platz 1, 80539 Munich
gregory.wheeler@lrz.uni-muenchen.de

In this tutorial we will introduce several topics in the foundations of imprecise probabilities through a review of key historical figures, including John Maynard Keynes, B.O. Koopman, and I.J. Good, Henry Kyburg, Terrence Fine and Isaac Levi, and their reactions to the subjectivist-rationalist tradition associated with Ramsey, de Finetti, and Savage, and the later developments associated with Peter Williams and Peter Walley. We will end with a short overview of Epistemic Decision Theory, which aims to reinterpret the machinery of strictly proper scoring rules as measures of "epistemic accuracy," and the issues which arise from impossibility theorems which indicate that there are no strictly IP proper scoring rules.

Papers

The Multilabel Naive Credal Classifier

Alessandro Antonucci and **Giorgio Corani**
IDSIA SUPSI/USI
Lugano (Switzerland)
{alessandro,giorgio}@idsia.ch

Abstract

We present a credal classifier for multilabel data. The model generalizes the naive credal classifier to the multilabel case. An imprecise-probabilistic quantification is achieved by means of the imprecise Dirichlet model in its global formulation. A polynomial-time algorithm to compute whether or not a label is optimal according to the maximality criterion is derived. Experimental results show the importance of robust predictions in multilabel problems.

Keywords. Credal classification, imprecise Dirichlet model, multilabel classification.

1 Introduction

A classifier represents the relationship between the characteristics of an object (*features*) and its category (*class*). A traditional *classifier* predicts the *class* variable given the value of the features. *Credal classifiers* generalize traditional classifiers, allowing for set-valued predictions of classes. A credal classifier drops the non-optimal classes returning the classes that are potentially optimal given the information available. Depending on the data, there can be one or multiple optimal classes. Credal classifiers are thus less informative but more reliable than traditional classifiers [8]. Both credal and traditional classifiers assume the classes to be mutually *exclusive*.

Multilabel classification is a modern type of classification, in which an object is allowed to have multiple *relevant* classes (or *labels*). Multilabel classification arises naturally in many domains. A news article discussing EU treaties could be labeled for instance as politics *and* finance *and* environment. Similarly, tagging of photos and videos are natural multilabel problems. In bioinformatics, the identification of the best mix of drugs for curing HIV has been addressed as a multilabel problem [14].

The simplest approach for multilabel classification is *binary relevance*. Given q labels, binary relevance develops q independent single-label classifiers. The main shortcoming of binary relevance is that it ignores the dependencies among the different classes, which in many cases are important [12]. The algorithm of classifier chain [17] is a state-of-the-art approach to model dependencies among classes. Although it achieves good empirical performance, it has no direct probabilistic interpretation.

To model the dependence among classes in a probabilistically sound way, probabilistic graphical models are typically used [1, 3, 5, 18]. Each label is represented by a Boolean variable. The i-th Boolean variable represents whether the i-th label is relevant or not for the current instance. The inference task is to detect the most probable joint configuration of the labels. A joint configuration of the labels is a *sequence* of zeros and ones. Given q labels, there are 2^q possible sequences. Evaluating the robustness of the prediction, already important in traditional classification, is even more important in multilabel classification. There is however little work on this subject.

In this paper, we tackle this problem by means of *imprecise probabilities* [19]. We propose a graphical model which generalizes the naive Bayes to the multilabel setting. We learn the model using the *imprecise Dirichlet model* (IDM) [4, 20]. We discuss two types of inferences based on the criterion of *maximality*. The joint model detects the *maximal sequences*, among the 2^q possible ones. This inference is exact but is feasible only when q is limited, for instance smaller than 10. The marginal inference detects separately the maximal states of each label. We provide an approximated algorithm to solve this inference which scales to tens of labels.

The only other example of credal multilabel classifier currently available is the recent work of Destercke [13] which devises a framework similar to binary relevance but based on credal classifiers.

The paper is organized as follows. We review some basics about Bayesian networks and the IDM in Sect. 2. We indeed show how the IDM applies to Bayesian networks in Sect. 3. The (single-label) classical naive credal classifier is reviewed in Sect. 4. The new model we present for multilabel data is described in Sect. 5.1. Classification with this model is addressed in Sect. 5.2 and the technical theorems behind the inference algorithms are in Sect. 5.3. Simulations and conclusions are in Sects. 6 and 7, while the proofs of the technical results are in the Appendix.

2 Preliminaries

We denote random variables by uppercase letters, generic values by lowercase letters and the sets of possible values by calligraphic letters. For instance X is a variable whose generic value is $x \in \mathcal{X}$. For a Boolean variable X, $\mathcal{X} := \{0, 1\}$; given a generic value $x \in \mathcal{X}$, its negation is $\neg x$.

We denote by $P(X)$ the probability mass function over X. Given a set of variables $\boldsymbol{X}$, arranged into a directed acyclic graph, a *Bayesian network* is a set of conditional tables $P(X_i|\mathrm{Pa}(X_i))$ where $\mathrm{Pa}(X_i)$ are the parents of X_i, i.e., the immediate predecessors of X_i within the graph. This defines a joint mass function $P(\boldsymbol{x}) = \prod_i P(x_i|\mathrm{pa}(X_i))$ [15].

A credal set over X is a (convex) set of probability mass functions over X. Given a credal set, the *maximality* criterion allows to choice the optimal (i.e., most probable) states as follows: $x'' \in \mathcal{X}$ is *maximal* if and only if there is no $x' \in \mathcal{X}$ s.t. $P(x') > P(x'')$ for each $P(X)$ in the credal set [19].

The *imprecise Dirichlet model* [20] (IDM) is a standard approach to learn credal sets from multinomial data. Given a variable X, a Dirichlet prior $P(\theta_x) \propto \theta^{st(x)-1}$ would induce a probability $\theta_x = \frac{n(x)+st(x)}{N+s}$. Thus, considering all the priors s.t. $\sum_x t(x) = 1$, would make θ_x to vary between $\frac{n(x)}{N+s}$ and $\frac{n(x)+s}{N+s}$.

3 IDM-Based Learning with Independence

In this section we discuss the particular problem of learning a set of multivariate distributions through the IDM under specific independence assumption. This is done in the special case where the independence relations can be described within the framework of Bayesian networks. We extend Zaffalon's ideas stated in [23].

To begin the discussion let us consider the following example.

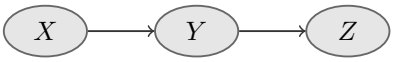


Figure 1: A chain topology.

Example 1. *Consider a Bayesian network over three Boolean variables X, Y, and Z with the topology in Fig. 1. This models the conditional independence between X and Z given Y, with the joint distribution factorizing as $P(x, y, z) = P(x) \cdot P(y|x) \cdot P(z|y)$. The likelihood of a set of observations $\mathcal{D}$ is:*

$$L(\boldsymbol{\theta}) := P(\mathcal{D}|\boldsymbol{\theta}) = \prod_x \theta_x^{n(x)} \left[\prod_y \theta_{y|x}^{n(x,y)} \left[\prod_z \theta_{z|y}^{n(y,z)} \right] \right], \tag{1}$$

where $\theta_x := P(x)$, $\theta_{y|x} := P(y|x)$, and $\theta_{z|y} := P(z|y)$, for each x, y, z, and $n(\cdot)$ is the counting function. A conjugate prior over the parameters $\boldsymbol{\theta}$ is:

$$P(\boldsymbol{\theta}) \propto \prod_x \theta_x^{st(x)-1} \left[\prod_y \theta_{y|x}^{st(x,y)-1} \left[\prod_z \theta_{z|y}^{st(y,z)-1} \right] \right], \tag{2}$$

where s and the $t(\cdot)$ are nonnegative parameters. The first term in Eq. (2) is proportional to a Dirichlet prior. We set $\sum_x t(x) = 1$. Considering the corresponding (structural) constraint for the counts in the likelihood, i.e., $\sum_x n(x) = N$, we can regard s as the equivalent sample size *(ESS) of this prior distribution.*

Let us identify the additional constraints required to regard s as an ESS even for the prior in Eq. (2). We just identify the (again, structural) constraints on the likelihood $\sum_{xy} n(x, y) = \sum_{yz} n(y, z) = N$, which correspond to:

$$\sum_{xy} t(x, y) = \sum_{yz} t(y, z) = 1. \tag{3}$$

The updated parameters become therefore:

$$\theta_x = \frac{n(x) + st(x)}{N + s}, \tag{4}$$

$$\theta_{y|x} = \frac{n(x, y) + st(x, y)}{n(x) + st(x)}, \tag{5}$$

$$\theta_{z|y} = \frac{n(y, z) + st(y, z)}{n(y) + st(y)}, \tag{6}$$

with $t(x) = \sum_y t(x, y)$ and $t(y) := \sum_z t(y, z)$.

An IDM-based model is therefore obtained by considering all *the specifications of the parameters in Eqs. (4-6) consistent with the above constraints over $t(x)$,*

$t(x,y)$, *and* $t(y,z)$*:*

$$\sum_x t(x) = 1 \tag{7}$$

$$\sum_y t(x,y) = t(x), \forall x \tag{8}$$

$$\sum_z t(y,z) = \sum_x t(x,y), \forall y. \tag{9}$$

Such a model can be regarded as induced by a set of priors made of Dirichlet components and with ESS s*. This is the way we generalize the IDM to multivariate models with independence. To check that the constraints are sufficient, consider all the (structural and not all independent) constraints satisfied by the count function* $n(\cdot)$ *in Eq.* (1)*, i.e.,* $\sum_x n(x) = \sum_{xy} n(x,y) = \sum_{yz} n(y,z) = N$*,* $\sum_y n(x,y) = n(x)$*,* $\sum_z n(y,z) = n(y)$*,* $\sum_x n(x,y) = n(y)$*. It is a trivial exercise to check that the* $t(\cdot)$ *parameters satisfy the analogous relations (with one replacing* N*).*

The example deals with a node which is a child of a child of another variable. This situation does not appear in Zaffalon's original work for the naive topology, neither in other papers about more connected topologies [24].

This approach can be easily extended to general Bayesian networks. The specifications over X apply to parentless nodes with Y replaced by the whole children set, the specifications over Z apply to any childless node with Y replaced by the whole parents set, and those for Y apply to any non-root non-leaf node with the parents and children playing the role of X and Z.

This section provides guidelines for learning the parameters of Bayesian networks based on the IDM. The resulting model is a *credal network* [9], with the local parameters taking their values from different credal sets, but with the constraints over the parameters of the prior inducing a *non-separate* specification [2].

4 The Naive Credal Classifier

In this section we briefly review the credal version of the naive Bayes classifier as proposed by Zaffalon in [23]. We denote the class variable as C and the feature variables as $\boldsymbol{F} := (F_1, \ldots, F_m)$. A dataset of N complete i.i.d. joint observations of $(C, \boldsymbol{F})$ is available together with a counting function $n(\cdot)$.

The features are assumed to be conditionally independent given the class. This corresponds to the topology in Fig. 2 and induces the factorization $P(c, \boldsymbol{f}) = P(c) \cdot \prod_{i=1}^m P(f_i|c)$, for each $c \in \mathcal{C}$ and $\boldsymbol{f} := (f_1, \ldots, f_m) \in \prod_{i=1}^m \mathcal{F}_i$.

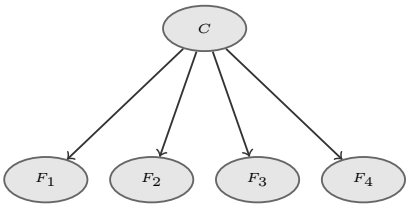


Figure 2: An example of naive topology.

By proceeding as in Ex. 1, we have:

$$P(c) = \frac{n(c) + st(c)}{N + s}, \tag{10}$$

$$P(f_i|c) = \frac{n(c, f_i) + st(c, f_i)}{n(c) + st(c)}, \tag{11}$$

for each $f_i \in \mathcal{F}_i$, $c \in \mathcal{C}$, $i = 1, \ldots, m$. The class labels assigned to an unannotated instance $\boldsymbol{f}$ of the features are those s.t. $\arg\max_{c \in \mathcal{C}} P(c, \boldsymbol{f})$.

The IDM constraints on the above positive parameters are: $\sum_c t(c) = 1$ and $\sum_{f_i} t(c, f_i) = t(c)$, for each $i = 1, \ldots, m$ and $c \in \mathcal{C}$.[1] We denote as $\boldsymbol{t}$ a generic value for the joint variable of these parameters and by $\mathcal{T}$ the corresponding feasible region.

The class labels assigned to $\boldsymbol{f}$ by this credal classifier are the *undominated* ones according to the maximality criterion. Given $c', c'' \in \mathcal{C}$, c' dominates c'' if $P(c', \boldsymbol{f}) > P(c'', \boldsymbol{f})$ for any specification consistent with the IDM constraints. This is equivalent to check:

$$\inf_{\boldsymbol{t} \in \mathcal{T}} \left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{m-1} \prod_{i=1}^m \frac{n(c', f_i) + st(c', f_i)}{n(c'', f_i) + st(c'', f_i)} > 1. \tag{12}$$

The optimization of the second term can be achieved independently. The objective function rewrites as:

$$\left[\frac{n(c'') + st(c'')}{n(c') + st(c')}\right]^{m-1} \prod_{i=1}^m \frac{n(c', f_i)}{n(c'', f_i) + st(c'')}, \tag{13}$$

with the constraints being simply now $t(c') + t(c'') = 1$, with $t(c'), t(c'') > 0$. In other words, we can express the objective function as a function of a single variable. Its logarithmic derivative is a linear fractional variable, and the second derivative is always positive. Overall the minimization can be efficiently achieved by bracketing (see [23] for the details).

5 The Multilabel Credal Classifier

5.1 Model Specification

In this section we extend the setup of the previous section to multilabel classification. The class variable C is

[1] The strict positivity is required because otherwise the corresponding prior would be improper.

replaced by q (Boolean) class labels $\boldsymbol{C} := (C_1, \ldots, C_q)$, where q is the cardinality of $\mathcal{C}$. This is standard way to cope with non-exclusivity: if the j-th label of $\mathcal{C}$ is active $C_j = 1$, otherwise $C_j = 0$.

We call C_1 the *superclass*, and the other class labels *subclasses*. We assume the conditional independence of the subclasses given the superclass. Simplistically we set as superclass the class which is more frequently observed as active. The dependencies between classes can be learned in more sophisticated way, optimizing for instance the Bayesian scores [7] of the graph which connects the classes.

A dataset of N joint observations of $(\boldsymbol{C}, \boldsymbol{F})$ is available together with a counting function $n(\cdot)$.

Each feature is *replicated* q times. For each $k = 1, \ldots, m$, $\{F_k^j\}_{j=1}^q$ are replicas of F_k. For each $j = 1, \ldots, q$, the replicated features $\{F_k^j\}_{k=1}^m$ are assumed to be independent given C_j. This is a simplifying assumption, already formulated in other papers [3]. Strictly speaking, an additional dummy child modeling the fact that all the replicas corresponds to the same variable should have been added.

Accordingly, the joint factorizes as follows:

$$P(\boldsymbol{c}, \boldsymbol{f}) = P(c_1) \left[\prod_{i=2}^{q} P(c_i|c_1) \right] \prod_{j=1}^{q} \prod_{k=1}^{m} P(f_k^j|c_j), \tag{14}$$

where the values of the class labels and of the features are those consistent with $\boldsymbol{c}$ and $\boldsymbol{f}$. Parameters in Eq. (14) can be learned from the data through a procedure similar to that in the previous sections, i.e.,

$$P(c_1) = \frac{n(c_1) + st(c_1)}{n + s}, \tag{15}$$

$$P(c_i|c_1) = \frac{n(c_1, c_i) + st(c_1, c_i)}{n(c_1) + st(c_1)}, \tag{16}$$

$$P(f_k^j|c_j) = \frac{n(c_j, f_k) + st(c_j, f_k^j)}{n(c_j) + st(c_j)}. \tag{17}$$

An IDM-like version is obtained by considering all the models consistent with the following constraints:[2]

$$\sum_{c_1} t(c_1) = 1, \tag{18}$$

$$\sum_{c_i} t(c_1, c_i) = t(c_1), \forall c_i \tag{19}$$

$$\sum_{f_k^j} t(c_j, f_k^j) = \sum_{c_1} t(c_1, c_j) = t(c_j), \forall c_j, \tag{20}$$

[2]Here and in the following, if there is no risk of ambiguity, the arguments of the sums and the products are omitted for sake of notation. E.g., $\sum_{c_1}$ is a shortcut for $\sum_{c_1 \in \mathcal{C}_1}$.

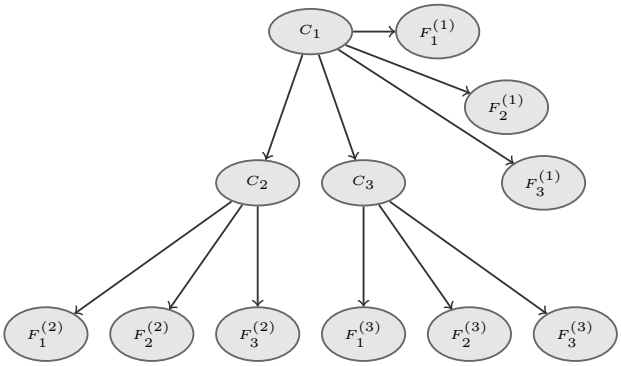


Figure 3: The multilabel naive topology.

together with the strict positivity of all the parameters. Even in this case we denote by $\boldsymbol{t}$ the generic value of the joint variable including all these parameters and by $\boldsymbol{\mathcal{T}}$ the corresponding feasible region. The imprecision in this model can be regarded as induced by s missing observations, which we are completely ignorant about.

5.2 Maximal Sequences and Maximal Labels

Consider a complete observation $\boldsymbol{f}$ of the features and two sequences of labels $\boldsymbol{c}'$ and $\boldsymbol{c}''$. According to maximality, the second sequence is undominated by the first if and only if there is (at least) a prior consistent with the constraints s.t. the first sequence is less (or equally) probable than the second, i.e.,[3]

$$\inf_{\boldsymbol{t} \in \boldsymbol{\mathcal{T}}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \leq 1. \tag{21}$$

In Section 5.3 we discuss how to ascertain whether sequence $\boldsymbol{c}'$ dominates $\boldsymbol{c}''$, in linear time with respect to the number of classes and features.

A more complex problem is to ascertain whether sequence $\boldsymbol{c}''$ is optimal. This happens if the condition (21) is satisfied for each possible specifications of $\boldsymbol{c}'$, i.e.,

$$\max_{\boldsymbol{c}'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \leq 1. \tag{22}$$

To detect the non-dominated sequences it is in principle necessary to compare each possible sequence $\boldsymbol{c}'$ against each possible alternative sequence $\boldsymbol{c}''$. This implies running $2^q \cdot 2^q = 2^{2q}$ tests of the same type as Eq. (21). In Section 5.3 we present a more efficient procedure, which detects the maximal sequences by running the test of Eq. (22) only once for each candidate sequence $\boldsymbol{c}''$ (i.e., 2^q times), with a substantial computational saving. We call this model the *joint* model, as it makes inference on the joint probability

[3]This is an alternative formulation w.r.t. that in Eq. (12).

of the labels. Yet the complexity of the joint is exponential in the number of labels; thus the identification of the optimal sequences is feasible only of the number of classes is limited, for instance $q < 10$.

We thus devise a different approach in order to deal with datasets containing many labels. It looks for the maximal states of *each label* rather than for the maximal sequences. We call this approach the *marginal* model. The marginal inference has polynomial complexity (see Section 5.3); it is however less informative than the detection of the maximal sequences. Consider having detected k labels whose maximal states are both *relevant* and *non-relevant*. The 2^k sequences obtained combining their states in all possible ways contain the maximal sequences and others non-maximal sequences. It is not possible to know which of the 2^k sequences is maximal and which is non-maximal.

This approach corresponds to the following optimization task:

$$\min_{\boldsymbol{c}'':c''_l=1} \max_{\boldsymbol{c}'} \inf_{\boldsymbol{t}} \frac{P_{\boldsymbol{t}}(\boldsymbol{c}', \boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'', \boldsymbol{f})} \leq 1, \tag{23}$$

for each $l = 1, \ldots, q$, with the minimum over all the specifications of the second sequence s.t. $c''_l = 1$. If the inequality is satisfied, then there is at least an optimal sequence whose l-th label is active. By replacing $c''_l = 1$ with $c''_l = 0$, we can decide if there is an optimal sequence with the l-th label inactive.[4]

By iterating the test in Eq. (23) and its analogous with $c_l = 0$ for each $l = 1, \ldots, q$, we can decide, for each label, which one of the following three options applies: (i) all the maximal sequences have that label active; or (ii) all the maximal sequences have the label inactive; or (iii) there are maximal sequences with the label active and others with the label inactive.

We call this approach based on the joint model in Eq. (14) and the IDM constraints in Eqs. (18-20) *multilabel naive credal classifier* (MNCC). The derivation uses ideas analogous to those proposed by De Bock and de Cooman to detect the maximal sequences in hidden Markov models [11].

5.3 Solving the Optimization

In this section we present the technical results behind our implementation of the MNCC and a possible direction for its development. Let us start from the maximality-based dominance test among two sequences, which can be performed as follows.

[4]By removing the constraints $c''_l = 1$ from Eq. (23) we test whether there is a maximal sequence. But this is true by definition. Thus, if the inequality in Eq. (23) is not satisfied for $c''_l = 1$, then it should be satisfied for $c''_l = 0$, and vice versa.

Theorem 1. *Given two sequences $\boldsymbol{c}'$ and $\boldsymbol{c}''$ and an instance of the features $\boldsymbol{f}$, the decision task in Eq. (21) is equivalent to:*

$$\prod_{i:c'_i=\neg c''_i} \frac{n(c'_1, c'_i) \cdot g_i(c'_i, c''_i, \boldsymbol{f})}{n(c''_1, c''_i) + s} \leq 1, \tag{24}$$

if $c'_1 = c''_1$, and to

$$\inf_{0<t_1<1} h(c'_1, c''_1, t_1, \boldsymbol{f}) \prod_i \frac{n(c'_1, c'_i)\tilde{g}_i(c'_i, c''_i, \boldsymbol{f})}{n(c''_1, c''_i) + st_1}, \tag{25}$$

if $c'_1 = \neg c''_1$, where

$$g_i(c'_i, c''_i, \boldsymbol{f}) := \inf_{0<t_i<1} \prod_k \frac{\frac{n(c'_i, f_k)}{n(c'_i)+s(1-t_i)}}{\frac{n(c''_i, f_k)+st_i}{n(c''_i)+st_i}}, \tag{26}$$

$\tilde{g}_i(c'_i, c''_i, \boldsymbol{f}) := g_i(c'_i, c''_i, \boldsymbol{f})$ if $c'_i = \neg c''_i$ and one otherwise, and $h(c'_1, c''_1, t_1, \boldsymbol{f})$ is defined as

$$\left[\frac{n(c''_1) + st_1}{n(c'_1) + s(1-t_1)}\right]^{q+m-2} \prod_k \frac{n(c'_1, f_k)}{n(c''_1, f_k) + st_1}. \tag{27}$$

Furthermore, the objective functions in Eq. (25) and Eq. (26) are convex.

The proof of this theorem is in the Appendix.

Th. 1 can be used to decide whether or not $\boldsymbol{c}'$ does not dominate $\boldsymbol{c}''$. Because of the convexity results, the optima in Eq. (25) and Eq. (26) can be evaluated by bracketing (e.g., bisection) in constant time (assuming that we work with finite precision). Thus, the dominance test only takes $O(qf)$ time.

To detect the set of maximal sequences, the test should be iterated over all the possible pairs. Alternatively, we can adopt the approach in Eq. (22), i.e., maximizing w.r.t. $\boldsymbol{c}'$. If we add the constraint $c'_1 = c''_1$, the maximization becomes trivial because of the factorization in Eq. (24). If $\tilde{\boldsymbol{c}}'$ is the value leading to the maximum, we have $\tilde{c}'_1 = c''_1$ and, for $i > 1$,

$$\tilde{c}'_i := \begin{cases} \neg c''_i & \text{if } \frac{n(c''_1, \neg c''_i) g_i(\neg c''_i, c''_i, \boldsymbol{f})}{n(c''_1, c''_i)+s} > 1, \\ c''_i & \text{otherwise.} \end{cases} \tag{28}$$

Thus, we perform the dominance test as in Th. 1 with $\tilde{\boldsymbol{c}}'$ and $\boldsymbol{c}''$. We similarly proceed for $c'_1 = \neg c''_1$ by considering Eq. (25) instead of Eq. (24). If t^*_1 is the value leading to the infimum, the task rewrites as:

$$\max_{c'_2,\ldots,c'_q} \left[h(\neg c''_1, c''_1, t^*_1, \boldsymbol{f}) \prod_i \frac{n(\neg c''_1, c'_i)\tilde{g}_i(c'_i, c''_i, \boldsymbol{f})}{n(c''_1, c''_i) + st^*_1} \right]. \tag{29}$$

The value of t^*_1 depends on $\boldsymbol{c}'$ and the maximization cannot be distributed over the product as in the previous case. Nevertheless, for the i-th term of the product,

a maximization w.r.t. $c'_i \in \{\neg c''_i, c''_i\}$ would be:

$$\max\left\{\frac{n(\neg c''_1,\neg c''_i)g_i(\neg c''_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+st_1^*},\frac{n(\neg c''_1,c''_i)}{n(c''_1,c''_i)+st_1^{**}}\right\}, \tag{30}$$

with the double star denoting the fact that the two optima w.r.t. t_1 can be different. Sufficient conditions for one of these two terms being the maximum irrespectively of the values of t_1^* and t_1^{**} can be used to determine $\tilde{\boldsymbol{c}}'$ as in the previous case, i.e.,

$$\tilde{c}'_i := \begin{cases} \neg c''_i & \text{if } \frac{n(\neg c''_1,\neg c''_i)g_i(\neg c''_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)+s} > \frac{n(\neg c''_1,c''_i)}{n(c''_1,c''_i)}, \\ c''_i & \text{if } \frac{n(\neg c''_1,\neg c''_i)g_i(\neg c''_i,c''_i,\boldsymbol{f})}{n(c''_1,c''_i)} < \frac{n(\neg c''_1,c''_i)}{n(c''_1,c''_i)+s}. \end{cases} \tag{31}$$

Yet, unlike the specification in Eq. (28), it might be that none of the two inequalities in Eq. (31) are satisfied, and the corresponding value of $\tilde{c}'_i$ remains undefined. If this is the case, we heuristically set the value of $\tilde{c}'_i$ corresponding to the limit of Eq. (31) for small values of $s > 0$.[5]

The above approach, whose complexity is the same as a single dominance test, i.e., $O(qf)$, can be used to decide whether or not a sequence $\boldsymbol{c}''$ is maximal. This is the case if the test in Th. 1 is satisfied for both the specifications of $\boldsymbol{c}'$ in Eq. (28) and Eq. (31).

To obtain the whole set of optimal sequences, we iterate this procedure over all the 2^q possible specifications of $\boldsymbol{c}''$. To avoid this exponential blow-up, the approach in Eq. (23), i.e., minimizing w.r.t. $\boldsymbol{c}''$ with a fixed value for c''_l, can be considered instead. In practice this corresponds to minimize the maximum between the above considered expressions for $c'_1 = c''_1$ and $c'_1 = \neg c''_1$. Although each one of the two expressions factorizes, moving the minimum w.r.t. the different factors inside the two arguments of the maximum might introduce an approximation, i.e.,

$$\min_{c''_1}\min_{c''_2,\ldots,c''_q}\max_{c'_1}\max_{c'_2,\ldots,c'_2}\inf_{\boldsymbol{t}}\frac{P_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'',\boldsymbol{f})} \geq \min_{c''_1}\max_{c'_1}\min_{c''_2,\ldots,c''_q}\max_{c'_2,\ldots,c'_2}\inf_{\boldsymbol{t}}\frac{P_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'',\boldsymbol{f})}, \tag{32}$$

where the constraint $c''_l = 1$ on both sides is left implicit for sake of readability. The above inequality trivially follows from the technical result here below.

Lemma 1. *Given two arrays $\vec{a}$ and $\vec{b}$ with the same length n, the following inequality holds:*

$$\min_i \max\{a_i, b_i\} \geq \max\{\min_i a_i, \min_i b_i\} \tag{33}$$

where a_i and b_i are the i-th elements of $\vec{a}$ and $\vec{b}$, and the minima are intended w.r.t. $i = 1,\ldots,n$.

[5] If $n(\neg c''_1,\neg c''_i)g_i(\neg c''_i,c''_i,\boldsymbol{f}) \neq n(\neg c''_1,c''_i)$, it is easy to check that the two inequalities cannot be simultaneously satisfied and, for sufficiently small s, one of them is always satisfied.

The proof of this lemma is in the Appendix. The right-hand side of Eq. (32) can be efficiently evaluated by reducing it to a single dominance test as we did in the first part of this section for the task in Eq. (22). If its value is (strictly) greater than one, Eq. (32) implies that also the left-hand side of Eq. (23) is greater than one, i.e., there is no maximal sequence with the l-th label active. If this is the case, we conclude that *all* the maximal sequences have the l-th label inactive. If the analogous optimization with the constraint $c''_l = 0$ instead of $c''_l = 1$ gives a result greater than one, we similarly conclude that all the maximal sequences have the l-th label active. Finally, if none of the above two is the case, we adopt a cautious approach by stating that there could either be maximal sequences with the l-th label active and inactive. The above approach can be considered to efficiently characterize the set of maximal sequences of the MNCC by means of an outer approximation.

6 Experiments

We compare the two variants of MNCC (joint model and marginal model) with the Bayesian graphical model, whose structure is as in Fig. 3. We adopt the BDeu prior [15, Chap.17] to learn the Bayesian model. This model is referred to in the following as the Bayesian model.

We consider four benchmark datasets, whose characteristics are reported in Tab. 1. *Emotions*, *Scene*, and *Slashdot* are classical benchmark datasets for multilabel classifiers. The *E-mobility* dataset is taken from a mobility study. It tracks which means of transport (car, train, bus, etc.) are used by a person for a given trip. The features are constituted by the length and duration of the trip, hour and day of the week, number of persons, reason of the trip, etc. [6].

Data set	Classes	Features	Instances
Emotions	6	44/72	593
Scene	6	224/294	2407
E-mobility	10	14/18	4226
Slashdot	22	496/1079	3782

Table 1: Benchmark datasets.

We validate the classifiers by a ten-folds cross-validation. Before training any classifier, we perform two pre-processing steps. First, we discretize numerical features into four bins. Then we perform feature selection as follows. We adopt the correlation-based feature selection (CFS) [21, Chap. 7.1], often used in traditional classification. We perform CFS q times, once for each different label. Eventually, we retain the *union* of the features selected in the q runs. This is a

useful pre-processing step which reduces the number of features, removing the non-relevant ones. As an example, Tab. 1 displays the number of features after and before this selection procedure when applied to the benchmark datasets considered in this paper. Feature selection for multilabel classification is however an open problem, and more sophisticated approaches can be designed to this end.

We start by assessing the joint model. We measure the *exact match* of the Bayesian model, namely the proportion of times in which the whole sequence of classes has been correctly predicted. For the MNCC we measure the *# of sequences*, namely the number of maximal sequences; moreover we measure the *credal match*, namely the proportion of times in which the actual sequence belongs to the set of optimal sequences.

Dataset	Bayesian	Credal (MNCC)	
	Exact match	# of seqs	Credal match
Emotions	.27	9.4	.80
Scene	.29	7.6	.80

Table 2: Experimental results of the joint model.

The sequence predicted by the Bayesian model is always recognized as maximal. The credal joint model is more robust than its Bayesian counterpart: the credal match is about three times larger than the total accuracy of the Bayesian multilabel classifier (see Tab.2). The number of maximal sequences is reasonably limited, considering that the presence of 6 classes implies 64 possible sequences. The exact match of the Bayesian classifier drops sharply on the instances which have many maximal sequences. On the Scene dataset, the total accuracy is 0.23 and 0.40 on the instances which have respectively less and more than nine maximal sequences. A similar pattern is observed also on the Emotions dataset. These results are obtained through the joint model, which enumerates all the 2^q possible sequences and checks whether they are maximal as in Eq. (22). They show the interesting potential of the credal approach to multilabel classification. Yet, the joint model can only cope with small q.

The marginal model can deal with larger q and thus can be tested on more challenging datasets. We adopt the outer approximation corresponding to the dominance test in Eq. (23). Results of a ten-folds cross validations are in Figs. 4–6. We evaluate the marginal model label-wise. In particular we measure for each label the accuracy of Bayesian model when MNCC returns a determinate and an indeterminate prediction. We also report the *determinacy*, i.e. the proportion of instances on which MNCC is determinate. On Scene

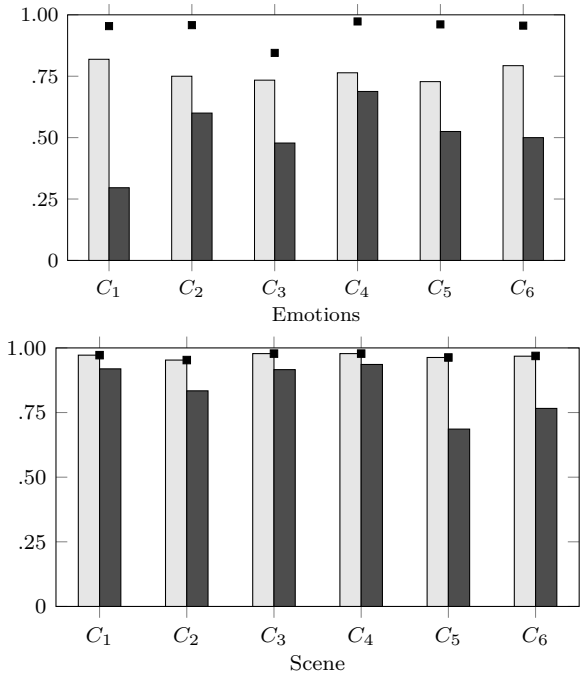


Figure 4: Accuracy of the Bayesian model on the instances on which the marginal MNCC model is determinate (light bars) and indeterminate (dark bars). The black squares denote the determinacy level. The results are presented label-wise.

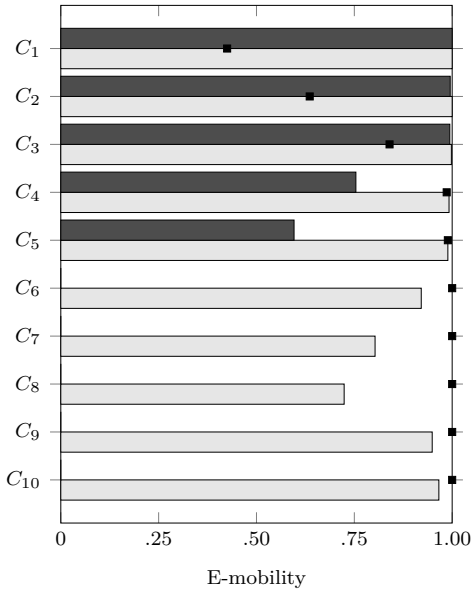


Figure 5: Accuracy of the Bayesian model on the E-mobility dataset. Light gray bars denote the accuracy when the marginal MNCC model is determinate. When determinacy (black squares) is one, the dark gray bar associated to the case when MNCC is indeterminate is not shown. The results are presented label-wise.

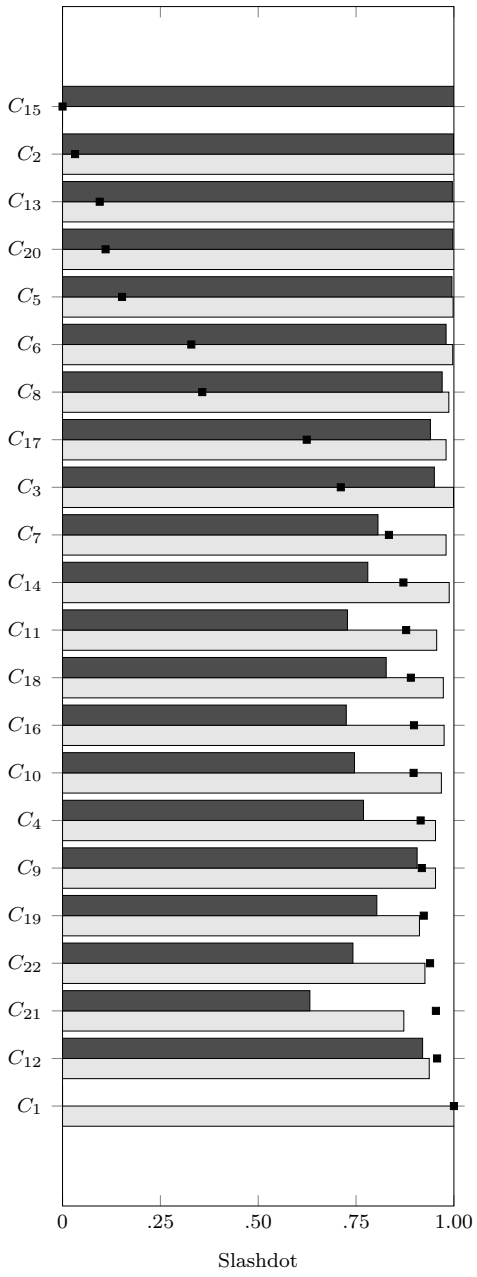


Figure 6: Accuracy of the Bayesian model on the Slashdot dataset. The dark gray bars denote the accuracy of the Bayesian model when the MNCC is indeterminate. If the determinacy (black squares) is zero, the light gray bar corresponding to the cases when the MNCC is determinate is undefined. Labels are sorted according to the determinacy level just for sake of readability. The results are presented label-wise.

and Emotions the accuracy of the Bayesian model sharply drops when the multilabel classifier becomes indeterminate. This confirms a well-known strength of credal classifiers compared to Bayesian classifiers [8]. This is generally confirmed also on E-mobility and Slashdot. However in these datasets there are also labels in which the Bayesian model is perfectly accurate when the credal model is indeterminate (see the first labels of both datasets). This suggests that the credal model is excessively indeterminate in some situations. This is a problem which is also known in traditional classification and which could be mitigated for instance by ϵ-contaminating the IDM with the uniform prior.

Future studies might inspect also further indicator of performance for multilabel classification, such as the F-metric. We focus on the exact match and on the label-wise accuracy as the inferences for this indicators are optimal. Optimal inferences for other indicators have still to be developed.

A Matlab software implementation of the MNCC is freely available at `http://ipg.idsia.ch/software`.

7 Conclusions

We have generalized the naive credal classifier to cope with multilabel data. The preliminary experiments are promising: the credal approach yields more robust predictions than the Bayesian approach. To scale to large number of labels it is necessary adopting the marginal model, whose inference is approximated.

As future work, it could be interesting to compare the inferences yielded by local and the global specification of the IDM (e.g., by exploiting some of the results in [10]). Moreover one could consider optimality criteria others than maximality (e.g., E-admissibility). A comparison with other methods possibly yielding multiple sequences (e.g., [16, 22]) could be also considered.

Acknowledgements

We thank Claudio Bonesana for support during the preparation of the datasets. We thank Jasper De Bock and Cassio de Campos for stimulating discussions about MAP tasks in credal networks. We also thank Denis Mauá for his suggestions about possible justification of models with the replicated features.

References

[1] A. Antonucci, G. Corani, D.D. Mauá, and S. Gabaglio. An ensemble of Bayesian networks for multilabel classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI-13)*, pages 1220–1225, 2013.

[2] A. Antonucci and M. Zaffalon. Decision-theoretic specification of credal networks: a unified language for uncertain modeling with sets of Bayesian networks. *International Journal of Approximate Reasoning*, 49(2):345–361, 2008.

[3] J. Arias, J. Gámez, T.D. Nielsen, and J.M. Puerta. A pairwise class interaction framework for multilabel classification. In L. van der Gaag and A. Feelders, editors, *PGM'14: Proceedings of the Seventh European Workshop on Probabilistic Graphical Models*, Lecture Notes in Artificial Intelligence, pages 17–32. Springer, 2014.

[4] J.M. Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2):123–150, 2005.

[5] C. Bielza, G. Li, and P. Larrañaga. Multidimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.

[6] F. Cellina, A. Förster, D. Rivola, L. Pampuri, R. Rudel, and A. Rizzoli. Using smartphones to profile mobility patterns in a living lab for the transition to e-mobility. In J. Hebiek, G. Schimak, M. Kubasek, and A. Rizzoli, editors, *Environmental Software Systems. Fostering Information Sharing*, volume 413 of *IFIP Advances in Information and Communication Technology*, pages 154–163. Springer, 2013.

[7] G. Corani, A. Antonucci, D. Mauá, and S. Gabaglio. Trading off Speed and Accuracy in Multilabel Classification. In L. van der Gaag and A. Feelders, editors, *PGM'14: Proceedings of the Seventh European Workshop on Probabilistic Graphical Models*, Lecture Notes in Artificial Intelligence, pages 145–159. Springer, 2014.

[8] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *The Journal of Machine Learning Research*, 9:581–621, 2008.

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

[10] J. De Bock, C.P. de Campos, and A. Antonucci. Global sensitivity analysis for MAP inference in graphical models. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

[11] J. De Bock and G. de Cooman. An efficient algorithm for estimating state sequences in imprecise hidden Markov models. *Journal of Artificial Intelligence Research*, 50:189–233, 2014.

[12] K. Dembczynski, W. Waegeman, and E. Hüllermeier. An analysis of chaining in multi-label classification. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pages 294–299, 2012.

[13] S. Destercke. Multilabel predictions with sets of probabilities: the Hamming and ranking loss cases. *Pattern Recognition*, 2015.

[14] D. Heider, R. Senge, W. Cheng, and E. Hüllermeier. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, 29(16):1946–1952, 2013.

[15] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[16] I. Pillai, G. Fumera, and F. Roli. Multi-label classification with a reject option. *Pattern Recognition*, 46(8):2256–2266, 2013.

[17] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[18] L.C. Van Der Gaag and P.R. De Waal. Multi-dimensional Bayesian network classifiers. In M. Studený and J. Vomlel, editors, *Proc. of the 3rd European Workshop on Probabilistic Graphical Models (PGM '06)*, pages 107–114. Action M, 2006.

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

[20] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society: Series B*, 58:3–34, 1996.

[21] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[22] Z. Younes, F. Abdallah, and T. Denoeux. Fuzzy multi-label learning under veristic variables. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE, 2010.

[23] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T.L. Fine, and T. Seidenfeld, editors, *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393, The Netherlands, 2001. Shaker.

[24] M. Zaffalon and E. Fagiuoli. Tree-based credal networks for classification. *Reliable Computing*, 9(6):487–509, 2003.

A Proofs

Proof of Theorem 1. *We consider the objective function in Eq.* (21) *by distinguishing whether or not the two sequences* $\boldsymbol{c}'$ *and* $\boldsymbol{c}''$ *share the first label, i.e.,*

$$\frac{P_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{f})}{P_{\boldsymbol{t}}(\boldsymbol{c}'',\boldsymbol{f})}=\begin{cases} G_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f}), & \text{if } c'_1=c''_1,\\ H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f}), & \text{if } c'_1=\neg c''_1. \end{cases} \tag{34}$$

Because of Eq. (14)*, function* $G_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$ *writes as:*

$$\prod_{i:c'_i=\neg c''_i}\left[\frac{n(c'_1,c'_i)+st(c'_1,c'_i)}{n(c''_1,c''_i)+st(c''_1,c''_i)}\prod_{k=1}^{m}\frac{\frac{n(c'_i,f_k)+st(c'_i,f_k)}{n(c'_i)+st(c'_i)}}{\frac{n(c''_i,f_k)+st(c''_i,f_k)}{n(c''_i)+st(c''_i)}}\right], \tag{35}$$

where the restriction in the outer product is possible because of the contribution of the other terms is one (remember that $c'_1=c''_1$*). A preliminary optimization w.r.t. the constraints can be achieved as in Sect. 4 by setting* $t(c'_i,f_k)\to 0$ *and* $t(c''_i,f_k)\to t(c''_i)$ *(remember that* $c'_i=\neg c''_i$*). Similarly,* $t(c'_1,c'_i)\to 0$ *and* $t(c''_1,c''_i)\to t(c''_1)$*. After these operations, the result rewrites as:*

$$\prod_{i}{}'\left[\frac{n(c'_1,c'_i)}{n(c''_1,c''_i)+st(c''_1)}\prod_{k}\frac{\frac{n(c'_i,f_k)}{n(c'_i)+st(c'_i)}}{\frac{n(c''_i,f_k)+st(c''_i)}{n(c''_i)+st(c''_i)}}\right], \tag{36}$$

where the prime in the product is a shortcut for the restriction. The optimization w.r.t. $t(c''_1)$ *is achieved in the limit* $t(c''_1)\to 1$*. Even the remaining optimization tasks can be achieved independently of the others. The result is the left-hand side of Eq.* (24)*, where, in Eq.* (26)*, we have set* $t_i:=t(c''_i)$*, and hence* $t(c'_i)=1-t_i$ *(remember that, for these terms,* $c'_i=\neg c''_i$*).*

We similarly proceed for $H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$*, i.e., because of Eq.* (34) *and Eq.* (14)*:*

$$\left[\frac{n(c''_1)+st(c''_1)}{n(c'_1)+st(c'_1)}\right]^{q+m-2}\prod_{k}\frac{n(c'_1,f_k)+st(c'_1,f_k)}{n(c''_1,f_k)+st(c''_1,f_k)}$$
$$\prod_{i}\frac{n(c'_1,c'_i)+st(c'_1,c'_i)}{n(c''_1,c''_i)+st(c''_1,c''_i)}\prod_{j}{}'\prod_{k}\frac{\frac{n(c'_j,f_k)+st(c'_j,f_k)}{n(c'_j)+st(c'_j)}}{\frac{n(c''_j,f_k)+st(c''_j,f_k)}{n(c''_j)+st(c''_j)}}. \tag{37}$$

As in the previous case, we perform some optimization, rename the remaining variables, and independently optimize w.r.t. t_i $(i>1)$*. Afterwards, we optimize w.r.t.* t_1 *and* $\inf_{\boldsymbol{t}} H_{\boldsymbol{t}}(\boldsymbol{c}',\boldsymbol{c}'',\boldsymbol{f})$ *becomes as in Eq.* (25)*.*

Finally, we prove that the objective functions in the right-hand side of Eq. (26) *and in Eq.* (25) *are convex. The derivative of the logarithm of the objective function in the right-hand side of Eq.* (26) *divided by the positive constant* s *is equal to:*

$$\frac{m}{n(c'_i)+s(1-t_i)}-\sum_{k}\frac{1}{n(c''_i,f_k)+st_i}+\frac{m}{n(c_i)''+st_i}. \tag{38}$$

The second derivative, again divided by s*, is:*

$$\frac{m}{[n(c'_i)+s(1-t_i)]^2}+\sum_{k}\frac{1}{[n(c''_i,f_k)+st_i]^2} \tag{39}$$
$$-\frac{m}{[n(c_i)''+st_i]^2}, \tag{40}$$

and its nonnegativity easily follows from $n(c''_i)\geq n(c''_i,f_k)$*. Similarly, the second derivative of the logarithm of the objective function in Eq.* (25) *is:*

$$-\frac{q+m-2}{[n(c''_1)+st_1]^2}+\frac{q+m-2}{[n(c'_1)+s(1-t_1)]^2}$$
$$+\sum_{k}\frac{1}{[n(c''_1,f_k)+st_1]^2}+\sum_{i}\frac{1}{[n(c''_1,c''_i)+st_1]^2} \tag{41}$$

As in the previous case, the nonnegativity follows from $n(c''_i)\geq n(c''_i,f_k)$*.* □

Proof of Lemma 1. *We prove the result by contradiction. Thus, we assume that:*

$$\min_i\max\{a_i,b_i\}<\max\{\min_i a_i,\min_i b_i\}. \tag{42}$$

Let i^* *denote the* $\arg\min$ *of the left-hand side. If, without any lack of generality, we assume* $\min_i a_i\geq\min_i b_i$*, Eq.* (42) *rewrites as:*

$$\max\{a_{i^*},b_{i^*}\}<\min_i a_i. \tag{43}$$

If $a_{i^*}>b_{i^*}$*, we obtain the contradiction* $a_{i^*}<\min_i a_i$*. Otherwise, we have:*

$$a_{i^*}\leq b_{i^*}<\min_i a_i \tag{44}$$

which is also a contradiction. □

Efficient L1-Based Probability Assessments Correction: Algorithms and Applications to Belief Merging and Revision

Marco Baioletti and **Andrea Capotorti**
Dip. Matematica e Informatica - Università degli Studi di Perugia
{marco.baioletti, andrea.capotorti}@unipg.it

Abstract

In this article we define a procedure which corrects an incoherent probability assessment on a finite domain by exploiting a geometric property of L1-distance (known also as Manhattan distance) and mixed integer programming. L1-distance minimization does not produce, in general, a unique solution but rather a corrected assessment that could result an imprecise probability model. We propose a correction method for the merging of two separate assessments whose direct juxtaposition could be incoherent, and for the revision of beliefs where the core of the assessment must remain unchanged. A prototypical example on antidoping analysis guides the reader through this article to explain the various procedures.

Keywords. coherence, mixed-integer optimization, probability merging and revision, imprecise probability.

1 Introduction

The problem of correcting probability evaluations, especially on finite settings, has a long history and has been largely debated. Considering the significant amount of research on this subject, we can just mention two main "streams": one is the "right way" of assessing probability values, whose roots can be found in [4, 17, 20] while the other is the so called "calibration question" that stems from the seminal paper [31] and subsequent developments [15, 16]. More recently these two streams have been joined and faced with a unifying view by de Finetti's notion of coherence ([18] and in particular [27, pag. 361]). Hence several approaches have been proposed to deal with "incoherent" probabilities, for both unconditional and conditional values and by adopting different notions of "distances" and "scoring rules" (among the many, refer, e.g., to [6, 7, 8, 9, 28, 30]).

The risk of dealing with incoherent probability assessments is significantly present when the numerical evaluation comes from different sources of information and/or structural constraints limit the possible states (see, e.g., [5, 11, 12, 24, 33]). In this paper we come back to the fore of this argument leaving aside the more probabilistic approaches based on scoring rules that have a forecasting perspectives, by adopting the more aseptic approach based on geometrical distance minimization. In particular we will deal with the simple and easily understood L_1-distance, known also as "Manhattan" or "taxi-cab" metric. The main reason for using such metric is because we are able to propose an effective procedure (presented in Sec. 3), which is based on integer linear programming and hence is much more efficient than the correction procedures needed for other distances, for instance the quadratic programming for L_2-distance. L_1-distance minimization has moreover a simple interpretation, since it implies a direct minimal modification of each single value, permitting to use it for different purposes like the merging between two separate assessments (described in Sec. 4) and the revision of beliefs (depicted in Sec. 5).

The peculiarity of using L_1 minimization is the non-uniqueness, in general, of the solution and this could represent an alternative way of legitimating the adoption of imprecise probability models, in addition to the historical ones as stemming from buying/selling prices or desirability of gambles [35], or from extensions of coherent precise initial assessments [13, Chap.15]. In this paper, we assume that the initial assessments are precise, but this assumption could be easily generalized to initial imprecise probability assessments. However, assuming initial assessments as being precise is reasonable as it is consistent with usual estimate techniques which tend to express precise values.

In order for this paper to be as self-contained as possible, the next Section 2 briefly introduces the notion of probability assessments and formalizes the problem of their coherence. As already stated, the subsequent

Section 3 contains a proposal of a new correction algorithm based on L_1-distance minimization via mixed integer programming and on properties of convex polytopes, while Sections 4 and 5 legitimate its usefulness. A short concluding Section 6 closes the contribution.

2 Probability Assessments

A probability assessment on a finite domain can be expressed through a quadruple $\pi = (V, U, p, \mathfrak{C})$, where $V = \{X_1, \ldots, X_n\}$ is a finite set of propositional variables, representing any potential event of interest, U is a subset of V that contains the effective events taken into consideration, $p : U \to [0, 1]$ is a function which assigns a probability value to each variable in U, and $\mathfrak{C}$ is a finite set of logical constraints which lie among all the variables in V.

Note that the explicit presence of the set of variables V, even if the numerical assessment is given on the subset U, permits to extend an initial assessment to a larger domain without redefine the whole model, allowing a dynamical analysis. In this paper we will use it only on the merging application of Sec.4, but it is a good practice to allow this distinction also in static descriptions.

Since the Boolean logical setting in which we embed the assessment, in the sequel we will adopt the usual logical notation, with $\neg$, $\wedge$ and $\vee$ denoting the negation, disjunction and conjunction connectives, respectively; $\Rightarrow$ the material implication; $=$ the logical equivalence; $\top$ and $\bot$ the universal tautology and contradiction (sure and impossible events), respectively.

Usually some possible forms of logical constraints are: $\phi = \psi$, $\phi \implies \psi$ and $\phi = \bot$, where ψ and ϕ are boolean expressions involving the variables of V. But without loss of generality, we suppose that $\mathfrak{C}$ is expressed in conjunctive normal form, i.e., each element of $\mathfrak{C}$ is a disjunction of literals formed with variables in V, i.e., each element can be written as disjunctive clause

$$\left(\bigvee_{h \in H} X_h\right) \vee \left(\bigvee_{l \in L} \neg X_l\right)$$

for some $H, L \subseteq \{1, \ldots, n\}$, so that $\mathfrak{C}$ results as their conjunction.

For example, the constraint $X_i \implies X_j$ is expressed in $\mathfrak{C}$ by the clause $\neg X_i \vee X_j$.

A truth-value assignment α is a function from V to $\{0, 1\}$. Given a proposition ϕ, we write $\alpha \models \phi$ when α satisfies ϕ, otherwise we write $\alpha \not\models \phi$.

There are different, but equivalent, ways to define the coherence, i.e., the "rationality", of an assessment π: from semantical, syntactical or operational point of views (see, e.g., [12, 13, 18, 27]). Here we adopt the pragmatic way already used in [1], where a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is coherent if there exists a probability distribution $\mu : 2^V \to [0, 1]$ on the set of all truth-value assignments 2^V which satisfies the following properties

1. for each $\alpha \in 2^V$, if there exists a constraint $c \in \mathfrak{C}$ such that $\alpha \not\models c$, then $\mu(\alpha) = 0$;
2. $\displaystyle\sum_{\alpha \in 2^V} \mu(\alpha) = 1$;
3. for each $X \in U$, $\displaystyle\sum_{\alpha \in 2^V, \alpha \models X} \mu(\alpha) = p(X)$.

The coherence of a probability assessment, called shortly CPA, has been already studied in [1, 2, 3, 32], albeit in a slightly different form, showing that checking if π is coherent is a NP-complete problem, even when the constraints in $\mathfrak{C}$ are binary (i.e., each of them involves only two variables).

The computational problem CPA is strictly related to the Probabilistic Satisfiability problem (PSAT [23]), where the probability assessment is defined on some finite set of propositions, instead that on the propositional variables. It can be proved that every instance of CPA can be easily translated as a PSAT instance, and that every PSAT can be formulated in a normal form, which is essentially a CPA instance [14].

There exist several algorithms to solve CPA and PSAT problems:

- A column-generation [23, 25] approach, where the problem is solved using linear programming techniques which exploit the sparsity of the solutions;
- CPA algorithm [1, 2], which is based on a symbolic manipulation which, in some cases, needs a further linear programming procedure;
- SAT-based approach [19], in which the problem is translated in a pure propositional satisfiability form (SAT);
- MIP-based approach [14], in which the problem is formulated as a mixed integer programming problem (MIP).

3 Correcting Probability Assessments

When a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is not coherent, then it is possible to "correct" it in different ways, in order to obtain a coherent probability assessment π' which is as close as possible to π, according

to a distance or a pseudo-distance function between probability assessments.

One possibility is to revise only the probability values, i.e., $\pi' = (V, U, p', \mathfrak{C})$, and to use a distance between probability assessments which is defined only in terms of p and p'.

Another possibility, which will not be taken into account in this paper, could be to revise (also) the logical constraints.

Since p and p' correspond to vectors of $\mathbb{R}^n$, where $n = |U|$, it is possible to use a distance d in $\mathbb{R}^n$. Then, chosen a distance d, a d-correction of a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is a vector p' such that the probability assessment $\pi' = (V, U, p', \mathfrak{C})$ is coherent and $d(p, p')$ is minimized. We denote $\mathcal{C}_d(\pi)$ the sets of all the d-correction of π.

Clearly if π is coherent, then $\mathcal{C}_d(\pi) = \{p\}$, for any distance d of $\mathbb{R}^n$.

In general, given a probability assessment π, $\mathcal{C}_d(\pi)$ could have more than one element and in this case the operation of correcting a probability assessment leads to an imprecise probability model, the so called "credal set".

As already stated in te Introduction, several distance choice are possible. Among the many, in this paper we focus on the L_1 distance defined as

$$d_1(p, p') = \sum_{i=1}^{n} |p(X_i) - p'(X_i)|$$

and we denote $\mathcal{C}_{d_1}(\pi)$ as $\mathcal{C}(\pi)$. Whether this could be the best distance and how it performs with respect to the others is not directly considered. Rather its use as a tool is considered as it is reasonable and easily interpretable by users so that technical aspects connected with its adoption are addressed. Our interest in L_1 distance is that with its adoption translating the optimization problem into a linear problem by using both integer and real variables is possible. This last represents a computational advantage compared to other distances that imply implementation of non linear (quadratic, logarithmic, etc.) optimizations tools.

The resulting mixed integer program $\mathcal{P}1$ is built similarly to the method described in [14]. Let us suppose that $U = \{X_1, \ldots, X_n\}$. Moreover let $m = |\mathfrak{C}|$.

The real variables of $\mathcal{P}1$ are

- b_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$.
- q_j, for $j = 1, \ldots, n+1$
- r_i, s_i, for $i = 1, \ldots, n$

all of them are non-negative (as usual in linear programming).

The program $\mathcal{P}1$ also has the integer variables

- a_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$

which are constrained to 0 or 1.

The constraints of $\mathcal{P}1$ are

1. for each $i = 1, \ldots, n$,
$$\sum_{j=1}^{n+1} b_{ij} = p(X_i) + (r_i - s_i)$$
2. for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$,
$$0 \leq b_{ij} \leq a_{ij}, \quad a_{ij} - 1 + q_j \leq b_{ij} \leq q_j$$
3. $$\sum_{i=1}^{n+1} q_j = 1$$
4. for $i = 1, \ldots, n$,
$$r_i \leq 1, \quad s_i \leq 1$$

Moreover, for each clause c_i (for $i = 1, \ldots, m$), where $c_i = \bigvee_{h \in H_i} X_h \vee \bigvee_{l \in L_i} \neg X_l$, and for each $j = 1, \ldots, n+1$ the following linear constraint is added

$$\sum_{h \in H_i} a_{h,j} + \sum_{l \in L_i} (1 - a_{l,j}) \geq 1$$

Finally, the objective function to be minimized is

$$\sum_{i=1}^{n} (r_i + s_i)$$

Solving the linear program $\mathcal{P}1$ is equivalent to correcting the probability p, because every correction p' of p is a coherent probability assessment, hence it can be written as a convex combination of at most $n+1$ atoms, i.e., truth assignments which satisfy the logical constraint $\mathfrak{C}$ (for more details refer, e.g., to [25]). The binary variables a_{ij} are a representation of these atoms, because of the constraint 4., while the real variables q_j are the coefficients of the convex combination. The role of constraint number 2. is to set $b_{ij} = a_{ij} \cdot q_j$ (for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$), without using the multiplication, otherwise $\mathcal{P}1$ would not be a linear problem. The variables r_i, s_i are slack variables, which represent, respectively, the positive and the negative difference between $p(X_i)$ and $p'(X_i)$, as implied by

the constraint 1. Finally, the objective function corresponds to minimize the L_1-distance between p and p', i.e., $\sum_{i=1}^{n} |p(X_i) - p'(X_i)|$.

From a theoretical point of view, to find the correction is a computational hard problem. Indeed given a probability assessment $\pi = (V, U, p, \mathfrak{C})$ and a real non-negative number D, it is a NP-complete problem to check if there exists a coherent probability assessment $\pi' = (V, U, p', \mathfrak{C})$ such that $d_1(p, p') \leq D$. The proof of NP-containment is easy because any solution of $\mathcal{P}1$ provides a succinct certificate for the existence of π' i.e., the values of a_{ij}, r_i and s_i's. While the NP–hardness derives from the fact that the coherence of π (which is a NP-complete problem) can be tested by posing $D = 0$.

Anyway, the actual implementations of MIP solvers make possible to solve probability correction problems of reasonable size in a feasible amount of time.

The optimal value δ for the objective function corresponds to the minimum possible correction on p and any coherent probability assessment $\pi' = (V, U, p', \mathfrak{C})$ such that $d_1(p, p') = \delta$ is a possible solution i.e., p' is an element of $\mathcal{C}(\pi)$.

In many situations $\mathcal{C}(\pi)$ has more than one element and the MIP problem is able to find just one solution, which could not be a good representative of all the elements of $\mathcal{C}(\pi)$, as it happens when it is an extreme value. Hence the following procedure to generate all the elements of $\mathcal{C}(\pi)$ is proposed.

Let $\mathcal{Q}$ be the set of all vectors $q \in \mathbb{R}^n$ such that the probability assessments $(V, U, q, \mathfrak{C})$ are coherent. $\mathcal{Q}$ forms a convex polytope whose extremal points are exactly the atoms, i.e., all truth-value assignments α which satisfy the logical constraints $\mathfrak{C}$.

Let $\mathcal{B}_\pi(\delta)$ be the ball of all vectors $q \in \mathbb{R}^n$ such that $d(p, q) \leq \delta$, with p the numerical probability assessment present in π. Such ball $\mathcal{B}_\pi(\delta)$ is a convex set whose extremal points are the points $p \pm \delta e_i$, where e_i is the i-th vector of the canonical basis, for $i = 1, \ldots, n$.

Then $\mathcal{C}(\pi)$ is a convex set of $\mathbb{R}^n$, because it is the intersection (see [29]) between the convex sets $\mathcal{Q}$ and $\mathcal{B}_\pi(\delta)$.

It is possible to describe $\mathcal{C}(\pi)$ in terms of its extremal points $q_1, \ldots, q_s$, indeed any element of $q \in \mathcal{C}(\pi)$ can be expressed as

$$q = \sum_{i=1}^{s} \lambda_i q_i$$

for some coefficients $\lambda_1, \ldots, \lambda_s \in \mathbb{R}$, such that $0 \leq \lambda_i \leq 1$, for $i = 1, \ldots, s$, and $\sum_{i=1}^{s} \lambda_i = 1$.

As a starting point, let us find a particular element $\bar{p} \in \mathcal{C}(\pi)$, which has the property that

$$\max_{i=1,\ldots,n} |\bar{p}(X_i) - p(X_i)| \tag{1}$$

is minimum, among all the coherent assessments such that

$$d_1(\bar{p}, p) = \delta. \tag{2}$$

This optimization problem can be formulated as a MIP problem $\mathcal{P}2$. All the constraints and the variables of $\mathcal{P}1$ are reported in $\mathcal{P}2$. Moreover, $\mathcal{P}2$ contains a new real variable z, which is subject to the constraints $r_i + s_i \leq z$, for $i = 1, \ldots, n$ (hence $z \geq \max_{i=1,\ldots,n} (r_i + s_i)$), and the new additional constraint $\sum_{i=1}^{n} (r_i + s_i) = \delta$ (which represents the equality (2)). In this way, the $\mathcal{P}2$ objective function to be minimized is simply z, since it equates (1).

The corrected assessment $\bar{\pi} = (V, U, \bar{p}, \mathfrak{C})$ differs from π by δ and tries to spread this difference as much as possible among the variables of U. Moreover, $\bar{p}$ is, in some sense, the most "central" point of $\mathcal{C}(\pi)$.

Using $\bar{p}$, it is possible to find the face F_1 of the polytope $\mathcal{Q}$ where $\mathcal{C}(\pi)$ lies. The face F_1 is itself a convex set with at most $n + 1$ atoms as extremal points, which can be found as a part of the solutions of $\mathcal{P}2$ (i.e., the optimal values of a_{ij}).

By looking at the signs of $\bar{p}(X_i) - p(X_i)$, for $i = 1, \ldots, n$, it is also possible to determine the face F_2 of $\mathcal{B}_\pi(\delta)$ which contains $\mathcal{C}(\pi)$. Indeed, F_2 is a convex set with at most n extremal points of the form

$$p + sign(\bar{p}(X_j) - p(X_j)) \cdot \delta \cdot e_j. \tag{3}$$

The extremal points $Q = \{q_1, \ldots, q_s\}$ of $\mathcal{C}(\pi)$ can be easily found by means of the following procedure.

- let E_1 be the extremal points of F_1 and E_2 be the extremal points of F_2
- compute H_1 as the H-representation of F_1
- compute H_2 as the H-representation of F_2
- let $H = H_1 \cup H_2$, the H-representation of $F_1 \cap F_2 = \mathcal{C}(\pi)$
- compute Q as the V-representation of H

where the V-representation of a convex set C is the set of its extremal points, while the H-representation of C is a set H of half-spaces such that $C = \bigcap_{h \in H} h$. It is possible to convert from the V-representation of C to its H-representation by means of a face enumeration algorithm, while the inverse conversion is performed by a vertex enumeration algorithm [21].

Both steps can be computed in polynomial time as shown, for instance, in [21].

Let us summarize the whole process with the following pseudo-code where FaceEnum and VertexEnum are suitable procedures to compute the H and V representations.

procedure Correct
Input: assessment $(V, U, p, \mathfrak{C})$
Output: extr. points W and min. distance δ
begin
prepare MIP program $\mathcal{P}1$
solve it and extract the optimal value δ
if $\delta = 0$ **then**
return $(\{p\}, 0)$
else
prepare MIP program $\mathcal{P}2$
solve it
extract the values a_{ij}, r_i, s_i
$E1$:=columns of matrix a_{ij}
compute $\bar{p}$ from r_i, s_i
compute $E2$ with formula 3
$H1$:= FaceEnum($E1$)
$H2$:= FaceEnum($E2$)
Q := VertexEnum($H1 \cup H2$)
return (Q, δ)
endif
end

3.1 A Simple Numerical Example

Let us illustrate a simplified example that can help one to show the previous procedure step by step.

Example 1. *Consider a statistical analysis of doping in sports and how it improves performance while simultaneously damaging health. So let us consider the binary variables (i.e., events) $X_1 = D \equiv$"the athlete uses banned performance-enhancing drugs" (i.e., "doping"), $X_2 = E \equiv$"the athlete is showing a performance-enhancing in the last period" and $X_3 = H \equiv$"the athlete is showing a significant change in his/her biological profile".*

Hence the domain U of our assessment will be $U = \{D, E, H\}$, while, at the moment, the universal set V can be any, not better specified, superset $V \supseteq U$.

Suppose one obtains the probability values $p(D) = 0.9$, $p(E) = 0.8$ and $p(H) = 0.9$ by collecting information from disparate sources of information (e.g., public health registers, drugs consumption's and physicians' files, trainer interviews, etc.) on athletes showing significant increases in their performances or health alterations. At a first look the numerical evaluation $p = (0.9, 0.8, 0.9)$ on U, except from the extremely high values, could seems "acceptable". On the contrary, since doping causes both an enhancing in the performance and an alteration of the health and furthermore and since information is collected only among athletes already showing at least one of the two "symptoms", the assessment must be endowed with the logical constraints:

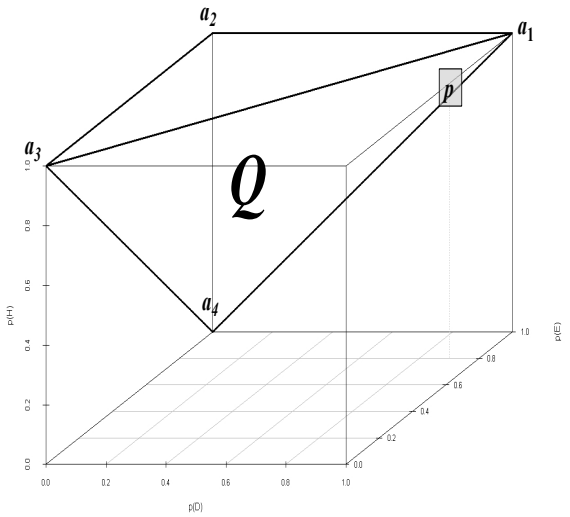


Figure 1: Configuration of the assessment of the doping example: the coherent assessments polytope $\mathcal{Q}$ is delimited by vertexes a_1, a_2, a_3, a_4; the initial incoherent assessment $p = (0.9, 0.8, 0.9)$ is at L_1 distance $\delta = 0.2$ from $\mathcal{Q}$.

$$D \Rightarrow E \wedge H; \quad (4)$$
$$\neg(E \vee H) = \bot, \quad (5)$$

or, equivalently, with the set of clauses

$$\mathfrak{C} = \{E \vee H, \neg D \vee E, \neg D \vee H\}. \quad (6)$$

Consequently the assessment $\pi = (V, U, p, \mathfrak{C})$ is incoherent since the set of coherent values $\mathcal{Q}$ is characterized by the probability inequalities

$$\begin{cases} p'(D) \leq p'(E) \\ p'(D) \leq p'(H) \\ p'(E) + p'(H) - p'(D) \geq 1 \end{cases},$$

and consist of the convex 0-1 polytope with vertexes $a1 = (1, 1, 1)$, $a2 = (0, 1, 1)$, $a3 = (0, 0, 1)$, $a4 = (0, 1, 0)$ and, as it also apparent from Fig. 1, p is outside it and hence incoherent. The first step of the previously described procedure, through MIP program $\mathcal{P}1$, returns that $\delta = 0.2$, while the second MIP program $\mathcal{P}2$ finds $\bar{p} = (0.833, 0.867, 0.967)$ so that one can

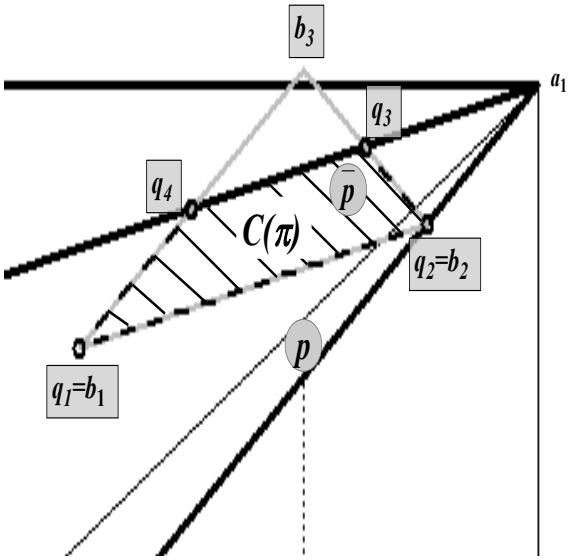


Figure 2: Zooming on the doping example assessment: facet F2 of $\mathcal{B}_\pi(\delta)$ is the grey triangle (b_1, b_2, b_3), the set of corrected assessment $\mathcal{C}(\pi)$ is the dashed polygon with vertexes $q_1 = b1$, $q_2 = b2$, q_3 and q_4.

derive the V-representations of the two facets as

$$E1 = \left\{ \begin{array}{l} a_{1\cdot} = a_1, \\ a_{2\cdot} = a_3, \\ a_{3\cdot} = a_4 \end{array} \right\} \tag{7}$$

$$E2 = \left\{ \begin{array}{l} b_1 = (0.7, 0.8, 0.9), \\ b_2 = (0.9, 1, 0.9), \\ b_3 = (0.9, 0.8, 1.1) \end{array} \right\}. \tag{8}$$

With them the intermediate calls to FaceEnum produce the H-representations (not reported here because of an hard reading) of the facets F1 and F2. The final call to the VertexEnum returns the V-representation of the whole set $C(\pi)$ of corrected values (the dashed polygon in Fig. 2) that is

$$Q = \left\{ \begin{array}{l} q_1 = b1 = (0.7, 0.8, 0.9), \\ q_2 = b2 = (0.9, 1, 0.9), \\ q_3 = (0.9, 0.9, 1), \\ q_4 = (0.8, 0.8, 1) \end{array} \right\}. \tag{9}$$

□

4 Merging Probability Assessments

Given two coherent probability assessments $\pi_1 = (V, U, p, \mathfrak{C})$ and $\pi_2 = (V, W, q, \mathfrak{D})$, on the same propositional variables V, one can say that π_1 and π_2 are compatible if for each variable $x \in U \cap W$, $p(x) = q(x)$. In other words, p and q coincide on the variables in common among π_1 and π_2. The compatibility of two probability assessments means that they do not assign, in an apparent way, different probability values to the same variable. Nevertheless, they can be contradictory by assigning different values to same proposition in an *implicit way*, thus the assessment formed by joining together π_1 and π_2 could be incoherent.

Example 2. *Take for example as $\pi_1 = (V, U, \overline{p}, \mathfrak{C})$ the "barycentric" correction of π in Ex.1 given by the first $\mathcal{P}1$ MIP program, and let $\pi_2 = (V, W, q, \mathfrak{D})$ be a further investigation over official training data that, agreeing with the percentages of enhancing performers and of biological perturbations, claims moreover that the percentage of athletes that naturally, i.e., without doping, are able to enhance significantly their performance showing biological modifications is of the 1%. Hence one has*

$$W = \{X_2 = E, X_3 = H, X_4 = (\neg D \wedge E \wedge H)\}; \tag{10}$$

$$q = \left(\begin{array}{l} q(E) = \overline{p}(E) = 0.867, \\ q(H) = \overline{p}(H) = 0.967, \\ q(X_4) = 0.01 \end{array} \right); \tag{11}$$

$$\mathfrak{D} \equiv \mathfrak{C} \cup \{\neg D \vee \neg X_4, E \vee \neg X_4, H \vee \neg X_4\}. \tag{12}$$

By construction π_1 and π_2 are "compatible" since they give the same probabilities to the common subdomain $U \cap W = \{E, H\}$, but they disagree on X_4 since the all coherent extensions of $\overline{p}$ give zero probability to X_4.□

If two probability assessments $\pi_1 = (V, U, p, \mathfrak{C})$ and $\pi_2 = (V, W, q, \mathfrak{D})$ are compatible, one can denote $\pi_1 + \pi_2$ as the probability assessment $(V, U \cup W, r, \mathfrak{C} \cup \mathfrak{D})$, where $r : U \cup W \to [0, 1]$ is defined by joining together p and q, i.e.,

$$r(x) = \left\{ \begin{array}{ll} p(x) & \text{if } x \in U \\ q(x) & \text{if } x \in W \end{array} \right.$$

The compatibility condition assures that the value of $r(x)$, when $x \in U \cap W$, is uniquely defined.

Given two compatible probability assessments $\pi_1 = (V, U, p, \mathfrak{C})$ and $\pi_2 = (V, W, q, \mathfrak{D})$, the merging operation of π_1 and π_2 is defined by

$$\pi_1 \oplus \pi_2 = \text{Correct}(\pi_1 + \pi_2)$$

Example 2. *(continues) If one takes the juxtaposition of the two assessments π_1 and π_2 one gets an assessment $\pi_1 + \pi_2$ with components V, $U \cup W = (D, E, H, X_4)$, $r = (0.8333, 0.8667, 0.9667, 0.01)$ and logical constraints $\mathfrak{D}$. It is, as explained before, incoherent, with an L_1 minimal distance of $\delta = 0.01$ and its correction $\pi_1 \oplus \pi_2$ is the credal set with extremal*

numerical values

$$q_1 = (0.8333, 0.8767, 0.9667, 0.01)$$
$$q_2 = (0.8333, 0.8667, 0.9767, 0.01)$$
$$q_3 = (0.8233, 0.8667, 0.9667, 0.01)$$
$$q_4 = (0.8333, 0.8667, 0.9667, 0.00). \qquad \square$$

Compatibility is not always present between different assessments, especially if the two sources of information stem from disparate contexts. When the probability assessments to be merged are non compatible, i.e., they assign different probability values to some common variables, it is not possible to join directly them into a unique assessment. Anyhow two different solutions are possible: a "weighted combination" of the two assessments, or a "assignment to duplicates", as detailed in the next paragraphs.

The first approach requires one to create a non contradictory probability assessment from π_1 and π_2, by choosing a unique probability value for each variable in $U \cap W$. A possible solution is to use a weighted average of p and q, i.e., chosen a weighting coefficient $\omega \in [0, 1]$, where $\pi_1 +_\omega \pi_2$ defines the probability assessment $(V, U \cup W, r, \mathfrak{C} \cup \mathfrak{D})$, where $r : U \cup W \to [0, 1]$ is now defined

$$r(x) = \begin{cases} p(x) & \text{if } x \in U \setminus W \\ q(x) & \text{if } x \in W \setminus U \\ \omega p(x) + (1-\omega) q(x) & \text{if } x \in U \cap W \end{cases}$$

Finally, the merging operation of π_1 and π_2 is

$$\pi_1 \oplus_\omega \pi_2 = \text{Correct}(\pi_1 +_\omega \pi_2)$$

When $\omega = \frac{1}{2}$, equal importance is given to π_1 and π_2 and $\oplus_{\frac{1}{2}}$ becomes commutative. While the extreme values $\omega = 0$ and $\omega = 1$ correspond to the cases where the values of π_2 (or π_1, respectively), are used for contradictory situations. In some sense $\frac{\omega}{1-\omega}$ is a measure of the relative reliability of π_1 with respect to π_2.

Example 3. *If one renders explicit the contradiction on X_4 of the two assessment π_1 and π_2 of Ex.2, i.e., by considering $p(X_4) = 0$, and one chooses $\omega = \frac{1}{2}$, one has the starting weighted assessment $\pi_1 +_{\frac{1}{2}} \pi_2$ with components V, $U \cup W = (D, E, H, X_4)$, $r = (0.8333, 0.8667, 0.9667, 0.005)$ and logical constraints again expressed through the same set of clauses $\mathfrak{D}$. It is anyhow incoherent with an L_1 minimal distance of $\delta = 0.01$ but its correction $\pi_1 \oplus_{\frac{1}{2}} \pi_2$ is now the credal set with extremal values*

$$q_1 = (0.8333, 0.8742, 0.9667, 0.0075)$$
$$q_2 = (0.8308, 0.8642, 0.9667, 0.00)$$
$$q_3 = (0.8333, 0.8667, 0.9742, 0.0075)$$
$$q_4 = (0.8308, 0.8667, 0.9642, 0.00)$$
$$q_5 = (0.8358, 0.8692, 0.9667, 0.00)$$
$$q_6 = (0.8258, 0.8667, 0.9667, 0.0075). \qquad \square$$

Of course, by varying the weight ω in $\pi_1 \oplus_\omega \pi_2$, one obtains a class of new coherent (imprecise) assessments over the domain $U \cup W$ that have the peculiarity of being "compromises" of the two original π_1 and π_2, but with the same weight for each event with associated different values.

A different approach is to create a probability assessment which maintains both numerical values and to solve the apparent contradiction by adding a new logical variable X_i', for each event $X_i \in U \cap W$ such that $p(X_i) \neq q(X_i)$, and assigning the values $r(X_i) = p(X_i)$ and $r(X_i') = q(X_i)$. Moreover, the logical constraint $X_i = X_i'$ is added to $\mathfrak{C} \cup \mathfrak{D}$.

Indeed, the assessment so obtained $\pi_1 + \pi_2$ is obviously incoherent and the merging operation of π_1 and π_2 is computed as

$$\pi_1 \oplus_I \pi_2 = \text{Correct}(\pi_1 + \pi_2).$$

Note that, whenever the two assessments π_1 and π_2 are compatible, this merging operator $\pi_1 \oplus_I \pi_2$ coincides with the previous $\pi_1 \oplus \pi_2$ since no duplication of variables is needed in such a case.

The main difference between the two approaches is that the latter $\oplus_I$ tries to automatically solve the contradiction, while the operator $\oplus_\omega$ needs an explicit way of solving it. The approach of $\oplus_\omega$ is in some sense a supervised one, because the user must explicitly provide a weight ω, while $\oplus_I$ adopts an unsupervised approach, and these difference can leads to very different final results, as the following example shows.

Example 4. *Let us proceed as in Ex.3 but maintaining the two distinct values associated to X_4, i.e., let us start with the assessment $\pi_1 + \pi_2$ with components V, $U' = (D, E, H, X_4, X_4')$, $r = (0.8333, 0.8667, 0.9667, 0.00, 0.01)$ and with logical constraints augmented to $\mathfrak{D} \cup \{\neg X_4 \vee X_4', X_4 \vee \neg X_4'\}$. This further assessment has again a minimal L_1 distance of $\delta = 0.01$ from the polytope $\mathcal{Q}$ of coherent assessments (note anyhow the different cardinality of the space $n = 5$), but whose correction leads now to a precise assessment with numerical values*

$$(0.8333, 0.8667, 0.9667, 0.00, 0.00). \qquad \square$$

Anyway, the idea behind these two definitions is the same, i.e., the merging of two information sources can be performed in two steps. First, put together all the information $\mathcal{I}$, and then find the smallest number of corrections on $\mathcal{I}$ such that the new information $\mathcal{I}'$ is consistent. The choice of which merging operator adopt should be based on the availability or not of relevance, or better of the reliability, of the sources of information. If a reliability grade is available, or reasonably assessed, the $\oplus_\omega$ should be preferred, if not the $\oplus_I$ operator avoids the use of unrealistic assumptions.

Thinking the probability assessments as belief states, the merging operators are a belief merging functions (see, e.g., [22]).

Our approach is different from usual imprecise probability technique "a la Walley" (see in particular [34]), where usually the convex hull of incompatible assessments is considered. This is a so called "least commitment" procedure, while our proposal can be dually thought as "maximal commitment". In fact, in our merging operators, values which are exogenous to the initial assessments (like those appearing by doing the convex hull) are avoided as much as possible, and original opinions are maintained fixed and crisp as much as possible. Moreover the convex hull of initial assessments is not guaranteed to at least "avoid sure loss", so that the Walley's "natural extension" procedure is not always applicable. On the contrary, our approach is always applicable.

5 Revising Probability Assessments

In this section we propose how the correction procedure can be used to revise a probability assessment.

Suppose that the coherent probability assessment $\pi_1 = (V, U, p, \mathfrak{C})$ represents our current belief state and a new reliable information arrives, represented by the probability assessment $\pi_2 = (V, W, q, \mathfrak{D})$.

One could merge π_1 and π_2 as described in the previous section, but suppose that one would rather update our belief state with the new available information, with the idea that

- one assumes that the new information is correct
- one allows to revise, as less as possible, our current state in order to adapt it to the new information

The revision can be performed as follows. First, π_1 and π_2 are merged together with the operator $+_0$, thus in the case of contradiction, the values from π_2 are used. Second, the resulting assessment is corrected by forbidding any change the probabilities of the variables in W. This can be achieved with the procedure Correct2 which is a small modification of the procedure Correct. Correct2 has a further parameter, the set T of the variables whose probability value cannot be corrected, and when the MIP systems $\mathcal{P}1$ and $\mathcal{P}2$ are built, the constraint 1 for the variables of T reduces to

$$\sum_{j=1}^{n+1} b_{ij} = p(X_i)$$

and their corresponding variable r_i and s_i are not created.

The revision of π_1 with π_2 is then computed as

$$\pi_1 \star \pi_2 = Correct2(\pi_1 +_0 \pi_2, W)$$

Note that any probability assessment $(V, U \cup W, r', \mathfrak{C} \cup \mathfrak{D})$ resulting from $\pi_1 \star \pi_2$ is such that it agrees with q, i.e., $r'(x) = q(x)$ for all $x \in W$.

Example 5. *If in Ex.2 one wants to inevitably maintain as valid the latter investigation π_2 one starts with an adjoined initial assessment $\pi_1 +_0 \pi_2$ with components V, $U \cup W = (D, E, H, X_4)$, $W = (E, H, X_4)$, $r = (0.8333, 0.8667, 0.9667, 0.01)$ and logical constraints $\mathfrak{D}$. The only possibility to correct it is to reduce the numerical evaluation $r(D) = 0.8333$ to $r'(D) = 0.823$, so that the result of the revision is the precise assessment $\pi_1 \star \pi_2$ with components V , $U \cup W = (D, E, H, X_4)$, $r' = (0.8233, 0.8667, 0.9667, 0.01)$ and the same logical constraints $\mathfrak{D}$.* □

Such revising methodology, that in general leads to an imprecise model, could be thought as an analogous of the famous Jeffrey's rule of combination [26]. The main difference between the two is that our proposal minimize the probability mass dislocation from the original assessment, maintaining as much as possible the magnitude of the values, hence working in an "additive" way, while Jeffrey's rule maintains as much as possible the odds ratios, hence working in a "multiplicative" way.

Moreover the Jeffrey's rule produces a final probability assessment which could be too different from π since it inevitably alters all the values of p on $U \setminus W$, while our approach tries to modify p as less as possible, in line with the belief revision methodology [22].

6 Conclusions

In this article, a preliminary proposal for a correction of incoherent probability assessments on finite domains through L_1 distance minimization was presented. The proposal's novelty is reflected in a new procedure that uses mixed integer programming while profiting from

geometrical properties of the convex sets involved, makes such a method easily applicable.

Apart from the applicability of the direct incoherent correction "per se," we have stressed that such a method can be tailored to naturally implement the merging of disparate assessments or reasonable belief revisions. We focused on the combination of two different assessments, but the generalization to the merging or revision of several assessments is straightforward: it simply requires the generalization of the weighted combination $+_\omega$ by allowing convex combination of several values and to iterate the juxtaposition $+$ which duplicates several times.

Our procedure can be seen as a reasonable way to generate lower-upper probability models from precise, but incoherent, probabilities estimates.

Future research should systematically analyze the procedure through simulation studies and investigate formal properties of the correction operator. We are confident that the revising $\star$ operator satisfies some properties which are the probabilistic counter-parts of the Katsuno-Mendelzon axiom for belief revision operators.

Acknowledgments

This work has been partially supported by Italian Ministry of Education, University and Research funding of Research Projects of National Interest (PRIN 2010-11) under grant `2010FP79LR_003` *"Logical methods of information management"* and by Italian Ministry of Health under grant `J52I14001640001` *"Sistemi intelligenti di ausilio alle decisioni per l'identificazione precoce e la dissuasione all'utilizzo del doping"*.

References

[1] M. Baioletti, A. Capotorti, S. Tulipani, B. Vantaggi. Elimination of Boolean variables for probabilistic coherence. *Soft Computing*, 4(2):81–88, 2000.

[2] M. Baioletti, A. Capotorti, S. Tulipani, B. Vantaggi. Simplification Rules for the Coherent Probability Assessment Problem. *Ann. of Math. and Artif. Intell.*, 35:11–28, 2002.

[3] M. Baioletti, A. Capotorti, S. Tulipani. An empirical complexity study for a 2CPA solver. In: *Modern Information Processing: From Theory to Applications*. B. Bouchon-Meunier, G. Coletti and R.R. Yager, Eds. 1–12, 2005.

[4] C.E. Bonferroni. Teorie e probabilità. Discorso Inaugurale annuario 1925-1926. *Regio Istituto Superiore di Scienze Economiche e Commerciali*, Bari, Casa Editrice Cressati.

[5] A. Capotorti. Benefits of embedding structural constraints in coherent diagnostic processes. *Int. J. of Approx. Reasoning*, 39(2-3):211–233, 2005.

[6] A. Capotorti. A Further Empirical Study on the Over-Performance of Estimate Correction in Statistical Matching. In: *Advances in Computational Intelligence*. S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, R.R. Yager Eds. Springer Berlin Heidelberg. 124–133, 2012.

[7] A. Capotorti, G. Regoli, Coherent correction of inconsistent conditional probability assessments, in: L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds), Proceeding of IPMU'08, Malaga (ES), 891–898, 2008.

[8] A. Capotorti, G. Regoli, F. Vattari, Correction of incoherent conditional probability assessments, *Int. J. of Approx. Reasoning* 51(6): 718–727, 2010.

[9] A. Capotorti, B. Vantaggi, Incoherence correction strategies in statistical matching, In: *Proceedings of ISIPTA 2011*, Innsbruck (Austria), 109–118, 2011.

[10] A. Capotorti, B. Vantaggi. Correction of Incoherences in Statistical Matching. In: *Statistical Models for Data Analysis*. P. Giudici, and S. Ingrassia, and M. Vichi Eds. Springer International Publishing, 73–80, 2013.

[11] G. Coletti. Numerical and Qualitative Judgments in Probabilistic Expert Systems. In *Proc. of the International Workshop on Probabilistic Methods in Expert Systems*, Romano Scozzafava Ed. Roma: SIS, 37–55, 1993.

[12] G. Coletti. Coherent numerical and Ordinal probabilistic assessments. *IEEE Transaction on Systems, Man, and Cybernetics*, 24:1747–1754, 1994.

[13] G. Coletti, R. Scozzafava, *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", 2002.

[14] F.G. Cozman, L. Fargoni di Ianni. Probabilistic Satisfiability and Coherence Checking through Integer Programming. *Lecture Notes in Computer Science*, 7958, 145–156, 2013.

[15] A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605-613, 1982.

[16] A. P. Dawid. Calibration-Based Empirical Probability. *Ann. Statist.* 13(4): 1251–1274, 1985.

[17] B. de Finetti. La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré. Section B*, 7:l–68, 1937.

[18] B. de Finetti. *Teoria della Probabilità*. Torino: Einaudi, 1970 (Engl. transl. *Theory of probability*, London: Wiley & Sons, 1974).

[19] M. Finger, G. De Bona. Probabilistic satisfiability: Logic based algorithms and phase transition. *IJCAI*, 528-533, 2011.

[20] M. Fréchet. *Les Mathématiques et le Concret*, Paris. P.U.F.

[21] K. Fukuda. Lecture: Polyhedral Computation, Spring 2011. Institute for Operations Research and Institute of Theoretical Computer Science. ETH Zurich. Available online at `http://stat.ethz.ch/ifor/teaching/lectures/poly_comp_ss11/lecture_notes`.

[22] P. Gärdenfors and H. Rott. Belief revision. In Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 4, pages 35–132. Oxford University Press. (1995)

[23] G.Georgakopoulos, D.Kavvadias and C.H.Papadimitriou. Probabilistic Satisfiability. *Journal of Complexity*. 4:1–11, 1988.

[24] A. Gilio. Probabilistic Consistency of Knowledge Bases in Inference Systems. *Lecture Notes in Computer Science*, M. Clarke, R. Kruse, S. Moral Eds. Springer–Verlag, 747:160–167, 1993.

[25] B. Jaumard, P. Hansen, M. Poggi de Aragão. Column Generation Methods for Probabilistic Logic. *ORSA Journal on Computing*, 3(2): 135-148, 1991.

[26] R.C. Jeffrey. *The logic of decision*. McGraw-Hill, 1965, 2nd Ed. Univ. Chicago Press, Chicago, 1983.

[27] F. Lad, Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction, Wiley, New York, 1996.

[28] D. V. Lindley, A. Tversky And R. V. Brown. On the Reconciliation of Probability Assessments. *J. R. Statist. Soc. A*, 142(2):146–180, 1979.

[29] K.G. Murty, K.S. Al-Sultan. On Relationship between L_1, L_2, L_∞ Minima. *Technical Report 89-26*, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, USA, 1989.

[30] D. Osherson, D. Lane, P. Hartley, and R. R. Batsell. Coherent Probability from Incoherent Judgment. *Journal of Experimental Psychology: Applied*, 7(1):3–12, 2001.

[31] J. Pratt. Must subjective probabilities be realized as relative frequencies? *Unpublished seminar paper*, Harvard University Graduate School of Business Administration, *Technical report* 1962.

[32] D. Pretolani. Probability Logic and Optimization SAT: the PSAT and CPA Models. *Ann. of Math. and Artif. Intell.*, 43:211–221, 2005.

[33] B. Vantaggi. Statistical matching of multiple sources: A look through coherence. *Int. J. of Approx. Reasoning* 49:701–711, 2008.

[34] P. Walley, The elicitation and aggregation of beliefs. *Technical Report University of Warwick*, 1982.

[35] P. Walley, *Statistical reasoning with Imprecise Probabilities*. Chapman and Hall, London 1991.

The Geometry of Imprecise Inference

Miķelis Bickis
Department of Mathematics and Statistics
University of Saskatchewan
bickis@snoopy.usask.ca

Abstract

A statistical model can be constructed from a null probability measure by defining a set of statistics representing log-likelihood ratios of alternative measures to the null measure. Conversely, any model consisting of equivalent measures can be so expressed. A linear combination of statistics will also define a log-likelihood ratio if the normalizing constant is finite. In this way, any such model can be naturally extended to a convex subset of the linear span of these statistics. A finite dimensional subset defines an exponential family with the canonical parameters of a measure defined by coordinates relative to a set of basis functions.

Given a base measure on the parameter space, one can implement a similar structure with a set of parametric functions. The log-likelihood itself being a parametric function, the set of all possible log-likelihoods thus defines a space of measures conjugate to the statistical model. The conjugate space will have one more dimension spanned by the above-mentioned parameter-dependent normalizing constant.

If the base measure is considered a prior distribution, then the translation by the observed log-likelihood defines the posterior. An imprecise prior defined by a set of measures is in the same manner translated to a set of posterior measures. Upper and lower previsions can then be computed as extrema over this posterior set.

Keywords. Information geometry, exponential family, sets of measures.

1 Introduction

Statistical inference deals with observations that are realizations of a random process whose probability law is postulated to be one of a set of probability laws. We call this set the *model space*. Bayesian inference also requires a probability measure defined on the model space indexed by a set of *parameters* such that the distribution of the observations is viewed as being conditional on an unobserved realized parameter. Bayes' rule is then used to combine the *prior distribution* on the model space with the observation to give a *posterior distribution* on the model space, which will hopefully be more informative than the prior. This procedure is called *Bayesian updating*, but in the computer science community it is also known as *learning from data*, a terminology that is more descriptive of what is actually happening.

While Bayesian inference is based on a solid mathematical foundation, its use has been much criticized as being an improper method for scientific investigation (see Mayo [10] for an overview). One of the criticisms relates to the arbitrariness of the prior distribution. The subjectivity reflected in the prior seems out of place in the objectiveness of science. Even if one acknowledges that all inference relies on prior assumptions that are inherently subjective, there remains the practical issue of enunciating these assumptions sufficiently precisely to define a probability distribution on the model space.

These criticisms were addressed in Walley's fundamental treatise [14]. Walley introduces the concepts of lower and upper *previsions* on a set of *gambles*. In more conventional language, gambles are just random variables, and the term prevision (borrowed from de Finetti [6]), is essentially an expectation. Walley's novelty is in allowing the prevision to be defined on only a subset of random variables, thus providing for an incomplete description of a prior probability distribution which is more realistic than the classical Bayesian requirement. Moreover, Walley posits so-called *upper* and *lower previsions* which are merely bounds on the expectations, thereby further providing for incomplete knowldege, freeing one from having to specify a precise number as the prior expectation of any random variable. When applied to indicator variables, upper and lower previsions define upper and lower probabilities. Walley's development however is constrained

by the assumption that gambles are bounded. The case of unbounded gambles is discussed by Troffaes and de Cooman [13].

Walley's lower envelope theorem [14, Section 3.3.3] shows that if the upper and lower previsions satisfy coherence axioms, then they can be expressed in terms of conventional expectations: One can find a set of probability measures (dubbed *credal set* by Levi [9]) with corresponding expectation functionals, such that the lower prevision is the infimum of all expectations over the set, and the upper prevision is the supremum. Thus working with upper and lower previsions is equivalent to replacing probability measures with sets of probability measures.

Inference can now be based on such imprecise prior probabilities. Walley proposed a *generalized Bayes' rule* in which imprecise prior probabilities are updated to imprecise posterior probabilities. The posterior probabilities would then be expected to be more precise than the priors in the sense that the difference between upper and lower probabilities is reduced. Walley [15] also introduced the *imprecise Dirichlet model* (IDM) for learning from multinomial data, in which the priors are defined as a set of Dirichlet distributions with a fixed concentration parameter s, and the posteriors are Dirichlet distributions with s increased by the sample size.

Diaconis and Ylvisaker [5] discussed the process of Bayesian updating in exponential families. When the model space is an exponential family, then one can define a conjugate exponential family of prior distributions (indexed by *hyperparameters*) on the model parameters such that Bayesian updating can be expressed as a data induced change in the hyperparameters. Moreover, under certain regularity conditions, the predictive expectations of the canonical sufficient statistics can be expressed as a weighted average of prior expectation and sample mean.

Since the multinomial and Dirichlet distributions are conjugate in the sense of Diaconis and Ylvisaker, Walley's IDM can be viewed as an imprecise probability version of their setup. Imprecise versions of other exponential families have been proposed by Quaeghebeur and de Cooman [12], Quaeghebeur [11], Bickis [4], Benavoli and Zaffalon [3], Bataineh [2], and Lee [8]. The problematic step in all these situations is determining a set of priors. One wants a set sufficiently large such that previsions are near-vacuous *a priori* but not so large that learning from data is not possible. Such a set of priors will be said to have the *Benavoli-Zaffalon (BZ) property* as discussed in their paper [3].

In this paper, we consider a geometric representation of model and prior probabilities in which the idea of conjugacy is extended beyond that considered by Diaconis and Ylvisaker. Using canonical parameterizations, Bayes' rule can be seen as a data-dependent translation of a point representing the prior distribution. The generalized Bayes rule can similarly be seen as a translation of an entire set. We can thus visualize how various choices of prior set affect the process of learning from data. We present several examples to illustrate various situations that arise in this paradigm. In most of these examples we consider the effect of a single observation. The effect of i.i.d. samples should then be viewed as iterations of the updating paradigm, illustrating the effect of accumulating information.

2 Geometry of Probability Measures

Let $\mathcal{Y}$ be an observation space whose elements represent possible empirical observations. We make few assumptions about the structure of this space; elements may be numeric or nominal, scalar or vector of finite or infinite dimension. All we require is that we are able to specify a probability measure P_0 on some σ-algebra of events defined on $\mathcal{Y}$. We are interested in making an inference about the probabilistic nature of $\mathcal{Y}$ and may think of P_0 as a null model which we wish to compare with some other putative measure P_1. We will assume that no deterministic inference is possible, i.e., that any event that is possible (with positive probability) under one measure is similarly possible under another. In the language of probability theory, P_0 and P_1 are equivalent measures: $P_0 \sim P_1$.

2.1 One-Dimensional Case

The likelihood principle implies that any inference concerning P_1 vs. P_0 is based solely on the likelihood ratio, which is convenient to express in its logarithmic form:

$$\ell = \log \frac{dP_1}{dP_0}, \tag{1}$$

from which it follows that we can write

$$P_1(A) = \int \mathbf{1}_A e^{\ell}\, dP_0 \tag{2}$$

where $\mathbf{1}_A$ represents the indicator function of a measurable subset A of $\mathcal{Y}$. By introducing a scalar parameter θ, we can define one-dimensional exponential family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where

$$P_\theta(A) = \int \mathbf{1}_A \exp(\theta\ell - \phi(\theta))\, dP_0, \tag{3}$$

$$\phi(\theta) = \log \int e^{\theta\ell}\, dP_0, \tag{4}$$

and

$$\Theta = \{\theta \in \mathbb{R} : \phi(\theta) < \infty\}. \tag{5}$$

Theorem 1 *Θ is a convex set.*

Proof: If $\theta_1, \theta_2 \in \Theta$ and $0 < \alpha < 1$ then

$$e^{\phi(\alpha\theta_1+(1-\alpha)\theta_2)} = \int e^{\alpha\theta_1\ell} e^{(1-\alpha)\theta_2\ell}\, dP_0.$$

By Hölder's inequality, this is less than

$$\left(\int \left(e^{\alpha\theta_1\ell}\right)^{1/\alpha} dP_0\right)^{\alpha} \left(\int \left(e^{(1-\alpha)\theta_2\ell}\right)^{1/(1-\alpha)} dP_0\right)^{1-\alpha}$$
$$= \phi(\theta_1)^{\alpha}\phi(\theta_2)^{1-\alpha}.$$

Since $\phi(\theta_1)$ and $\phi(\theta_2)$ are both finite by definition, so is $\phi(\alpha\theta_1 + (1-\alpha)\theta_2)$, and the result follows. ∎

Instead of postulating an alternative probability model P_1, we may start with a random variable (i.e., measurable function) v on $\mathcal{Y}$ that we think encapsulates the inference we are interested in making. In the same fashion we may define a one-dimensional exponential family

$$P_\theta(A) = \int \mathbf{1}_A \exp(\theta v - \phi(\theta))\, dP_0, \quad \theta \in \Theta \tag{6}$$

where ϕ and Θ are defined as before in (4) and (5).

Definition 1 *The family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ defined by ((6)) will be called the family generated by v (over P_0).*

Definition 2 *For any random variable T, $E_\theta(T)$ is defined as*

$$E_\theta(T) = \int T\, dP_\theta = \int T e^{\theta v - \phi(\theta)}\, dP_0.$$

Theorem 2 *If $\theta_1 \neq \theta_2$ then $P_{\theta_1} \neq P_{\theta_2}$ iff v is not almost surely constant.*

Proof:

$$P_{\theta_1} = P_{\theta_2} \iff P_0\left\{e^{\theta_1 v - \phi(\theta_1)} = e^{\theta_2 v - \phi(\theta_2)}\right\} = 1$$

which is equivalent to

$$(\theta_1 - \theta_2)v = \phi(\theta_1) - \phi(\theta_2) \quad \text{a.s.} \tag{7}$$

Since the right side of (7) is constant, this equality can hold only if v is almost surely constant or if $\theta_1 = \theta_2$. On the other hand, if v is almost surely constant, then

$$\phi(\theta) = \log \int e^{\theta v}\, dP_0 = \theta v \quad \text{a.s} \tag{8}$$

and thus (7) holds. ∎

If v is almost surely constant, then $P_\theta = P_0$ for all θ. On the other hand, if

$$\int e^{\theta v}\, dP_0 = \infty \quad \text{for all } \theta \neq 0, \tag{9}$$

then Θ consists of a single point. In either case the family generated by v has but a single probability measure and provides no prospect for inference. In the following we will assume that v is not constant and that $\int \exp(\theta_1 v)\, dP_0 < \infty$ for at least one $\theta_1 \neq 0$. By Theorems 1 and 2, Θ will then include an interval with endpoint θ_1, with distinct θ's corresponding to distinct probability measures.

Theorem 3 *If v_1 and v_2 are random variables on $\mathcal{Y}$ such that $v_1 - v_2$ is almost surely constant, then for any $\theta \in \Theta$, v_1 and v_2 define the same probability measure and hence v_1 and v_2 generate the same family.*

Proof: Let

$$\phi_i(\theta) = \log \int e^{\theta v_i}\, dP_0, \quad i = 1, 2.$$

Then

$$\phi_2(\theta) = \log \int e^{\theta v_1} e^{\theta(v_2 - v_1)}\, dP_0$$
$$= \phi_1(\theta) + \theta(v_2 - v_1) \quad \text{a.s.} \tag{10}$$

For any event A, $\theta \in \Theta$, the probability defined by v_1 is

$$\int \mathbf{1}_A e^{\theta v_1 - \phi_1(\theta)}\, dP_0 = \int \mathbf{1}_A e^{\theta v_2 - (\phi_1(\theta) + \theta(v_2 - v_1))}\, dP_0, \tag{11}$$

$$= \int \mathbf{1}_A e^{\theta v_2 - \phi_2(\theta)}\, dP_0, \tag{12}$$

by (10). ∎

The random variable v may thus differ from a log likelihood ratio by an arbitrary constant. We can make the representation (6) unique by requiring that

$$\int v\, dP_0 = 0. \tag{13}$$

Since

$$\theta v = \log \frac{dP_\theta}{dP_0} + \phi(\theta),$$

the convention (13) implies that

$$\phi(\theta) = \int \log \frac{dP_0}{dP_\theta}\, dP_0. \tag{14}$$

The right side of (14) was described by Kullback [7] as the *mean information for discrimination in favour of P_0 against P_θ* and is one way of quantifying the ease with which a probability measure P_θ can be distinguished from P_0. It is commonly called the *Kullback-Leibler information* or *divergence* [1] and denoted by $I(P_0|P_\theta)$. The divergence may be viewed as the distance from P_0 to P_θ, although it does not satisfy the axioms of a metric.[1] A significant property of

[1] While $I(P_0|P_\theta) > 0$ iff $P_0 \neq P_\theta$, it is not symmetric, does not satisfy the triangle inequality and may even be infinite. However, it can be shown that $I(P_0|P_\theta) < \infty$ when θ is in the *interior* of the set (5).

divergence is additivity over independent observations. Let $P_\theta^{(1,2)} = P_\theta^{(1)} \times P_\theta^{(2)}$ be the joint distribution of two independent observations with distributions $P_\theta^{(1)}$ and $P_\theta^{(2)}$. Then

$$I(P_0^{(1,2)}|P_\theta^{(1,2)}) = I(P_0^{(1)}|P_\theta^{(1)}) + I(P_0^{(2)}|P_\theta^{(2)}). \quad (15)$$

The requirement (13) makes the representation unique, and relates the normalizing constant ϕ to the divergence.

The set of random variables forms a vector space, and the representation (6) identifies a family of probability measures with a convex subset of a one-dimensional subspace, the origin representing the null measure P_0. The function v is a basis vector such that all probability measures in the family can be represented as scalar multiples of v, the scalar being the parameter θ. Because of the need of a normalizing constant $\phi(\theta)$, the log-likelihood ratios actually do not lie in a one-dimensional subspace, but in a two-dimensional subspace spanned by v and the constant function $\mathbf{1}$ equal to 1 everywhere. A probability measure P_θ actually corresponds to an equivalence class of vectors differing by a multiple of $\mathbf{1}$. The convention (13) picks a particular representative of the equivalence class.

We illustrate these ideas with an almost trivial example.

Example 1. Let $\mathcal{Y} = \{0, 1\}$ with $P_0\{0\} = P_0\{1\} = \frac{1}{2}$ and $P_1\{0\} = 1 - P_1\{1\} = 1 - p$ for some $p \in (0, 1)$. Then

$$\frac{dP_1}{dP_0}(0) = (1-p)/\tfrac{1}{2} \qquad \frac{dP_1}{dP_0}(1) = p/\tfrac{1}{2}$$

so that

$$\begin{aligned}\frac{dP_1}{dP_0}(y) &= 2(1-p)^{1-y}p^y \\ \log\frac{dP_1}{dP_0} &= \log 2 + (1-y)\log(1-p) + y\log p \\ &= \log\frac{p}{1-p}\,y + \log 2 + \log(1-p).\end{aligned}$$

putting $v(y) = y - \frac{1}{2}$ and $\theta = \log(p/(1-p))$ we have that

$$\begin{aligned}\log\frac{dP_1}{dP_0} &= \theta v + \theta/2 + \log 2 - \log(1+e^\theta) \\ &= \theta v - \log\left(\frac{1+e^\theta}{2e^{\theta/2}}\right) \\ &= \theta v - \log\cosh(\theta/2). \quad (16)\end{aligned}$$

The family of binary distributions is thus displayed in the form (6) parametrized by the log-odds $\theta = \log(p/(1-p))$ with

$$\phi(\theta) = I(P_0|P_1) = \log\cosh(\theta/2).$$

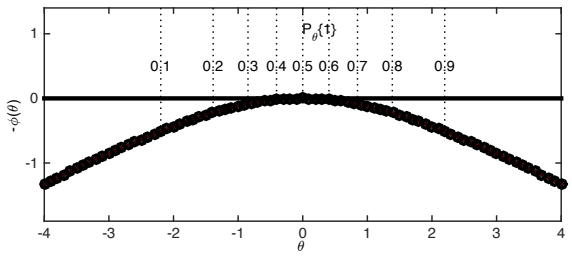


Figure 1: Probability manifold for binary distributions. The set of measures forms a one-dimensional manifold in the plane. The distance in the vertical directions represents the divergence from the uniform distribution. The points in the manifold can be projected in this direction onto the tangent plane. Location along the plane is linear in the canonical parameter θ, but non-linear in the success probability p.

The set of *all* functions on $\{0, 1\}$ is two-dimensional, being isomorphic to $\mathbb{R}^2$. The representation (16) identifies those functions that are log-likelihood ratios relative to uniform probabilities. Using a basis consisting of the functions $v_0(y) = 1$ and $v_1(y) = y - \frac{1}{2}$, the set of log-likelihood ratios (equivalently, probability measures) can be visualized as in Figure 1. In this figure, the equivalence classes correspond to vertical lines.

Example 2. Suppose now that we have n i.i.d. observations $y_1, \ldots, y_n$ from Example 1. Let $P_{0,n}$ (resp. $P_{1,n}$) represent the joint distribution of n i.i.d. binary observations with success probability $\frac{1}{2}$ (esp. p). Again, let $\theta = \log(p/(1-p))$. The joint log likelihood of independent observations is the sum of the log likelihoods, and by (15) the same will be true for the divergences. Thus adding terms of the form (16) we get

$$\log\frac{dP_{1,n}}{dP_{0,n}}(y_1, \ldots, y_n) = \theta\left(\sum_{i=1}^n y_i - \frac{n}{2}\right) - n\log\cosh\frac{\theta}{2}, \quad (17)$$

which is the canonical form of the binomial family. Alternatively, let $\mathcal{Y}$ be the set of all 2^n binary sequences and P_0 be the uniform measure on this set. Then if we decide that inferences are to made solely on the basis of the function $v(y_1, \ldots, y_n) = \sum_i y_i$, the family generated by v is again binomial. The picture of this family is just a rescaling of Figure 1 and thus has the same intrinsic geometry. This geometric equivariance under repeated sampling is characteristic of exponential families.

Example 3. Let $\mathcal{Y} = \mathbb{R}^+$, and define P_0 by the cumulative distribution function

$$P_0((0, y]) = 1 - e^{-y}, \quad y > 0,$$

then with $v_1(y) = y - 1$ the one-dimensional exponential family is

$$\log \frac{dP_\theta}{dP_0} = \theta v_1 - \phi(\theta) \tag{18}$$

where

$$\phi(\theta) = I(P_0|P_\theta) = -\theta - \log(1-\theta).$$

The natural parameter space is $\Theta = (-\infty, 1)$ which defines the family of exponential distributions with expectation $(1-\theta)^{-1}$.

2.2 Multidimensional Case

The inference of interest may not be expressible in terms of a single function v; we may require a family of functions $\mathcal{L}_0$, in which case a construction as in (3) is possible for any $v \in \mathcal{L}_0$. Indeed, for any finite number of functions $v_1, \ldots, v_k \in \mathcal{L}_0$ and scalar parameters $\theta_1, \ldots, \theta_k$ we can construct a probability measure

$$\begin{aligned} P_{\boldsymbol{\theta}}(A) &= P_{\theta_1,\ldots,\theta_k}(A) \\ &= \int \mathbf{1}_A \exp\left(\sum_{i=1}^{k} \theta_i v_i - \phi(\boldsymbol{\theta})\right) dP_0 \end{aligned} \tag{19}$$

provided that

$$\phi(\boldsymbol{\theta}) = \log \int \exp\left(\sum_i \theta_i v_i\right) dP_0 < \infty. \tag{20}$$

Thus a given set $\mathcal{L}_0$ of functions can be augmented by their linear combinations, the set $\mathcal{L}$ of all such linear combinations forming a vector space. In that case we have a generalization of Definition 1:

Definition 3 *Given a set $\mathcal{L}_0$ of random variables, the set of probability measures defined by* (19) *and* (20) *will be called the family generated by $\mathcal{L}_0$.*

If for a fixed set of functions $v_1, \ldots, v_k$ every function in $\mathcal{L}$ can be *uniquely* expressed as a linear combination of $v_1, \ldots, v_k$, then $\mathcal{L}$ will be a k-dimensional vector space and $v_1, \ldots, v_k$ will be basis vectors. The vector space will be infinite dimensional if no such finite basis can be found.[2] We focus on the finite-dimensional case. Here it is convenient to fix a basis $v_1, \ldots, v_k$ and consider $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_k)$ representing the measure $P_{\boldsymbol{\theta}}$ as a row vector and the values $v_1(y), \ldots, v_k(y)$ as a column vector $\mathbf{v}$. Then the vectors of parameters and statistics act on each other via matrix multiplication. Thus, the family generated by $v_1, \ldots, v_k$ can be represented as

$$\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \quad \text{where} \tag{21}$$

$$P_\theta(A) = \int \mathbf{1}_A e^{\boldsymbol{\theta}^\top \mathbf{v} - \phi(\boldsymbol{\theta})}\, dP_0 \tag{22}$$

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^k : \phi(\boldsymbol{\theta}) = \int e^{\boldsymbol{\theta}^\top \mathbf{v}}\, dP_0 < \infty\}. \tag{23}$$

As in the one-dimensional case, the log likelihood ratio may differ by a constant from a function in $\mathcal{L}$. Thus if $\mathcal{L}$ is k-dimensional with basis $v_1, \ldots, v_k$, the set of log-likelihood ratios lies in a $k+1$-dimensional space spanned by $v_0, v_1, \ldots, v_k$, where $v_0 = \mathbf{1}$. Again, two functions that differ by a scalar multiple of $\mathbf{1}$ will define the same probability measure, and we can consider probability measures to correspond to equivalence classes of functions. To make the representation (22) unique we add the additional constraint that $E_0(v_i) = 0$ for every $i \geq 1$, which again will specify a representative of the equivalence class. In that case the normalizing constant $\phi(\boldsymbol{\theta}) = I(P_0|P_{\boldsymbol{\theta}})$ as discussed before. Uniqueness also requires that the functions $v_0, v_1, \ldots, v_k$ are linearly independent *when restricted to the support of* P_0.

With these additional conditions, for each $P_{\boldsymbol{\theta}} \in \mathcal{P}$, $\log dP_{\boldsymbol{\theta}}/dP_0$ corresponds to a unique point $(-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k)$ in $\mathbb{R}^{k+1}$. The set of probability measures thus defines a k-dimensional manifold

$$\mathcal{M} = \{(-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k) : I(P_0|P_{\boldsymbol{\theta}}) < \infty\} \tag{24}$$

embedded in $\mathbb{R}^{k+1}$. This manifold can be projected one-to-one onto its tangent plane at the origin, giving the *natural parameter space*[3]

$$\Theta = \{\boldsymbol{\theta} : I(P_0|P_{\boldsymbol{\theta}}) < \infty\} \tag{25}$$

which is a convex subset of $\mathbb{R}^k$.

The family of normal distributions is a well-known example:

Example 4. Let $\mathcal{Y} = \mathbb{R}$, and let P_0 be the standard normal distribution.

$$P_0(A) = \frac{1}{\sqrt{2\pi}} \int \mathbf{1}_A e^{-y^2/2}\, dy,$$

and define $v_1(y) = y$, $v_2(y) = y^2 - 1$. The representation (19) gives

$$\log \frac{dP_{\boldsymbol{\theta}}}{dP_0} = \theta_1 y + \theta_2(y^2 - 1) - I(P_0|P_{\boldsymbol{\theta}})$$

[2] If $\mathcal{L}$ spans an infinite-dimensional space, then a basis might be impossible to find, even if its existence is implied by the axiom of choice.

[3] This is slightly more restrictive than the usual definition, which only requires the finiteness of $\phi(\theta)$ and not of its particular version $I(P_0|P_{\boldsymbol{\theta}})$.

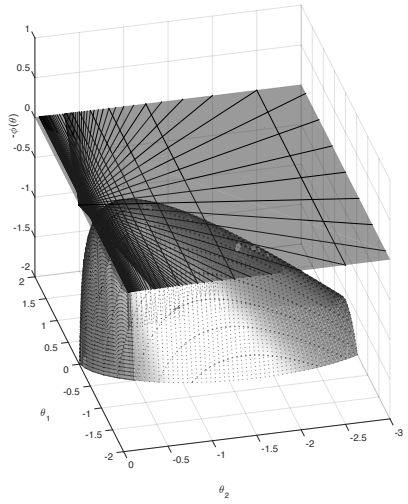


Figure 2: Probability manifold for the Gaussian family, with tangent plane at $P_0 = N(0,1)$. The tangent plane is ruled with coordinate lines corresponding to mean and variance.

where $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_2 < 0\}$. This can be seen to be a Gaussian distribution with mean $\mu = -\theta_1/(2\theta_2)$ and variance $\sigma^2 = -1/(2\theta_2)$.

Example 5. Consider now the setup of Example 3 but with the observation is right-censored at T. This means that if $y > T$ then one actually observes $y = T$. Now

$$P_0((0,y]) = \begin{cases} 1 - e^{-y} & 0 < y < T, \\ 1 & y \geq T. \end{cases}$$

so that the distribution is no longer continuous but has an atom, i.e., point of positive probability, at T.

Now let P_ϑ be an exponential distribution with mean $(1-\vartheta)^{-1}$ also censored at T. Then the log likelihood ratio is

$$\ell = \log \frac{dP_{\theta'}}{dP_0} = \begin{cases} \theta' y + \log(1-\theta') & 0 < y < T, \\ \theta' T & y \geq T. \end{cases}$$

The one-dimensional family generated by ℓ now has natural parameter space $(-\infty, \infty)$, but only P_0 and P_φ represent censored exponential distributions.[4] To model a family of censored exponential distributions, we need to introduce a second function $\delta = \mathbf{1}_{y<T}$. For any exponential distribution censored at T we can now write

$$\log \frac{dP_\theta}{dP_0} = \theta_1 v_1 + \theta_2 v_2 - \phi(\theta_1, \theta_2). \tag{26}$$

[4]Each of the members of the family is a mixture of a truncated exponential distribution and a point mass at T, but the probability of the point mass in most cases is different from that given by censoring.

The canonical representation with $\phi(\theta_1, \theta_2) = I(P_0|P_{\theta_1,\theta_2})$ would require that

$$v_1(y) = y - E_0(y) = y - (1 - e^{-T}) \tag{27}$$
$$v_2(y) = \delta - E_0(\delta) = \delta - (1 - e^{-T}). \tag{28}$$

$$\begin{aligned} \phi(\theta_1, \theta_2) &= I(P_0|P_{\theta_1,\theta_2}) \\ &= \log\left(e^{(\theta_1-1)T} + (\theta_1 - 1)e^{(\theta_2-1)T-\theta_2}\right) \\ &\quad - \log(\theta_1 - 1) + \theta_2 \\ &\quad - (\theta_1 + \theta_2)\left(1 - e^{-T}\right). \end{aligned}$$

Exponential distributions censored at T form a one-dimensional non-linear manifold, defined by

$$\{(\theta_1, \theta_2) : \theta_2 = \log(1 - \theta_1)\},$$

in this two-dimensional exponential family. Such a family is called a *curved exponential family*[1]. Restricted to this submanifold, we have

$$I(P_0|P_{\theta_1,\theta_2}) = -(\theta_1 + \theta_2)(1 - e^{-T}),$$

which in this instance is a *linear* function of the canonical parameters.

3 Geometry of Inference

3.1 Precise Priors

Suppose now that we express our prior uncertainty about the model by a probability measure Π_0 defined on a suitable σ-algebra of subsets of $\mathcal{P}$.

Denote by π_0 the density of Π_0 (considered as a measure on $\mathcal{P}$) with respect to some dominating measure λ. Then if the likelihood is given by (21) and an observation y is observed, Bayes' rule will give the posterior density

$$\pi_y(v) = \frac{\pi_0(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}}))}{\int \pi_0(\vartheta) \exp(\vartheta^\top \mathbf{v}(y) - I(P_0|P_\vartheta))\, d\lambda(\vartheta)}, \tag{29}$$

where $\mathbf{v}(y)$ is the vector $(v_1(y), \ldots, v_k(y))^\top$. If we take the log ratio of posterior to prior, we get

$$\log \frac{d\Pi_y}{d\Pi_0}(v) = \boldsymbol{\theta}^\top \mathbf{v} - I(P_0|P_{\boldsymbol{\theta}}) - \psi(y) \tag{30}$$

where

$$\psi(y) = \log \int \exp\left(\boldsymbol{\theta}^\top \mathbf{v} - I(P_0|P_{\boldsymbol{\theta}})\right)\, d\Pi_0(v). \tag{31}$$

The set of possible posteriors (30) is of the same exponential form as (19) where the roles of parameter and function are reversed.

Let $\mathcal{L}^*$ be the vector space of functions $v^* : \mathcal{P} \to \mathbb{R}$ spanned by

$$v^0 : P \mapsto -I(P_0|P) \quad \text{and} \tag{32}$$

$$v^i : (P_{\theta_1,\ldots,\theta_k}) \mapsto \theta_i \quad i = 1, \ldots, k. \tag{33}$$

For brevity, denote by $\mathbf{v}^*$ the *row* vector $(v^0(P), v^1(P), \ldots, v^k(P))^\top$.

Now given a vector $\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_k)$ of *hyperparameters* we can now define analogously to (19) for any measurable set W of measures in $\mathcal{P}$

$$\Pi_{\boldsymbol{\eta}}(W) = \int \mathbf{1}_W \exp\left(\sum_{i=0}^{k} v^i \eta_i - \psi(\boldsymbol{\eta})\right) d\Pi_0. \tag{34}$$

This will define a probability measure on $\mathcal{P}$ provided that

$$\psi(\boldsymbol{\eta}) = \int e^{\mathbf{v}^* \boldsymbol{\eta}}\, d\Pi_0 < \infty. \tag{35}$$

Definition 4 *The conjugate hyperparameter space* Θ^* *is the set of all* $\boldsymbol{\eta} \in \mathbb{R}^{k+1}$ *such that* (35) *holds.*

Definition 5 *The space of measures* $\mathcal{P}^*$ *conjugate* [5] *to the family* $\mathcal{P}$ *is the set* $\{\Pi_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Theta^*\}$.

By definition, $\mathcal{P}^*$ includes the prior distribution and all possible posteriors (but is generally much larger). Moreover, if a posterior distribution is in $\mathcal{P}^*$, then a proper prior from which it was updated must also be in $\mathcal{P}^*$.

Theorem 4 *If a prior distribution* $\Pi_{\boldsymbol{\eta}}$ *in* $\mathcal{P}^*$ *has hyperparameters*

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_k)$$

then after observing y *the posterior distribution will have hyperparameters*

$$(\eta_0 + 1, \eta_1 + v_i(y), \ldots, \eta_k + v_k(y)).$$

Proof: The density of the prior Π is $d\Pi/d\Pi_0 = \exp(\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta}))$. By Bayes' theorem, the posterior density is

$$\frac{d\Pi_y}{d\Pi_0} = \frac{e^{\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta})} e^{\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}})}}{\int e^{\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta})} e^{\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}})}\, dP_{\boldsymbol{\theta}}} \tag{36}$$

By definition,

$$\mathbf{v}^*(P_{\boldsymbol{\theta}}) = (-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k)$$

so the numerator of (36) becomes $\exp\big(\mathbf{v}^*(\boldsymbol{\eta} + (1, \mathbf{v}(y))\big)$, and the denominator becomes $\psi(\boldsymbol{\eta} + (1, \mathbf{v}(y))$. ∎

[5] This definition is more general than that of Diaconis and Ylvisaker. Their construction would follow from using a (possibly improper) Lebesgue prior for Π_0.

The transformation from prior to posterior by an observation y can be represented in Θ^* as a translation by the vector $(1, v_1(y), \ldots, v_k(y))$ is a translation by the vector $y^* - v_0$. Note that the translation is the same for all priors. Even improper priors can be accommodated by going outside Θ^*.

3.2 Imprecise Priors

Because the translation in Θ^* is the same for all priors (proper or improper), one can update a *set* of priors simply by translating the whole set. This provides a convenient way of representing updating of imprecise priors, as the set of hyperparameters for the posteriors is congruent to the set of prior hyperparameters.

It is often of interest to predict the value of some future observation, by the posterior expectation of a random variable $v \in \mathcal{L}$. With a precise prior distribution Π_0, this would be computed as

$$\hat{v} = \int \int v(z)\, dP_\theta(z)\, d\Pi_y(\theta).$$

If instead of a precise prior, we have a set of priors $\boldsymbol{\Pi}_0$ leading to a set of posteriors $\boldsymbol{\Pi}_y$, then we compute *lower and upper previsions* as

$$\underline{v} = \inf_{\Pi \in \boldsymbol{\Pi}_y} \int \int v(z)\, dP_\theta(z)\, d\Pi(\theta) \tag{37}$$

$$\overline{v} = \sup_{\Pi \in \boldsymbol{\Pi}_y} \int \int v(z)\, dP_\theta(z)\, d\Pi(\theta). \tag{38}$$

If the conjugate family is of the type discussed by Diaconis and Ylvisaker, and if $v \in \mathcal{L}$, then the sets of constant predictive expectation $\hat{v}$ form hyperplanes in $\mathcal{L}^*$ that intersect in a subspace containing the improper Lebesgue prior. In this case the lower and upper previsions (37) and (38) are given by the supporting hyperplanes of the convex hull of the posterior set, which thus can, without loss of generality, be taken to be convex. If the prior set intersects all of these diverging hyperplanes, then the prior prediction is vacuous. As data are observed, the prior set is shifted such that it no longer intersects all the hyperplanes, and non-vacuous prediction can be made.

Definition 6 *A set of priors will be said to have the Benavoli-Zaffalon (BZ) property relative to the function* v *if* $\underline{v} > \inf v$ *and* $\overline{v} < \sup v$ *in* (37) *and* (38) *for some observation* y*, but* $\underline{v} = \inf v$ *and* $\overline{v} = \sup v$ *when* $\boldsymbol{\Pi}_y$ *is replaced by the prior set* $\boldsymbol{\Pi}_0$.

Example 6. Consider the setup in Example 1. For Π_0 take a logistic distribution of θ (which is equivalent to a uniform on $p = (1 + \exp(-\theta))^{-1}$):

$$\frac{d\Pi_0}{d\lambda}(\theta) = \frac{e^\theta}{(1 + e^{-\theta})^2}. \tag{39}$$

Define $v_0^*, vii_1^* \in \mathcal{L}^*$ by

$$v_0^*(\theta) = -\log\cosh(\theta/2)$$
$$v_1^*(\theta) = \theta/2$$

It can be shown that

$$\Theta^* = \{\eta_1 f_1^* + \eta_0 f_2^* : \ |\eta_1| < 1 + \eta_2/2\}.$$

Plotting the basis vector v_0^* horizontally and v_1^* vertically, the set of proper priors and posteriors is defined by the wedge-shaped region in Figure 3. The update rule for a single binary observation y can be expressed as

$$\eta_0 \mapsto \eta_0 + 1 \tag{40}$$
$$\eta_1 \mapsto \eta_1 + y - \tfrac{1}{2} \tag{41}$$

Given any point representing a prior, the posterior after a single observation is obtained by moving one step to the right, a half-step up for a success, a half-step down for a failure. A sequence of independent observations then traces a path in the hyperparameter space.

Sets of constant prediction of $v = y - \frac{1}{2}$ form rays emanating from $\eta_0 = -2, \eta_1 = 0$ (Figure 3). (The intersection of these rays is not in Θ^* but represents an improper prior.) From this picture, one can visualize which sets of priors will have the Benavoli-Zaffalon property. For example, Walley's imprecise beta model (IBM) gives a prior set corresponding to

$$\{(\eta_0, \eta_1) : \eta_0 = s, |\eta_1| < s/2\}, \tag{42}$$

where s is taken to be 1 or 2. The prior predictions are thus $\underline{v} = 0$ and $\overline{v} = 1$. After taking observations, the prior set has moved such that it is contained in a narrow cone of rays, leading to informative upper and lower previsions.

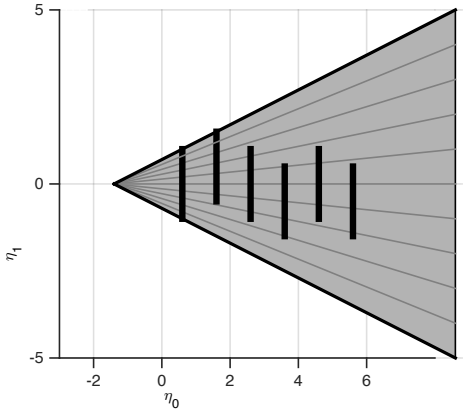


Figure 3: Path of sets of posteriors from IDM

Example 7. If the data are $N(\mu, 1)$, then the conjugate prior family would be $N(\nu, \sigma^2)$ which can be reparametrized in canonical exponential form by $\eta_0 = 1/2\sigma^2$ and $\eta_1 = \nu/\sigma^2$. If we choose $\Pi_0 \sim N(0, 1)$ then $\Theta^* = (-1, \infty) \times (-\infty, \infty)$. Sets of constant predictive expectation are again rays emanating from $(-1, 0)$ (Figure 4). Note that η_0 again represents a concentration parameter. Unlike the case of the IBM, fixing the set of priors by fixing the concentration parameter does not allow for learning from data, as the interval of posterior predictions remains infinite. Benavoli and Zaffalon [3] suggested using a set of priors which in the present parametrization is the rectangular region in Figure 4 which satisfies the BZ-property.

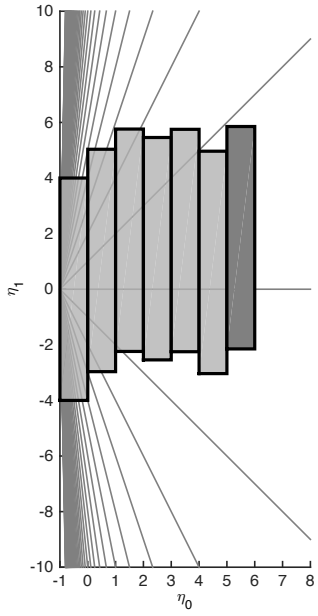


Figure 4: Set of posteriors from Normal distribution, using prior set suggested by Benavoli

Example 8. Let the model space be as in Example 6 but define Π_0 as a Gaussian distribution on Θ.

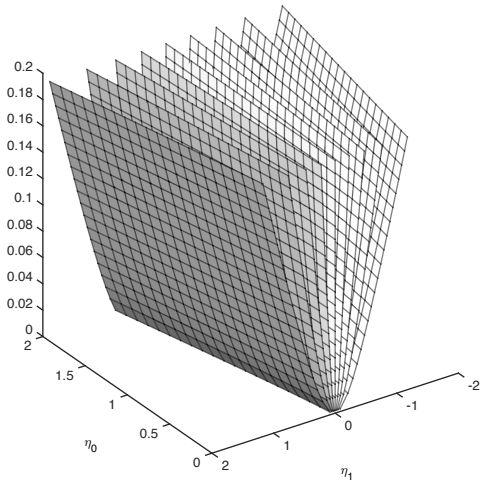


Figure 5: Contour sheets showing sets of constant predictions for the logit-normal model.

The same construction applies, but in this case the conjugate family is not conjugate in the sense of Diaconis and Ylvisaker. The 2-dimensional exponential family created from a $N(\mu, \sigma^2)$ prior is spanned by θ and $\cosh(\theta/2)$. Nonetheless, similar arguments still allow for learning from data. If we start with a set of $N(\mu, \sigma^2)$ priors for various σ^2, we obtain a 3-dimensional family of posteriors spanned by $\cosh(\theta/2)$, θ and $-\theta^2$. The update rules for η_0 and η_1 are the same as in (41), but η_2, the coefficient of $-\theta^2$, is not changed.

There seems to be no explicit formula for the normalizing constant ψ, nor for the predictive expectations. Nonetheless, such quantities can be computed numerically. As shown in Figure 8, the level sets of predictive expectations appear as a set of almost flat sheets pinched together at the origin. The limiting case $\eta_2 = 0$ is equivalent to the conjugate family in Example 6. The path traced by a sequence of observations is as in Figure 3, raised by $\eta_2 = 1/(2\sigma^2)$ in the prior distribution. A set of priors with the Benavoli-Zaffalon property can be obtained by including in its boundary a set of the type in (42).

Example 9. Consider now the censored exponential model of example (5). While the "natural" parameter space is two-dimensional, we are only concerned with models on the one-dimensional manifold $\theta_2 = \log(1 + \theta_1)$. We thus take as Π_0 the singular distribution concentrated on this manifold such that that θ_1 has an exponential distribution with mean 1. (Note that in this case the dominating measure λ is not Lebesgue measure.) The conjugate space of posteriors then takes the form

$$\log \frac{dP_0}{d\Pi_{\eta_1,\eta_2}} = \theta_1\eta_1 + \theta_2\eta_2 - \psi(\eta_1, \eta_2)$$

where

$$\psi(\eta_1, \eta_2) = (\eta_2 + 1)\log(\eta_1 + 1) - \Gamma(\eta_2 + 1)$$

The natural hyperparameter space is $\{\eta_1 > -1, \eta_2 > -1\}$. In this case the family is only two-dimensional because of the linear dependence between ϕ and (θ_1, θ_2).

The Bayesian updating rule is

$$\eta_1 \mapsto \eta_1 + y$$
$$\eta_2 \mapsto \eta_2 + \delta,$$

moving to the right by the observed lifetime and one step up if the lifetime is not censored. This setup still works if we allow T itself to vary with time. The hyperparameter keeps moving right while the individual is alive (i.e., censored) and then jumps up one step once a death is observed.

The posterior predictive expectation of the *uncensored* lifetime is $(\eta_1 + 1)/\eta_2$. To create an imprecise inference, we can start with the hyperparameter set $\{\eta_2 > 0, \eta_1 + \eta_2\} = 0$. Initially, the predictive lower prevision is 0, and the predictive upper prevision is ∞. If an individual is observed to be alive at time y, the lower prevision rises to y, but the upper prevision remains at ∞. Once the individual is observed to die at y, the upper prevision drops to $1 + y$ and the lower prevision drops to $y/2$. If one observes a set of independent lifetimes, then this process compounds. If t is the total of observed lifetimes and d is the total number of observed deaths, then the lower prevision is $t/(d + 1)$ and the upper prevision is $(t + 1)/d$. This set of priors again has the Benavoli-Zaffalon property (Figure 3.2).

4 Conclusions

In this paper we have shown how an exponential family of probability measures is generated by postulating a null distribution and a set of inferential functions. If the set of functions is k-dimensional, then the family of probability measures forms a k-dimensional manifold embedded in $k + 1$-dimensional Euclidean space. This manifold can be uniquely projected onto a tangent plane whose coordinates parametrize the model. If a prior distribution is defined on the set of probability distribution, then the above development can be repeated with the parametric functions, thus giving an exponential family that includes all possible posteriors.

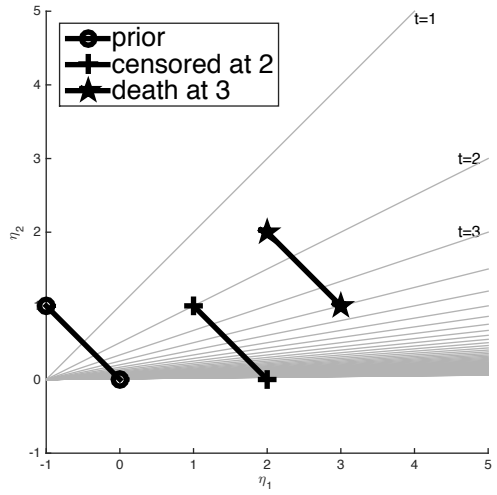


Figure 6: Imprecise updating of censored exponential survival times, showing the prior set, the set after an observation censored at time 2, and the set after observing a death after another time unit. The rays are level sets for predicted uncensored lifetimes.

This family can again be projected onto a tangent space of hyperparameters.

In this representation, Bayesian updating of a hyperparameter is expressed as a translation by a data-dependent vector. This same translation can be applied to a set of hyperparameters, demonstrating the updating of imprecise priors to imprecise posteriors. The geometric perspective allows one to see when a set of priors would enjoy the Benavoli-Zaffalon property of near vacuous priors that allow for learning from data.

This paper concentrates on the linear aspects of the space of measures, and does not further explore the metric aspects of the geometry implied by the Kullback-Leibler information measure. These topics will be examined in future papers.

Acknowledgments

The author is grateful to several anonymous referees for their detailed comments that helped to improve the paper. This research has been supported by grants from the Natural Science and Engineering Research Council of Canada and from the Office of Vice-President Research at the University of Saskatchewan. Part of this research was done while the author was the Alan Richards Mathematics Fellow at Grey College, Durham University.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka, *Methods of Information Geometry*, American Mathematical Society, 2000

[2] Osama Bataineh, *Imprecise Probability Models for Logistic Regression.* PhD Thesis, University of Saskatchewan, 2012.

[3] Alessio Benavoli and Marco Zaffalon, A model of prior ignorance for inferences in the one-parameter exponential family, *Journal of Statistical Planning and Inference*, 2012, 1960–1979.

[4] M.G. Bickis, The imprecise logit-normal model and its application to estimating hazard functions, *Journal of Statistical Theory and Practice* **3** (2009), 183–195.

[5] P. Diaconis and D. Ylvisaker, Conjugate priors for exponential families. *Ann. Statist.* **7** (1979), 269–281.

[6] Bruno de Finetti, *Theory of probability*, Wiley, New York, 1974.

[7] Solomon Kullback, *Information Theory and Statistics*, Wiley,1959.

[8] Chel Hee Lee, *Imprecise Prior for Imprecise Inference on Poisson Sampling Model.* PhD Thesis, University of Saskatchewan, 2014.

[9] Isaac Levi, *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Change*, MIT Press, 1980.

[10] Deborah Mayo, *Error and the Growth of Experimental Knowledge*, University of Chicago Press, 1996.

[11] Erik Quaeghebeur, *Learning from samples using coherent lower previsions.* PhD thesis, University of Ghent, 2009.

[12] Erik Quaeghebeur and Gert de Cooman, *Imprecise probability models for inference in exponential families*, ISIPTA '05: Proc. 4th Int. Symp. on Imprecise Probabilities and Their Applications (Fabio G. Cozman, Robert Nau, and Teddy Seidenfeld, eds.), July 2005, pp. 287–296.

[13] Matthias C. M. Troffaes and Gert de Cooman, *Lower Previsions*, Wiley, 2014.

[14] Peter Walley, *Statistical reasoning with imprecise probabilities*, Chapman and Hall, London, 1991.

[15] Peter Walley, *Inferences from multinomial data: Learning about a bag of marbles*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 3–34.

How to Choose Among Choice Functions

Seamus Bradley
Munich Centre for Mathematical Philosophy
LMU, Munich
seamus.bradley@lmu.de

Abstract

If one models an agent's degrees of belief by a set of probabilities, how should that agent's choices be constrained? In other words, what choice function should the agent use? This paper summarises some suggestions, and outlines a collection of properties of choice functions that can distinguish between different functions.

Keywords. decision making, choice functions, sets of probabilities

1 Basics

This first section outlines some basic formalism. We have a finite set of states Ω and we take the set of events to be the power set of that: 2^Ω.

We define a *probability function* over 2^Ω as a function $\mathbf{pr} : 2^\Omega \to \mathbb{R}$ with the following properties:

- $\mathbf{pr}(\emptyset) = 0$ and $\mathbf{pr}(\Omega) = 1$
- $\mathbf{pr}(\emptyset) \leq \mathbf{pr}(X) \leq \mathbf{pr}(\Omega)$ for all X
- $\mathbf{pr}(X \cup Y) + \mathbf{pr}(X \cap Y) = \mathbf{pr}(X) + \mathbf{pr}(Y)$ for all $X, Y \subseteq \Omega$

An agent's degrees of belief are represented by a set of probability functions, $\mathcal{P}$. Call this set your *representor*. With a little abuse of notation, we can define a function $\mathcal{P}(H)$ which maps event H to the set of values that the probability functions in $\mathcal{P}$ give to H. So $\mathcal{P}(H) = \{\mathbf{pr}(H) : \mathbf{pr} \in \mathcal{P}\}$. We can then define $\overline{\mathcal{P}}(H)$ and $\underline{\mathcal{P}}(H)$ as the minimal and maximal values that the probabilities in $\mathcal{P}$ assign to H.[1] These "summary functions" give us objects that somehow represent the belief and are easier to handle than the full set of probability functions. It is sometimes convenient to think of each $\mathbf{pr} \in \mathcal{P}$ as a member of a "credal committee" who collectively represent your opinions and make your choices.

The objects of choice are *gambles*: real valued functions from the set of states. A gamble φ wins $\varphi(w)$ if w turns out to be the true state. Let's say we have acts φ and ψ. Say we have some kind of random device that outputs a 1 with probability p and a 0 otherwise. $p\varphi + (1-p)\psi$ is the act "get whatever φ gets you with probability p, get whatever ψ gets you otherwise". If A is a set of acts, $pA + (1-p)\psi$ is the set of acts of the form $p\varphi + (1-p)\psi$ for $\varphi \in A$. Let A^* be the set of mixed acts over A. Note that the gambles have real valued outcomes, so I am implicitly assuming that your utility function is precise. I use "act" and "gamble" interchangeably.

For probability function $\mathbf{pr}$ we define its *expectation* $\mathrm{E}_{\mathbf{pr}}(\varphi) = \sum_{w\in\Omega} \mathbf{pr}(w)\varphi(w)$. That is, the expectation – or expected value – for an act is a weighted sum of what the act gets you in each state, weighted by how likely $\mathbf{pr}$ considers that state. Orthodox decision making is aimed at maximising this expected value.

We can define an imprecise expectation by taking the set of the expectations for each $\mathbf{pr} \in \mathcal{P}$. That is, $\mathcal{E}_{\mathcal{P}}(\varphi) = \{\mathrm{E}_{\mathbf{pr}}(\varphi), \mathbf{pr} \in \mathcal{P}\}$. We often drop the subscript and just talk about $\mathcal{E}$ when it is obvious what $\mathcal{P}$ is at issue. We can define $\underline{\mathcal{E}}(\varphi)$ and $\overline{\mathcal{E}}(\varphi)$ as the smallest and largest expectations assigned to φ by members of $\mathcal{P}$. How are we to choose with imprecise expectations? The first thing to note is that we can't simply "choose the biggest". The $\mathcal{E}$s for the various acts will typically be sets of numbers: there's no obvious sense in which one collection of numbers is *bigger* than the other. The sets can overlap. So we need to think a little more carefully about what imprecise choice involves.

We consider two kinds of gambles: those whose outcome depends on the throw of a fair die, where the probability of its landing even is fixed $\mathcal{P}(E) = \{1/2\}$; and those whose outcome depends on the toss of a coin of unknown bias, where the probability of the

[1] More properly, these should be the greatest lower bound and the least upper bound, since we aren't sure that the extrema are attained. Nothing hangs on this.

coin landing heads is unknown $\mathcal{P}(H) = [0, 1]$.

The main object of study in this paper will be various forms of *choice function*. A choice function will take a set of available acts and output a subset of choiceworthy acts. A choice function is a function $\mathcal{C}\colon 2^{\mathcal{A}} \to 2^{\mathcal{A}}$ such that for all $A \subset \mathcal{A}$ we have $\mathcal{C}(A) \subseteq A$ and $\mathcal{C}(\mathcal{C}(A)) = \mathcal{C}(A)$. That is, the function outputs a subset of the acts available: it would be unhelpful if the choice function gave you the advice to perform some act that wasn't available to you. We also require that the choice function is stable in a certain sense. That is, applying the function a second time has no effect. Call the set that the choice function outputs – $\mathcal{C}(A)$ – the *choice set*. The majority of this paper will be about what properties we can impose on choice functions, and which of those properties it is reasonable to demand in the imprecise case. We will explore some well-known imprecise choice functions and discover which properties they do or do not satisfy.

$\mathcal{C}(A)$ is meant to represent or encode what it is that rationality requires of you when you must make a choice among the members of A. There are many ways of interpreting $\mathcal{C}(A)$. A "Strong" interpretation would say that acts in $\mathcal{C}(A)$ are all equally the best act: there is nothing to choose between the acts in $\mathcal{C}(A)$ and you should be equally happy to take any of them. $\varphi \in \mathcal{C}(A)$ is here considered an endorsement of act φ. A weaker interpretation might be to say that all the acts in $\mathcal{C}(A)$ are better[2] than the acts not in $\mathcal{C}(A)$. This interpretation does not preclude there being strict preference between the acts in $\mathcal{C}(A)$. $\varphi \in \mathcal{C}(A)$ isn't now such a strong endorsement of φ; but $\psi \notin \mathcal{C}(A)$ is still considered a real flaw in ψ. Consider the "vegetarian choice function" that rejects all menu items containing meat. It is not the case that all elements that survive this rejection criterion are necessarily on a par.

In short, we can think of standards of rationality as giving sufficient conditions for being acceptable, or we can think of the standards of rationality as giving necessary conditions for being acceptable. The former accords with the positive understanding of rationality: endorsing elements in $\mathcal{C}(A)$. The latter accords with the negative understanding of rationality: those elements outside $\mathcal{C}(A)$ are advised against.

Consider the *reject set*[3] for a given choice rule: $R(A) = A \backslash \mathcal{C}(A)$. $R(A)$ is the set of options that the choice rule rejects. The weak interpretation of the choice function amounts to endorsing the rejection of elements of $R(A)$, while the strong interpretation amounts to endorsing the choice of elements in $\mathcal{C}(A)$. Call these reject-R and endorse-$\mathcal{C}$, respectively. The aim of this paper is to suggest that there might not be a strong (endorse-$\mathcal{C}$) choice function for IP decision making, and that we might have to make do with weak (reject-R) choice functions. The contribution of the paper is primarily philosophical, rather than mathematical. I further want to present a case for preferring the "Maximality" choice rule to the "E-admissibility" choice rule, and while at least some of the properties of E-admissibility that I mention are already known, I don't know of anyone who turns them into an argument against E-admissibility. Finally, I mention a new "regret-based" choice rule, although I don't have space to do much more than present it.

2 How to Constrain Choice Functions

What does a reasonable imprecise choice rule look like? There are many places in the literature where enterprises like this have been developed. There are a great many ways we could approach the question of how best to settle on an imprecise decision rule. I survey some ways here.

I take inspiration from the classic discussions of choice under complete ignorance, such as Milnor's important "Games Against Nature" [17] and Chapter 13 Luce and Raiffa's classic textbook [16]. I also look to social choice theory: if we think of each probability in your representor as a member of a credal committee that has to vote on what you should do, then the parallel between imprecise decision and social choice becomes clear. Here I will draw on Arrow's theorem [8] and the work of Amartya Sen [26, 24].

There are two ways one might frame the discussion: in terms of an ordering over the acts (Arrow, Milnor), or in terms of a choice rule (Luce and Raiffa, Sen). I will talk in terms of choice rules, but we will see that relations will also play an important role.

There are several ways we could describe conditions on the choice function. One is just to put conditions on the functional form of the choice function. That is, we could impose intuitive conditions on the function with respect to how it interacts with unions and intersections of sets of acts. For example consider the condition we built into the definition of choice function: $\mathcal{C}(\mathcal{C}(A)) = \mathcal{C}(A)$. This is a property that constrains what kind of functions count as choice rules.

There is another way we might want to impose constraints on reasonable choice functions. This is by

[2] Note that such "betterness" needn't determine an order on the acts. Consider the case where φ is better than ψ just in case that φ doesn't have some obvious flaw that ψ does. A choice rule that returned the set of acts without this flaw would be an example of this weaker sort of choice rule.

[3] Note that a reject set in this sense is not the same as what [19] call a "reject statement".

restricting various kinds of relation associated with the choice function.

For this, we need some definitions. For reflexive relation $\succeq$, let $\sim$ and $\succ$ be its symmetric and irreflexive parts respectively. A choice function $\mathcal{C}$ *pairwise satisfies* a relation $\succeq$ when, for all $\varphi, \psi \in \mathcal{A}$:

- If $\varphi \succeq \psi$ then $\varphi \in \mathcal{C}(\{\varphi, \psi\})$
- If $\varphi \succ \psi$ then $\{\varphi\} = \mathcal{C}(\{\varphi, \psi\})$

If $\succeq$ is understood as preference relation then pairwise satisfying a relation means never picking a dispreferred option in pairwise choices. A choice function $\mathcal{C}$ *satisfies* a relation $\succeq$ when, for all $\varphi, \psi \in A \subseteq \mathcal{A}$:

- If $\varphi \succ \psi$ then $\psi \notin \mathcal{C}(A)$
- If $\varphi \sim \psi$ then $\varphi \in \mathcal{C}(A) \Leftrightarrow \psi \in \mathcal{C}(A)$

Satisfying a relation can be understood as never picking a dispreferred option in *any* choice. We could then constrain reasonable choice by demanding that the choice function (pairwise) satisfies some particular relation defined on the acts. If $\mathcal{C}(A)$ is nonempty for all nonempty A[4] and satisfies $\succeq$ then it pairwise satisfies it, but the converse need not be true.

A relation can also determine a kind of choice function. The *maximal set* for a relation $\succeq$ is $\mathcal{M}_\succeq$:

$$\mathcal{M}_\succeq(A) = \{\varphi \in A : \neg\exists \psi \in A, \psi \succ \varphi\}$$

Interpreting the "$\succeq$" as a relation of preference, this $\mathcal{M}_\succeq$ is the set of acts that aren't strictly dispreferred to anything else in the set.[5] Here are some facts about $\mathcal{M}_\succeq$.

(i) $\mathcal{M}_\succeq$ is a choice function
(ii) $\mathcal{M}_\succeq$ pairwise satisfies $\succeq$
(iii) If $\succeq$ is acyclic[6] on A where A is finite then $\mathcal{M}_\succeq(A)$ is non-empty
(iv) If $\succeq$ is transitive, then $\mathcal{M}_\succeq$ satisfies $\succeq$.

These are proved in the appendix (Theorem 2).

Going the other way, a choice function determines a relation by

$$\varphi \succeq_\mathcal{C} \psi \Leftrightarrow \varphi \in \mathcal{C}(\{\varphi, \psi\})$$

$\mathcal{C}$ pairwise satisfies $\succeq_\mathcal{C}$. Under certain conditions $\mathcal{C}$ satisfies $\succeq_\mathcal{C}$ [25]. Say that $\mathcal{C}$ is determined by pairwise comparisons when this is the case.

Call a choice rule $\mathcal{C}$ *more discriminating* than $\mathcal{C}'$ when $\mathcal{C}(A) \subseteq \mathcal{C}'(A)$ for all A. $\mathcal{M}_\succeq$ is the least discriminating choice function that satisfies $\succeq$. That is, if $\mathcal{C}$ satisfies $\succeq$ then $\mathcal{C}(A) \subseteq \mathcal{M}_\succeq(A)$ for all A. This is also proved in the appendix (Theorem 3). We can think of relations as pairs of elements of the domain of the relation,[7] so it makes sense to talk about the intersection and union of relations, and of one relation being a subset of another.

Sometimes we will talk about the relation generated by a function F into an ordered set (normally the reals), $\succeq_F$. We understand this to be the relation such that $\varphi \succeq_F \psi$ iff $F(\varphi) \geq F(\psi)$. For instance, $\varphi \succeq_{\mathrm{E}_\mathbf{pr}} \psi$ iff $\mathrm{E}_\mathbf{pr}(\varphi) \geq \mathrm{E}_\mathbf{pr}(\psi)$. We will sometimes write $\mathcal{M}_F$ where more properly we should write $\mathcal{M}_{\succeq_F}$. For example, when your credences are precise, your choice rule is $\mathcal{M}_{\mathrm{E}_\mathbf{pr}}$. That is, you choose among the things that do best by the criterion of expected value. Note that $\varphi \in \mathcal{M}_{\mathrm{E}_\mathbf{pr}}(A)$ means (by definition) that there does not exist a $\psi \in A$ such that $\psi \succ_{\mathrm{E}_\mathbf{pr}} \varphi$. This means that for all $\psi \in A$, $\mathrm{E}_\mathbf{pr}(\varphi) \geq \mathrm{E}_\mathbf{pr}(\psi)$. Which is just to say that φ maximises expectation.

What if, instead of talking about maximality, we talked about *optimality*? The *optimal set* for a relation $\succeq$ is:

$$\mathrm{Opt}_\succeq(A) = \{\varphi \in A : \forall \psi \in A, \varphi \succeq \psi\}$$

What we will find is that optimality – which is stronger than maximality – is too strong a property. That is, $\mathrm{Opt}_\succeq$ is often empty. Consider the set $\{\varphi, \psi\}$ where no relation holds between the two options. For this set, there are no optimal acts – although both acts are maximal in the sense of $\mathcal{M}_\succeq$. If the relation is complete, reflexive and acyclic then $\mathrm{Opt}_\succeq$ is nonempty [26, p. 55]. When $\mathrm{Opt}_\succeq(A) \neq \emptyset$, and $\succeq$ is transitive then $\mathrm{Opt}_\succeq(A) = \mathcal{M}_\succeq(A)$ (Theorem 4). This means that talking about optimality is superfluous. Maximality is the more interesting concept in general. The two happen to coincide for complete, transitive relations but when we have incomplete relations, optimality can be empty while maximality won't be. See [27] for more on the relationship between optimality and maximality (in particular, theorems 5.2 and 5.3).

In summary, we want to analyse what sort of choice rule makes sense for imprecise decision. We are going to proceed by imposing certain intuitive constraints on choice and showing that certain decision rules violate these principles. The principles will come in two flavours: restrictions on the functional form of $\mathcal{C}$, and relations that $\mathcal{C}$ must satisfy.

One might think that given the material I'm taking inspiration from, I would be aiming at a representation theorem (Luce and Raiffa, Milnor) or an impossibility theorem (Arrow, Sen). I am doing neither. I don't think the conditions I discuss below are enough to

[4] We will call this property Decisive later.

[5] [3] makes a distinction between maximality (as defined above) and strong maximality. The distinction won't matter in the current project since the relations I discuss are transitive, and thus the two concepts overlap (see his Theorem 2).

[6] Meaning for all $\varphi_1 \ldots \varphi_n$, if $\varphi_1 \succ \varphi_2, \ldots \varphi_{n-1} \succ \varphi_n$ then $\varphi_n \not\succ \varphi_1$.

[7] That is, define $X_\succeq \subseteq \mathcal{A} \times \mathcal{A}$ by: $(\varphi, \psi) \in X_\succeq$ iff $\varphi \succeq \psi$.

generate an impossibility, nor do I think they are sufficient for any interesting kind of representation (although the extremely general theorems of [7] or [5, 4] might apply). Some of the decision rules I discuss have been characterised. For example, E-admissibility [23]. And using $\mathcal{M}_{\succeq}$ and axiomatising $\succeq$, Maximality [22]. Perhaps also Gamma-maximin [11]. My main focus is not on impossibility or representation, but on what we can say about *rational constraints on choice*. Note that in what follows I am presupposing some expected utility evaluations of the gambles.

3 Properties of Choice Functions

3.1 Dominance Principles

Consider the choice function defined by

$$\mathcal{C}_{\text{ID}}(A) = \{\varphi \in A : \forall \psi \in A \ \underline{\mathcal{E}}(\varphi) \geq \overline{\mathcal{E}}(\psi)\}$$

This is a decision rule that Henry Kyburg [13] discussed. He calls it "Principle III".[8] It has also been called "Interval Dominance". Unfortunately, $\mathcal{C}_{\text{ID}}$ is often empty. A choice rule that fails to give us advice is not particularly helpful. This suggests a property of choices rules that we might like to endorse.

Decisiveness: If $\mathcal{C}(A) = \emptyset$ then $A = \emptyset$.

But consider the set of gambles that consists of the set of gambles $f_n = n$ for all natural n. Or consider $g_n = -1/n$ for all natural n. Arguably, no act in either set is best, since there's always a larger n (and thus a smaller loss). I will focus my attention on closed and bounded – often finite – sets of gambles.[9]

Despite failing as a choice rule, we can use this ID idea to further restrict reasonable choice rules: when some act does interval dominate all others, then the dominating act should be in the choice set. Define the relation $\varphi \succ_{\text{ID}} \psi$ iff $\underline{\mathcal{E}}(\varphi) \geq \overline{\mathcal{E}}(\psi)$.[10] This gives us another core condition.

Interval Dominance: $\mathcal{C}$ satisfies $\succ_{\text{ID}}$

$\succ_{\text{ID}}$ is transitive and thus acyclic, so $\mathcal{M}_{\succ_{\text{ID}}}$ is decisive. Often $\succ_{\text{ID}}$ is empty, so this condition will put no restrictions on choice (i.e. $\mathcal{M}_{\succ_{\text{ID}}}(A) = A$). However, when $\mathcal{C}_{\text{ID}}$ is not empty, the restrictions it puts on choice are reasonable.

There is a stronger dominance property we can impose on our choice rule. Imagine if every member of the credal committee thought that $\varphi \succ_{\text{E}_{\mathbf{pr}}} \psi$. Surely in such a case, your choice rule should respect this unanimity. Let's consider the relation of dominance, $\succeq_{\text{Dom}}$, as a relation that we want our choice rule to satisfy. Define:

$$\succeq_{\text{Dom}} = \bigcap_{\mathcal{P}} \succeq_{\text{E}_{\mathbf{pr}}}$$

That is, the relation of dominance is the intersection of all the relations of higher expectation. φ dominates ψ if and only if every relation of expectation (in your representor) ranks φ and least as high as ψ. $\varphi \sim_{\text{Dom}} \psi$ means that the gambles have the same expectation for each **pr**. One sometimes considers the logically stronger (thus less constraining) relation of strict dominance, which amounts to the existence of an everywhere positive gamble ε such that $\varphi \succeq_{\text{Dom}} \psi + \varepsilon$, or uniform strict dominance where ε is also constant. Since I think even weak dominance (as captured by $\succeq_{\text{Dom}}$) is enough to make an act unchoiceworthy, I won't say more about this subtlety.

This motivates another important desideratum for imprecise choice.

Non-domination: $\mathcal{C}$ satisfies $\succeq_{\text{Dom}}$

Note that this is a stronger condition than Interval Dominance. That is, whenever φ interval dominates ψ, φ dominates ψ. Put another way, $\succeq_{\text{Dom}} \supseteq \succ_{ID}$.

This expectation-dominance relation also subsumes another kind of dominance, namely *state-wise dominance*. φ state-wise dominates ψ if, for every $w \in \Omega$, $\varphi(w) \geq \psi(w)$. Clearly this entails that $\varphi \succeq_{\text{Dom}} \psi$.

3.2 Contraction Consistency

Consider the following scenario. You go to a restaurant and see that the menu consists of Fish, Steak or Chicken. You decide on Chicken. The waiter comes to take your order and tells you there is no more Fish. So you decide to have the Steak. This story seems a little odd. Why should the availability of an option you don't choose cause a switch in choice like the one exhibited in the move from Chicken to Steak? It seems like a reasonable choice rule should be somewhat consistent under various kinds of expansion or contraction of the option set. This motivates the following principle:

Contraction Consistency: $C(A \cup B) \subseteq \mathcal{C}(A) \cup C(B)$

[8] In response to Teddy Seidenfeld's comments (pp. 259–61), Kyburg changes his mind (p. 271). We will discuss this in due course.

[9] These restrictions are made for convenience, rather than because more general sets of gambles, or more general spaces of gambles (infinite dimensional, non-Archimedean, etc) are not amenable to study [31, 1].

[10] Note this is defined directly as an irreflexive relation, since it doesn't lend itself to having a reflexive part. But $\varphi \succ_{\text{ID}} \psi$ and $\psi \succ_{\text{ID}} \varphi$ implies φ and ψ have the same precise expectation. So the second condition of the definition of "satisfies" is still reasonable in this odd case.

This rule is more normally seen in one of these equivalent forms:

$$\text{If } \varphi \in \mathcal{C}(A), B \subseteq A, \varphi \in B \text{ then } \varphi \in \mathcal{C}(B) \quad (1)$$
$$\text{If } \varphi \notin \mathcal{C}(B), B \subseteq A, \varphi \in B \text{ then } \varphi \notin \mathcal{C}(A) \quad (2)$$

So in the preceding story, $\mathcal{C}(S, C, F) = C$ but $\mathcal{C}(S, C) = S$. This violates the above property. This property is also known as Sen's alpha condition [26, 24]. I am following [8] in calling it "contraction consistency", but it also somewhat restricts expansion of the option set. Luce and Raiffa have a version of (2) as their Axiom 7.

There is a property that is slightly stronger than Contraction Consistency that is known as Path Independence:

Path Independence: $\mathcal{C}(A \cup B) = \mathcal{C}(\mathcal{C}(A) \cup \mathcal{C}(B))$

It is obvious that this entails Contraction Consistency since $\mathcal{C}(X) \subseteq X$ for all X. In fact, Path Independence is equivalent to Contraction Consistency and the property that Sen [26] calls "epsilon":

If $A \subset B$ then it is not the case that $\mathcal{C}(B) \subset \mathcal{C}(A)$

See [26, p. 69] for a proof.

3.3 Independence

We can cash out independence as:

Independence: $\mathcal{C}(pA + (1-p)\varphi) = p\mathcal{C}(A) + (1-p)\varphi$

Perhaps the best way to understand independence is with an example.

> Example 1: I am going to ask you to choose c or d. Then I'm going to roll a fair die and flip a coin of unknown bias. If the die lands even, you gain £6 if $\neg H$, nothing otherwise. If the die lands odd, c and d pay out as set out here:
>
> - c: Gain £10 if H, nothing otherwise
> - d: Gain £2 if H, £8 otherwise

The idea is that since what you choose – c or d – doesn't make a difference if the die lands even, then you should choose in order to get the better of the options when it matters (in the odd branch of the game). One can further justify independence in a sequential choice setting: agents who violate independence pay to avoid free information [20].

3.4 Union Consistency

Recall that Contraction Consistency puts a sort of "upper bound" on $\mathcal{C}(A \cup B)$ by requiring that it be a subset of $\mathcal{C}(A) \cup \mathcal{C}(B)$. Union Consistency puts a *lower* bound on $\mathcal{C}(A \cup B)$.

Union consistency: $\mathcal{C}(A) \cap \mathcal{C}(B) \subseteq \mathcal{C}(A \cup B)$

This is Sen's gamma condition. It is sometimes seen in this equivalent form:

$$\text{If } \varphi \in \mathcal{C}(A), \varphi \in \mathcal{C}(B) \text{ then } \varphi \in \mathcal{C}(A \cup B) \quad (3)$$

The motivation here is that if you would choose Steak out of Steak or Fish, and you'd choose Steak out of Steak or Chicken, then you should choose Steak when all three options are on the menu.

3.5 Other Properties of Choice

Let's consider some properties whose violation I don't consider a flaw at all.

The first property appears in many contexts. Understanding why I think imprecise choice rules should be allowed to violate this property will point to an important difference between imprecise choice and precise choice. I shall call this property "all-or-nothing expansion consistency". It is called 7″ by Luce and Raiffa and "beta" by Sen. This says that if an old choiceworthy act is made non-choiceworthy by the addition of new acts, then *all* old choiceworthy acts are made non-choiceworthy.

> All-or-nothing: If $\varphi \in \mathcal{C}(A)$ but $\varphi \notin \mathcal{C}(A \cup B)$ then, for all $\psi \in \mathcal{C}(A)$, we have $\psi \notin \mathcal{C}(A \cup B)$

As Luce and Raiffa show All-or-nothing makes sense only when you are evaluating the acts on a single scale. Sugden [28] discusses an example where one race car is faster and another is more manoeuvrable: the first will win in a head to head race, but the second will win if there are other cars on the track. Thus the "race winning function", if you like, does not satisfy All-or-nothing. Such a choice function can't be given a strong interpretation. That is, each member of the choice set is better than all acts outside the choice set in some sense; but it is not the case that all members of the choice set are equally good. They are merely good in different ways. I claim that imprecise decision can be a little like this, and thus that All-or-nothing should not be required. It is a property that makes sense only for *strong* choice functions. Single criterion choice (as characterised by All-or-nothing) and the strong interpretation of the choice set go hand in hand.

Two further properties that I don't endorse as constraints on rational choice are the following:

MIXING: $\mathcal{C}(A) \subseteq \mathcal{C}(A^*)$
CONVEXITY: $\mathcal{C}(A)^* \cap A = \mathcal{C}(A)$

MIXING says that if φ is not choiceworthy among the mixtures of A, then φ should not be choiceworthy in A itself. This seems an odd requirement of rationality: if you are choosing among the members of A, why should the fact that an act is not choiceworthy in some larger set of acts be relevant? CONVEXITY says that mixtures of choiceworthy acts should be choiceworthy. This property seems to be trading on the same "single-criterion choice" idea as I discussed above.

4 Examples of Choice Functions

4.1 Non-Domination

What about just taking $\mathcal{M}_{\succeq_{\text{Dom}}}$ as our choice rule? That is, any acts that are not dominated are in the choice set. It is, perhaps, too permissive a rule.

Consider the following example.

> EXAMPLE 2: There is a coin of unknown bias. You are offered the choice between these bets:
>
> - a: win £1 if the next toss lands heads
> - b: win £1 if the next ten tosses all land heads
> - b': win £$1+\varepsilon$ if the next ten tosses land heads, win £ε otherwise

It seems right that $\mathcal{M}_{\succeq_{\text{Dom}}}$ rules out act b. However, it seems unfortunate that it doesn't rule out the "almost dominated" act b'.

Also, this rule does not satisfy the ALL-OR-NOTHING property. Here is an example of how $\mathcal{M}_{\succeq_{\text{Dom}}}$ fails all-or-nothing expansion consistency.

> EXAMPLE 3: Consider the choice between g and h, and the choice between g, h and k.
>
> - g: Gain £10 if H, nothing otherwise
> - h: Gain nothing if H, £10 otherwise
> - k: Gain £11 if H, £1 otherwise

k dominates g, so in the expanded decision problem, g is not choiceworthy. However, h is still undominated, so this violates ALL-OR-NOTHING. As I said above, I don't think violating ALL-OR-NOTHING is a mark against an imprecise choice rule.

A mixture of undominated acts can be dominated (see Table 1). Each of a_1 and a_2 are undominated, but the mixture is dominated by a_3. So CONVEXITY is not true for $\mathcal{M}_{\succeq_{\text{Dom}}}$. This choice rule also violates MIXING [23].

	s_1	s_2
a_1	2	-2
a_2	-2	2
a_3	1	1
$0.5a_1 + 0.5a_2 = a_4$	0	0

Table 1: A mixture of undominated acts can be dominated

4.2 E-Admissibility

The main problem with $\mathcal{M}_{\succeq_{\text{Dom}}}$ is that it isn't really discriminating enough. That is, the choice sets that that rule generates will often contain many acts. We would really like choice to be more constrained. Let's consider a more discriminating choice rule. Another restriction of the act set – "E-admissibility" – is due to Isaac Levi [14, 15]. An act is E-admissible if there is some probability in your representor such that that act maximises expectation with respect to that probability function. E-admissible acts are the ones that some credal committee member thinks are best (by that member's standard of $\mathrm{E}_{\mathbf{pr}}$). Levi argues that you should only choose among E-admissible acts. A first attempt at cashing out this choice rule is:

$$L(A) = \bigcup_{\mathbf{pr} \in \mathcal{P}} \mathcal{M}_{\succeq_{\mathrm{E}_{\mathbf{pr}}}}(A) \tag{4}$$

This might be more perspicuously rephrased as:[11]

$$L(A) = \{\varphi \in A : \exists\, \mathbf{pr} \in \mathcal{P}, \forall \psi \in A, \mathrm{E}_{\mathbf{pr}}(\varphi) \geq \mathrm{E}_{\mathbf{pr}}(\psi)\} \tag{5}$$

The intuition is that we ask each credal committee member to pick their favourite act(s): we then take the collection of each of these favourites. Compare with $\mathcal{M}_{\succeq_{\text{Dom}}}$ where we take out all the acts where the committee unanimously prefers some other act.

As it stands, the definition of E-admissible isn't quite good enough. Recall Example 2 where we had the choice between a bet on heads and a bet on ten heads in a row. The latter maximises expectation for $\mathbf{pr}(H) = 0$ and $\mathbf{pr}(H) = 1$ so it is E-admissible. This act is, however, weakly dominated.[12] To fix this, consider $\mathcal{M}_{\succeq_{\text{Dom}}} \circ L(A)$ where "$\circ$" is composition of functions. We shall call this $\mathcal{L}(A)$.

We know that $\mathcal{L}(A) \subseteq \mathcal{M}_{\succeq_{\text{Dom}}}(A)$. There are undominated acts that are not E-admissible.[13] So we in fact know that $\mathcal{L}(A) \subsetneq \mathcal{M}_{\succeq_{\text{Dom}}}(A)$ for some A.

[11] This rephrasing makes it clear that non-domination and E-admissibility differ only in the order of quantification and the strictness of the inequality. That is, non-domination becomes: $\varphi \in A, \forall \psi \in A, \exists\, \mathbf{pr} \in \mathcal{P}, \mathrm{E}_{\mathbf{pr}}(\varphi) > \mathrm{E}_{\mathbf{pr}}(\psi)$.

[12] L never contains *strongly* dominated acts.

[13] For example gamble n in Example 4 below.

So $\mathcal{L}$ is more discriminating than $\mathcal{M}_{\succeq_{\text{Dom}}}$. Given that E-admissibility is more discriminating and given that non-domination is arguably too permissive (not discriminating enough), one might think that E-admissibility is obviously the better rule. However, $\mathcal{L}$ doesn't help solve any of the problems with $\mathcal{M}_{\succeq_{\text{Dom}}}$.

$\mathcal{L}$ violates union consistency, as can be seen from considering Example 4.

> EXAMPLE 4: You are betting on a coin of unknown bias. You can choose among these bets:
>
> - l: Gain £10 if H, lose 5 otherwise
> - m: Lose £5 if H, gain 10 otherwise
> - n: Gain 0 whatever happens

$\mathcal{L}(\{l, n\}) = \{l, n\}$ and $\mathcal{L}(\{m, n\}) = \{m, n\}$, but $\mathcal{L}(\{l, m, n\}) = \{l, m\}$. That is, n is choiceworthy in both pairwise choices, but if all three options are offered together, then n is ruled out.

Seidenfeld et al. point out that it follows from Lemma 3 of [18] that E-admissibility satisfies MIXING [23]. It also means that if $A = A^*$ then $\mathcal{L}(A) = \mathcal{M}_{\succeq_{\text{Dom}}}(A)$. It is also worth noting that $\mathcal{L}$ is not determined by pairwise comparisons, while $\mathcal{M}_{\succeq_{\text{Dom}}}$ is. I don't think either of these features tells in favour of the rule's rationality.

Despite being more discriminating, E-admissibility does not seem like an improvement on non-domination. I doesn't help with almost dominated acts, or with CONVEXITY, and it adds violations of a further intuitive property: UNION CONSISTENCY.

4.3 Valuing Acts

The standard approach to decision making with precise probabilities is to assign to each act a number representing how much that act is valued: $\mathrm{E}_{\mathbf{pr}}$. Let's try to do the same thing here: can we find some number that represents how valuable a certain gamble is? A first attempt at valuing acts in the imprecise case would be to look at $\underline{\mathcal{E}}$. That is, consider the decision rule that says "act to maximise the worst-case expected value". Is $\mathcal{M}_{\underline{\mathcal{E}}}$ a good decision rule? This rule is sometimes described as "gamma-maximin" [21]. It is also the rule that [9] advocate.[14]

$\mathcal{M}_{\underline{\mathcal{E}}}$ does not satisfy non-domination. That is, $\mathcal{M}_{\underline{\mathcal{E}}}$ sometimes contains acts that are weakly dominated, as Example 2 shows. The above problem isn't just a problem for $\mathcal{M}_{\underline{\mathcal{E}}}$, but for any rules that focus only on the set of expectations. For example, instead of maximising $\underline{\mathcal{E}}$, consider maximising $\mathcal{H}_\alpha(\varphi) = \alpha\,\underline{\mathcal{E}}(\varphi) + (1-\alpha)\,\overline{\mathcal{E}}(\varphi)$ for some real number α between 0 and 1. This is an "imprecise analogue" of the *Hurwicz criterion* for choice under complete ignorance [12, 17]. This is actually a whole class of different decision rules depending on choice of α. If $\alpha = 1$ then we recover maximise minimum expectation ($\mathcal{M}_{\underline{\mathcal{E}}}$). If a precise α value seems arbitrary, perhaps consider looking for acts that do well for many different values of α. [2] suggests a rule that, effectively, amounts to preferring φ to ψ just in case φ is better according to all values of α. Sadly, none of these rules can avoid making (weakly) dominated acts permissible: none of these rules can make b inadmissible in Example 2.[15] That is, since $\underline{\mathcal{E}}(a) = \underline{\mathcal{E}}(b)$ and $\overline{\mathcal{E}}(a) = \overline{\mathcal{E}}(b)$, any rule that values acts as some function of these values must treat the two bets the same.

As well as violating the rationally compelling NON-DOMINATION principle, the $\mathcal{M}_{\underline{\mathcal{E}}}$ rule also violates independence. Consider Example 1: $\mathcal{M}_{\underline{\mathcal{E}}}$ chooses d over c in the odd branch. But when you mix with the even branch, c ends up looking better. That is, the payouts of c and d for the "mixed" decision problem are "5 if H, 3 otherwise" and "1 if H, 4 otherwise" respectively.

If we focus on strict dominance $\succeq_{\text{SDom}}$ rather than weak dominance $\succeq_{\text{Dom}}$, then $\mathcal{M}_{\underline{\mathcal{E}}}(A) \subseteq \mathcal{M}_{\text{SDom}}$ [30].[16]

4.4 Composite Rules

Since the problem with $\mathcal{M}_{\underline{\mathcal{E}}}$ (and similar rules) is that it allows weakly dominated acts to be choiceworthy, why not just compose it with $\mathcal{M}_{\succeq_{\text{Dom}}}$ to make a better rule? Consider $\mathcal{M}_{\underline{\mathcal{E}}} \circ \mathcal{M}_{\succeq_{\text{Dom}}}$: this is the rule that maximises minimum expectation among the acts that are undominated. This rule obviously satisfies NON-DOMINATION. It still fails independence, however.

What about composing $\mathcal{M}_{\underline{\mathcal{E}}}$ with $\mathcal{L}$? Isaac Levi, for instance, advocated using $\mathcal{M}_{\underline{\mathcal{E}}}$ as a tie-breaker among E-admissible acts. We have seen that both choice functions have problems as decision rules. The composite rule still violates UNION CONSISTENCY and INDEPENDENCE. Combining them in the way Levi suggests leads to further problems. This composite rule violates CONTRACTION CONSISTENCY, as [21] points out.

> EXAMPLE 5: Consider the choice between t, u and the choice between t, u, v.
>
> - t: £10 if H, nothing otherwise

[14] Their decision rule is slightly more complex in that it takes into account the "reliability" of the functions in your representor, but if all probabilities are equally reliable, then their rule reduces to gamma-maximin.

[15] Indeed, b' is uniquely admissible for $\mathcal{M}_{\underline{\mathcal{E}}}$.

[16] This paper also discusses several other interesting connections between imprecise choice rules.

- u: £3 if H, £3 otherwise
- v: £-1 if H, £8 otherwise

In a choice between t and u, it is u that does best by $\mathcal{M}_{\underline{\mathcal{E}}}$. However, adding v means that u is no longer E-admissible and of t and v, t does better.

4.5 Aggregate Value

Perhaps we have been approaching this the wrong way, and what we should be doing is looking for some way to aggregate $\mathcal{P}$ or $\mathcal{E}$ to get a (precise) aggregate expected utility and maximise that in the standard way? There is a large literature on aggregating probability judgements [10]; might this not provide new insight on IP decision making? First, I'm not sure that such an approach is in the spirit of IP. Second, it isn't clear that such an aggregate value approach will be able to rationalise ambiguity aversion in the Ellsberg game [6] which is, after all, a desideratum for IP decision making.

In one sense, we would like to have some all-things-considered aggregate value to attach to acts. We would like to have some notion of value that rational agents seek to maximise, some concept of rational choice that can be given a strong interpretation. But when your attitudes about the expected goodness are *conflicted* in the way they are in IP models, I'm not sure why we should think that such reasonable aggregation is possible.

We can aggregate the credal committee's opinions about the probabilities ($\mathcal{P}$), but this doesn't seem to be true to the goals of IP models. We can aggregate the credal committee's opinions about the expected values ($\mathcal{E}$), but the previous two subsections show that this leads to some problematic consequences. Or we can aggregate the credal committee's preferences (the $\succeq_{\mathrm{E}_{\mathbf{pr}}}$ relations), but the choice rules we get ($\mathcal{M}_{\succeq_{\mathrm{Dom}}}, \mathcal{L}$) can't be given the strong interpretation we would like.

4.6 Regret

We have seen imprecise analogues of maximin and Hurwicz criterion rules for decision under ignorance. What about an imprecise analogue of minimax-regret? Consider:

$$\mathcal{R}(\varphi) = -\max_{\mathbf{pr}\in\mathcal{P}} \left\{ \max_{\psi\in A} \{\mathrm{E}_{\mathbf{pr}}(\psi)\} - \mathrm{E}_{\mathbf{pr}}(\varphi) \right\} \qquad (6)$$

And consider the choice rule $\mathcal{M}_{\mathcal{R}}$. This rule violates Union consistency and Contraction consistency.[17] On the other hand, it satisfies Non-domination and also rules out "almost dominated" acts like b' in Example 2. This rule deserves further attention, although note that it is computationally demanding. It's also unclear under what conditions it is decisive.

[17] Interestingly, in Example 4, $\mathcal{M}_{\mathcal{R}}$ chooses l out of l, n and m out of m, n, but makes n uniquely admissible in the three way choice, which is a very different profile of choices from $\mathcal{L}$.

5 Conclusion

We have explored a number of different kinds of choice rule. None is entirely satisfactory. So how should we act? I think we can at least take Non-domination as a requirement on rational choice. So $\mathcal{M}_{\succeq_{\mathrm{Dom}}}$ serves to rule out some bad acts. This means that $\varphi \in \mathcal{M}_{\succeq_{\mathrm{Dom}}}(A)$ is acting as a necessary *but not sufficient* condition on imprecise choice. A variety of options for going beyond this – to attempt to find sufficient conditions for rational choice – have failed. All the more discriminating rules we have looked at seem to violate one or more intuitively compelling properties of rational choice.

We can understand $\mathcal{M}_{\succeq_{\mathrm{Dom}}}(A)$ as a weak kind of choice set. That is, it is reasonable to rule out all the acts that $\mathcal{M}_{\succeq_{\mathrm{Dom}}}$ rules out. But it seems like some acts that make it into $\mathcal{M}_{\succeq_{\mathrm{Dom}}}$ that we would not consider to be reasonable choices. The various attempts to come up with a choice rule that can be given a stronger interpretation have failed. That is, every attempt to construct a choice rule that positively endorses all the acts in the choice set have come up short. $\mathcal{C}_{\mathrm{ID}}$ is such a rule, but it is often empty.

In summary, rules like $\mathcal{M}_{\underline{\mathcal{E}}}$ and $\mathcal{M}_{\mathcal{H}_\alpha}$ violate Non-domination and so are not good rules. They also violate Independence. $\mathcal{L}$ violates Union Consistency which might be considered a problem. Levi's suggestion of using $\underline{\mathcal{E}}$ to break ties among elements of $\mathcal{L}$ is doubly bad: it violates Contraction Consistency and Independence. In short, $\mathcal{M}_{\succeq_{\mathrm{Dom}}}$ seems hard to improve on: every proposed improvement, every more discriminating choice rule, has some flaw or other.

What I take myself to have shown here is that we can make some progress on the problem of imprecise choice. It is not the case that when your credences become imprecise, all constraint on choice falls away. In many cases of "moderate" imprecision, the above constraints on choice (in particular Non-domination) will be enough to fix your choice.

When your credences are imprecise, then it's difficult to know how you should act. Put another way: weaken the theory of decision and it's not surprising that the constraints on choice aren't as strong. Perhaps the conclusion to draw from this is that there is no rationally compelling IP choice function that admits

of a strong interpretation. Obviously, we can't take $\mathcal{M}_{\succeq_{\text{Dom}}}$ to cash out all there is to rationality, since it doesn't rule out "almost dominated" acts like b' in Example 2, as we would like. But it does seem to capture a necessary condition on rational choice. This makes it clear that even when we expect rationality to be silent on some questions in this area, it is not the case that imprecise choice is unconstrained.

A Proofs

Theorem 1 *If $\mathcal{C}$ satisfies $\succeq$ and $\mathcal{C}(A)$ is nonempty for nonempty A, then $\mathcal{C}$ pairwise satisfies $\succeq$.*
Proof: *Assume $\varphi \succ \psi$ and $\mathcal{C}$ satisfies $\succeq$ and is nonempty. Then $\psi \notin \mathcal{C}(\{\varphi, \psi\})$. $\mathcal{C}(\{\varphi, \psi\})$ is a subset of $\{\varphi, \psi\}$, does not contain ψ and is nonempty. Therefore $\mathcal{C}(\{\varphi, \psi\}) = \{\varphi\}$.*

Assume $\varphi \succeq \psi$ and $\mathcal{C}$ satisfies $\succeq$ and is nonempty. Now, either $\varphi \succ \psi$ and the above argument shows that $\mathcal{C}(\{\varphi, \psi\}) = \{\varphi\}$, or $\varphi \sim \psi$. Therefore, since $\mathcal{C}$ satisfies $\succeq$, $\varphi \in \mathcal{C}(\{\varphi, \psi\})$ if and only if $\psi \in \mathcal{C}(\{\varphi, \psi\})$. Since $\mathcal{C}$ can't be empty, and must be a subset of $\{\varphi, \psi\}$, $\mathcal{C}(\{\varphi, \psi\}) = \{\varphi, \psi\}$. In either case, $\varphi \in \mathcal{C}(\{\varphi, \psi\})$ as required.

Theorem 2 *(i) $\mathcal{M}_{\succeq}$ is a choice function and (ii) $\mathcal{M}_{\succeq}$ pairwise satisfies $\succeq$. (iii) If $\succeq$ is acyclic on A where A is finite then $\mathcal{M}_{\succeq}(A)$ is non-empty. (iv) Furthermore, if $\succeq$ is transitive, then $\mathcal{M}_{\succeq}$ satisfies $\succeq$.*
Proof: *(i) $\mathcal{M}_{\succeq}(A) \subseteq A$ by definition. It is equally obvious that $\mathcal{M}_{\succeq}(\mathcal{M}_{\succeq}(A)) = \mathcal{M}_{\succeq}(A)$.*

(ii) We need to show that if $a \succeq b$ then $a \in \mathcal{M}_{\succeq}(\{a, b\})$. The only way a could fail to be in $\mathcal{M}_{\succeq}(\{a, b\})$ is if $b \succ a$. But this is ruled out by definition of $\succ$. If $a \succ b$ then $a \succeq b$, so by the above, we have that $a \in \mathcal{M}_{\succeq}(\{a, b\})$, and by definition, $b \notin \mathcal{M}_{\succeq}(\{a, b\})$.

(iii) Let $\succeq$ be acyclic on some finite A. If the size of A, $|A| = 1$, then that singleton element is maximal. Assume $\mathcal{M}_{\succeq_{\text{Dom}}}(A)$ is non-empty for $|A| \leq n$. Consider A of size $n+1$. We need to find an element $\varphi \in \mathcal{M}_{\succeq_{\text{Dom}}}(A)$. Take an arbitrary $\varphi_0 \in A$. If $\varphi_o \in \mathcal{M}_{\succeq_{\text{Dom}}}(A)$ then we are done. Otherwise, let $A_0 = A \setminus \{\varphi_0\}$. By hypothesis, $\mathcal{M}_{\succeq_{\text{Dom}}}(A_0) \neq \emptyset$. Say $\varphi^* \in \mathcal{M}_{\succeq_{\text{Dom}}}(A_0)$. If φ^* is maximal in A then we are done. If not, then we must have $\varphi_0 \succ \varphi^*$. If φ_0 is not maximal then there must be some φ_1 such that $\varphi_1 \succ \varphi_0$. And since $\succeq$ is acyclic, φ_1 can't be equal to φ^*. This procedure will eventually pick out an element that is maximal in A [29, Theorem A(3), p.14].

(iv) If $a \succ b$ then $b \notin \mathcal{M}_{\succeq}(A)$ by definition. Finally, assume for contradiction that $a \sim b$ and $a \in \mathcal{M}_{\succeq}(A)$ but $b \notin \mathcal{M}_{\succeq}(A)$. This means there exists some $c \succ b$. But $b \succeq a$ so by transitivity[18] $c \succ a$, contradicting $a \in \mathcal{M}_{\succeq}(A)$.

Theorem 3 *If $\mathcal{C}$ satisfies $\succeq$ then $\mathcal{C}(A) \subseteq \mathcal{M}_{\succeq}(A)$ for all A.*
Proof: *Let $a \in \mathcal{C}(A)$. Assume for contradiction that there is some $b \in A$ such that $b \succ a$. If there were such a b, then a would not have been in $\mathcal{C}(A)$ by definition of "satisfies". Thus $\neg\exists b \in A, b \succ a$. This is exactly the condition required for inclusion in $\mathcal{M}_{\succeq}$.*

For the next theorem we will need a little bit more notation. We will use $\varphi \bowtie \psi$ to mean $\neg\varphi \succeq \psi$ and $\neg\psi \succeq \varphi$. That is, $\varphi \bowtie \psi$ if and only if the two acts are incomparable. We will also need this fact about $\bowtie$.

Lemma 1 *For transitive $\succeq$: if $\varphi \sim \psi$ and $\psi \bowtie \rho$ then $\varphi \bowtie \rho$*
Proof: *Assume $\varphi \sim \psi \bowtie \rho$. Assume for contradiction that $\varphi \succeq \rho$. Then $\psi \sim \varphi \succeq \rho$ which implies $\psi \succeq \rho$ which contradicts our assumptions.*[19] *Likewise for $\rho \succeq \varphi$. Thus $\varphi \bowtie \rho$.*

Theorem 4 *When $\text{Opt}_{\succeq}(A) \neq \emptyset$, and $\succeq$ is transitive then $\text{Opt}_{\succeq}(A) = \mathcal{M}_{\succeq}(A)$.*
Proof: *We first show that $\text{Opt}_{\succeq}(A) \subseteq \mathcal{M}_{\succeq}(A)$. We then show that if φ is maximal but not optimal, then no act is optimal.*

Assume $\varphi \in \text{Opt}_{\succeq}(A)$. Assume for contradiction that there is some ψ such that $\psi \succ \varphi$. Therefore $\neg\varphi \succeq \psi$, which contradicts our assumption. Thus $\neg\exists\psi \in A, \psi \succ \varphi$. This is exactly the criterion for inclusion in $\mathcal{M}_{\succeq}(A)$.

Assume now that $\varphi \in \mathcal{M}_{\succeq}(A)$ but, $\varphi \notin \text{Opt}_{\succeq}(A)$. For φ not to be optimal, this means there is some ψ such that $\neg\varphi \succeq \psi$. φ is maximal, so φ and ψ must be incomparable. Assume there is some $\rho \in \text{Opt}_{\succeq}(A)$. So $\rho \succeq \varphi$, but since φ is maximal, this must mean $\varphi \sim \rho$. $\rho \sim \varphi \bowtie \psi$, therefore $\rho \bowtie \psi$ by the above lemma. In particular $\neg\rho \succeq \psi$ which contradicts our assumption. Therefore $\text{Opt}_{\succeq}(A)$ is empty.

References

[1] Thomas Augustin, Frank P.A. Coolen, Gert de Cooman, and Matthias C.M. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley and Sons, 2014.

[2] Prasanta S. Bandyopadhyay. In search of a pointless decision principle. *Philosophy of Science Association Proceedings*, pages 260–269, 1994.

[3] Richard Bradley. A note on transitivity, completeness and Suzumura consistency. LSE Choice Group Working Papers, 2013.

[4] Francis Chu and Joseph Y. Halpern. Great expectations. Part II: Generalized expected utility as a universal decision rule. *Artificial intelligence*, 159:207–230, 2004.

[5] Francis Chu and Joseph Y. Halpern. Great expectations. Part I: On the customizability of general expected utility. *Theory and Decision*, 64:1–36, 2008.

[6] Daniel Ellsberg. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75:643–696, 1961.

[18] Strictly speaking, we don't really need transitivity here: we only need that $\psi \sim \varphi$ and $\rho \succ \psi$ imply $\rho \succ \varphi$.

[19] Strictly speaking we only need something slightly weaker than transitivity: if $\psi \sim \varphi$ and $\varphi \succeq \rho$ then $\psi \succeq \rho$.

[7] Özgür Evren and Efe Ok. On the multi-utility representation of preference relations. *Journal of Mathematical Economics*, 47:554–563, 2011.

[8] Wulf Gaertner. *A Primer in Social Choice Theory*. Oxford University Press, 2009.

[9] Peter Gärdenfors and Nils-Eric Sahlin. Unreliable probabilities, risk taking and decision making. *Synthese*, 53:361–386, 1982.

[10] Christian Genest and James V. Zidek. Combining probability distributions: A critique and annotated bibliography. *Statistical Science*, 1:114–135, 1986.

[11] Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.

[12] Leonid Hurwicz. Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper Statistics 370, 1951.

[13] Henry E. Kyburg. Rational belief. *The Brain and Behavioural Sciences*, 6:231–273, 1983.

[14] Isaac Levi. On indeterminate probabilities. *Journal of Philosophy*, 71:391–418, 1974.

[15] Isaac Levi. *Hard Choices: Decision Making Under Unresolved Conflict*. Cambridge University Press, 1986.

[16] R. Duncan Luce and Howard Raiffa. *Games and Decisions*. Dover, 1989.

[17] John Milnor. Games against nature. Technical report, RAND corporation, 1951.

[18] David Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1986.

[19] Erik Quaeghebeur, Gert de Cooman, and Filip Hermans. Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning*, 57:69–102, 2015.

[20] Teddy Seidenfeld. Decision theory without "independence" or without "ordering". what's the difference? *Economics and Philosophy*, pages 267–290, 1988.

[21] Teddy Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: Gamma-maximin versus E-admissibility. *Synthese*, 140:69–88, 2004.

[22] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. A representation of partially ordered preferences. *Annals of Statistics*, 23:2168–2217, 1995.

[23] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172:157–176, 2010.

[24] Amartya Sen. *Collective choice and social welfare*. Holden-Day, 1970.

[25] Amartya Sen. Choice functions and revealed preference. *The Review of Economic Studies*, 38:307–317, 1971.

[26] Amartya Sen. Social choice theory: A re-examination. *Econometrica*, 45:53–89, 1977.

[27] Amartya Sen. Maximisation and the act of choice. *Econometrica*, 65:745–779, 1997.

[28] Robert Sugden. Why be consistent? a critical analysis of consistency requirements in choice theory. *Economica*, 52:167–183, 1985.

[29] Kotaro Suzumura. *Rational Choice, Collective Decisions and Social Welfare*. Cambridge University Press, 1983.

[30] Matthias Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45:17–29, 2007.

[31] Matthias Troffaes and Gert de Cooman. *Lower Previsions*. Wiley, 2014.

The Generalization of the Conjunctive Rule for Aggregating Contradictory Sources of Information Based on Generalized Credal Sets

Andrey G. Bronevich
National Research University
Higher School of Economics
Moscow, Russia
brone@mail.ru

Igor N. Rozenberg
JSC "Research, Development and Planning Institute for Railway
Information Technology, Automation and Telecommunication"
Moscow, Russia
I.Rozenberg@gismps.ru

Abstract

In the paper we consider the generalization of the conjunctive rule in the theory of imprecise probabilities. Let us remind that the conjunction rule, produced on credal sets, gives their intersection and it is not defined if this intersection is empty. In the last case the sources of information are called contradictory[1]. Meanwhile, in the Dempster-Shafer theory it is possible to use the conjunctive rule for contradictory sources of information having as a result a non-normalized belief function that can be greater than zero at empty set. In the paper we try to exploit this idea and introduce into consideration so called generalized credal sets allowing to model imprecision (non-specificity), conflict, and contradiction in information. Based on generalized credal sets the conjunctive rule is well defined for contradictory sources of information and it can be conceived as the generalization of the conjunctive rule for belief functions. We also show how generalized credal sets can be used for modeling information when the avoiding sure loss condition is not satisfied, and consider coherence conditions and natural extension based on generalized credal sets.

Keywords. Imprecise probabilities, conjunctive rule, generalized credal sets, contradictory sources of information.

1 Introduction

In the theory of imprecise probabilities [18, 7, 1] there are many models for describing uncertainty: credal sets, upper and lower probabilities, lower and upper coherent previsions, sets of desirable gambles, etc. But in any case, we can equivalently represent the information with the help of sets of probability measures. As one can check, up to now there are no many works concerning the case when the available information is contradictory, i.e. the avoiding sure loss condition is not satisfied.

[1] We will use next the term "contradictory" because the traditional term "conflict" is also used by identifying another type of uncertainty described by probability measures.

By the way, in evidence theory [15, 8, 16] there is a possible way to describe contradiction based on transferable belief model. In this model we can describe contradictory information by assigning non-zero values to the corresponding belief function at empty set[2]. In this paper we will try to exploit this idea that leads to some generalizations of the theory of imprecise probabilities, in particular based on this idea it is possible to extend the conjunctive rule (C-rule) for aggregating belief functions for more general theories of imprecise probabilities [3, 4].

Let us notice that in the literature one can find results concerning the aggregation rules for imprecise probabilities [17, 9, 14, 13]. The rule from [17] deals with lower previsions and generalizes the pooling method for aggregation of probability measures. In [9] the aggregation rule is based on an idea that non-conflicting information should be aggregated in conjunctive manner and conflicting information should be aggregated in disjunctive manner. In [14] the proposed aggregation rules are based on modeling the interaction among expert's opinions. Authors of [13] try to get the aggregation rule for credal sets with properties close to the C-rule but their rule is based on some heuristic algorithmic procedure.

The paper has the following structure. Sections 2 and 3 remind some definitions from the theory of monotone measures, belief functions and the theory of imprecise probabilities. Then in Sections 4 and 5 we describe the basic rules of aggregation in general theories of imprecise probabilities and investigate the connection of these rules to the combination rules in evidence theory. After that we try to generalize the C-rule firstly (Section 6) for probability measures, and secondly (Section 7) for general models of imprecise probabilities using so-called generalized credal sets. Based on generalized credal sets it is possible to model contradiction in information and introduce analogous notions and constructions as in traditional theory of imprecise probabilities like coherence and natural extension, as shown in Section 8.

[2] This statement will be clarified in the next sections.

2 Some Definitions and Notations from the Theory of Non-additive Measures

Let X be a non-empty finite set and let 2^X be the power set of X. We will consider set functions on the algebra 2^X of various types: monotone measures, probability measures, lower and upper probabilities. A set function $\mu : 2^X \to [0,1]$ is called

1) *normalized* if $\mu(\varnothing) = 0$ and $\mu(X) = 1$;
2) *monotone* if $A, B \in 2^X$ and $A \subseteq B$ implies $\mu(A) \leq \mu(B)$;
3) *additive* if $\mu(A) + \mu(B) = \mu(A \cap B) + \mu(A \cup B)$ for all $A, B \in 2^X$;
4) *2-monotone* if $\mu(A) + \mu(B) \leq \mu(A \cap B) + \mu(A \cup B)$ for all $A, B \in 2^X$;
5) *2-alternating* if $\mu(A) + \mu(B) \geq \mu(A \cap B) + \mu(A \cup B)$ for all $A, B \in 2^X$;
6) a *monotone measure* if it is monotone and normalized;
7) a *probability measure* if it is additive and normalized;
8) a *belief function* if there is non-additive set function $m : 2^X \to [0,1]$ called the *basic belief assignment* (bba) such that $\sum_{A \in 2^X} m(A) = 1$ and $\mu(B) = \sum_{A \subseteq B} m(A)$.

The following operations on set functions are defined:

a) convex sum: $\mu = a\mu_1 + (1-a)\mu_2$, where $a \in [0,1]$, and $\mu(A) = a\mu_1(A) + (1-a)\mu_2(A)$ for all $A \in 2^X$;
b) $\mu_1 \leq \mu_2$ if $\mu_1(A) \leq \mu_2(A)$ for all $A \in 2^X$;
c) μ^d is the dual of μ if $\mu^d(A) = 1 - \mu(\overline{A})$ for all $A \in 2^X$, and $\overline{A}$ denotes the compliment of A.

Let us remind that the theory of evidence models uncertainty with the help of belief functions. In this theory (e.g. transferable belief model) we describe contradiction using non-normalized belief functions, i.e. it is possible that $Bel(\varnothing) > 0$ for belief function Bel. Let Bel be a belief function with the bba m. Then

- the set $A \in 2^X$ is a *focal element* for Bel if $m(A) > 0$;
- the set of all focal elements is called the *body of evidence*;
- Bel is called *categorical* if its body of evidence contains only one focal element. Any categorical belief function $\eta_{\langle B \rangle}$ with focal element B can be computed as

$$\eta_{\langle B \rangle}(A) = \begin{cases} 1, & B \subseteq A, \\ 0, & \text{otherwise.} \end{cases}$$

- Bel is a probability measure iff $m(A) = 0$ for all $A \in 2^X$ with $|A| \geq 2$. In this paper we also consider non-normalized probability measures P for which $P(\varnothing) > 0$.
- any belief function μ has the following representation through categorical belief functions:

$$Bel = \sum_{B \in 2^X} m(B)\eta_{\langle B \rangle}.$$

In the sequel we will use the following notations:

- M_{pr} is the set of all probability measures on 2^X and $\overline{M}_{pr}$ be the set of all probability measures including also non-normalized probability measures.
- M_{bel} and $\overline{M}_{bel}$ are sets of all belief functions on 2^X and the bar indicates that belief functions from $\overline{M}_{bel}$ may be non-normalized.
- M_{mon} is the set of all monotone measures on 2^X.
- $M_{2\text{-}mon}$ is the set of all 2-monotone measures on 2^X.
- if M is a family of set functions, then we denote $M^d = \{\mu^d | \mu \in M\}$. For example, M^d_{bel} denotes the set of all plausibility functions, which are dual to belief functions, or $M^d_{2\text{-}mon}$ is the set of all 2-alternating measures on 2^X.

3 Models of Imprecise Probabilities: Lower and Upper Probabilities and Credal Sets

Assume that $\mu : 2^X \to [0,1]$ is a set function that gives us lower bounds of probabilities. Then this function *avoids sure loss* iff there is a probability measure $P \in M_{pr}$ such that $\mu \leq P$. If the avoiding sure loss condition is not fulfilled, then the information described by μ is *contradictory*. Any non-contradictory lower probability function μ defines the non-empty set of probability measures

$$\mathbf{P}(\mu) = \{P \in M_{pr} | P \geq \mu\}$$

called the credal set. Generally, a set $\mathbf{P}$ of probability measures is called a *credal set* if it is convex and closed.

Analogously the model of upper probabilities is introduced. Let us suppose that $\nu : 2^X \to [0,1]$ gives us the upper bounds of probabilities. Then this function avoids sure loss iff there is a probability measure $P \in M_{pr}$ such that $\nu \geq P$. In this case we call an upper probability function non-contradictory and describe it by a credal set

$$\mathbf{P}(\nu) = \{P \in M_{pr} | P \leq \nu\}.$$

We can equivalently replace the model based on lower probabilities by the model based on upper probabilities. For this purpose we transform any lower probability μ to the upper probability μ^d. It easy to show that

$$\left\{P \in M_{pr} | P \leq \mu^d\right\} = \{P \in M_{pr} | P \geq \mu\},$$

i.e. the corresponding credal sets coincide.

Let us introduce also coherent lower and upper probabilities. A non-contradictory lower probability μ is called *coherent* if for any $A \in 2^X$ there exists $P \in M_{pr}$ such that $\mu(A) = P(A)$ and $\mu \leq P$, in other words,

$$\mu(A) = \inf\{P(A) | P \in \mathbf{P}(\mu)\},$$

where $\mathbf{P}(\mu) = \{P \in M_{pr} | P \geq \mu\}$.

Analogously, a non-contradictory upper probability ν is called *coherent* if for any $A \in 2^X$ there exists $P \in M_{pr}$ such that $\nu(A) = P(A)$ and $\nu \geq P$, in other words,

$$\nu(A) = \inf\{P(A) | P \in \mathbf{P}(\nu)\},$$

where $\mathbf{P}(\nu) = \{P \in M_{pr} | P \geq \nu\}$.

Coherent lower probabilities and coherent upper probabilities are connected with the dual relation, i.e. if μ is a coherent lower probability then μ^d is the coherent upper probability. We can also generate a coherent lower probability μ and coherent upper probability ν using a credal set $\mathbf{P}$ by formulas:

$$\mu(A) = \inf\{P(A) | P \in \mathbf{P}\},$$
$$\nu(A) = \sup\{P(A) | P \in \mathbf{P}\},$$

where $A \in 2^X$, and obviously, $\nu = \mu^d$ in this case.

Let μ be a non-contradictory lower probability. Then we can improve lower bounds of probabilities using the natural extension. It is defined as

$$\mu_{coh}(A) = \inf\{P(A) | P \in \mathbf{P}(\mu)\},$$

where $A \in 2^X$. Clearly, μ_{coh} is a coherent lower probability.

Let us remind that any credal set can be equivalently defined with the help of lower previsions. Let K' be a subset of the set K of all real functions of the type $f : X \to \mathbb{R}$. In some cases we assume that $K' = K$. Then *lower previsions* on K' are defined by the functional $\underline{E} : K' \to \mathbb{R}$. This functional defines the credal set

$$\mathbf{P}(\underline{E}) = \left\{ P \in M_{pr} \middle| \forall f \in K' : \sum_{x \in X} f(x)P(\{x\}) \geq \underline{E}[f] \right\}.$$

If the credal set $\mathbf{P}(\underline{E})$ is empty then lower previsions do not satisfy the avoiding sure loss condition and we say that lower previsions contain contradiction. In some sense lower previsions can be understood as lower bounds of expectations of random variables in K'. The model based on lower previsions is more general than the model based on lower probabilities because we obtain the last model if we assume that $K' = \{1_A\}_{A \in 2^X}$, where

$$1_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases}$$

is the characteristic function of the set A. We can improve the lower bounds of expectations using the procedure called the natural extension

$$\underline{E}'[f] = \inf\left\{ \sum_{x \in X} f(x)P(\{x\}) \middle| P \in \mathbf{P}(\underline{E}) \right\}.$$

Note that this procedure is not defined if $\mathbf{P}(\underline{E}) = \varnothing$. Let us remind that the functional $\underline{E}$ defines coherent lower previsions if $\underline{E}'[f] = \underline{E}[f]$ for all $f \in K'$.

Analogously, upper previsions are introduced. Any functional $\overline{E} : K' \to \mathbb{R}$ can be conceived as upper previsions. The upper previsions are not contradictory (or avoid sure loss) iff the credal set

$$\mathbf{P}(\overline{E}) = \left\{ P \in M_{pr} \middle| \forall f \in K' : \sum_{x \in X} f(x)P(\{x\}) \leq \overline{E}[f] \right\}.$$

is not empty. We can improve the upper bounds of expectations using the natural extension:

$$\overline{E}'[f] = \sup\left\{ \sum_{x \in X} f(x)P(\{x\}) \middle| P \in \mathbf{P}(\overline{E}) \right\}.$$

If $\overline{E}'[f] = \overline{E}[f]$ for all $f \in K'$, then $\overline{E}$ is a coherent lower prevision. Let us notice that we can equivalently describe uncertain information by lower or upper previsions. If the functional $\underline{E} : K' \to \mathbb{R}$ describes the lower previsions then we can equivalently describe the same information by upper previsions defined by

$$\overline{E}[f] = -\underline{E}[-f]$$

for all $-f \in K'$.

4 The Conjunctive and Disjunctive Rules for Aggregating Sources of Information

Consider n sources of information described by credal sets $\mathbf{P}_1, \ldots, \mathbf{P}_n$. Then there are several possible ways for aggregating this information that depends on prior assumptions. If we suppose that each source of information is reliable then we can aggregate them using intersection of the corresponding sets:

$$\mathbf{P} = \mathbf{P}_1 \cap \cdots \cap \mathbf{P}_n.$$

This rule of aggregation is called the *conjunctive rule (C-rule)*. It is easy to see that if we describe credal sets with the help of lower probability functions $\mu_1, \ldots, \mu_n$, then C-rule can be represented as

$$\mu = \mu_1 \vee \cdots \vee \mu_n,$$

where $\vee$ is the maximum operation. The last formula is justified because in this case

$$\mathbf{P}(\mu) = \mathbf{P}(\mu_1) \cap \cdots \cap \mathbf{P}(\mu_n)$$

If we describe sources of information by upper probabilities $\mu_1, \ldots, \mu_n$, then the C-rule is clearly expressed with the minimum operation $\wedge$ as

$$\mu = \mu_1 \wedge \cdots \wedge \mu_n$$

Analogously, the conjunctive rule is expressed in models based on lower previsions $E_i : K' \to \mathbb{R}$, $i = 1, \ldots, n$, or upper previsions $\overline{E}_i : K' \to \mathbb{R}$, $i = 1, \ldots, n$, as

$$\underline{E} = \underline{E}_1 \vee \cdots \vee \underline{E}_n, \quad \overline{E} = \overline{E}_1 \wedge \cdots \wedge \overline{E}_n. \tag{1}$$

We would like to emphasize that there are other rules for aggregation of information sources. If we know that at least one source of information is reliable and all sources of information are represented by credal sets $\mathbf{P}_1,\dots,\mathbf{P}_n$, then we can use the *disjunctive rule*, in which the result is the minimal credal set $\mathbf{P}$ that contains the corresponding credal sets $\mathbf{P}_i$, $i=1,\dots,n$. This disjunctive rule is expressed through lower previsions $E_i: K'\to\mathbb{R}$, $i=1,\dots,n$, or upper previsions $\overline{E}_i: K'\to\mathbb{R}$, $i=1,\dots,n$, as

$$\underline{E}=\underline{E}_1\wedge\dots\wedge\underline{E}_n,\quad \overline{E}=\overline{E}_1\vee\dots\vee\overline{E}_n.$$

The mixture rule can be used if we can evaluate the reliability of information. Let us assume this reliability is given by non-negative numbers a_i, $i=1,\dots,n$, such that $\sum_{i=1}^n a_i=1$. Then we can aggregate sources of information described by credal sets $\mathbf{P}_i$, $i=1,\dots,n$, as

$$\mathbf{P}=\left\{\sum_{i=1}^n a_iP_i\middle|P_i\in\mathbf{P}_i,i=1,\dots,n\right\}.$$

The counterparts of this rule for lower previsions $E_i: K'\to\mathbb{R}$, $i=1,\dots,n$, or upper previsions $\overline{E}_i: K'\to\mathbb{R}$, $i=1,\dots,n$, are

$$\underline{E}=\sum_{i=1}^n a_i\underline{E}_i\quad\text{or}\quad\overline{E}=\sum_{i=1}^n a_i\overline{E}_i.$$

Let us notice that other possible rules of aggregation have properties that more or less similar to the considered rules.

Let us observe that applying the C-rule is possible if the resulting credal set is not empty. In the opposite case we say that there is contradiction among sources of information. Meanwhile, in evidence theory the C-rule is also applicable if the sources of information are contradictory. In the next section we will introduce such C-rules, considered in [4], and give some hints about how they can be generalized in the theory of imprecise probabilities.

5 Conjunctive Rules of Aggregation in Evidence Theory, the Order of Specialization

Let $Bel_1=\sum_{A\in2^X}m_1(A)\eta_{\langle A\rangle}$ and $Bel_2=\sum_{B\in2^X}m_2(B)\eta_{\langle B\rangle}$ be belief functions. Then the *conjunctive combination rule* (C-rule)[3] [10, 4] is defined by

$$Bel=\sum_{A,B\in2^X}m(A,B)\eta_{\langle A\cap B\rangle},$$

where $m: 2^X\times2^X\to[0,1]$ is such that

$$\begin{cases}\sum_{B\in2^X}m(A,B)=m_1(A), & A\in2^X,\\ \sum_{A\in2^X}m(A,B)=m_2(B), & B\in2^X.\end{cases}\qquad(2)$$

[3] In [4] such combination rules are called generalized Dempster-Shafer rules.

Observe that we get the classical C-rule [8] if $m(A,B)=m_1(A)m_2(B)$ for any $A,B\in2^X$. The use of such general rule can be explained using the interpretation of belief functions through random sets. A random set ξ is a random variable taking its values in 2^X. Any such random variable can be defined by probabilities $P(\xi=A)$ and these probabilities can be identified with values $m(A)$ in evidence theory. Given two random sets ξ_1 and ξ_2 with values in 2^X. If we assume that these random sets are independent, then

$$P(\xi_1=A,\xi_2=B)=P(\xi_1=A)P(\xi_2=B).$$

The using of the classical C-rule means that from two sources of information described by independent random sets ξ_1 and ξ_2 we obtain a new random set ξ defined by

$$P(\xi=C)=\sum_{A\cap B=C}P(\xi_1=A)P(\xi_2=B).$$

Thus, the generalization of the classical C-rule can be got if we suppose that random sets ξ_1 and ξ_2 can be dependent. In this case we can only guarantee that the non-negative set function $m(A,B)=P(\xi_1=A,\xi_2=B)$ obeys (2).

Let us notice that the C-rule is not uniquely defined and it can be also applied in a case, when the sources of information are contradictory. The ways of choosing optimal conjunctive combination rules according to several justified criteria can be found in [4]. The main conclusion from [4] is that an optimal C-rule should be chosen among Pareto optimal C-rules w.r.t. the partial order on belief functions called specialization.

Let Bel_1 and Bel_2 be belief functions with bbas m_1 and m_2. We write $Bel_1\preceq Bel_2$ if Bel_2 can be obtained from Bel_1 using a linear contraction transform $\Phi: 2^X\times2^X\to[0,1]$, i.e. $m_2(B)=\sum_{A\in2^X}\Phi(A,B)m_1(A)$, and the set function $\Phi: 2^X\times2^X\to[0,1]$ has the following properties: $\sum_{B\in2^X}\Phi(A,B)=1$ for any $B\in2^X$ and $\Phi(A,B)=0$ if $B\not\subseteq A$. The partial order $\preceq$ is called *specialization.* It is easy to show [11] that $Bel_1\preceq Bel_2$ implies $Bel_1\le Bel_2$, but the opposite is not true in general. The main results [4] showing the connections of C-rules and the order $\preceq$ are given in the next propositions.

Proposition 1 *If Bel is the result of a C-rule applied to $Bel_1,Bel_2\in\overline{M}_{bel}$, then $Bel_1\preceq Bel$ and $Bel_2\preceq Bel$. Furthermore, each minimal element of the set*

$$\mathbf{Bel}(Bel_1,Bel_2)=\left\{Bel\in\overline{M}_{bel}|Bel_1\preceq Bel,Bel_2\preceq Bel\right\}$$

w.r.t. the order $\preceq$ for arbitrary $Bel_1,Bel_2\in\overline{M}_{bel}$ can be obtained by a C-rule.

This result shows that the optimal choice of a C-rule should be made to get the best approximation of the set function $\max\{Bel_1,Bel_2\}$ and this choice should obviously be made in the set of minimal elements of $\mathbf{Bel}(Bel_1,Bel_2)$ w.r.t. $\preceq$ that can be obtained by so called Pareto optimal C-rules.

Proposition 2 *The order $\preceq$ is equivalent to the order $\leq$ on the set $\overline{M}_{pr}$. In addition if $Bel \leq P$ for $P \in \overline{M}_{pr}$ and $Bel \in \overline{M}_{Bel}$, then $Bel \preceq P$. Furthermore,*

$$Bel(A) = \inf\{P(A) | P \in \mathbf{P}(Bel)\},$$

where $\mathbf{P}(Bel) = \left\{P \in \overline{M}_{pr} \middle| Bel \preceq P\right\}$.

Remark 1 Proposition 2 shows that in evidence theory any belief function can be equivalently represented by $\mathbf{P}(Bel)$ that may be called a generalized credal set. Such a construction with a slightly different definition will be introduced in the next section. Clearly, the above proposition allows us to write

$$\mathbf{P}(Bel) = \left\{P \in \overline{M}_{pr} \middle| Bel \leq P\right\}.$$

Let $Bel_1, Bel_2 \in \overline{M}_{bel}$. Then we denote by $R(Bel_1, Bel_2)$ the set of all possible belief measures that can be obtained by C-rules applied to Bel_1 and Bel_2. Then the amount of contradiction between Bel_1 and Bel_2 by C-rules can be computed as

$$Con(Bel_1, Bel_2) = \inf\left\{Bel(\varnothing) \middle| Bel \in R(Bel_1, Bel_2)\right\}.$$

Let us observe that this measure of contradiction (or conflict) is considered in many papers [4, 5, 6, 10], where authors show that $Con(Bel_1, Bel_2)$ has better properties than a measure of contradiction based on the classical C-rule.

Proposition 3 *Let $\mathbf{P}(Bel_i) = \left\{P \in \overline{M}_{pr} \middle| Bel_i \leq P\right\}$, where $Bel_i \in \overline{M}_{bel}$, $i = 1, 2$. Then*

$$Con(Bel_1, Bel_2) = \inf\{P(\varnothing) | P \in \mathbf{P}(Bel_1) \cap \mathbf{P}(Bel_2)\}.$$

Thus, in this section we has shown that it is possible to extend the model of non-normalized belief functions on more general theories of imprecise probabilities using generalized credal sets, and this problem will be investigated in the next sections.

6 The Conjunctive Rule for Probability Measures Admitting Contradiction

Let us consider the case when we have two sources of information described by probability measures P_1 and P_2. These sources of information are absolutely contradictory if we can divide the space X on two disjoint subsets A and B such that $P_1(A) = 1$ and $P_2(B) = 1$. In other words, sources of information support that events A and B are certain, but it is not possible because these events are disjoint. In classical logic false implies anything, thus we can write

$$P_1 \wedge P_2 = \bigwedge_{P_i \in M_{pr}} P_i = \eta^d_{\langle X \rangle},$$

where $\eta^d_{\langle X \rangle}$ describes the result of conjunction of all possible probability measures on 2^X. Now we will try to generalize the above rule for two probability measures that are not absolutely contradict each other. In this case we can divide probability measures on 2 parts:

$$P_1 = (1-a)P_1^{(1)} + aP_1^{(2)}, \quad P_2 = (1-a)P_2^{(1)} + aP_2^{(2)},$$

where $a \in [0,1]$, $P_k^{(i)} \in M_{pr}$, $i = 1, 2$, $k = 1, 2$, and $P_1^{(1)}$, $P_2^{(1)}$ are parts of probability measures that don't contradict each other, i.e. $P_1^{(1)} = P_2^{(1)}$, and probability measures $P_1^{(2)}$, $P_2^{(2)}$ are absolutely contradict each other. The value

$$Con(P_1, P_2) = a = 1 - \sum_{x_i \in X} \min\{P_1(\{x_i\}), P_2(\{x_i\})\}$$

is called the *amount of contradiction* and the above measures are defined by the following formulas:

$$P_1^{(1)}(\{x_i\}) = P_2^{(1)}(\{x_i\}) = \frac{1}{1-a} \min\{P_1(\{x_i\}), P_2(\{x_i\})\},$$

where $x_i \in X$ and $a < 1$ (if $a = 1$, then a measure $P_1^{(1)} = P_2^{(1)}$ is defined arbitrary);

$$P_1^{(2)}(\{x_i\}) = \frac{1}{a}\left(P_1(\{x_i\}) - (1-a)P_1^{(1)}(\{x_i\})\right),$$

$$P_2^{(2)}(\{x_i\}) = \frac{1}{a}\left(P_2(\{x_i\}) - (1-a)P_2^{(1)}(\{x_i\})\right),$$

where $x_i \in X$ and $a > 0$ (if $a = 0$, then absolutely contradictory measures $P_1^{(2)}$, $P_2^{(2)}$ are defined arbitrary).

Example 1 Assume that $X = \{x_1, x_2, x_3\}$. In this example any probability measure P can be described by a vector $\big(P(\{x_1\}), P(\{x_2\}), P(\{x_3\})\big)$. Let the probability measures P_1 and P_2 be defined by the following vectors: $P_1 = (0.4, 0.2, 0.4)$ and $P_2 = (0.2, 0.4, 0.4)$. Then $a = 0.2$, $P_1^{(1)} = P_2^{(1)} = (0.25, 0, 25, 0.5)$, $P_1^{(2)} = (1, 0, 0)$, and finally $P_2^{(2)} = (0, 1, 0)$.

Let us observe that measures $P_1^{(2)}$, $P_2^{(2)}$ are absolutely contradictory, because $P_1^{(2)}(\{x_1\}) = 1$ and $P_2^{(2)}(\{x_2\}) = 1$ for disjoint sets $\{x_1\}$ and $\{x_2\}$.

Summarizing we introduce the following definition.

Definition 1 The C-rule for probability measures $P_1, P_2 \in M_{pr}$ is defined as

$$P_1 \wedge P_2 = \sum_{x_i \in X} \min\{P_1(\{x_i\}), P_2(\{x_i\})\}\eta_{\langle\{x_i\}\rangle} + a\eta^d_{\langle X \rangle},$$

where $a = 1 - \sum_{x_i \in X} \min\{P_1(\{x_i\}), P_2(\{x_i\})\}$.

Example 2 Consider probability measures P_1 and P_2 from Example 1. Then

$$\begin{aligned} P_1 \wedge P_2 &= 0.8P_1^{(1)} + 0.2\eta^d_{\langle X \rangle} \\ &= 0.2\eta_{\langle\{x_1\}\rangle} + 0.2\eta_{\langle\{x_2\}\rangle} + 0.4\eta_{\langle\{x_1\}\rangle} + 0.2\eta^d_{\langle X \rangle}. \end{aligned}$$

Let $X = \{x_1, \dots, x_n\}$. In the next we will describe the contradiction in information using measures of the type

$$P = \sum_{i=1}^{n} a_i \eta_{\langle\{x_i\}\rangle} + a_0 \eta^d_{\langle X\rangle}, \tag{3}$$

where $a_i \geq 0$, $i = 0, \dots, n$, and $\sum_{i=0}^{n} a_i = 1$. Observe that $P \in M_{pr}$ if $a_0 = 0$ [4], and P is understood as a contradictory lower probability. If $a_0 > 0$, then the value a_0 gives us the amount of contradiction. The set of all possible measures, represented by (3), is denoted by $\overline{M}_{cpr}$. Let us notice that $M_{pr} \subseteq \overline{M}_{cpr}$.

Remark 2 Note that the set functions in $\overline{M}_{cpr}$ are plausibility functions with bba m such that $m(A) = 0$ if $1 < |A| < |X|$.

It is possible to describe the C-rule with the order $\leq$ on $\overline{M}_{cpr}$ considered as a partially ordered set.

Lemma 1 *Let* $P_1, P_2 \in \overline{M}_{cpr}$ *and* $P_1 = \sum_{i=1}^{n} a_i \eta_{\langle\{x_i\}\rangle} + a_0 \eta^d_{\langle X\rangle}$, $P_2 = \sum_{i=1}^{n} b_i \eta_{\langle\{x_i\}\rangle} + b_0 \eta^d_{\langle X\rangle}$. *Then* $P_1 \leq P_2$ *iff* $a_i \geq b_i$, $i = 1, \dots, n$.

Corollary 1 *Let* $P_1, \dots, P_m \in \overline{M}_{cpr}$ *and defined by*

$$P_k = \sum_{i=1}^{n} a_i^{(k)} \eta_{\langle\{x_i\}\rangle} + a_0^{(k)} \eta^d_{\langle X\rangle}$$

for $k = 1, \dots, m$, *then the exact upper bound of* $P_1, \dots, P_m$ *in* $\overline{M}_{cpr}$ *is*

$$P = \sum_{i=1}^{n} c_i \eta_{\langle\{x_i\}\rangle} + c_0 \eta^d_{\langle X\rangle},$$

where $c_i = \min\{a_i^{(1)}, \dots, a_i^{(m)}\}$ *for* $i = 1, \dots, n$, *and* $c_0 = 1 - \sum_{i=1}^{n} c_i$.

Remark 3 Corollary 1 implies that the C-rule of probability measures $P_1, P_2 \in M_{pr}$ is the exact upper bound of the set $\{P_1, P_2\}$. Thus, we define next the C-rule for arbitrary measures $P_1, \dots, P_m \in \overline{M}_{cpr}$ as the exact upper bound of the set $\{P_1, \dots, P_m\}$ in $\overline{M}_{cpr}$. This bound is denoted as $P_1 \wedge \dots \wedge P_m$.

Example 3 Let we take probability measures P_1 and P_2 from Example 1, and the probability measure $P_3 = (0.4, 0.4, 0.2)$, then

$$P_1 \wedge P_2 \wedge P_3 = 0.2\eta_{\langle\{x_1\}\rangle} + 0.2\eta_{\langle\{x_2\}\rangle} + 0.2\eta_{\langle\{x_3\}\rangle} + 0.4\eta^d_{\langle X\rangle}.$$

7 Generalized Upper and Lower Credal Sets

Observe that using measures from $\overline{M}_{cpr}$ we can describe contradictory and conflicting information. Let us remind (see e.g. [12, 2] for details) that pure conflict is described by probability measures, and the theory of imprecise probabilities allows us to model conflict and non-specificity (imprecision) in information, and non-specificity is caused by uncertainty in choosing a "true probability measure" among possible alternatives. If we try to describe imprecise information with some contradiction and conflict we should consider subsets of $\overline{M}_{cpr}$. Let us observe the following. Let $P_1 \in \overline{M}_{cpr}$, then $P_2 \in \overline{M}_{cpr}$ with $P_2 \geq P_1$ can be used for describing the same information but with a greater amount of contradiction. Thus, the subset $\mathbf{P}$ in $\overline{M}_{cpr}$ describing imprecise information has to satisfy the following property:

a) $P_1 \in \mathbf{P}$, $P_2 \in \overline{M}_{pr}$, $P_1 \leq P_2$ implies that $P_2 \in \mathbf{P}$.

The next two properties are essential for the most models of imprecise probabilities (cf. credal sets).

b) If $P_1, P_2 \in \mathbf{P}$ then $aP_1 + (1-a)P_2 \in \mathbf{P}$ for any $P_1, P_2 \in \mathbf{P}$ and $a \in [0,1]$.

c) The set $\mathbf{P}$ is closed in a sense that it can be considered as a subset of Euclidean space (any $P = a_0 \eta^d_{\langle X\rangle} + \sum_{i=1}^{n} a_i \eta_{\langle\{x_i\}\rangle}$ is a vector $(a_0, a_1, \dots, a_n)$ in $\mathbb{R}^{n+1}$).

Summarizing we can introduce the following definition.

Definition 2 A subset $\mathbf{P} \subseteq \overline{M}_{cpr}$ is called an *upper generalized credal set* if it satisfies conditions a), b), c).

The C-rule for generalized upper credal sets can be defined analogously as for usual credal sets.

Definition 3 Let $\mathbf{P}_1, \dots, \mathbf{P}_m$ be non-empty credal sets in $\overline{M}_{cpr}$. Then the credal set $\mathbf{P}$ produced by the C-rule is defined as $\mathbf{P} = \mathbf{P}_1 \cap \dots \cap \mathbf{P}_m$.

Let us introduce new concepts that help to understand this definition. Let $\mathbf{P}$ be a credal set in $\overline{M}_{cpr}$. A subset consisting of all minimal elements in $\mathbf{P}$ is called the *profile* of $\mathbf{P}$ and it is denoted by $profile(\mathbf{P})$. Evidently, any profile uniquely defines the corresponding credal set. If $\mathbf{P}$ describes information without contradiction, then $profile(\mathbf{P})$ is a credal set in usual sense, i.e. $profile(\mathbf{P}) \subseteq M_{pr}$. In particular, if we have two credal sets $\mathbf{P}_1, \mathbf{P}_2$ in $\overline{M}_{cpr}$ with $profile(\mathbf{P}_i) \in M_{pr}$, $i = 1, 2$, then applying the C-rule gives us the profile:

$$profile(\mathbf{P}_1 \cap \mathbf{P}_2) = profile(\mathbf{P}_1) \wedge profile(\mathbf{P}_2).$$

Observe that any upper generalized credal set give us many lower possible bounds of probabilities and each possible value is characterized by contradiction. Let us denote the amount of contradiction in $P \in \overline{M}_{cpr}$ by $Con(P)$. Then to characterize the possible lower bounds of probabilities computed by an upper generalized credal set $\mathbf{P}$ we introduce into consideration the set function

$$\mu^r(A) = \inf\{P(A) | P \in \mathbf{P}, Con(P) \leq r\},$$

where $A \in 2^X$ and $r \in [0,1]$ is the level of contradiction. The set function μ^r can be interpreted as a lower probability for the credal set $\mathbf{P}$ with a level of contradiction r.

Lemma 2 *For any upper generalized credal set* $\mathbf{P}$*:*

$$\mu^r(A) = \inf\{P(A) | P \in profile(\mathbf{P}), Con(P) \leq r\}.$$

[4] Observe that if $P \in M_{pr}$, then $P = \sum_{i=1}^{n} P(\{x_i\}) \eta_{\langle\{x_i\}\rangle}$.

Remark 4 We can consider the generalized upper credal sets whose profiles are credal sets in usual sense. In a case, when profiles of upper generalized credal sets are credal sets in usual sense, μ^r does not depend on r, and the considered model coincides with the model of imprecise probabilities based on usual credal sets.

Example 4 Let $X = \{x_1, x_2, x_3\}$. Then any

$$P = a_1\eta_{\langle\{x_1\}\rangle} + a_2\eta_{\langle\{x_2\}\rangle} + a_3\eta_{\langle\{x_3\}\rangle} + a_0\eta^d_{\langle X\rangle}$$

in $\overline{M}_{cpr}$ can be defined by the vector $P = (a_1, a_2, a_3, a_0)$. Consider upper generalized credal sets $\mathbf{P}_i$, $i = 1,2,3$, whose profiles are credal sets in usual sense:

$$profile(\mathbf{P}_1) = \{aP_1 + (1-a)P_2 | t \in [0,1]\},$$
$$profile(\mathbf{P}_2) = \{P_3\}, \quad profile(\mathbf{P}_3) = \{P_4\},$$

where

$$P_1 = (2/3, 0, 1/3, 0), \qquad P_2 = (0, 2/3, 1/3, 0),$$
$$P_3 = (1/3, 1/3, 1/3, 0), \qquad P_4 = (1/3, 1/2, 1/6, 0).$$

Let us find the profile of $\mathbf{P}_1 \cap \mathbf{P}_2$. It obviously consists of minimal elements in the set

$$\{P' \wedge P'' | P' \in profile(\mathbf{P}_1), P'' \in profile(\mathbf{P}_2)\} =$$
$$\{P | P = (1/3, t, 1/3, 1/3 - t), t \in [0, 1/3]\} \cup$$
$$\{P | P = (t, 1/3, 1/3, 1/3 - t), t \in [0, 1/3]\}.$$

The above set has only one minimal element, namely, $P_5 = (1/3, 1/3, 1/3, 0)$. Therefore, $profile(\mathbf{P}_1 \cap \mathbf{P}_2) = \{P_5\}$.

Analogously, let us find the profile of $\mathbf{P}_1 \cap \mathbf{P}_3$. It consists of minimal elements in the set

$$\{P' \wedge P'' | P' \in profile(\mathbf{P}_1), P'' \in profile(\mathbf{P}_3)\} =$$
$$\{P | P = (2t/3, 1/2, 1/6, 1/3 - 2t/3), t \in [0, 1/4)\} \cup$$
$$\{P | P = (2t/3, 2(1-t)/3, 1/6, 1/6), t \in [1/4, 1/2]\} \cup$$
$$\{P | P = (1/3, 2(1-t)/3, 1/6, 2t/3 - 1/6), t \in (1/2, 1]\}.$$

The minimal elements of this set are $tP_6 + (1-t)P_7$, where $t \in [0,1]$, and

$$P_6 = (1/6, 1/2, 1/6, 1/6), \quad P_7 = (1/3, 1/3, 1/6, 1/6).$$

Thus, $profile(\mathbf{P}_1 \cap \mathbf{P}_3) = \{tP_6 + (1-t)P_7 | t \in [0,1]\}$.

Let us show next how it is possible to define lower bounds of expectation. Consider first expectations w.r.t. measures in $\overline{M}_{cpr}$. If $P \in M_{pr}$, then for any function $f : X \to \mathbb{R}$ we define the expectation $E_P(f)$ as

$$E_P(f) = \sum_{x \in X} f(x)P(\{x\}).$$

We can extend the functional E_P to the set of all measures in $\overline{M}_{cpr}$, using the considered interpretation of a measure $P \in \overline{M}_{cpr}$ through the C-rule. Obviously, $P = \wedge_{P_i \in M_{pr} | P_i \le P} P_i$. Then this C-rule is expressed through expectations E_{P_i}, $P_i \le P$, as (cf. formula (1))

$$\underline{E}_P = \bigvee_{P_i \in M_{pr} | P_i \le P} E_{P_i}.$$

Lemma 3 *For any* $P = a_0\eta^d_{\langle X\rangle} + \sum_{i=1}^n a_i\eta_{\langle\{x_i\}\rangle}$ *and* $f : X \to \mathbb{R}$ *the value* $\underline{E}_P(f)$ *can be computed as*

$$\underline{E}_P(f) = a_0 \max_{x \in X} f(x) + \sum_{i=1}^n a_i f(x_i).$$

Let $\mathbf{P}$ be a credal set in $\overline{M}_{cpr}$. We will define first the lower expectation $\underline{E}_{\mathbf{P}}(f)$ for non-negative functions $f : X \to \mathbb{R}$. Let the set of all such functions be denoted by K^+. Because $\underline{E}_{\mathbf{P}}(f)$ is the lower expectation, we can define this value for any $f \in K^+$ as

$$\underline{E}_{\mathbf{P}}(f) = \inf_{P \in \mathbf{P}} \underline{E}_P(f).$$

Example 5 Let $\mathbf{P} = \mathbf{P}_1 \cap \mathbf{P}_3$, where $\mathbf{P}_1 \cap \mathbf{P}_3$ is defined in Example 4, then

$$\underline{E}_{\mathbf{P}}(f) = \min\left\{\underline{E}_{P_6}(f), \underline{E}_{P_7}(f)\right\},$$

where

$$\underline{E}_{P_6}(f) = \frac{1}{6}f(x_1) + \frac{1}{2}f(x_2) + \frac{1}{6}f(x_3) + \frac{1}{6}\max_{x_i \in X} f(x_i),$$
$$\underline{E}_{P_7}(f) = \frac{1}{3}f(x_1) + \frac{1}{3}f(x_2) + \frac{1}{6}f(x_3) + \frac{1}{6}\max_{x_i \in X} f(x_i).$$

Let us indicate some properties of $\underline{E}_{\mathbf{P}}$ on K^+. In the next we denote by $\mathbb{R}^+$ the set of all non-negative real numbers. The function in K^+ with values equal to $a \in \mathbb{R}^+$ is denoted also by a. We write $f_1 \le f_2$ for $f_1, f_2 \in K^+$ if $f_1(x) \le f_2(x)$ for all $x \in X$.

Lemma 4 *The functional* $\underline{E}_{\mathbf{P}}$ *on* K^+ *has the following properties:*

1) $\underline{E}_{\mathbf{P}}(0) = 0$*;* $\underline{E}_{\mathbf{P}}(1) = 1$*;*

2) $\underline{E}_{\mathbf{P}}(f + a) = \underline{E}_{\mathbf{P}}(f) + a$ *for any* $f \in K^+$ *and* $a \in \mathbb{R}^+$*;*

3) $\underline{E}_{\mathbf{P}}(af) = a\underline{E}_{\mathbf{P}}(f)$ *for any* $f \in K^+$ *and* $a \in \mathbb{R}^+$*;*

4) $\underline{E}_{\mathbf{P}}(f_1) \le \underline{E}_{\mathbf{P}}(f_2)$ *for* $f_1, f_2 \in K^+$ *if* $f_1 \le f_2$.

Let us consider also the dual concept of generalized upper credal sets. In this case we describe uncertainty by set functions from the set $\overline{M}^d_{cpr}$. Any measure P in $\overline{M}^d_{cpr}$ is represented as

$$P = a_0\eta_{\langle X\rangle} + \sum_{i=1}^n a_i\eta_{\langle\{x_i\}\rangle},$$

where $a_i \ge 0$, $i = 0, \ldots, n$, and $\sum_{i=0}^n a_i = 1$, and it is conceived as an upper probability. The value a_0 shows the

amount of contradiction. If $a_0 = 0$, then P is a probability measure. Evidently, measures from $\overline{M}^d_{cpr}$ describe conflict and contradiction in information and we can define the upper expectation $\overline{E}_P(f)$ for any $f \in K$ w.r.t. arbitrary P in $\overline{M}^d_{cpr}$ through the Choquet integral:

$$\overline{E}_P(f) = \int_X f(x)dP = a_0 \min_{x\in X} f(x) + \sum_{i=1}^{n} a_i f(x_i).$$

For describing conflict, contradiction and non-specificity with the help of measures in $\overline{M}^d_{cpr}$, we introduce the notion of lower generalized credal set.

Definition 4 A *lower generalized credal set* $\mathbf{P}$ is a non-empty subset of $\overline{M}^d_{cpr}$ with the following properties:

a) $P_1 \in \mathbf{P}$, $P_2 \in \overline{M}^d_{cpr}$, $P_1 \geq P_2$ implies that $P_2 \in \mathbf{P}$.

b) if $P_1, P_2 \in \mathbf{P}$, then $aP_1 + (1-a)P_2 \in \mathbf{P}$ for any $P_1, P_2 \in \mathbf{P}$ and $a \in [0,1]$.

c) $\mathbf{P}$ is closed set if we consider it as a subset of Euclidean space (any $P = a_0\eta^d_{\langle X\rangle} + \sum_{i=1}^n a_i \eta_{\langle\{x_i\}\rangle}$ is a vector $(a_0, a_1, \ldots, a_n)$ in $\mathbb{R}^{n+1}$).

The set of all maximal elements in a generalized lower credal set $\mathbf{P}$ is called the *profile* of $\mathbf{P}$ and it is denoted by *profile*$(\mathbf{P})$. Emphasize that generalized lower and upper credal sets are dual concepts, for instance, if $\mathbf{P}$ is a credal set in $\overline{M}_{cpr}$, then $\mathbf{P}^d$ is a credal set in $\overline{M}^d_{cpr}$; profiles of $\mathbf{P}$ and $\mathbf{P}^d$ are also connected with the dual relation: *profile*$(\mathbf{P})^d =$ *profile*$(\mathbf{P}^d)$; if $\mathbf{P}_1, \ldots, \mathbf{P}_m$ are credal sets in $\overline{M}_{cpr}$, then the expression for the C-rule is defined by the same way for the credal sets in $\overline{M}_{cpr}$ and $\overline{M}^d_{cpr}$, and

$$(\mathbf{P}_1 \cap \cdots \cap \mathbf{P}_m)^d = \mathbf{P}_1^d \cap \cdots \cap \mathbf{P}_m^d.$$

The upper expectation $\overline{E}_{\mathbf{P}}(f)$ of $f \in K^+$ w.r.t. the credal set $\mathbf{P}$ in $\overline{M}^d_{cpr}$ is defined as follows:

$$\overline{E}_{\mathbf{P}}(f) = \sup_{P\in\mathbf{P}} \overline{E}_P(f).$$

It is easy to check that the functional $\overline{E}_{\mathbf{P}}$ obeys the same properties as $\underline{E}_{\mathbf{P}}$ described in Lemma 4. The duality property of functionals $\underline{E}_{\mathbf{P}}$ and $\overline{E}_{\mathbf{P}}$ on K^+ is described in the following lemma.

Lemma 5 *$\overline{E}_{\mathbf{P}^d}(f) = a - \underline{E}_{\mathbf{P}}(a-f)$, where $\mathbf{P}$ is a credal set in $\overline{M}_{cpr}$, $f \in K^+$, and $a = \max_{x\in X} f(x)$.*

Remark 5 In the next we will extend functionals $\underline{E}_{\mathbf{P}}$ and $\overline{E}_{\mathbf{P}}$ on the set K of all real valued functions, assuming that the property 2) from Lemma 4 is valid for functions in K. Then for any $f \in K$ the values $\underline{E}_{\mathbf{P}}(f)$ and $\overline{E}_{\mathbf{P}}(f)$ are computed by

$$\underline{E}_{\mathbf{P}}(f) = \underline{E}_{\mathbf{P}}(\underline{f}) + a, \quad \overline{E}_{\mathbf{P}}(f) = \overline{E}_{\mathbf{P}}(\underline{f}) + a,$$

where $a = \min_{x\in X} f(x)$, and $\underline{f} = f - a$. Clearly $\underline{f} \in K^+$ and there exists $x \in X$ such that $\underline{f}(x) = 0$. We will call such functions *normalized* and keep the notation $\underline{f}$ (using lower bar).

Example 6 Let us consider the lower generalized credal set $\mathbf{P}^d = (\mathbf{P}_1 \cap \mathbf{P}_3)^d$, where $\mathbf{P}_1 \cap \mathbf{P}_3$ is defined in Example 4. Then we can compute $\overline{E}_{\mathbf{P}^d}(f)$ for any $f \in K^+$ as

$$\overline{E}_{\mathbf{P}^d}(f) = \max\left\{\overline{E}_{P_6^d}(f), \overline{E}_{P_7^d}(f)\right\},$$

where

$$\overline{E}_{P_6^d}(f) = \frac{1}{6}f(x_1) + \frac{1}{2}f(x_2) + \frac{1}{6}f(x_3) + \frac{1}{6}\min_{x_i\in X} f(x_i),$$
$$\overline{E}_{P_7^d}(f) = \frac{1}{3}f(x_1) + \frac{1}{3}f(x_2) + \frac{1}{6}f(x_3) + \frac{1}{6}\min_{x_i\in X} f(x_i).$$

Observe that for normalized functions $\frac{1}{6}\min_{x_i\in X} \underline{f}(x_i) = 0$. Let us compute $\overline{E}_{\mathbf{P}^d}(f)$ if $f = (f(x_1), f(x_2), f(x_3)) = (1,1,-3)$. Then $\min_{x_i\in X} f(x_i) = -3$, $\underline{f} = (4,4,0)$,

$$\overline{E}_{P_6^d}(f) = \overline{E}_{P_6^d}(\underline{f}) - 3 = \frac{4}{6} + \frac{4}{2} + 0 - 3 = -\frac{1}{3}.$$

Let us notice that all properties formulated in Lemma 4 remain valid for functionals $\underline{E}_{\mathbf{P}}$ and $\overline{E}_{\mathbf{P}}$ on K. The dual relation between $\underline{E}_{\mathbf{P}}$ and $\overline{E}_{\mathbf{P}}$ can be reformulated as $\overline{E}_{\mathbf{P}^d}(f) = -\underline{E}_{\mathbf{P}}(-f)$ for any credal set in $\overline{M}_{cpr}$ and $f \in K$.

The next lemma gives us the additional characteristic property of $\overline{E}_{\mathbf{P}}$, which, we will see later, helps us to describe the whole set of functionals $\underline{E}_{\mathbf{P}}$ and $\overline{E}_{\mathbf{P}}$.

Lemma 6 *Let $\underline{f}_1, \underline{f}_2, \underline{f}_3$ be normalized functions in K^+ such that $\underline{f}_1 + \underline{f}_2 = \underline{f}_3$. Then for any credal set $\mathbf{P}$ in $\overline{M}^d_{cpr}$ it is valid $\overline{E}_{\mathbf{P}}(\underline{f}_1) + \overline{E}_{\mathbf{P}}(\underline{f}_2) \geq \overline{E}_{\mathbf{P}}(\underline{f}_3)$.*

Theorem 1 *A functional $\Phi : K^+ \to \mathbb{R}$ coincides with $\overline{E}_{\mathbf{P}}$ on K^+ for some credal set $\mathbf{P}$ in $\overline{M}^d_{cpr}$ iff it has the following properties:*

1) $\Phi(0) = 0$; $\Phi(1) = 1$;

2) $\Phi(f + a) = \Phi(f) + a$ for any $f \in K^+$ and $a \in \mathbb{R}^+$;

3) $\Phi(af) = a\Phi(f)$ for any $f \in K^+$ and $a \in \mathbb{R}^+$;

4) $\Phi(f_1) \leq \Phi(f_2)$ for $f_1, f_2 \in K^+$ if $f_1 \leq f_2$;

5) $\Phi(\underline{f}_1) + \Phi(\underline{f}_2) \geq \Phi(\underline{f}_3)$ for any normalized functions $\underline{f}_1, \underline{f}_2, \underline{f}_3$ in K^+ such that $\underline{f}_1 + \underline{f}_2 = \underline{f}_3$.

8 Generalized Coherent Upper Previsions

Let $K' \subseteq K$, where K is the set of all functions of the type $f : X \to \mathbb{R}$, and let $\overline{E} : K' \to \mathbb{R}$ be the functional that defines the upper previsions, that may not satisfy the avoiding sure loss condition. Then $\overline{E}$ defines the non-empty lower generalized credal set $\mathbf{P}$ in $\overline{M}^d_{cpr}$ as follows:

$$\mathbf{P} = \left\{P \in \overline{M}^d_{cpr} | \forall f \in K' : \overline{E}_P(f) \leq \overline{E}(f)\right\} \qquad (4)$$

iff $\inf_{x\in X} f(x) \leq \overline{E}(f)$ for all $f \in K'$. Based on generalized credal set $\mathbf{P}$, we can define the natural extension of $\overline{E}$ by

$$\overline{E}'(f) = \sup\left\{\overline{E}_P(f) | P \in \mathbf{P}\right\} = \overline{E}_{\mathbf{P}}(f)$$

for all $f \in K$.

Theorem 2 *Let $\overline{E}: K' \to \mathbb{R}$ be the functional that defines the upper previsions. Then its natural extension $\overline{E}': K \to \mathbb{R}$ based on generalized credal sets can be computed as*

$$\overline{E}'(\underline{f}) = \inf\left\{\sum_k a_k \overline{E}(\underline{f_k}) + a \,\middle|\, \sum_k a_k \underline{f_k} + a\mathbf{1} \geq \underline{f}, f_k \in K', a_k, a \geq 0\right\},$$

where $\underline{f}$ and $\underline{f_k}$ are normalized functions, and $\overline{E}'(\underline{f}) = \overline{E}'(f) - b$, $\overline{E}(\underline{f_k}) = \overline{E}(f_k) - b_k$, $b = \min_{x\in X} f(x)$, and $b_k = \min_{x\in X} f_k(x)$.

9 Conclusion

In this paper we generalize the C-rule for general theories of imprecise probabilities using the way of modeling contradiction (conflict) in evidence theory. This allows us to introduce upper and lower generalized credal sets and represent the C-rule as the intersection of corresponding generalized credal sets. The paper contains also some insights of how this model can be used in the theory of imprecise probabilities admitting contradiction.

Appendix[5]

Proof (Lemma 1) *Necessity.* Let $P_1 \leq P_2$, then in particular, $P_1(X \setminus \{x_i\}) \leq P_2(X \setminus \{x_i\})$, $i = 1,\dots,n$, or equivalently, $1 - a_i \leq 1 - b_i$, or $a_i \geq b_i$, $i = 1,\dots,n$.

Sufficiency. Let $a_i \geq b_i$, $i = 1,\dots,n$, then

$$\begin{aligned} P_1 &= \sum_{i=1}^n \left(b_i \eta_{\langle\{x_i\}\rangle} + (a_i - b_i)\eta_{\langle\{x_i\}\rangle}\right) + a_0 \eta^d_{\langle X\rangle} \\ &\leq \sum_{i=1}^n \left(b_i \eta_{\langle\{x_i\}\rangle} + (a_i - b_i)\eta^d_{\langle X\rangle}\right) + a_0 \eta^d_{\langle X\rangle} = P_2. \end{aligned}$$

Proof (Lemma 2) Because the set $\mathbf{P}$ is closed, we have $\mathbf{P} = \{P \in \overline{M}_{cpr} | \exists P' \in profile(\mathbf{P}) : P \geq P'\}$. This implies the required result.

Proof (Lemma 3) Because P is a plausibility function (2-alternating measure), the value $\underline{E}_P(f)$ is expressed through the Choquet integral:

$$\begin{aligned} \underline{E}_P(f) = \int_X f(x) dP &= a_0 \int_X f(x) d\eta^d_{\langle X\rangle} + \sum_{i=1}^n a_i \int_X f(x) d\eta_{\langle\{x_i\}\rangle} \\ &= a_0 \max_{x\in X} f(x) + \sum_{i=1}^n a_i f(x_i). \end{aligned}$$

In the last expression we use the additivity of the Choquet integral w.r.t. the sum of measures, and also that $\int_X f(x) d\eta_{\langle\{x_i\}\rangle} = f(x_i)$ and $\int_X f(x) d\eta^d_{\langle X\rangle} = \max_{x\in X} f(x)$.

Proof (Lemma 5) Notice that the validity of $\overline{E}_{P^d}(f) = a - \underline{E}_P(a - f)$ for $P \in \overline{M}_{cpr}$ follows from the properties of the Choquet integral. By definition

$$\begin{aligned} \overline{E}_{\mathbf{P}^d}(f) &= \sup_{P^d \in \mathbf{P}^d} \overline{E}_{P^d}(f) = \sup_{P\in\mathbf{P}} (a - \underline{E}_P(a - f)) \\ &= a - \inf_{P\in\mathbf{P}} \underline{E}_P(a - f) = a - \underline{E}_{\mathbf{P}}(a - f). \end{aligned}$$

[5] Straightforward proofs are omitted.

Proof (Lemma 6) Because by definition the credal set $\mathbf{P}$ is closed, there exists $P \in \mathbf{P}$ such that $\overline{E}_P\left(\underline{f_3}\right) = \overline{E}_{\mathbf{P}}(\underline{f_3})$. Assume that $P = a_0 \eta_{\langle X\rangle} + \sum_{i=1}^n a_i \eta_{\langle\{x_i\}\rangle}$. Notice that in this case

$$\overline{E}_P(\underline{f_k}) = \sum_{i=1}^n a_i \underline{f_k}(x_i), \quad k = 1,2,3,$$

since $\min_{x\in X} \underline{f_k}(x) = 0$. Thus, $\overline{E}_P(\underline{f_1}) + \overline{E}_P(\underline{f_2}) = \overline{E}_P(\underline{f_3})$. In addition, clearly $\overline{E}_{\mathbf{P}}(\underline{f_k}) \geq \overline{E}_P(\underline{f_k})$, $k = 1,2$. This implies the inequality from the lemma.

Proof (Theorem 1) Necessity follows from Lemma 4 (see Remark 5) and Lemma 6. Let us prove sufficiency. It is sufficient to show that for any normalized function $\underline{f}$ there is a $P \in \overline{M}^d_{cpr}$ such that $\Phi(\underline{f}) = \overline{E}_P(\underline{f})$ and $\Phi \geq \overline{E}_P$. Because $\underline{f}$ is normalized there is $x_k \in X$ such that $\underline{f}(x_k) = 0$. Let us consider the set K' of all functions f in K^+ with $f(x_k) = 0$. Let us notice that the monotone functional Φ on K' is sublinear, and by Hahn-Banach's Theorem there is a linear functional on K'

$$\alpha(f) = \sum_{i=1}^n a_i f(x_i)$$

such that $a_i \geq 0$, $i = 1,\dots,n$, $\sum_{i=1}^n a_i \leq 1$, $\alpha \leq \Phi$ and $\alpha(\underline{f}) = \Phi(\underline{f})$. Obviously, we can assume that $a_k = 0$. Introduce into consideration

$$P = a_0 \eta_{\langle X\rangle} + \sum_{i=1}^n a_i \eta_{\langle\{x_i\}\rangle},$$

where $a_0 = 1 - \sum_{i=1}^n a_i$ and show that $\Phi(\underline{f}) = \overline{E}_P(\underline{f})$ and $\Phi \geq \overline{E}_P$. The equality $\Phi(\underline{f}) = \overline{E}_P(\underline{f})$ is obvious. Let us show that $\Phi(g) \geq \overline{E}_P(g)$ for any $g \in K^+$. Obviously, $\Phi(g) \geq \overline{E}_P(g)$ iff $\Phi(\underline{g}) \geq \overline{E}_P(\underline{g})$. Notice that $\overline{E}_P(\underline{g}) = \overline{E}_P(\underline{g'})$, where $\underline{g'}(x_i) = \underline{g}(x_i)$ for $i \neq k$ and $\underline{g'}(x_i) = 0$ otherwise. Since $\underline{g'} \leq \underline{g}$, we get $\overline{E}_P(\underline{g}) = \overline{E}_P(\underline{g'}) \leq \Phi(\underline{g'}) \leq \Phi(\underline{g})$. The theorem is fully proved.

Proof (Theorem 2) Let us show first that functionals $\overline{E}$ and $\overline{E}'$ define the same credal set, i.e. the credal set $\mathbf{P}$ defined by (4) is equal to

$$\mathbf{P}' = \left\{P \in \overline{M}^d_{cpr} \middle| \forall f \in K : \overline{E}_P(f) \leq \overline{E}'(f)\right\}.$$

The inclusion $\mathbf{P}' \subseteq \mathbf{P}$ is obvious. Let $P \in \mathbf{P}$, then by our assumption $\overline{E}_P(\underline{f_k}) \leq \overline{E}(\underline{f_k})$ for $f_k \in K'$ and

$$\begin{aligned} \overline{E}_P(\underline{f}) = \sum_{i=1}^n P(\{x_i\}) \underline{f}(x_i) &\leq \sum_{i=1}^n P(\{x_i\}) \left(\sum_k a_k \underline{f_k}(x_i) + a\right) \\ &\leq \sum_{i=1}^n P(\{x_i\}) \sum_k a_k \underline{f_k}(x_i) + a \\ &= \sum_k a_k \overline{E}_P(\underline{f_k}) + a \\ &\leq \sum_k a_k \overline{E}(\underline{f_k}) + a. \end{aligned}$$

Thus, $\mathbf{P} \subseteq \mathbf{P}'$, i.e. $\mathbf{P}' = \mathbf{P}$. Let us show that the functional $\overline{E}'$ obeys all properties on K^+ for functional Φ given in Theorem 1. It is easy to check that properties 1), 2), 3), 5) are valid. Let us show that the monotonicity property 4) is also satisfied. For this purpose introduce into consideration the functional

$$\Phi(f) = \inf\left\{\sum_k a_k \overline{E}(\underline{f_k}) + a \,\middle|\, \sum_k a_k \underline{f_k} + a\mathbf{1} \geq f, f_k \in K', a_k, a \geq 0\right\}$$

on K^+. Evidently, $\overline{E}'(\underline{f}) = \Phi(\underline{f})$ for every $f \in K^+$. It is easy to check that this functional on K^+ has the following properties:

1) $\Phi(\mathbf{0}) = 0$, $\Phi(\mathbf{1}) \le 1$;
2) $\Phi(af) = a\Phi(f)$ for any $f \in K^+$ and $a \in \mathbb{R}^+$;
3) $\Phi(f_1) \le \Phi(f_2)$ for $f_1, f_2 \in K^+$ if $f_1 \le f_2$;
4) $\Phi(f_1)+\Phi(f_2) \ge \Phi(f_3)$ for any functions f_1, f_2, f_3 in K^+ such that $f_1 + f_2 = f_3$.

By Hahn-Banach's Theorem for every $f \in K^+$ there is a linear functional on K^+, $\alpha(f) = \sum_{i=1}^n a_i f(x_i)$, such that $a_i \ge 0$, $i = 1,\dots,n$, $\sum_{i=1}^n a_i \le 1$, $\alpha \le \Phi$ and $\alpha(f) = \Phi(f)$. We will use next this functional for proving monotonicity of $\overline{E}'$. Consider an arbitrary $f, g \in K^+$ such that $f \le g$. Let $f = \underline{f} + c$. Then inequality $\overline{E}'(f) \le \overline{E}'(g)$ is equivalent to $\overline{E}'(\underline{f}) \le \overline{E}'(g')$, where $g' = g - c$. Obviously, $\overline{E}'(\underline{f}) = \Phi(\underline{f}) \le \Phi(g')$. By previous conclusions, there is a linear functional $\alpha(f) = \sum_{i=1}^n a_i f(x_i)$ on K^+ such that $a_i \ge 0$, $i = 1,\dots,n$, $\sum_{i=1}^n a_i \le 1$, $\alpha \le \Phi$ and $\alpha(g') = \Phi(g')$. Let $P = a_0\eta_{\langle X\rangle} + \sum_{i=1}^n a_i \eta_{\langle\{x_i\}\rangle}$, where $a_0 = 1 - \sum_{i=1}^n a_i$. It is easy to see that $P \in \mathbf{P}$ and $\Phi(g') \le \overline{E}_P(g') \le \overline{E}'(g')$, i.e. $\overline{E}'(\underline{f}) \le \overline{E}'(g')$ and $\overline{E}'(f) \le \overline{E}'(g)$.

Thus, we prove that the functional $\overline{E}'$ obeys all properties from Theorem 1. This means that it is the natural extension of $\overline{E}$.

Acknowledgements

This study (Research Grant No.14-01-0015) was supported by The National Research University Higher School of Economics Academic Fund Program in 2014/2015.

Authors express their sincerely thanks to the anonymous reviewers for detailed and helpful comments.

References

[1] T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, eds. *Introduction to Imprecise Probabilities*. New York: Wiley, 2014.

[2] A. G. Bronevich and G. J. Klir, Measures of uncertainty for imprecise probabilities: an axiomatic approach. *International Journal of Approximate Reasoning* 51: 365-390, 2010.

[3] A. G. Bronevich and I. N. Rozenberg. The choice of generalized Dempster-Shafer rules for aggregating belief functions based on imprecision indices. *Belief Functions: Theory and Applications. Lecture Notes in Computer Science*, vol. 8764, Springer Verlag, Berlin, 2014, pp 21-28.

[4] A. G. Bronevich and I. N. Rozenberg. The choice of generalized Dempster-Shafer rules for aggregating belief functions. *International Journal of Approximate Reasoning* 56: 122-136, 2015.

[5] M. E. G. V. Cattaneo. Combining belief functions issued from dependent sources, in: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (Eds.), ISIPTA '03, Proceedings in Informatics, vol. 18, Carleton Scientific,Waterloo, 2003, pp. 133-147.

[6] M. E. G. V. Cattaneo. Belief functions combination without the assumption of independence of the information sources. *International Journal of Approximate Reasoning* 52: 299-315, 2011.

[7] G. de Cooman and M. C. M. Troffaes. *Lower Previsions*. New York: Wiley, 2014.

[8] T. Denoeux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artificial Intelligence* 172: 234-264, 2008.

[9] S. Destercke and V. Antoine. Combining imprecise probability masses with maximal coherent subsets: Application to ensemble classification. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis Advances in Intelligent Systems and Computing,* Springer, vol. 190, 2013, pp 27-35.

[10] S. Destercke and T. Burger. Toward an axiomatic definition of conflict between belief functions, IEEE Trans. Syst. Man Cybern. 43: 585-596, 2013.

[11] D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12:193-226, 1986.

[12] G. J. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory.* Hoboken, NJ: Wiley-Interscience, 2006.

[13] S. Moral and J. Sagrado. Aggregation of imprecise probabilities. In B. Bouchon Meunier, ed., *Aggregation and Fusion of Imperfect Information*, pages 162-188. Physica-Verlag, Heidelberg, 1997.

[14] R. Nau. The aggregation of imprecise probabilities. *Journal of Statistical Planning end Inference* 105: 265-282, 2002.

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

[16] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion* 8: 387-412, 2007.

[17] M. C. M. Troffaes. Generalising the conjunction rule for aggregating conflicting expert opinions. *International Journal of Intelligent Systems* 21: 361-380, 2006.

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

Decisions under Risk and Partial Knowledge Modelling Uncertainty and Risk Aversion

Giulianella Coletti and **Davide Petturiti**
University of Perugia, Italy
{coletti,davide.petturiti}@dmi.unipg.it

Barbara Vantaggi
"La Sapienza" University of Rome, Italy
barbara.vantaggi@sbai.uniroma1.it

Abstract

We deal with decisions under risk starting from a partial preference relation on a finite set of generalized convex lotteries, that are random quantities equipped with a convex capacity. A necessary and sufficient condition (Choquet rationality) is provided for its representability as a Choquet expected utility of a strictly increasing utility function. The restriction to concave utility functions is discussed. Moreover, we show that this condition, with or without the constraint of concavity for the utility function, assures the extension of the preference relation and it actually guides the decision maker in the extension process.

Keywords. Preference, Choquet rationality, Concave utility, Choquet expected utility.

1 Introduction

The classical axioms of the von Neumann-Morgenstern decision theory under risk [26] assure that a preference relation on lotteries, i.e., random quantities endowed with a probability distribution, is representable by an expected utility (EU). In this setting the decision maker behaves like an expected utility maximizer.

The assumptions behind the EU theory implicitly rely on a common probability measure which determines the lotteries.

Nevertheless, in situations of incomplete and revisable information, uncertainty cannot always be handled through a probability, but it is often unavoidable to refer to a class of probabilities and so to its lower envelope, which is a non-additive uncertainty measure (such as a belief function, a convex capacity or a lower probability [8, 22, 28]).

For example, in situations like that considered in the Ellsberg paradox [11], a convex capacity is obtained as lower envelope of the probabilities extending a partial probabilistic assessment. Note that, as is well-known, a lower envelope could not be convex in general [27]. The lower envelope is indeed surely convex (actually it is a belief function) when the probability is defined on an algebra and it is extended to a super-algebra [7].

In the following we restrict to convex capacities, which are used to express "objective" uncertainty on the prizes of lotteries, thus they are assumed to be part of the decision environment.

The decision maker is asked to specify a possibly partial preference relation on the resulting generalized convex lotteries (*gc-lotteries* for short). The aim is to provide a rationality principle for the existence of a utility function on the set of prizes whose Choquet expected utility (CEU) represents the preference relation.

This leads to a generalization of the von Neumann-Morgenstern decision theory under risk and imprecise information in the spirit of [17] (see also [12, 13] and [18] for a different generalization). Note that this setting distinguishes from that of [2, 23, 16] which relies on the Anscombe-Aumann framework, where the capacities are endogenous, i.e., they are not part of the decision environment. The maximization of the CEU functional consists in a maxmin criterion of choice under risk and imprecise probability information. Thus, a decision maker acting like a CEU maximizer [15, 16] realizes a form of *uncertainty aversion* for decisions under risk.

Another relevant aspect that must be recalled is that in the classical expected utility framework, as well as in the CEU model, it can be difficult to construct the utility function on prizes, only by taking into account the "few" available preferences expressed on the "few" available lotteries. The classical methods essentially rely on the totality of the preference relation, thus the decision maker is often forced to make comparisons among some lotteries that are not easy to compare since they have nothing to do with the given problem (for example, comparisons between risky prospects and

certainty). Not to mention that the set of lotteries to consider is "automatically" infinite.

In [6], referring to the EU model, a different approach based on a "rationality principle" is proposed: it does not need all these non-natural comparisons but, instead, it can work by considering only the "few" lotteries and comparisons of interest. In [4, 5] a similar approach for the CEU model has been introduced by generalizing the usual definition of lottery in a way to consider a random quantity endowed with a belief function.

Here, taking the CEU model as reference, we consider a partial preference relation on an arbitrary finite set of gc-lotteries. The *Choquet rationality principle* is introduced and is proven to be equivalent to the existence of a strictly increasing utility function, whose CEU represents the given preference relation. This principle relies, for each gc-lottery, on a probability distribution (namely, *aggregated Möbius inversion*) realizing the lower expected utility with respect to the probabilities dominating the convex capacity of the gc-lottery. Such principle requires that it is not possible to obtain the same probability distribution through the same convex combination of the aggregated Möbius inversions of two groups of gc-lotteries, if every gc-lottery of the first group is not preferred to the corresponding one of the second group, and at least a preference is strict. Moreover, a (not necessarily unique) utility function can be explicitly determined by solving a linear system. It is straightforward that once a utility function has been chosen, a complete preference relation extending the one provided by the decision maker is induced by the corresponding CEU functional.

Qualitative conditions are provided on the given preference relation that, together with the Choquet rationality principle, imply the existence of a strictly increasing concave continuous (or strictly concave twice continuously differentiable) utility function whose CEU represents the given preference. This allows to model the *risk aversion* of the decision maker under imprecise information.

The non-uniqueness of the utility function singled out by the Choquet rationality principle implies that different complete preference relations can arise, thus any choice of a utility function causes a loss of information, moreover, it is not clear why one should choose a utility function in place of another. For this reason we deal with the extension of the preference relation in a qualitative setting by considering the entire class of utility functions whose CEU represents the preference relation. This leads to an algorithm for a step by step extension of the given preference relation which guides the decision maker in assessing his new preferences. The aforementioned algorithm is shown to work independently of the concavity constraints for the utility function.

2 Numerical Model of Reference

Let $X = \{x_1, \ldots, x_n\}$ be a finite set and denote by $\wp(X)$ the power set of X. We recall that a *(normalized) capacity* is a function $\varphi : \wp(X) \to [0,1]$ such that $\varphi(\emptyset) = 0$, $\varphi(X) = 1$ and $\varphi(A) \leq \varphi(B)$ when $A, B \in \wp(X)$ and $A \subseteq B$.

A capacity φ on $\wp(X)$ is said *convex* if it satisfies the further property for every $A, B \in \wp(X)$,

$$\varphi(A \cup B) \geq \varphi(A) + \varphi(B) - \varphi(A \cap B). \tag{1}$$

As is well-known (see [3]) a convex capacity φ on $\wp(X)$ is completely characterized by its *Möbius inversion*, defined for every $A \in \wp(X)$ as

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \varphi(B), \tag{2}$$

and for every $A \in \wp(X)$ it holds

$$\varphi(A) = \sum_{B \subseteq A} m(B). \tag{3}$$

The Möbius inversion of a convex capacity is a function $m : \wp(X) \to \mathbb{R}$ such that $m(\emptyset) = 0$, $\sum_{B \in \wp(X)} m(B) = 1$, $m(\{x_i\}) \geq 0$ for every $x_i \in X$, and for every $A \in \wp(X)$ with $|A| \geq 2$ and every $\{x_i, x_j\} \subseteq A$, it satisfies $\sum_{\{x_i, x_j\} \subseteq B \subseteq A} m(B) \geq 0$ (see [3, 2]). Notice that m can be negative on sets of cardinality greater than 1.

Given a set $X = \{x_1, \ldots, x_n\}$ and a normalized capacity φ on $\wp(X)$ (not necessarily convex), the *Choquet integral* of a function $f : X \to \mathbb{R}$, with $f(x_1) \leq \ldots \leq f(x_n)$ is defined as

$$\oint f \, \mathrm{d}\varphi = \sum_{i=1}^{n} f(x_i)(\varphi(E_i) - \varphi(E_{i+1})) \tag{4}$$

where $E_i = \{x_i, \ldots, x_n\}$ for $i = 1, \ldots, n$, and $E_{n+1} = \emptyset$ [9].

In the classical von Neumann-Morgenstern theory [26] a *lottery* L consists of a *probability distribution* on a finite *support* X_L, which is an arbitrary finite set of *prizes* or *consequences*.

In this paper, following the idea of Jaffray [17] involving belief functions, we deal with generalized convex lotteries, by assuming that a convex capacity φ_L is assigned on the power set $\wp(X_L)$ of X_L.

Definition 1. *A* **generalized convex lottery**, *or* **gc-lottery** *for short, on a finite set X_L is a pair $L = (\wp(X_L), \varphi_L)$ where φ_L is a convex capacity on $\wp(X_L)$.*

Obviously, a gc-lottery $L = (\wp(X_L), \varphi_L)$ could be equivalently defined as $L = (\wp(X_L), m_L)$, where m_L is the Möbius inversion of φ_L. The following simple gc-lottery L on $X_L = \{x_1, x_2\}$ expressed in terms of φ_L

$$L = \begin{pmatrix} \{x_1\} & \{x_2\} & X_L \\ \varphi_L(\{x_1\}) & \varphi_L(\{x_2\}) & \varphi_L(X_L) \end{pmatrix}$$

has an equivalent representation through the Möbius inversion m_L of φ_L

$$L = \begin{pmatrix} \{x_1\} & \{x_2\} & X_L \\ m_L(\{x_1\}) & m_L(\{x_2\}) & m_L(X_L) \end{pmatrix}.$$

We notice that gc-lotteries generalize classical lotteries, in which $m_L(A) = 0$ for every $A \in \wp(X_L)$ with $|A| > 1$, and those introduced in [17], where $m_L(A) \geq 0$ for every $A \in \wp(X_L)$.

Given a finite set $\mathcal{L}$ of gc-lotteries, let $X = \bigcup\{X_L : L \in \mathcal{L}\}$. Then, any gc-lottery L on X_L with convex capacity φ_L can be rewritten as a gc-lottery on X by defining a suitable extension φ'_L of φ_L.

Proposition 1. *Let $L = (\wp(X_L), \varphi_L)$ be a gc-lottery on X_L and m_L the Möbius inversion of φ_L. Then for any finite $X \supseteq X_L$ there exists a unique convex capacity φ'_L extending φ_L to $\wp(X)$, whose Möbius inversion m'_L coincides with m_L on $\wp(X_L)$ and is 0 on $\wp(X) \setminus \wp(X_L)$.*

Note that φ'_L on $\wp(X)$ coincides with the inner measure induced by φ_L on $\wp(X_L)$ and the convexity of φ'_L follows from a result in [28].

Given $L_1, \ldots, L_t \in \mathcal{L}$, all rewritten on X, and a real vector $\mathbf{k} = (k_1, \ldots, k_t)$ with $k_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{t} k_i = 1$, the *convex combination* of $L_1, \ldots, L_t$ according to $\mathbf{k}$ is defined as

$$\mathbf{k}(L_1, \ldots, L_t) = \begin{pmatrix} A \\ \sum_{i=1}^{t} k_i m_{L_i}(A) \end{pmatrix}_{A \in \wp(X) \setminus \{\emptyset\}}. \quad (5)$$

It is readily verified that the convex combination of Möbius inversions $m_{L_1}, \ldots, m_{L_t}$ of convex capacities on $\wp(X)$ is itself a Möbius inversion of a convex capacity on $\wp(X)$.

For every $A \in \wp(X) \setminus \{\emptyset\}$, we denote with δ_A the *degenerate gc-lottery* on X such that $m_{\delta_A}(A) = 1$.

3 Rational Preferences over a Set of Generalized Convex Lotteries

Consider a set $\mathcal{L}$ of gc-lotteries with $X = \bigcup\{X_L : L \in \mathcal{L}\}$ and assume that a total preorder $\leq^*$ is given on X. This is a quite natural condition thinking at elements of X as prizes. Denote with $<^*$ and $=^*$ the asymmetrical and the symmetrical parts of $\leq^*$, respectively. Moreover, denote with $X^* = X_{/=^*}$ the set of equivalence classes of elements of X according to $=^*$, for which $<^*$ is a total strict order.

In what follows the set X is always assumed to be finite, i.e., $X = \{x_1, \ldots, x_n\}$ with $x_1 \leq^* \ldots \leq^* x_n$. This implies $X^* = \{[x_{i_1}], \ldots, [x_{i_m}]\}$ with $[x_{i_1}] <^* \ldots <^* [x_{i_m}]$, where $m \leq n$. Under previous assumption, we can define the *aggregated Möbius inversion* of a gc-lottery L, for every $[x_{i_j}] \in X^*$, as

$$M_L([x_{i_j}]) = \sum_{x_i \in [x_{i_j}]} \sum_{\{x_i\} \subseteq B \subseteq E_i} m_L(B), \quad (6)$$

where $E_i = \{x_i, \ldots, x_n\}$ for $i = 1, \ldots, n$. Note that $M_L([x_{i_j}]) \geq 0$ for every $[x_{i_j}] \in X^*$ and $\sum_{j=1}^{m} M_L([x_{i_j}]) = 1$, thus M_L determines a probability distribution on X^*.

The following example shows the computation of the aggregated Möbius inversion given a gc-lottery.

Example 1. *Let $X = \{x_1, x_2, x_3\}$ be totally preordered by $\leq^*$ as $x_1 =^* x_2 <^* x_3$ and consider the gc-lottery $L = (\wp(X), m_L)$ where $m_L(\{x_1\}) = m_L(\{x_3\}) = \frac{1}{4}$, $m_L(\{x_2, x_3\}) = \frac{1}{2}$ and 0 otherwise.*

It holds $X^* = \{[x_1], [x_3]\}$ with $[x_1] = \{x_1, x_2\}$ and $[x_3] = \{x_3\}$, and the aggregated Möbius inversion on X^* corresponding to L is

$$\begin{aligned} M_L([x_1]) &= m_L(\{x_1\}) + m_L(\{x_1, x_2\}) \\ &\quad + m_L(\{x_1, x_3\}) + m_L(X) \\ &\quad + m_L(\{x_2\}) + m_L(\{x_2, x_3\}) = \frac{3}{4}, \\ M_L([x_3]) &= m_L(\{x_3\}) = \frac{1}{4}. \end{aligned}$$

Let $\mathcal{R}$ be a possibly partial binary relation on $\mathcal{L}$. For every $(L, L') \in \mathcal{R}$ denote by $L \precsim L'$ the assertion L *is not preferred to* L'. The assertion L *is indifferent to* L', denoted by $L \sim L'$, summarizes the two assertions $L \precsim L'$ and $L' \precsim L$, so $\mathcal{R}$ determines the symmetric relation $\mathcal{I} = \{(L, L') \in \mathcal{R} : (L', L) \in \mathcal{R}\}$. An additional strict preference relation $\mathcal{R}'$ can be elicited by assertions such as L' *is strictly preferred to* L, denoted by $L \prec L'$. Let $\mathcal{R}^*$ be the asymmetric relation formally deduced from $\mathcal{R}$, namely $\mathcal{R}^* = \mathcal{R} \setminus \mathcal{I}$.

Since the pair of relations $(\mathcal{R}, \mathcal{R}')$ represents the opinion of the decision maker, it is natural to have $\mathcal{R}' \subseteq \mathcal{R}^*$:

in fact, it is possible that, in the first approach to the decision problem, the decision maker is not able to evaluate yet whether $L \prec L'$ or $L \sim L'$ and he/she expresses his/her opinion only by $L \precsim L'$.

If $\mathcal{R}$ is total on the set of gc-lotteries $\mathcal{L}$ then $\mathcal{R}' = \mathcal{R}^*$ and for every $L, L' \in \mathcal{L}$: $L \prec L'$ or $L' \prec L$ or $L \sim L'$.

We call a pair $(\mathcal{R}, \mathcal{R}')$ a *strengthened preference relation* if $\emptyset \neq \mathcal{R}' \subseteq \mathcal{R}$ and $\mathcal{I} \cap \mathcal{R}' = \emptyset$, moreover, in the following it will be simply denoted by $(\precsim, \prec)$.

Since the set X is totally preordered by $\leq^*$, it is natural to require that the partial preference relation $(\precsim, \prec)$ agrees with $\leq^*$ on degenerate gc-lotteries $\delta_{\{x\}}$, for $x \in X$, that correspond to decisions under certainty. For this the preference $(\precsim, \prec)$ is asked to satisfy the following assumption

(A0) $\mathcal{L}$ contains the set of degenerate gc-lotteries on singletons $\mathcal{L}_0 = \{\delta_{\{x\}} : x \in X\}$ and $x \leq^* x'$ if and only if $\delta_{\{x\}} \precsim \delta_{\{x'\}}$, for $x, x' \in X$.

Remark 1. *Note that the decision maker is not required to provide comparisons among degenerate gc-lotteries and the gc-lotteries of interest, but just to accept the set of (natural) preferences considered in condition* **(A0)**. *When X is not "naturally" preordered, one can require that the restriction to $\mathcal{L}_0$ of the preference relation $(\precsim, \prec)$ given by the decision maker is a total preorder. Then, by* **(A0)**, *we can induce a total preorder on X.*

The next rationality axiom requires that it is not possible to obtain the same probability distribution on X^* through the same convex combination of the aggregated Möbius inversions of two groups of gc-lotteries, if every gc-lottery of the first group is not preferred to the corresponding one of the second group, and at least a preference is strict.

Definition 2. *A strengthened preference relation $(\precsim, \prec)$ on a set $\mathcal{L}$ of gc-lotteries is said to be* **Choquet rational** *if it satisfies the following condition:*

(gc-CR) *For all $h \in \mathbb{N}$ and $L_i, L_i' \in \mathcal{L}$ with $L_i \precsim L_i'$ $(i = 1, \ldots, h)$, if*

$$\mathbf{k}(M_{L_1}, \ldots, M_{L_h}) = \mathbf{k}(M_{L_1'}, \ldots, M_{L_h'})$$

with $\mathbf{k} = (k_1, \ldots, k_h)$, $k_i > 0$ $(i = 1, \ldots, h)$ and $\sum_{i=1}^h k_i = 1$, then it cannot be $L_i \prec L_i'$ for any $i = 1, \ldots, h$.

Note that the convex combination referred to in condition **(gc-CR)** is the usual one involving probability distributions on X^*. Moreover, it is easily proven that if $\mathbf{k}(L_1, \ldots, L_h) = \mathbf{k}(L_1', \ldots, L_h')$, then it also holds $\mathbf{k}(M_{L_1}, \ldots, M_{L_h}) = \mathbf{k}(M_{L_1'}, \ldots, M_{L_h'})$ but the converse is generally not true.

4 Representability of Rational Preferences over gc-Lotteries

Given a finite set of gc-lotteries $\mathcal{L}$, in what follows we assume that all gc-lotteries are rewritten as gc-lotteries on $X = \bigcup\{X_L : L \in \mathcal{L}\}$. We say that a function $U : \mathcal{L} \to \mathbb{R}$ *represents* (or *agrees with*) $(\precsim, \prec)$ if, for every $L, L' \in \mathcal{L}$

$$\begin{cases} L \precsim L' \Longrightarrow U(L) \leq U(L'), \\ L \prec L' \Longrightarrow U(L) < U(L'). \end{cases} \tag{7}$$

In analogy with [6], given $(\precsim, \prec)$ on $\mathcal{L}$, our aim is to find a necessary and sufficient condition for the existence of a utility function $u : X \to \mathbb{R}$ such that the *Choquet expected utility* of gc-lotteries in $\mathcal{L}$, defined for every $L \in \mathcal{L}$ as

$$\mathrm{CEU}(L) = \oint u \, \mathrm{d}\varphi_L, \tag{8}$$

represents $(\precsim, \prec)$. In particular, since X is totally preordered by $\leq^*$ and $\mathrm{CEU}(\delta_{\{x\}}) = u(x)$ for every $x \in X$, we search for a *strictly increasing* u, i.e., satisfying, for $x, x' \in X$, $x <^* x' \Longrightarrow u(x) < u(x')$.

This implies that such a u is constant over the elements of X^*, so for $L \in \mathcal{L}$ the CEU functional reduces to

$$\mathrm{CEU}(L) = \sum_{[x_{i_j}] \in X^*} u(x_{i_j}) M_L([x_{i_j}]). \tag{9}$$

Let us stress that every gc-lottery L determines a family of probabilistic lotteries on X whose probability distributions form the closed and convex family $\mathcal{P}_L = \{\tilde{P} : \wp(X) \to [0,1] : \varphi_L \leq \tilde{P}\}$. The CEU functional turns out to be the minimum of expected utilities computed with respect to the family $\mathcal{P}_L$, i.e.,

$$\mathrm{CEU}(L) = \min_{\tilde{P} \in \mathcal{P}_L} \int u \, \mathrm{d}\tilde{P},$$

(see [24]) and this expresses a kind of *uncertainty aversion* of the decision maker [23, 15]. For this, a CEU maximiser decision maker acts according to a maxmin criterion of choice.

The following theorem shows that **(gc-CR)** is a necessary and sufficient condition for the existence of a strictly increasing utility function u whose Choquet expected value on gc-lotteries represents $(\precsim, \prec)$.

Theorem 1. *Let $\mathcal{L}$ be a finite set of g-lotteries, $X = \bigcup\{X_L : L \in \mathcal{L}\} = \{x_1, \ldots, x_n\}$ and let $\leq^*$ be a total preorder on X. For a strengthened preference relation $(\precsim, \prec)$ on $\mathcal{L}$ satisfying* **(A0)** *the following statements are equivalent:*

(i) $(\precsim, \prec)$ is Choquet rational (i.e., it satisfies **(gc-CR)***);*

(ii) there exists a strictly increasing function $u : X \to \mathbb{R}$, whose CEU *functional on $\mathcal{L}$ represents $(\precsim, \prec)$.*

Proof. Let $X^* = X_{/=^*} = \{[x_{i_1}], \ldots, [x_{i_m}]\}$. Introduce the collections $S = \{(L_j, L'_j) : L_j \prec L'_j, L_j, L'_j \in \mathcal{L}\}$ and $R = \{(G_h, G'_h) : G_h \precsim G'_h, G_h, G'_h \in \mathcal{L}\}$ with $s = \operatorname{card} S$ and $r = \operatorname{card} R$. Then condition **(gc-CR)** is equivalent to the *non-existence* of a row vector $\mathbf{k}$ of size $(1 \times s + r)$ with $k_i > 0$ for at least a pair $(L_i, L'_i) \in S$ and $\sum_{i=1}^{s+r} k_i = 1$ such that $\mathbf{k}(M_{L_1}, \ldots, M_{L_s}, M_{G_1}, \ldots, M_{G_r}) = \mathbf{k}(M_{L'_1}, \ldots, M_{L'_s}, M_{G'_1}, \ldots, M_{G'_r})$.

In turn, setting $\mathbf{k} = (\mathbf{y}, \mathbf{z})$, previous condition is equivalent to the *non-solvability* of the following linear system (in which $||\cdot||_1$ denotes the L^1-norm)

$$\mathcal{S}' : \begin{cases} \mathbf{y}A + \mathbf{z}B = \mathbf{0} \\ \mathbf{y}, \mathbf{z} \geq \mathbf{0} \\ \mathbf{y} \neq \mathbf{0} \\ ||(\mathbf{y}, \mathbf{z})||_1 = 1 \end{cases} \tag{10}$$

where $A = (a^j)$ and $B = (b^h)$ are, respectively, $(s \times m)$ and $(r \times m)$ real matrices with rows $a^j = M_{L'_j} - M_{L_j}$ for $j = 1, \ldots, s$, and $b^h = M_{G'_h} - M_{G_h}$ for $h = 1, \ldots, r$, and $\mathbf{y}$ and $\mathbf{z}$ are, respectively, $(1 \times s)$ and $(1 \times r)$ unknown row vectors.

By a well-known alternative theorem (see, e.g., [14]), the non-solvability of $\mathcal{S}'$ is equivalent to the *solvability* of the following system

$$\mathcal{S} : \begin{cases} A\mathbf{w} > \mathbf{0} \\ B\mathbf{w} \geq \mathbf{0} \end{cases} \tag{11}$$

where $\mathbf{w}$ is a $(m \times 1)$ unknown column vector. Setting $u(x_i) = w_j$ for $x_i \in [x_{i_j}]$ and $j = 1, \ldots, m$, the solution $\mathbf{w}$ induces a utility function u on X which by **(A0)** is strictly increasing and whose CEU functional on $\mathcal{L}$ represents $(\precsim, \prec)$. □

Notice that Theorem 1 implies that condition **(gc-CR)** is equivalent to the existence of a (not necessarily unique) total relation $\precsim'$ on $\mathcal{L}$ extending $(\precsim, \prec)$: such $\precsim'$ is simply induced by the CEU functional once a utility u is fixed.

Consider now the particular case in which a strengthened preference $(\precsim, \prec)$ is defined on a finite set of gc-lotteries $\mathcal{L}$ satisfying **(A0)**, where $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ with $\leq^*$ coinciding with the usual total order $\leq$, for which it holds $x_1 < \ldots < x_n$. In this case we have $[x_i] = \{x_i\}$ for $i = 1, \ldots, n$, so X^* can be identified with X and for every $L \in \mathcal{L}$, the corresponding basic assignment M_L can be simply viewed as a probability distribution on X.

As is well-known, the *risk aversion* of the decision maker can be expressed by means of an increasing concave utility function. In order to get a concave utility function we consider the following assumptions, where

$$\mathcal{L}_1 = \left\{\alpha_i \delta_{\{x_{i-1}\}} + (1 - \alpha_i)\delta_{\{x_{i+1}\}} : i = 2, \ldots, n-1\right\} \tag{12}$$

with $\alpha_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}$ and, for $i = 2, \ldots, n-1$:

(A1) $\mathcal{L}_1 \subseteq \mathcal{L}$ and $\alpha_i \delta_{\{x_{i-1}\}} + (1 - \alpha_i)\delta_{\{x_{i+1}\}} \prec \delta_{\{x_i\}}$ or $\alpha_i \delta_{\{x_{i-1}\}} + (1 - \alpha_i)\delta_{\{x_{i+1}\}} \sim \delta_{\{x_i\}}$.

(A1*) $\mathcal{L}_1 \subseteq \mathcal{L}$ and $\alpha_i \delta_{\{x_{i-1}\}} + (1 - \alpha_i)\delta_{\{x_{i+1}\}} \prec \delta_{\{x_i\}}$.

Notice that condition **(A1*)** implies condition **(A1)**.

Proposition 2. *Let $(\precsim, \prec)$ be a strengthened preference relation on a finite set of gc-lotteries $\mathcal{L}$ with $X = \bigcup\{X_L : L \in \mathcal{L}\} = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ such that $x_1 < \ldots < x_n$. Assume $(\precsim, \prec)$ satisfies* **(A0)** *and* **(gc-CR)** *and let u be the utility function in (ii) of Theorem 1. The following statements hold:*

(i) if **(A1)** *holds then u extends to a strictly increasing concave function $v \in C^0([x_1, x_n])$;*

(ii) if **(A1*)** *holds then u extends to a strictly increasing strictly concave function $w \in C^2([x_1, x_n])$.*

Proof. If **(A1)** is satisfied then we have $x_1 < \ldots < x_n$ and $s_1 \geq \ldots \geq s_{n-1}$ where $s_i = \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}$, for $i = 1, \ldots, n-1$. Thus it is sufficient to take as v the piecewise linear function connecting the points $\{(x_i, u(x_i)) : i = 1, \ldots, n\}$. In particular, if **(A1*)** holds, we have $s_1 > \ldots > s_{n-1}$ so the main theorem in [10] implies that the set of points $\{(x_i, -u(x_i)) : i = 1, \ldots, n\}$ can be interpolated by a strictly convex function f in $C^2([x_1, x_n])$ which must be strictly decreasing. Thus, the proof follows taking $w = -f$. □

Assuming $X \subset \mathbb{R}$, every gc-lottery L induces a cumulative probability distribution function F_L on $\mathbb{R}$ through the corresponding aggregated Möbius inversion M_L, defined for every $x \in \mathbb{R}$ as

$$F_L(x) = \sum_{x_i \leq x} M_L(x_i). \tag{13}$$

The function F_L will be referred to as *cumulative aggregated Möbius inversion*. It can be used to express the following kind of second order stochastic dominance.

Proposition 3. *Let $(\precsim, \prec)$ be a strengthened preference relation on a finite set of gc-lotteries $\mathcal{L}$ with $X = \bigcup\{X_L : L \in \mathcal{L}\} = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ such that $x_1 < \ldots < x_n$. Assume* **(A0)** *and* **(A1)** *are satisfied. If $(\precsim, \prec)$ satisfies* **(gc-CR)** *then for every complete preference relation $\precsim'$ on $\mathcal{L}$ extending $(\precsim, \prec)$ and satisfying* **(gc-CR)** *the following condition holds for every $L_1, L_2 \in \mathcal{L}$:*

(S2) *if* $\int_{-\infty}^{x} F_{L_1}(t)\,\mathrm{d}t \leq \int_{-\infty}^{x} F_{L_2}(t)\,\mathrm{d}t$ *for every* $x \in \mathbb{R}$*, it cannot be* $L_1 \prec' L_2$.

Proof. For every $L_1, L_2 \in \mathcal{L}$, $\int_{-\infty}^{x} F_{L_1}(t)\,\mathrm{d}t \leq \int_{-\infty}^{x} F_{L_2}(t)\,\mathrm{d}t$ for every $x \in \mathbb{R}$ is equivalent to $\int_{x_1}^{x_n} v(t)\,\mathrm{d}F_{L_1}(t) \geq \int_{x_1}^{x_n} v(t)\,\mathrm{d}F_{L_1}(t)$ for every increasing concave utility function v on $[x_1, x_n]$.

By Theorem 1, condition **(gc-CR)** is equivalent to the existence of a strictly increasing utility function on X. Every such utility function on X determines through the corresponding CEU functional a complete preference relation on $\mathcal{L}$ extending $(\precsim, \prec)$ and satisfying **(gc-CR)**. Moreover, statement *(i)* of Proposition 2 implies that the utility function on X extends to a strictly increasing concave utility function belonging to $C^0([x_1, x_n])$.

Let u be a utility function on X determining the complete preference relation $\precsim'$ on $\mathcal{L}$ which extends $(\precsim, \prec)$ and satisfies **(gc-CR)**. Let v be a strictly increasing concave function in $C^0([x_1, x_n])$ extending u. For every $L \in \mathcal{L}$ it holds

$$\begin{aligned}\int_{x_1}^{x_n} v(t)\,\mathrm{d}F_L(t) &= \sum_{i=1}^{n} u(x_i) M_L(x_i) \\ &= \oint u\,\mathrm{d}\varphi_L = \mathrm{CEU}(L).\end{aligned}$$

Hence, $\int_{-\infty}^{x} F_{L_1}(t)\,\mathrm{d}t \leq \int_{-\infty}^{x} F_{L_2}(t)\,\mathrm{d}t$ for every $x \in \mathbb{R}$ implies $\mathrm{CEU}(L_1) \geq \mathrm{CEU}(L_2)$ and so it cannot be $L_1 \prec' L_2$. □

The following example shows the construction of a concave utility function whose CEU functional represents a strengthened preference relation $(\precsim, \prec)$.

Example 2. *Let* $X = \{0, 10, 20\}$ *be a set of money payoffs and consider the following gc-lotteries expressed in terms of their Möbius inversions*

	$\{0\}$	$\{10\}$	$\{20\}$	$\{0,10\}$	$\{0,20\}$	$\{10,20\}$	X
L_1	0.4	0.1	0.2	0.1	0.1	0.2	−0.1
L_2	0.5	0.5	0	0	0	0	0
L_3	0.2	0	0.2	0	0.6	0	0

whose corresponding aggregated Möbius inversions (viewed as probability distributions on X*) are*

X	0	10	20
M_{L_1}	0.5	0.3	0.2
M_{L_2}	0.5	0.5	0
M_{L_3}	0.8	0	0.2

Consider the following strengthened preference relation $(\precsim, \prec)$ *satisfying* **(A0)** *and* **(A1*)**, *and such that*

$$L_2 \prec L_1 \quad \text{and} \quad L_3 \prec L_1.$$

To prove that $(\precsim, \prec)$ *satisfies* **(gc-CR)** *we search for a utility function* $u : X \to \mathbb{R}$ *whose* CEU *represents* $(\precsim, \prec)$. *Setting* $w_1 = u(0)$, $w_2 = u(10)$, $w_3 = u(20)$, *the following system must be solvable*

$$\begin{cases} 0.5w_1 + 0.5w_2 < 0.5w_1 + 0.3w_2 + 0.2w_3 \\ 0.8w_1 + 0.2w_3 < 0.5w_1 + 0.3w_2 + 0.2w_3 \\ w_1 < w_2 < w_3 \\ 0.5w_1 + 0.5w_3 < w_2 \end{cases}$$

for which a solution is $w_1 = 0$, $w_2 = 3$ *and* $w_3 = 4$. *A strictly increasing concave utility function* $v \in C^0([0, 20])$ *extending* u *is the function* $v(x) = (0.3x)\mathbf{1}_{[0,10]}(x) + (0.1x + 2)\mathbf{1}_{(10,20]}(x)$. *A strictly increasing strictly concave utility function* $w \in C^2([0, 20])$ *extending* u *is* $w(x) = (0.4x - 0.01x^2)\mathbf{1}_{[0,20]}(x)$. *Figure 1 shows the plots of utility functions* u, v *and* w.

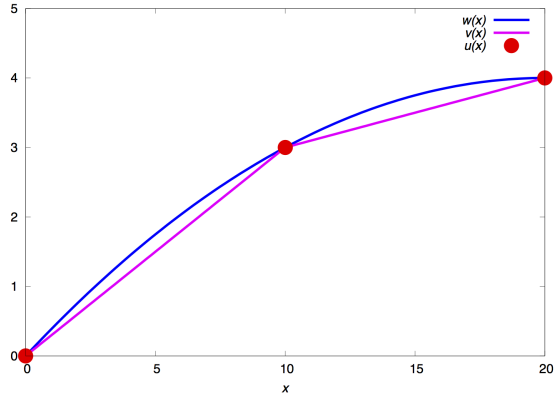


Figure 1: Plots of utility functions u, v and w

5 Extension of Choquet Rational Preferences

In previous section it has been shown that condition **(gc-CR)** is equivalent to the existence of a strictly increasing utility function u on X, whose CEU represents $(\precsim, \prec)$, moreover, such a u can be explicitly determined by solving the linear system $\mathcal{S}$ defined in (11). It is straightforward that once a utility u has been chosen, a complete preference relation on $\mathcal{L}$ (or on any finite superset $\mathcal{L}'$ of gc-lotteries on the same finite set X) extending $(\precsim, \prec)$ is induced by the corresponding CEU functional.

Nevertheless, system $\mathcal{S}$ has generally infinite solutions which can give rise to possibly very different complete preference relations, thus any choice of a utility function causes a loss of information, moreover, it is not clear why one should choose a utility function in place of another.

For this reason it is preferable to face the extension in a qualitative setting by considering the entire class of utility functions, whose CEU represents the preference $(\succsim, \prec)$, and suggesting to the decision maker those pairs of gc-lotteries where all the utility functions unanimously agree in the order induced by the corresponding CEU functional. In this view, the following Theorem 2 proves the extendibility of a Choquet rational relation and shows how condition **(gc-CR)** guides the decision maker in assessing his preferences.

Theorem 2. *Let $X = \{x_1, \ldots, x_n\}$ be a finite set with a total prorder $\leq^*$, $\mathcal{L}$ and $\mathcal{L}'$ finite sets of gc-lotteries on X, with $\mathcal{L} \subseteq \mathcal{L}'$, and $(\succsim, \prec)$ a strengthened preference relation on $\mathcal{L}$ satisfying* **(A0)**. *Then if $(\succsim, \prec)$ satisfies condition* **(gc-CR)** *there exists a family $\{\succsim^\gamma : \gamma \in \Gamma\}$ of complete relations on $\mathcal{L}'$ satisfying* **(gc-CR)** *which extend $(\succsim, \prec)$. Moreover, denoting with $\prec^\gamma$ and $\sim^\gamma$, respectively, the strict and symmetric parts of $\succsim^\gamma$, for $\gamma \in \Gamma$, condition* **(gc-CR)** *singles out the relations*

$$\prec^\star = \bigcap \{\prec^\gamma : \gamma \in \Gamma\} \quad \textit{and} \quad \sim^\star = \bigcap \{\sim^\gamma : \gamma \in \Gamma\}.$$

Proof. Let $X^* = X_{/=^*} = \{[x_{i_1}], \ldots, [x_{i_m}]\}$. By the proof of Theorem 1, $(\succsim, \prec)$ satisfies condition **(gc-CR)** if and only if system $\mathcal{S}$ defined in (11) admits a $(m \times 1)$ column vector $\mathbf{w}$ as solution. In turn, setting $u(x_i) = w_j$, for $x_i \in [x_{i_j}]$ and $j = 1, \ldots, m$, we get a strictly increasing utility function u on X whose Choquet expected value represents $(\succsim, \prec)$ on $\mathcal{L}$. Defining for every $L, L' \in \mathcal{L}'$

$$L \succsim^\gamma L' \Longleftrightarrow \mathrm{CEU}(L) \leq \mathrm{CEU}(L'),$$

we get a relation $\succsim^\gamma$ on $\mathcal{L}'$ which is complete and satisfies **(gc-CR)** by virtue of Theorem 1. This implies that the family $\{\succsim^\gamma : \gamma \in \Gamma\}$ is not empty and all its members are obtained varying the solution $\mathbf{w}$ of system $\mathcal{S}$. The correspondence between the set of solutions and the family $\{\succsim^\gamma : \gamma \in \Gamma\}$ is onto but not one-to-one, as every positive linear transformation of a solution $\mathbf{w}$ gives rise to the same relation $\succsim^\gamma$.

The relations $\prec^\star$ and $\sim^\star$ express, respectively, the pairs of gc-lotteries in $\mathcal{L}'$ on which all the strict $\prec^\gamma$ and symmetric $\sim^\gamma$ parts, for $\gamma \in \Gamma$, agree. It trivially holds that $\prec^\star$ and $\sim^\star$ extend the relations $\prec$ and $\sim$ obtained from $(\succsim, \prec)$, moreover, in order to determine $\prec^\star$ and $\sim^\star$, for every $F, G \in \mathcal{L}'$ such that $F \prec G$ or $G \prec F$ or $F \sim G$ does not hold, it is sufficient to test the solvability of the three linear systems

$$\mathcal{S}^{\prec^\star} : \begin{cases} A'\mathbf{w} > \mathbf{0} \\ B\mathbf{w} \geq \mathbf{0} \end{cases} \qquad \mathcal{S}^{\succ^\star} : \begin{cases} A''\mathbf{w} > \mathbf{0} \\ B\mathbf{w} \geq \mathbf{0} \end{cases}$$

$$\mathcal{S}^{\sim^\star} : \begin{cases} A\mathbf{w} > \mathbf{0} \\ B'\mathbf{w} \geq \mathbf{0} \end{cases}$$

where $\mathbf{w}$ is an unknown $(m \times 1)$ column vector, A and B are, respectively, $(s \times m)$ and $(r \times m)$ real matrices defined as in (10), A' is a $((s+1) \times m)$ real matrix obtained adding to A the $(s+1)$-th row $a^{(s+1)} = M_G - M_F$, A'' is a $((s+1) \times m)$ real matrix obtained adding to A the $(s+1)$-th row $a^{(s+1)} = M_F - M_G$, and B' is a $((r+2) \times m)$ real matrix obtained adding to B the $(r+1)$-th row $b^{(r+1)} = M_G - M_F$ and the $(r+2)$-th row $b^{(r+2)} = M_F - M_G$.

Depending on the solvability of systems $\mathcal{S}^{\prec^\star}, \mathcal{S}^{\succ^\star}, \mathcal{S}^{\sim^\star}$ we can have the following situations:

(a) $F \prec^\star G$ if and only if $\mathcal{S}^{\prec^\star}$ is solvable and $\mathcal{S}^{\succ^\star}, \mathcal{S}^{\sim^\star}$ are not, as this happens if and only if $\mathrm{CEU}(F) < \mathrm{CEU}(G)$ for every u given by a solution of $\mathcal{S}$;

(b) $G \prec^\star F$ if and only if $\mathcal{S}^{\succ^\star}$ is solvable and $\mathcal{S}^{\prec^\star}, \mathcal{S}^{\sim^\star}$ are not, as this happens if and only if $\mathrm{CEU}(G) < \mathrm{CEU}(F)$ for every u given by a solution of $\mathcal{S}$;

(c) $F \sim^\star G$ if and only if $\mathcal{S}^{\sim^\star}$ is solvable and $\mathcal{S}^{\prec^\star}, \mathcal{S}^{\succ^\star}$ are not, as this happens if and only if $\mathrm{CEU}(F) = \mathrm{CEU}(G)$ for every u given by a solution of $\mathcal{S}$.

In all the remaining cases, the Choquet expected utilities determined by solutions of $\mathcal{S}$ do not unanimously agree in ordering the pair F and G. □

Remark 2. *If $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$, $\leq^*$ coincides with $\leq$ and the initial preference $(\succsim, \prec)$ satisfies* **(A1)** *then every total preference $\succsim^\gamma$ extending $(\succsim, \prec)$ on $\mathcal{L}'$ still satisfies it, thus also the relations $\prec^\star$ and $\sim^\star$ must satisfy* **(A1)**. *An analogous observation holds for* **(A1*)**.

Relations $\prec^\star$ and $\sim^\star$, determined in the proof of previous theorem, express "forced" preferences that the decision maker has to accept in order to maintain Choquet rationality. On the other hand, pairs of gc-lotteries not ruled by $\prec^\star$ and $\sim^\star$ are subject to a choice by the decision maker. In the latter situation, a subjective elicitation is required or, in case of a software agent [20], a suitable automatic choice criterion can be adopted.

We stress that each choice made by the decision maker imposes a new constraint in system $\mathcal{S}$, thus the set of utility functions whose CEU represents the current strengthened preference $(\succsim, \prec)$ is possibly reduced. The present approach implicitly refers to the underlying set of utility functions privileging a direct treatment of qualitative information through $(\succsim, \prec)$. On the other hand, Theorem 1 allows to actually build the set of utility functions compatible with $(\succsim, \prec)$ giving

rise to a *demand extrapolation* for the CEU model in the spirit of [25].

Previous discussion suggests the following Algorithm 1 which is thought to guide the decision maker in enlarging a Choquet rational preference relation $(\precsim, \prec)$ to a (possibly new) pair of gc-lotteries F and G: the extended preference is still denoted as $(\precsim, \prec)$. In particular, Algorithm 1 returns to the decision maker what he must do or he cannot do in order to maintain **(gc-CR)**.

Algorithm 1 Extension of a Choquet rational relation

function EXTENSION($(\precsim, \prec)$, F, G)
if $\mathcal{S}^{\prec^\star}$ and $\mathcal{S}^{\succ^\star}$ are solvable **then** free preference between F and G
else if $\mathcal{S}^{\prec^\star}$ is solvable and $\mathcal{S}^{\sim^\star}$ is not **then** it must be $F \prec G$
else if $\mathcal{S}^{\succ^\star}$ is solvable and $\mathcal{S}^{\sim^\star}$ is not **then** it must be $G \prec F$
else if $\mathcal{S}^{\prec^\star}$ and $\mathcal{S}^{\sim^\star}$ are solvable **then** it cannot be $G \prec F$
else if $\mathcal{S}^{\succ^\star}$ and $\mathcal{S}^{\sim^\star}$ are solvable **then** it cannot be $F \prec G$
else it must be $F \sim G$
end function

Notice that possibly $F, G \in \mathcal{L}$, thus previous algorithm can be used to produce a step by step completion of the preference relation $(\precsim, \prec)$ on $\mathcal{L}$.

Algorithm 1 requires as input a Choquet rational preference relation $(\precsim, \prec)$ on a set of gc-lotteries $\mathcal{L}$, and two (possibly new) gc-lotteries F and G, all rewritten on $X = \{x_1, \ldots, x_n\}$ with $x_1 \leq^* \ldots \leq^* x_n$. The gc-lotteries in $\mathcal{L} \cup \{F, G\}$ can be simply regarded as Möbius inversions on $\wp(X)$, i.e., as real $(1 \times q)$ row vectors with $q = 2^n - 1$. The formation of matrices A, A', A'', B, B' requires the computation of the aggregated Möbius inversion M_L for every $L \in \mathcal{L} \cup \{F, G\}$, which can be done in polynomial time with respect to q.

The extension is faced through the solution of at most three linear programming problems, whose solution has time complexity which is polynomial in $m = O(\log_2(q+1))$ and the digital size of the coefficients in matrices A', B or A'', B or A, B', respectively [19].

6 Where do Generalized Convex Lotteries Come From?

One may ask how one can get a set of gc-lotteries. A typical situation is when an algebra of events $\mathcal{A}$ on a sample space S is considered and the decision maker has to decide among *acts*, i.e., measurable functions from S to a totally preordered set of prizes X as in [21]. So, the main question is how a convex capacity φ can be obtained on $\mathcal{A}$.

The first answer is obviously by situations similar to that considered in the Ellsberg paradox [11], where the convex capacity is obtained as lower envelope of the probabilities extending a partial probabilistic assessment on $\mathcal{A}$. Nevertheless, as is well-known, these lower envelopes could not be convex in general [28]. The lower envelope φ is indeed surely convex (actually it is totally monotone) when the assessment to be extended is given on a sub-algebra of $\mathcal{A}$ and it must be extended to the whole $\mathcal{A}$.

Another possible situation resulting in a convex capacity is when several experts assess each a probability measure on $\mathcal{A}$. Also in this case the lower probability is unique, but it could fail convexity.

We analyse here a different situation: an expert or the decision maker may have assigned on $\mathcal{A}$ only a comparative binary relation $\trianglelefteq$ which is a *comparative degree of belief*. It is well-known (see [1]) that a relation $\trianglelefteq$ on a finite algebra $\mathcal{A}$ is representable by a convex capacity φ if and only if it is a complete preorder satisfying the monotonicity with respect to the inclusion relation

(M) for every $A, B \in \mathcal{A}$ with $A \subseteq B$ one has $A \trianglelefteq B$,

together with Wong's condition [29]

(B) for every $A, B, C \in \mathcal{A}$ with $A \subseteq B$ and $C \cap B = \emptyset$ one has $A \triangleleft B \Longrightarrow A \cup C \triangleleft B \cup C$.

When $\trianglelefteq$ is representable, we have in general a (possibly infinite) class of convex capacities representing it. Suppose now to have a finite family of acts $\mathcal{F} = \{f_1, \ldots, f_t\}$ from S to a preordered set of consequences $X = \{x_1, \ldots, x_n\}$ together with a strengthened preference relation $(\precsim, \prec)$ on $\mathcal{F}$. Now, for every φ representing $\trianglelefteq$ we can construct a unique family $\mathcal{L} = \{L_i : f_i \in \mathcal{F}\}$ of gc-lotteries and transport $(\precsim, \prec)$ on $\mathcal{L}$. We can have one of the following situations:

1. for every φ representing $\trianglelefteq$ the preference relation $(\precsim, \prec)$ on $\mathcal{L}$ violates **(gc-CR)**;
2. for every φ representing $\trianglelefteq$ the preference relation $(\precsim, \prec)$ on $\mathcal{L}$ satisfies **(gc-CR)**;
3. for some φ representing $\trianglelefteq$ the preference relation $(\precsim, \prec)$ on $\mathcal{L}$ satisfies **(gc-CR)** and for the others φ, $(\precsim, \prec)$ violates it.

Notice that for every φ representing $\trianglelefteq$ and such that $(\precsim, \prec)$ satisfies **(gc-CR)**, we obtain a particular class

of utility functions $\{u_\varphi\}$ on X. Every pair (φ, u_φ) represents the same preference relation $(\precsim, \prec)$, nevertheless, as proved in the following example, different choices of φ can produce different extensions of $(\precsim, \prec)$ to new gc-lotteries.

Example 3. *Let $S = \{s_1, s_2\}$ be a finite set of states of nature and $X = \{x_1, x_2, x_3\}$ a finite set of prizes, totally preordered by $\leq^*$ as $x_1 <^* x_2 <^* x_3$. Consider the set of acts $\mathcal{F} = \{f_1, \ldots, f_5\}$ on S and ranging on X, defined as*

S	s_1	s_2
f_1	x_1	x_3
f_2	x_2	x_1
f_3	x_2	x_2
f_4	x_2	x_3
f_5	x_3	x_1

Consider on $\wp(S)$ the total preorder $\trianglelefteq$ with strict part $\triangleleft$, such that $\emptyset \triangleleft \{s_2\} \triangleleft \{s_1\} \triangleleft S$. This relation (trivially) satisfies the necessary and sufficient conditions for the existence of a convex capacity $\varphi : \wp(S) \to [0, 1]$ representing $\trianglelefteq$, i.e., such that $A \trianglelefteq B$ if and only if $\varphi(A) \leq \varphi(B)$, for every $A, B \in \wp(S)$ (see [1]). Obviously φ is not unique: every convex capacity φ representing $\trianglelefteq$ is such that $\varphi(\emptyset) = 0$, $\varphi(\{s_1\}) = \alpha$, $\varphi(\{s_2\}) = \beta$ and $\varphi(S) = 1$, with $0 < \beta < \alpha < 1$ and $\alpha + \beta \leq 1$. For fixed α and β, we get a set $\mathcal{L}$ of gc-lotteries corresponding to $\mathcal{F}$.

Introduce the preference relation $(\precsim, \prec)$ on $\mathcal{L}$ such that

$$L_1 \prec L_2 \prec L_3 \prec L_4.$$

In particular, assuming condition **(A0)**, *for a strictly increasing utility function $u : X \to \mathbb{R}$ we have*

$$\begin{aligned}
\mathrm{CEU}(L_1) &= (1-\beta)u(x_1) + \beta u(x_3),\\
\mathrm{CEU}(L_2) &= (1-\alpha)u(x_1) + \alpha u(x_2),\\
\mathrm{CEU}(L_3) &= u(x_2),\\
\mathrm{CEU}(L_4) &= (1-\beta)u(x_2) + \beta u(x_3),\\
\mathrm{CEU}(L_5) &= (1-\alpha)u(x_1) + \alpha u(x_3).
\end{aligned}$$

Let φ^1 and φ^2 be the convex capacities on $\wp(S)$ representing $\trianglelefteq$ and such that $\varphi^1(\{s_1\}) = \frac{2}{5}$, $\varphi^2(\{s_1\}) = \frac{4}{5}$, and $\varphi^1(\{s_2\}) = \varphi^2(\{s_2\}) = \frac{1}{5}$.

Simple computations show that both using φ^1 and φ^2, the preference relation $(\precsim, \prec)$ among the corresponding gc-lotteries satisfies **(gc-CR)**, *so in both cases there exists a strictly increasing utility function whose* CEU *functional on $\mathcal{L}$ represents $(\precsim, \prec)$.*

The aim now is to extend the preference $(\precsim, \prec)$ to the pair L_4 and L_5. If φ^1 is used, then simple computations show that for every strictly increasing $u^\gamma : X \to \mathbb{R}$ whose CEU *represents $(\precsim, \prec)$, the corresponding total preorder $\precsim^\gamma$ on $\mathcal{L}$ is such that $L_5 \prec^\gamma L_4$, and so $L_5 \prec^\star L_4$ according to Theorem 2.*

On the other hand, if φ^2 is used we get that the decision maker is completely free to express his/her preference among L_4 and L_5 as there are utilities $u^\gamma : X \to \mathbb{R}$ such that $L_4 \prec^\gamma L_5$ or $L_4 \sim^\gamma L_5$ or $L_5 \prec^\gamma L_4$.

7 Conclusions

A feature of the present approach to decisions under risk is the possibility to deal with a partial preference relation assessed on a finite set of gc-lotteries.

Under conditions analogous to those of the classical von Neumann-Morgenstern's theory, i.e., when a complete preference relation is given over the set of all gc-lotteries on X, the representability of the preference relation by a CEU functional coincides with the requirement that two gc-lotteries having the same aggregated Möbius inversion are indifferent and between the resulting equivalence classes the preference relation satisfies the von Neumann-Morgenstern's axioms.

In the same setting, the results in this paper together with Theorem 4.13 in [5] imply that, under the Archimedean axiom of the von Neumann-Morgenstern's theory, a decision maker behaving according to **(gc-CR)** accepts all the von Neumann-Morgenstern's axioms and judges indifferent the gc-lotteries with the same aggregated Möbius inversion.

Acknowledgements

This work was partially supported by GNAMPA - INdAM and by the Italian Ministry of Education, University and Research funding of Research Projects of National Interest (PRIN 2010-11) under grant `2010FP79LR_003`

References

[1] Capotorti, A., Coletti, G., Vantaggi, B.: Non additive ordinal relations representable by lower or upper probabilities. *Kybernetika*, 34(1), 79–90 (1998).

[2] Chateauneuf, A., Cohen, M.: Choquet expected utility model: a new approach to individual behavior under uncertainty and social choice welfare. *Fuzzy Meas. and Int.: Th. and App.*, pp. 289–314, Heidelberg: Physica (2000).

[3] Chateauneuf, A., Jaffray, J.Y.: Some characterizations of lower probabilities and other monotone

capacities through the use of Möbius inversion. *Math. Soc. Sci.*, 17, 263–283 (1989).

[4] Coletti, G., Petturiti, D., Vantaggi, B.: Choquet expected utility representation of preferences on generalized lotteries. *IPMU 2014*, *Part II*, *CCIS 443*, A. Laurent et al. (Eds.), pp. 444–453 (2014).

[5] Coletti, G., Petturiti, D., Vantaggi, B.: Rationality principles for preferences on belief functions. Accepted in *Kybernetika*.

[6] Coletti, G., Regoli, G.: How can an expert system help in choosing the optimal decision?. *Theory and Decisions*, 33(3), 253–264 (1992).

[7] Coletti, G., Scozzafava, R., Vantaggi, B.: Inferential processes leading to possibility and necessity. *Information Sciences*, 245, 132–145 (2013).

[8] Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. *Ann. of Math. Stat.*, 38(2), 325–339 (1967).

[9] Denneberg, D.: *Non-additive Measure and Integral*. Theory and Decision Library: Series B, Vol. 27, Kluwer Academic, Dordrecht, Boston (1994).

[10] Egerland, W.O.: Convex interpolation of convex data. Report no. 1952, USA BRL (1972).

[11] Ellsberg, D.: Risk, Ambiguity and the Savage Axioms. *Quart. Jour. of Econ.*, 75, 643–669 (1961).

[12] Gajdos, T., Tallon, J.M., Vergnaud, J.C.: Decision making with imprecise probabilistic information, *J. of Math. Ec.*, 40(6), 647–681 (2004).

[13] Gajdos, T., Hayashi, T., Tallon, J.M., Vergnaud, J.C.: Attitude toward imprecise information, *J. of Ec. Th.*, 140(1), 27 – 65 (2008).

[14] Gale, D.: *The Theory of Linear Economic Models*. McGraw Hill (1960).

[15] Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. *J. of Math. Econ.*, 18(2), 141–153 (1989).

[16] Gilboa, I., Schmeidler, D.: Additive representations of non-additive measures and the Choquet integral. *Ann. of Op. Res.*, 52, 43–65 (1994).

[17] Jaffray, J.-Y.: Linear utility theory for belief functions. *Op. Res. Let.*, 8 (2), 107–112 (1989).

[18] Quiggin, J.: A Theory of Anticipated Utility. *J. of Ec. Beh. and Org.*, 3, 323–343 (1982).

[19] Papadimitriou, C.H., Steiglitz, K.: *Combinatorial Optimization: Algorithms and Complexity*. Dover, New York (1998).

[20] Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*. Second edition. Prentice Hall, Upper Saddle River (2003).

[21] Savage, L.: *The foundations of statistics*. Wiley, New York (1954).

[22] Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press (1976).

[23] Schmeidler, D.: Subjective probability and expected utility without additivity. *Econometrica*, 57(3), 571–587 (1989).

[24] Schmeidler, D.: Integral representation without additivity. *Proc. of the Am. Math. Soc.*, 97(2), 255–261 (1986).

[25] Varian, H.R.: The Nonparametric Approach to Demand Analysis *Econometrica*, 50(4), 945–973 (1982).

[26] von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press (1944).

[27] Walley, P.: *Statistical reasoning with imprecise probabilities*. Chapman & Hall, London (1991).

[28] Walley, P.: Coherent lower (and upper) probabilities. *Technical Report n. 22*, Department of Statistics, University of Warwick (1981).

[29] Wong, S.K.M., Yao, Y.Y., Bollmann, P., Bürger, H.C.: Axiomatization of Qualitative Belief Structure. *IEEE Trans. on Sys., Man and Cyb.*, 21(4), 726–734 (1991).

Some Remarks on Sets of Lexicographic Probabilities and Sets of Desirable Gambles

Fabio Gagliardi Cozman
Escola Politécnica, Universidade de São Paulo, Brazil

Abstract

Sets of lexicographic probabilities and sets of desirable gambles share several features, despite their apparent differences. In this paper we examine properties of marginalization, conditioning and independence for sets of lexicographic probabilities and sets of desirable gambles.

1 Introduction

The standard theory of probabilities is widely used to represent situations that display uncertainty. In that theory, events form a field, and probabilities are real-valued, non-negative, and countably additive. There are *many* variants of this Kolmogorovian theory of probabilities [8, 10, 21, 29, 32], including proposals that abandon the real scale and focus on infinitesimal probabilities [16] or on lexicographic probabilities [3]. Other departures from probability theory attempt to represent imprecision in numeric values [33, 34]. For instance, the theory of credal sets uses sets of probability measures as its basic entities [20]. Yet another departure from probability theory is the theory of sets of desirable gambles [35]. Matters become even more involved if one allows a language with negation and conjunction of desirability statements [27].

The purpose of this paper is to examine some properties of sets of lexicographic probabilities and of sets of desirable gambles. We present these formalisms through a hopefully illuminating analysis, emphasizing their close connection (Section 2). Because both lexicographic probabilities and sets of desirable gambles represent the same sort of preference orderings, by studying one of them, we obtain insights about the other; perhaps contentious concepts and drawbacks can be clarified by such a study.

Sections 3 to 6 examine features of marginalization and conditioning. We first compare lexicographic and full conditional probabilities, and show they are not as similar as usually suggested by the literature. We then examine convexity, non-uniqueness and independence, always together with marginalization and conditioning. Several of the properties discussed here are well-known, but still they may be somewhat surprising as a whole, and call for further study concerning these formalisms.

2 Lexicographic Probabilities and Sets of Desirable Gambles

In this paper we only deal with finite objects, so that complications arising from infinity are entirely ignored. We assume there is a finite set of states, denoted by Ω, and that every subset of Ω is an event. A *gamble* is a function from states to real numbers. If Ω contains n states, a gamble can be thought of as an n-dimensional point. Hence we will treat sets of gambles as subsets of $\Re^n$.

The plan for this section is to emphasize the relationship between sets of lexicographic probability measures and sets of desirable gambles. Previously, Couso and Moral have studied this relationship in some restricted cases [5], and Quaeghebeur has dealt with this relationship in considerable detail [24, 25]. Most of the following discussion touches on topics that may be familiar to readers with background on imprecise probabilities.

Probabilities are often justified and derived by assuming axioms about preferences [1, 11, 30]. To simplify matters, we always take preferences over the set of all gambles. A popular way to do so is to take a preference relation $\succ$ between gambles, such that $f \succ g$ is interpreted as "f is preferred to g". Suppose $\succ$ is a (strict) partial order, meaning that it is irreflexive and transitive [11], and that $\succ$ satisfies a monotonicity condition: if $f(\omega) > g(\omega)$ for all ω, then $f \succ g$. Suppose additionally that it satisfies an "independence condition" such as: for any $\alpha \in (0, 1]$ and any f, g, h, we have $f \succ g$ if and only if $\alpha f + (1-\alpha)h \succ \alpha g + (1-\alpha)h$.

In this case the set of gambles that are preferred to the zero gamble is a cone that completely represents $\succ$. Suppose one assumes that this cone is open, an assumption that encodes an "Archimedean condition" on $\succ$ [31]. One then obtains the following representation: there is a unique maximal convex set $\mathbb{K}$ of probability measures such that $f \succ g$ if and only if $\forall \mathbb{P} \in \mathbb{K} : \mathbb{E}_{\mathbb{P}}[f] > \mathbb{E}_{\mathbb{P}}[g]$. Such a set of probability measures is called a *credal set*. Note that a preference profile may be completely characterized by more than one credal set, but there is a unique maximal credal set that offers such a representation, and this credal set is convex.

Suppose that $\succ$ is such that absence of preference is an equivalence (reflexive, transitive, symmetric); we then say that $\succ$ is a *strict weak order* [11]. If $\succ$ is a strict weak order, the credal set $\mathbb{K}$ is a singleton, so we obtain the usual representation by a single probability measure [11].

One might consider replacing the monotonicity condition by the following one: if $f(\omega) \geq g(\omega)$ for all ω and $f(\omega) > g(\omega)$ for some ω, then $f \succ g$. Following Blume et al., we refer to this property as *admissibility* [3, Definition 4.1]. Note that a standard probability measure may fail to represent admissibility (if $\mathbb{P}(\omega) = 0$, differences on this ω do not matter).

2.1 Lexicographic Probabilities

Now suppose $\succ$ is a strict partial order that satisfies the "independence condition" and admissibility, but no Archimedean condition. We then obtain a representation using lexicographic probabilities. A lexicographic probability measure is simply a sequence of standard probability measures $[\mathbb{P}_0, \ldots, \mathbb{P}_K]$, and the representation is of the form: $f \succ g$ if and only if there is $[\mathbb{P}_0, \ldots, \mathbb{P}_K]$ such that

$$[\mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{E}_{\mathbb{P}_K}[f]] >_{\mathrm{L}} [\mathbb{E}_{\mathbb{P}_1}[g], \ldots, \mathbb{E}_{\mathbb{P}_K}[g]],$$

where $>_{\mathrm{L}}$ denotes lexicographic comparison (for $a, b \in \Re^K$, $a >_{\mathrm{L}} b$ iff $a_j > b_j$ for some $j \leq K$ and $a_i = b_i$ for $1 < i < j$).

To produce the representation of $\succ$ in terms of lexicographic probability measures, note that $\succ$ can always be extended to a total order $\succ^*$ over gambles, such that $\succ^*$ satisfies the "independence condition" [31, Theorem 1] and admissibility. Every $\succ^*$ can be represented by a lexicographic linear utility [12, Chapter 4], and this lexicographic linear utility can be expressed as the expected value of a lexicographic probability measure (using the arguments by Blume et al. [3, Theorem 3.1]). Also the set of all extensions of $\succ$, and consequently $\succ$ itself, can be represented by a unique maximal convex set of lexicographic linear utilities [31, Theorem 2]. If $\succ$ is a strict weak order, the set of lexicographic linear utilities collapses to a single lexicographic probability measure [12, Chapter 4].

	H	T
layer 0	α	$(1-\alpha)$
layer 1	γ	$(1-\gamma)$

Table 1: Lexicographic probabilities; $\alpha, \gamma \in (0, 1)$.

Consider a lexicographic probability measure $[\mathbb{P}_0, \ldots, \mathbb{P}_K]$. Then $\mathbb{P}_i$ is called the ith *layer* of the lexicographic probability measure. One important fact is that each $\mathbb{P}_i$ is only unique up to linear combinations of $\mathbb{P}_0, \ldots, \mathbb{P}_i$ that assign positive weight to $\mathbb{P}_i$ [3, Theorem 3.1]. So in fact there is no intrinsic uniqueness in the lexicographic representation, as emphasized in the following example.

Example 1 Consider the lexicographic probability measure in Table 1, where each row contains a layer. The question is whether a gamble f such that $f(H) = a$ and $f(T) = b$ is preferred to the zero gamble. Using the first layer, $\mathbb{E}[f] = 0$ only if $a\alpha = -b(1-\alpha)$, so we might focus on the question of whether the gamble $(1-\alpha, -\alpha)$ is desirable or not. As the next layer gives value $\gamma - \alpha$ to this gamble, we only need to determine whether $\gamma > \alpha$ or $\gamma < \alpha$ to fix all preferences (if $\gamma = \alpha$, the second layer can be discarded). □

Admissibility requires each event to have positive probability with respect to at least one layer. This follows from the fact that any indicator function is nonnegative and positive for at least one ω; hence any indicator function if preferred to zero, and consequently for any event there is a probability measure that assigns it positive probability.

2.2 Sets of Desirable Gambles

For all preference orderings already discussed, axioms about preferences guarantee that $f \succ g$ if and only if $f - g \succ 0$. As noted already, we can then capture the preference relation by the set of gambles that are preferred to the zero gamble. This latter set is called the *set of desirable gambles* generated by the preference relation. But we can also start with sets of desirable gambles, properly axiomatized, and obtain preferences from them. For instance, here is a set of axioms that has been proposed for sets of desirable gambles [35]: a set of desirable gambles $\mathbb{D}$ is *coherent* if the zero gamble is not in $\mathbb{D}$; if all f such that $f \geq 0$ and $f \neq 0$ are in $\mathbb{D}$; if for any $\lambda > 0$ and any $f \in \mathbb{D}$, we have $\lambda f \in \mathbb{D}$; and if for any $f, g \in \mathbb{D}$, we have $f + g \in \mathbb{D}$. To obtain preferences from a given set of

desirable gambles, just say that $f \succ g$ if and only if $f - g \in \mathbb{D}$. By doing so, one notes that irreflexivity follows from the condition that the zero gamble is not in $\mathbb{D}$. Note also that admissibility follows from the second condition: if $f(\omega) \geq g(\omega)$ for all ω and $f(\omega) > g(\omega)$ for some ω, then $f \succ g$. Finally, transitivity and the independence condition follow from the other axioms. Hence a coherent set of desirable gambles can be completely represented by a (unique maximal convex) set of lexicographic linear utilities. From now on, every set of desirable gambles is assumed coherent, so we drop the qualifier "coherent" whenever possible.

2.3 Marginalization and Conditioning

Now consider a pair of random variables X and Y defined over Ω.

Marginalization of lexicographic probability measures is usually understood in a layer-wise fashion [14]. That is, if $[\mathbb{P}_0, \ldots, \mathbb{P}_K]$ are the layers of a lexicographic probability measure, then the marginal for X is a lexicographic probability measure where each layer is a probability measure over Ω_X with value (for X at x) $\sum_{\omega: X(\omega)=x} \mathbb{P}_i(\omega)$.

Given a set of desirable gambles $\mathbb{D}$, the *marginal* set of desirable gambles for X, denoted by $\mathbb{D}(X)$, is simply the set of all desirable gambles in $\mathbb{D}$ that are functions of X. For instance, if Ω is the Cartesian product of the set of values of X and the set of values of Y, respectively Ω_X and Ω_Y, then the Y-marginal $\mathbb{D}(Y)$ is $\{g : g \text{ is a function of } Y \text{ and } g \in \mathbb{D}\}$, with the understanding that $g \in \mathbb{D}$ means that the cylindrical extension of g belongs to $\mathbb{D}$ [26].

It should be apparent that marginalization means the same thing both for sets of desirable gambles and sets of lexicographic probability measures, given appropriate interpretation. If one starts with a set of desirable gambles, generates a set of lexicographic probability measures, marginalizes the latter, and generates the corresponding set of desirable gambles, one reaches the marginal of the original set of desirable gambles.

Conditioning of lexicographic probability measures has also received a layer-wise definition by Blume et al. [3]. That is, if we again have the lexicographic probability measure $[\mathbb{P}_0, \ldots, \mathbb{P}_K]$, then conditioning on A yields $[\mathbb{P}_0(\cdot|A), \ldots, \mathbb{P}_{K'}(\cdot|A)]$, where each layer that assigns positive probability to A is processed through Bayes rule, and all other layers are discarded. This sort of layer-wise Bayes rule is actually derived from preferences, as follows. From a preference relation $\succ$, obtain conditional preference given A, denoted by $\succ_A$, by saying that $f \succ_A g$ if and only if $Af \succ Ag$ [3, Definition 2.1]. Then $\succ_A$ is represented by a conditional lexicographic probability measure as just defined [3, Theorem 4.3].[1]

Given a set of desirable gambles $\mathbb{D}$ and an event A, the conditional set of desirable gambles $\mathbb{D}|A$ is simply $\{f : I_A f \in \mathbb{D}\}$, where I_A denotes the indicator function of A [26]. In fact, by using de Finetti's convention where an event and its indicator function are denoted by the same symbol, we have [35]:

$$\mathbb{D}|A = \{f : Af \in \mathbb{D}\}.$$

But this is clearly equivalent to representing the preferences $Af \succ Ag$. That is, both conditional lexicographic probability measures and conditional sets of desirable gambles represent the same operation.

In short, sets of (admissible) lexicographic probability measures and (coherent) sets of desirable gambles are equivalent representations for preferences under uncertainty.

3 Full Conditional Probabilities: Not Really

One of the attractive features of lexicographic probabilities and sets of desirable gambles is the fact that conditioning is well defined for any nonempty conditioning event (because every event has positive probability in some layer). Thus it is not surprising that lexicographic probability measures have been connected with the theory of full conditional probabilities [8, 19], because the latter also offers conditioning on every nonempty event.

In fact, there are some recurring themes in the connection between lexicographic and full conditional probabilities [3, 15, 16]. On the one hand, the structure of full conditional probabilities can be understood through lexicographic probabilities, and full conditional probabilities can be justified using the axioms of lexicographic probabilities. On the other hand, full conditional probabilities can be treated as if they were a class of lexicographic probabilities that are easy to specify, interpret, and handle. We now examine to which extent these intuitions are valid.

3.1 A Brief Review

To recap, a full conditional probability $\mathbb{P} : \mathcal{B} \times (\mathcal{B} \backslash \emptyset) \to \Re$, where $\mathcal{B}$ is a Boolean algebra, is a two-place set-function such that for every event $H \neq \emptyset$ [9]:
(1) $\mathbb{P}(H|H) = 1$;
(2) $\mathbb{P}(G|H) \geq 0$ for all G;
(3) $\mathbb{P}(G_1 \cup G_2|H) = \mathbb{P}(G_1|H) + \mathbb{P}(G_2|H)$ whenever

[1] Note that Blume et al. actually assumes that preference relations are reflexive, but their analysis of conditional probability is not affected by that.

$G_1 \cap G_2 = \emptyset$;
(4) $\mathbb{P}(G_1 \cap G_2|H) = \mathbb{P}(G_1|G_2 \cap H) \times \mathbb{P}(G_2|H)$ whenever $G_2 \cap H \neq \emptyset$.
Whenever the conditioning event H is equal to Ω, we suppress it and write the "unconditional" probability $\mathbb{P}(G)$.

The theory of coherent probabilities advocated by de Finetti adopts full conditional probabilities, and offers a justification for them that is based on betting (in fact de Finetti's original arguments were later formalized by Holzer [17] and Regazzini [28]). It should be noted that similar (but more general) axioms have been proposed by Renyi [29] and Popper [23]; there are also variants of those theories that we do not discuss for the sake of space.

Example 2 Take a coin with heads (H), tails (T), a sharp edge (S), and a blunt edge (B). We can have $\mathbb{P}(H) = \mathbb{P}(T) = 1/2$, hence $\mathbb{P}(S) = \mathbb{P}(B) = 0$, but still $\mathbb{P}(B|S \cup B) = 2/3$. □

A full conditional probability can always be represented as a sequence of standard probability measures $\mathbb{P}_0, \ldots, \mathbb{P}_K$ [3, 4, 16, 19]. To obtain this representation, we must partition Ω into several events $L_0, \ldots, L_K$. Take L_0 to be the set of elements of Ω that have positive unconditional probability. Then take L_1 to be the set of elements of Ω that have positive probability conditional on $\Omega \backslash L_0$. And then take L_i, for $i \in \{2, \ldots, K\}$, to be the set of elements of Ω that have positive probability conditional on $\Omega \backslash \cup_{j=0}^{i-1} L_j$. The event L_i denotes the support of the *layer* $\mathbb{P}_i$ of the full conditional probability. The *layer number* of layer $\mathbb{P}_i$ is i. For nonempty G, denote by L_G the support of the first layer such that $\mathbb{P}(G|L_G) > 0$. We then have $\mathbb{P}(G|H) = \mathbb{P}(G|H \cap L_H)$ [2, Lemma 2.1a]. Note that some authors use a different terminology, using instead the sequence $\cup_{j=i}^{K} L_j$ rather than L_i [4, 19].

Example 3 In Example 2, we have two layers. The first consists of H and T, with associated probabilities $\mathbb{P}_0(H) = \mathbb{P}_0(T) = 1/2$. The second layer consists of S and B, with associated probabilities $2\mathbb{P}_1(S) = \mathbb{P}_1(B) = 2/3$. □

3.2 Admissibility and Marginalization

Given the results enumerated in the previous section, it is natural to think that full conditional probabilities are just instances of lexicographic probability measures. However, strictly speaking, full conditional probabilities cannot pose as admissible lexicographic probabilities, as the theory of full conditional probabilities does not satisfy admissibility. For instance, consider the gamble f such that $f(H) = f(T) = f(S) = 0$, $f(B) = 1$ in Example 2. Computing expected value in the usual way with respect to this full conditional probability, we obtain zero; as far as preferences are to be extracted from expected values, this gamble is indistinguishable from the zero gamble. But if we were to interpret the layers of the full conditional probability as the layers of a lexicographic probability measure, then $f \succ 0$.

To obtain admissibility in actual decisions, one might use lexicographic expected values with respect to layers of a full conditional probability whenever necessary. For instance, in the previous paragraph one might say that $f \succ 0$ by looking at all layers of the full conditional probability. However, matters become even more delicate when we look at marginalization.

Example 4 Consider two variables X and Y, each with values $\{0, 1, 2\}$. Take a full conditional probability over (X, Y) with two layers (layer numbers are indicated by subscripts):

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$(1/5)_0$	$(1/5)_0$	$(1/5)_0$
$X=1$	$(1/5)_0$	$(1/4)_1$	$(1/4)_1$
$X=2$	$(1/5)_0$	$(1/4)_1$	$(1/4)_1$

We have marginal probabilities for X: $\mathbb{P}(X=0) = 3\mathbb{P}(X=1) = 3\mathbb{P}(X=2) = 3/5$. These marginal probabilities characterize a full conditional probability with a single layer. For this marginal full conditional probability, the gamble $f(X)$ such that $f(0) = -1$, $f(1) = 1$, $f(2) = 2$ has expected value equal to zero.

So, with respect to the marginal full conditional probabilities for X, there is not much to say about f; it is just indistinguishable from the zero gamble. There are no deeper layers to look at because the marginal full conditional probability does not assign zero probability to any event. So, there is no way to produce a lexicographic comparison if we first produce the full conditional probability that is the marginal of X.

However, as f can be obviously viewed as a function of X and Y, its expected value can be computed with respect to the joint full conditional probability. But now we see that we can have a lexicographic comparison: the expected value of f with respect to the second layer is $3/2$, hence $f \succ 0$.

That is, marginalization of full conditional probabilities loses information concerning layers, information that is needed if we were to compute lexicographic expected values. If we were to treat full conditional probabilities as lexicographic probabilities, we would need to have marginal full conditional probabilities that carry some extra information.

Indeed, if we took the joint full conditional probability

in our example as a lexicographic probability measure to begin with, and then marginalized it, we would obtain the following marginal lexicographic probabilities:

$X=0$	$X=1$	$X=2$
$(3/5)_0$	$(1/5)_0$, $(1/2)_1$	$(1/5)_0$, $(1/2)_1$

With respect to this marginal lexicographic probability measure, we obtain $f \succ 0$, as we must. □

So, if we wish to preserve admissibility by using lexicographic expectation with respect to full conditional probabilities, then the marginal of a full conditional probability must actually be represented as a lexicographic probability measure. The message is that lexicographic probabilities (and sets of desirable gambles) do offer conditioning on events of probability zero, but their solution is different from the one offered by full conditional probabilities. Lexicographic probabilities may be a representation for full conditional probabilities, but both behave differently.

4 Convexity?

When we adopt sets of lexicographic probability measures (or equivalently sets of desirable gambles), we seem to have convexity at hand. First, a set of desirable gambles is a convex object. Second, a strict ordering with independence and admissibility can be represented uniquely by a maximal convex set of lexicographic linear utilities.

However, convexity deserves further scrutiny. Again, it is useful to start this discussion with full conditional probabilities. Typically one assumes that, conditional on an event A, the set of probability measures $\mathbb{K}(\cdot|A)$ is convex [34]. But a set of full conditional probabilities cannot always be convex [13, 22], even if all probabilities are positive:

Example 5 Suppose $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathbb{P}_1(\omega_1) = \mathbb{P}_1(\omega_2) = \mathbb{P}_1(\omega_3) = 1/3$ and $\mathbb{P}_2(\omega_1) = 2\mathbb{P}_2(\omega_2) = 2\mathbb{P}_2(\omega_3) = 1/2$. Build the convex combination $\mathbb{P}_\alpha = \alpha\mathbb{P}_1 + (1-\alpha)\mathbb{P}_2$. There is no $\alpha \in (0,1)$ such that $\mathbb{P}_\alpha(\omega_1|\omega_1 \cup \omega_3) = 2(\alpha-3)/(\alpha-9)$ is equal to $\alpha\mathbb{P}_1(\omega_1|\omega_1 \cup \omega_3) + (1-\alpha)\mathbb{P}_2(\omega_1|\omega_1 \cup \omega_3) = \alpha/2 + 2(1-\alpha)/3$. That is, $\mathbb{P}_\alpha$ cannot be a convex combination of the functions $\mathbb{P}_1$ and $\mathbb{P}_2$. □

Consider a preference $\succ$ that can be extended to at least two orders, the former encoded by lexicographic linear utility u_1 and the latter by lexicographic linear utility u_2. On the one hand, any convex combination of these lexicographic linear utilities generates the same preference profile [31]. On the other hand, admissibility allows us to normalize each utility in the lexicographies, so as to obtain lexicographic probability measures [3]. However, suppose we wish to both normalize *and* do convex combinations. Apparently, matters are simple:

Example 6 Consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\alpha, \beta \in (0,1)$, and lexicographic probability measures $\mathbb{LP}_1$ and $\mathbb{LP}_2$:

	ω_1	ω_2	ω_3
$\mathbb{LP}_1(\omega_i)$	$(\alpha)_0$	$(1-\alpha)_0$	1_1

	ω_1	ω_2	ω_3
$\mathbb{LP}_2(\omega_i)$	$(1)_0$	$(\beta)_1$	$(1-\beta)_1$

Their half-half convex combination is:

ω_1	ω_2	ω_3
$((1+\alpha)/2)_0$	$((1-\alpha)/2)_0$, $(\beta/2)_1$	$(1-\beta/2)_1$

As a digression: $\mathbb{LP}_1$ and $\mathbb{LP}_2$ have disjoint layers, so they *could* be representations for full conditional probabilities. But their convex combination is certainly not the representation of a full conditional probability as the supports of the layers are not disjoint. □

The convex combination of lexicographic probability measures works perfectly if all lexicographic probability measures involved in the convex combination have the same number of layers. But suppose that modeling decisions have built two preference orderings with distinct depths; what to do?

Example 7 Consider $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\alpha, \beta, \gamma \in (0,1)$, all distinct, and lexicographic probability measures:

	ω_1	ω_2	ω_3
$\mathbb{LP}_1(\omega_i)$	$(\alpha)_0$, $(\gamma)_2$	$(1-\alpha)_0$, $(1-\gamma)_2$	1_1

	ω_1	ω_2	ω_3
$\mathbb{LP}_2(\omega_i)$	$(1)_0$	$(\beta)_1$	$(1-\beta)_1$

Note that $\mathbb{LP}_1$ reproduces Example 1, with one additional intervening layer. In fact $\mathbb{LP}_1$ defines a total order over gambles. And $\mathbb{LP}_2$ appeared in the previous example; for $\mathbb{LP}_2$ there are gambles that get zero expectation with respect to all layers (for instance, $h(\omega_1) = 0$, $f(\omega_2) = 1-\beta$, $f(\omega_3) = -\beta$).

Consider $\mathbb{LP}_{1/2}$, a half-half combination of $\mathbb{LP}_1$ and $\mathbb{LP}_2$. If we operate layer-wise,

	ω_1	ω_2	ω_3
$\mathbb{LP}_{1/2}(\omega_i)$	$(1+\alpha/2)_0$, $(\gamma/2)_2$	$((1-\alpha)/2)_0$, $(\beta/2)_1$ $((1-\gamma)/2)_2$	$(1-\beta/2)_1$

This is not a very satisfying result as probabilities in the last layer do not add up to one. □

One possible way to avoid the difficulties in this last example is always represent $\succ$ as a collection of total orders, all of which have the same depth. Indeed, the sets of lexicographic utilities by Seidenfeld et al. [31] are explicitly built from all such total orders, hence this sort of modeling decision makes sense conceptually.

However, there is a significant inconvenience. Suppose we wish to represent a set of preference orderings, some of which do display absence of preference. For instance, the ordering generated by $\mathbb{LP}_2$ does display absence of preferences (there are gambles that are not preferred nor dispreferred to the zero gamble). To represent such an ordering using total orders, we may need to introduce (possibly many) layers and measures that are apparently useless. To understand this, consider again $\mathbb{LP}_2$: to represent the strict weak order generated by $\mathbb{LP}_2$ using total orders, we might use a set consisting of two lexicographic probability measures:

	ω_1	ω_2	ω_3
$\mathbb{LP}_3(\omega_i)$	$(1)_0$	$(\beta)_1$ $(\delta_1)_2$	$(1-\beta)_1$ $(1-\delta_1)_2$

	ω_1	ω_2	ω_3
$\mathbb{LP}_4(\omega_i)$	$(1)_0$	$(\beta)_1$ $(\delta_2)_2$	$(1-\beta)_1$ $(1-\delta_2)_2$

It is particularly annoying that there is great latitude in selecting the probability values: as long as $(\delta_1-\beta)(\delta_2-\beta)<0$, we have the desired strict weak order collectively represented by $\mathbb{LP}_3$ and $\mathbb{LP}_4$ (and their convex combinations, if desired). The lack of control over the representation, given the ordering, is apparent.

One might look for alternative approaches. For instance, we might define the convex combination of two lexicographic probability measures so that a final normalization step is applied to each layer. Another possibility: adopt lexicographic "probability" measures that are not normalized below the first layer, and allow convex combinations without further concern. Whatever the solution, it seems that convexity deserves further analysis when applied to sets of lexicographic probability measures.

To a great extent, this discussion does not affect the theory of sets of desirable gambles. However, in practice one may be interested in representations for sets of desirable gambles that are based on probability values. When such representations are needed, the challenges in mixing sets of lexicographic probabilities and convexity are bound to surface.

5 Non-Uniqueness and Weakness

Some of the discussion in Section 3 concentrated on the fact that, given joint probabilities, marginals may not carry all necessary information. Now consider the reverse situation; that is, we have marginal and conditional lexicographic probabilities, and we wish to construct a joint lexicographic probability measure out of them. We find this not to be an easy problem. In fact, matters are difficult already for full conditional probabilities [6], as the next example shows. (Again, we resort to subscripts to indicate layer numbers.)

Example 8 Consider two binary variables X and Y. Suppose $\mathbb{P}(X=0)=1$ and $\mathbb{P}(Y=0|X=0)=\mathbb{P}(Y=0|X=1)=1$ (that is, the conditional probability of Y given X is actually not affected by X). The following joint full conditional probabilities:

	$Y=0$	$Y=1$
$X=0$	1_0	1_1
$X=1$	1_2	1_3

	$Y=0$	$Y=1$
$X=0$	1_0	1_2
$X=1$	1_1	1_3

satisfy all marginal and conditional assessments. □

The fact that marginal and conditional assessments cannot always uniquely characterize a joint full conditional probability has been noted before [6, 18]. In fact, one should take this phenomenon to suggest that as long as statistical modeling employs full conditional probabilities, one should not abide by any axiom that enforces uniqueness of probability values.

Lexicographic probabilities suffer from the same lack of uniqueness, only they suffer more deeply down their layers. Consider the following example.

Example 9 Suppose we have two variables X and Y, each with values $\{0,1,2\}$. Consider the following marginal assessments

$X=0$	$X=1$	$X=2$
$(1/2)_0$	$(1/2)_0$, $(1/2)_1$	$(1/2)_1$

and the following conditional assessments (for Y given X)

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$(1/2)_0$	$(1/2)_0$, $(1/2)_1$	$(1/2)_1$
$X=1$	$(1/2)_1$	$(1/2)_0$, $(1/2)_1$	$(1/2)_0$
$X=2$	$(1/2)_0$	1_1	$(1/2)_0$

There are *many* possible joint lexicographic probability measures that are compatible with these assessments. One possibility:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$(1/4)_0$	$(1/4)_0$, $(1/4)_1$	$(1/4)_1$
$X=1$	$(1/4)_1$, $(1/4)_3$	$(1/4)_0$, $(1/4)_1$, $(1/4)_2$, $(1/4)_3$	$(1/4)_0$, $(1/4)_2$
$X=2$	$(1/4)_2$	$(1/2)_3$	$(1/4)_2$

Another possible joint lexicographic probability measure is obtained, for instance, by exchanging the second and third layers of this latter lexicographic probability measure. But we can be more creative still, by adding layers in various ways; for instance, consider the following joint lexicographic probability measure, with *eight* layers, that satisfies all assessments. Here we use the notation $(\alpha)_{i:j}$ to indicate that value α appears in all layers between layer i (inclusive) and layer j (inclusive).

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$(1/4)_{0:1}$	$(1/4)_{0:3}$	$(1/4)_{2:3}$
$X=1$	$(1/4)_1$, $(1/4)_3$	$(1/4)_{0:7}$	$(1/4)_0$, $(1/4)_2$, $(1/4)_{4:7}$
$X=2$	$(1/4)_4$, $(1/4)_6$	$(1/2)_5$, $(1/2)_7$	$(1/4)_4$, $(1/4)_6$

We can produce many more joint lexicographic probabilities by combining marginal and conditional layers in various ways. □

To emphasize how information is lost through marginalization, consider one more example.

Example 10 Consider the following joint lexicographic probability measure.

	$Y=0$	$Y=1$	$Y=2$
$X=0$	1_0	$(1/3)_1$	$(3/8)_2$
$X=1$	$(1/6)_1$	$(1/6)_1$	$(1/8)_2$
$X=2$	$(1/6)_1$	$(1/2)_2$	$(1/2)_3$
$X=3$	$(1/6)_1$	$(1/2)_3$	1_4

To obtain the marginal lexicographic probability measure for Y, marginalize for each layer. We get $[1,0,0]$ for the first layer, $[1/2,1/2,0]$ for the second, $[0,1/2,1/2]$ for the third, $[0,1/2,1/2]$ for the fourth, and $[0,0,1]$ for the fifth. Note that the third and fourth layers collapse in the marginal; hence the "relative depth" of the fifth layer is lost.

One can interpret these facts as indicating that, once lexicographic probabilities are adopted, uniqueness of joint probabilities should be abandoned. So, one should be prepared to use sets of lexicographic probabilities (and the corresponding sets of desirable gambles) from the outset. This is a nice thought for anyone interested in imprecise and indeterminate probabilities; however, one can also interpret these examples as suggesting that marginalization and conditioning are quite weak when applied to lexicographic probabilities (and sets of desirable gambles). Consider this. If we start with a joint lexicographic probability measure, then its marginal and conditional probabilities contain some useful information, but not all the information needed to rebuild the joint. Specifically, we do not have information concerning which layers of marginal and conditional probabilities should be combined together to produce the joint. Similarly, if we start with marginal and conditional lexicographic probabilities, we do not have all the information to build a single joint. Should we really have all this indeterminacy?

6 Independence

In this section we briefly comment on the concept of independence in the context of lexicographic probabilities. To do so, first we must agree on what "independence" means here.

One might try to define independence by requiring the joint to be a product of the marginals. But a little reflection suggests this not to be easy: because a lexicographic probability does not fundamentally change if we transform linearly its layers, one can destroy an "independence" just by rewriting its terms through linear transformations. It seems wiser to define independence as a property of the preference orderings that are implied by conditional and marginal probabilities. This sort of definition is proposed by Blume et al. [3]. They use conditional preferences, denoted by $\succ_A$ (Section 2), as follows. Variables X and Y are independent when we have, first, $[f_1(X) \succ_{\{Y=y_1\}} f_2(X)] \Leftrightarrow [f_1(X) \succ_{\{Y=y_2\}} f_2(X)]$ for any f_1, f_2, y_1, y_2, and second, the same condition with X and Y exchanged.[2] A stronger condition is [7]: X and Y are independent when $[f_1(X) \succ_{\{Y=y\}} f_2(X)] \Leftrightarrow [f_1(X) \succ f_2(X)]$ for any f_1, f_2, y, and second, the same condition with X and Y exchanged. These concepts of independence fail

[2] The fact that X and Y are independent does not guarantee any factorization of lexicographic probabilities. Blume et al. show that even hyperreal representations of lexicographic probabilities fail to factorize under their definition [3].

	$W=0,Y=0$	$W=1,Y=0$	$W=0,Y=1$	$W=1,Y=1$
$X=0$	$(1/2)_0$	$(1/2)_0$	$(1)_2$	$(1)_3$
$X=1$	$(1/2)_1$	$(1/2)_1$	$(1/2)_4$	$(1/2)_4$

Table 2: Lexicographic probabilities in Example 11.

the Decomposition property [7]; that is, we may find that X and (W,Y) are independent but still X and W are *not* independent. An even stronger concept of independence has been proposed [7]: X and Y are independent when $[f_1(X) \succ_B f_2(X)] \Leftrightarrow [f_1(X) \succ f_2(X)]$ for any f_1, f_2, and any set B of values of Y, and second, the same condition with X and Y exchanged. But this fails the Contraction property [7]: we may have X and Y independent, and W and X independent given any value of Y, and yet X and (W,Y) fail to be independent. Failure of these properties reveal weaknesses of existing concepts and deserve further debate. Moreover, such concepts of independence do not guarantee a unique joint lexicographic measure for given marginals (consider again Example 8; X and Y are independent and there is no uniqueness). However, the purpose of this section is not to insist on these facts, but rather to examine a point that seems particularly hard to handle.

Example 11 Suppose we have three binary variables, W, X and Y, and joint lexicographic probabilities in Table 2. If we look at the marginal probabilities for (X,Y), we see that X and Y are independent according to all definitions above. Indeed, the preferences on (X,Y) are represented by:

	$Y=0$	$Y=1$
$X=0$	$(1)_0$	$(1)_2$
$X=1$	$(1)_1$	$(1)_3$

However, there is something intuitively strange about this independence. If we observe $\{Y=0\}$, the difference between $\{X=0\}$ and $\{X=1\}$ is a single "jump" between layers. We might interpret that $\{X=1\}$ is infinitesimally smaller than $\{X=0\}$. But given $\{Y=1\}$, the jump between them is twice as big as we go down two layers of the joint distribution. The interpretation should be that, given $\{Y=1\}$, $\{X=1\}$ is infinitesimally smaller than some event that is infinitesimally smaller than $\{X=0\}$. In a sense, one feels that the marginal for (X,Y) should be

	$Y=0$	$Y=1$
$X=0$	$(1)_0$	$(1)_2$
$X=1$	$(1)_1$	$(1)_4$

Now if all we had were these marginal lexicographic probabilities, it would be difficult to argue that X and Y should be considered independent, because there are different jumps given distinct conditioning events. But lexicographic probabilities do not let us keep the jumps between layers intact. In fact there seems to be no way to extract such differences in relative depth of layers by looking at preferences that only involve X and Y; by the same token, there seems to be no way to extract such differences from the corresponding set of desirable gambles. □

7 Discussion

This paper discussed properties of sets of lexicographic probability measures and sets of desirable gambles. Most of the discussion actually dealt with lexicographic probabilities and sets of them. However, any conclusions we reach for these objects should be easily transferred to the equivalent language of sets of desirable gambles. Even though sets of desirable gambles avoid some of the non-uniqueness inherent to lexicographic probabilities, most examples in this paper could also be expressed through sets of desirable gambles. Moreover, even if one wishes to focus on sets of desirable gambles, at some point their natural representation as lexicographic probabilities must be considered, and then the features of lexicographic probabilities must be properly understood.

In many ways, sets of desirable gambles offer an attractive formalism to handle uncertainty. We basically have to deal with cones of gambles; these are linear structures with clear geometric appeal. But this simplicity may be illusory; even though the geometry is simple, matters get complicated when we wish to represent in detail operations such as marginalization and conditioning. By playing with sets of desirable gambles and sets of lexicographic probabilities, we can better understand both operations.

To summarize, we have started by emphasizing the link between lexicographic probabilities and sets of desirable gambles. We have then examined the connection between lexicographic probabilities and full conditional probabilities; this connection seems to be weaker than sometimes assumed in the literature. We have emphasized the fact that modeling with full conditional probabilities and lexicographic probability measures leads one to deal with non-uniqueness of probability values. The move to non-uniqueness led us to con-

sider differences between full conditional probabilities and lexicographic probabilities concerning convexity. And we have examined some challenges in interpreting independence for lexicographic probabilities (and consequently for sets of desirable gambles).

Acknowledgements

The author is partially supported by CNPq (grant 305395/2010-6).

References

[1] Francis J. Anscombe and Robert J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205, 1963.

[2] Pierpaolo Battigalli and Pietro Veronesi. A note on stochastic independence without Savage-null events. *Journal of Economic Theory*, 70(1):235–248, 1996.

[3] Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 58(1):61–79, January 1991.

[4] Giulianella Coletti and Romano Scozzafava. *Probabilistic Logic in a Coherent Setting*. Trends in logic, 15. Kluwer, Dordrecht, 2002.

[5] Inés Couso and Serafín Moral. Sets of desirable gambles: Conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52:1034–1055, 2011.

[6] Fabio G. Cozman. Independence for full conditional probabilities: Structure, factorization, nonuniqueness, and Bayesian networks. *International Journal of Approximate Reasoning*, 54:1261–1278, 2013.

[7] Fabio G. Cozman. Independence for sets of full conditional measures, sets of lexicographic probabilities, and sets of desirable gambles. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 87–98, 2013.

[8] Bruno de Finetti. *Theory of Probability, vol. 1-2*. Wiley, New York, 1974.

[9] Lester E. Dubins. Finitely additive conditional probability, conglomerability and disintegrations. *Annals of Statistics*, 3(1):89–99, 1975.

[10] Terrence L. Fine. *Theories of Probability*. Academic Press, 1973.

[11] Peter C. Fishburn. *Utility Theory for Decision Making*. John Wiley and Sons, Inc., New York, 1970.

[12] Peter C. Fishburn. *The Foundations of Expected Utility*. D. Reidel Publishing Company, 1982.

[13] Angelo Gilio and Salvatore Ingrassia. Totally coherent set-valued probability assessments. *Kybernetica*, 34(1):3–15, 1998.

[14] Srihari Govindan and Tilman Klumpp. Perfect equilibrium and lexicographic beliefs. *International Journal of Game Theory*, 31:229–243, 2002.

[15] Joseph Y. Halpern. Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*, 68:155–179, 2010.

[16] Peter J. Hammond. Elementary non-Archimedean representations of probability for decision theory and games. In P. Humphreys, editor, *Patrick Suppes: Scientific Philosopher; Volume 1*, pages 25–59. Kluwer, Dordrecht, The Netherlands, 1994.

[17] S. Holzer. On coherence and conditional prevision. *Bolletino Un. Mat. Ital. Serie VI, Analisi Funzionale e Applicazioni*, IV-C(1):441–460, 1985.

[18] Elon Kohlberg and Philip J. Reny. Independence on relative probability spaces and consistent assessments in game trees. *Journal of Economic Theory*, 75:280–313, 1997.

[19] Peter Krauss. Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Academiae Scientiarum Hungaricae*, 19(3-4):229–241, 1968.

[20] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[21] Charles F. Manski. Learning and decision making when subjective probabilities have subjective domains. *Annals of Statistics*, 9(1):59–65, 1981.

[22] F. Matus. Conditional probabilities and permutahedron. *Annales de l'Institut H. Poincaré, Probabilités et Statistiques*, 39:687–701, 2003.

[23] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1975.

[24] Erik Quaeghebeur. The CONEstrip algorithm. In *Advances in Intelligent Systems and Computing*, volume 190, pages 45–54, 2013.

[25] Erik Quaeghebeur. A propositional CONEstrip algorithm. In *IPMU*, 2014.

[26] Erik Quaeghebeur. Desirability. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 1–27. Wiley, 2014.

[27] Erik Quaeghebeur, Gert de Cooman, and Filip Hermans. Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning*, 57:69–102, 2015.

[28] Eugenio Regazzini. Finitely additive conditional probability. *Rend. Sem. Mat. Fis.*, 55:69–89, 1985.

[29] Alfred Renyi. On a new axiomatic theory of probability. *Acta Math. Acad. Sci. Hungarian*, 6:285–335, 1955.

[30] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, Inc, New York, 1972.

[31] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Decisions without ordering. In W. Sieg, editor, *Acting and Reflecting*, pages 143–170. Kluwer Academic Publishers, 1990.

[32] Teddy Seidenfeld. Remarks on the theory of conditional probability: some issues of finite versus countable additivity. In *Statistics – Philosophy, Recent History, and Relations to Science*, pages 167–177. Kluwer Academic, 2001.

[33] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[34] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[35] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

On the Complexity of Propositional and Relational Credal Networks

Fabio Gagliardi Cozman
Escola Politécnica
Universidade de São Paulo, Brazil
fgcozman@usp.br

Denis Deratani Mauá
Instituto de Matemática e Estatística
Universidade de São Paulo, Brazil
denis.maua@gmail.com

Abstract

A credal network associates a directed acyclic graph with a collection of sets of probability measures. Usually these probability measures are specified through several tables containing probability values. Here we examine the complexity of inference in Boolean credal networks when probability measures are specified through formal languages, by extending a framework we have recently proposed for Bayesian networks. We show that sub-Boolean and relational logics lead to interesting complexity results. In short, we explore the relationship between language and complexity in credal networks.

Keywords. Credal networks, Propositional logic, Relational logic, Complexity, Data complexity.

1 Introduction

A credal network represents a set of probability distributions through a directed acyclic graph and an associated set of "local" credal sets [1, 6]. Usually these local credal sets are specified using tables containing probability values, possibly with some additional constraints between them. In practice, any elicitation strategy must adopt some specification language in which to encode probability assessments. For instance, one may allow inequalities such as $\mathbb{P}(A) \geq 1/2$, or perhaps interval-valued assessments such as $\mathbb{P}(A) \in [3/5, 7/10]$; of course, one may have a specification language with propositions and Boolean operators, or even relations and quantifiers.

In this paper we study properties of credal networks as parameterized by specification languages. We look at the balance of expressivity for specification languages and the complexity of inferences. We concentrate on Boolean variables, and focus on a particular semantics for credal networks (the semantics of "strong extensions"). To investigate the interplay between expressivity and complexity, we extend a framework we have recently developed to study the complexity of Bayesian networks [9].

We start with some necessary background in Section 2. We discuss our framework in Section 3, in particular looking at propositional languages. Sections 4 and 5 examine relational languages.

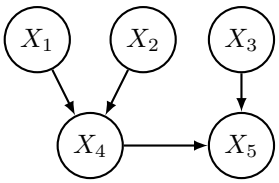
2 Credal networks and their strong extensions

In this paper every possibility space Ω is finite; a random variable is simply a function from Ω into the reals, and we consider only random variables taking on two values, 1 (meaning "true") and 0 (meaning "false"). A set of probability measures is called a *credal set* [18]. We abuse language by referring to sets of probability distributions, and also to sets of probability mass functions, as credal sets. A set of distributions for a variable X is denoted by $\mathbb{K}(X)$. Given a credal set, for any event A we have its *lower* and *upper* probabilities, denoted by $\underline{\mathbb{P}}(A)$ and $\overline{\mathbb{P}}(A)$ respectively: $\underline{\mathbb{P}}(A) = \inf \mathbb{P}(A)$ and $\overline{\mathbb{P}}(A) = \sup \mathbb{P}(A)$. In this paper W, X, Y and Z denote random variables, while A and B denote events or propositions.

A conditional credal set is obtained by applying Bayes rule to each possible distribution in a credal set; we also refer to sets of conditional distributions and conditional mass functions as conditional credal sets. We adopt regular conditioning; that is, $\mathbb{K}(X|A)$ is the set of all conditional distributions that are obtained from distributions such that $\mathbb{P}(A) > 0$ [30]. We denote by $\mathbb{K}(X|Y)$ the set containing a credal set $\mathbb{K}(X|Y = y)$ for each possible value of Y. The sets $\mathbb{K}(X|Y)$ are *separately specified* when there is no constraint on the conditional set $\mathbb{K}(X|Y = y_1)$ that is based on the properties of $\mathbb{K}(X|Y = y_2)$, for any $y_2 \neq y_1$. For events A and B, we define lower and upper conditional probabilities: $\underline{\mathbb{P}}(A|B) = \inf_{\mathbb{P}:\mathbb{P}(B)>0} \mathbb{P}(A|B)$ and $\overline{\mathbb{P}}(A|B) = \sup_{\mathbb{P}:\mathbb{P}(B)>0} \mathbb{P}(A|B)$.

Given some marginal and conditional credal sets, an *extension* of these sets is a joint credal set with the given marginal and conditional credal sets.

A credal network consists of a directed acyclic graph where each node is a random variable X_i, together with a set of constraints on probability values. The graph is assumed to encode independence relations amongst variables, and the constraints convey the probabilistic assessments. The independence relations are given by a Markov condition, soon to be explained. Such a structure is useful as a representation for beliefs, opinions, and statistical summaries that may be available when modeling a particular problem. For instance, suppose we have five variables, representing say economic indicators:



Here we have that X_1 and X_2 are *parents* of X_4; likewise, X_3 and X_4 are parents of X_5. The parents of X_i are denoted by $\text{pa}(X_i)$. The meaning of the graph is conveyed by the *Markov condition*: every X_i is independent of its nondescendants nonparents given its parents. So, X_5 is independent of X_1 and X_2 given X_3 and X_4. Hence by drawing the graph we are expressing our belief that, conditional on X_3 and X_4, no information about X_1 and X_2 can change our assessments on X_5.

To continue the example, we may have some constraints on probabilities. Even though one is free to impose say $\mathbb{P}(X_1 = 0|X_4 = 1) \geq 2/3$ and $\mathbb{P}(X_3 = 1 \wedge X_2 = 0) \leq 1/2$, usually applications constrain assessments to a few simple forms [1, 6]. Typically we have each variable X_i associated with separately specified sets $\mathbb{K}(X_i|\text{pa}(X_i))$. When every credal set $\mathbb{K}(X_i|\text{pa}(X_i) = \pi)$ is a singleton, the resulting model is equivalent to a *Bayesian network*.

Once assessments are given, we can construct their joint *extension*; that is, we can construct a credal set consisting of those joint distributions that satisfy the assessments. We have some freedom here, for we can interpret the "independence relations" in the Markov condition in various ways. There are several concepts of independence that apply to credal sets [7]; we might for instance consider extensions that interpret the Markov condition through *epistemic irrelevance* [11]. In this paper we adopt the most common concept of independence for credal sets; namely, we adopt *strong independence*: X and Y are strongly independent given Z if $\mathbb{K}(X, Y|Z = z)$ is the convex hull of a set of distributions that factorize; that is, if any $p(X, Y|Z = z)$ in this latter set satisfies $p(X, Y|Z = z) = p(X|Z = z)\, p(Y|Z = z)$.

We are always interested in the *largest* extension that satisfies given assessments and independence relations. We refer to such extensions, when strong independence is adopted, as *strong extensions*. Our results are *also* valid if one adopts *complete independence*, provided one always keeps the interest in the largest possible extension: X and Y are completely independent given Z if any probability mass $p(X, Y|Z = z)$ in $\mathbb{K}(X, Y|Z = z)$ satisfies $p(X, Y|Z = z) = p(X|Z = z)\, p(Y|Z = z)$. To simplify the presentation, we focus only on strong independence and strong extensions.

Given a credal network (graph and assessments) and its resulting extension, we are interested in computing conditional upper probabilities such as $\overline{\mathbb{P}}(X_1 = 0|X_2 = 1)$.

3 A Framework for Complexity Analysis

We now extend a framework for complexity analysis that we have recently developed for Bayesian networks [9], so as to include probability intervals. The basic idea is to restrict assessments to two simple forms that are inspired by probabilistic rules [22, 26] and structural models [21]. The framework lets one move down to sub-Boolean constructs and up to relations and quantifiers. In the context of credal networks and strong extensions, our framework is valuable as it imposes some regularity into the specification, for instance automatically implying that all local credal sets are separately specified. So, it offers a combination of flexibility and restraint that should be useful in practical elicitation scenarios.

We will refer to existing complexity classes in our results. To recap, the class PP consists of those problems that can be solved by a nondeterministic polynomially-bounded Turing machine where the acceptance condition is that more than half of computation paths accept [20]. And $\mathsf{NP}^{\mathsf{PP}}$ consists of those problems that can be solved by a nondeterministic polynomially-bounded Turing machine with an oracle that solves PP decision problems [19]. In proofs we use reductions from E-MAJSAT, an $\mathsf{NP}^{\mathsf{PP}}$-complete problem [19]. The E-MAJSAT decision problem is: given a pair (ϕ, k) where ϕ is a Boolean sentence with n propositions, and $k \in [1, n]$ is an integer, is there an assignment of the first k propositions such that the majority of assignments to the remaining propositions satisfies ϕ?

Returning to the specification framework: Consider a set of atomic propositions, $A_1, \ldots, A_n$, and take the

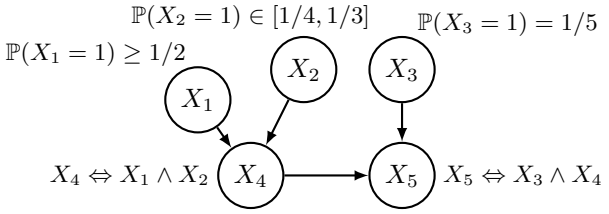


Figure 1: A simple credal network.

set Ω of 2^n truth assignments. Associate a binary variable X_i with atomic proposition A_i, such that $X_i(\omega) = 0$ when A_i is false, and $X_i(\omega) = 1$ when A_i is true, for $\omega \in \Omega$. Our credal networks are to be specified over $X_1, \ldots, X_n$; to simplify the presentation, we equate atomic propositions and their associated variables. That is, we write propositional sentences containing variables and their assignments, and we write probabilities for propositional sentences.

We assume that a directed acyclic graph is given, where each node is a variable X, and that each variable X is associated with either:

- an equivalence $X \Leftrightarrow F(Y_1, \ldots, Y_m)$, or
- a probabilistic assessment $\mathbb{P}(X = 1) \in [\alpha, \beta]$,

where F is a formula on propositions $Y_1, \ldots, Y_m$ that are parents of X, and where α and β are nonnegative rationals in $[0, 1]$. We call the former a *logical assessment*, and the latter a *probabilistic assessment*.

By adopting this restricted syntax, the graph is actually redundant. One can simply give a set of assessments, and as long as there are no cycles in the specification, the graph can be then constructed out of the assessments.

Note that we avoid direct assessments of conditional probability. First, such an assessment may essentially create negation (by imposing $\mathbb{P}(X = 1|Y = 1) = \mathbb{P}(X = 0|Y = 0) = 0$); we wish to control the use of negation. Second, by avoiding conditional probabilities we do not need to start by discussing conditioning on events that can have probability zero, a discussion that is always difficult for the novice [8].

To illustrate the framework, consider the specification in Figure 1. One might interpret this network as follows: X_4 is a health condition that is identified with the conjunction of two risk factors, and X_5 is an illness that depends probabilistic on X_4, with X_3 acting as "inhibitor".

The strong extension of this credal network is simply the convex hull of all extreme Bayesian networks, where an extreme Bayesian network is obtained by taking extreme (upper or lower) probabilities [12, 13]. Hence we have eight possible configurations of variables, and four extreme joint probability distributions. For instance, one such distribution assigns probability $1/2$ to $\{X_1 = 1\}$ and probabilty $1/4$ to $\{X_2 = 1\}$, while another distribution assigns probability 1 to $\{X_1 = 1\}$ and probability $1/4$ to $\{X_2 = 1\}$.

Denote by $\mathcal{C}(\mathcal{L})$ the set of credal networks that can be produced through the framework above, with formulas F from a language $\mathcal{L}$ (a language $\mathcal{L}$ is simply a set of well-formed formulas). Then $\mathsf{INF_d}(\mathcal{L})$ denotes the set of decision problems that yield YES if $\overline{\mathbb{P}}(Q|\mathbf{E}) > \gamma$ for an assignment Q, a conjunction $\mathbf{E}$ of assignments, a rational $\gamma \in [0, 1]$, and a credal network in $\mathcal{C}(\mathcal{L})$, and NO otherwise [10]. The set $\mathbf{E}$ is the *evidence*; we focus only on conjunctions of assignments, and leave for the future the study of more general languages in which to express evidence. To simplify the statement of some results, we denote by $\mathsf{INF_d}^+(\mathcal{L})$ the decision problems defined as in $\mathsf{INF_d}(\mathcal{L})$, with the additional constraint that all assignments are "positive" (that is, variables are only set to true).

Denote by $\mathsf{Prop}(\wedge, \neg)$ the language of well-formed propositional sentences with conjunction and negation. First note that $\mathsf{Prop}(\wedge, \neg)$ can specify any distribution over variables $X_1, \ldots, X_n$ that can be specified by a Bayesian network over these variables. To see why, suppose we have a Bayesian network over $X_1, \ldots, X_n$. Consider first a variable X with two parents Y_1 and Y_2. Impose:

$$X \Leftrightarrow (\neg Y_1 \wedge \neg Y_2 \wedge Z_{00}) \vee (\neg Y_1 \wedge Y_2 \wedge Z_{01}) \vee (Y_1 \wedge \neg Y_2 \wedge Z_{10}) \vee (Y_1 \wedge Y_2 \wedge Z_{11}),$$

where Z_{ab} are fresh binary variables (that do not appear anywhere else), associated with assessments $\mathbb{P}(Z_{ab} = 1) = \mathbb{P}(X = 1|Y_1 = a, Y_2 = b)$. Obviously we can always produce disjunction using conjunction and negation, so $\vee$ appears as syntactic sugar in this latter expression. Now for a variable X with many parents, we just repeat this structure, by taking into account any possible configuration of parents. The marginal distribution of $X_1, \ldots, X_n$ is exactly the distribution specified by the original Bayesian network.

By allowing interval-valued assessments in our framework, we obtain a similar result for credal networks: $\mathsf{Prop}(\wedge, \neg)$ allows us to specify any (separately specified) strong extension over variables $X_1, \ldots, X_n$. To see why, suppose we have a separately specified credal network over $X_1, \ldots, X_n$. Consider again a variable X with two parents Y_1 and Y_2, and suppose $\mathbb{K}(X|Y_1, Y_2)$ is such that each $\mathbb{K}(X|Y_1 = a, Y_2 = b)$ has two extreme points, $p_0(X|Y_1 = a, Y_2 = b)$ and $p_1(X|Y_1 = a, Y_2 = b)$. Introduce fresh variables W_{ab}

and Z_{abc}, and let

$$X \Leftrightarrow \bigvee_{\substack{a\in\{0,1\}\\ b\in\{0,1\}\\ c\in\{0,1\}}} (Y_1 = a)\wedge(Y_2 = b)\wedge(W_{ab} = c)\wedge(Z_{abc} = 1),$$

and assessments $\mathbb{P}(Z_{abc} = 1) = p_c(X = 1|Y_1 = a, Y_2 = b)$ and $\mathbb{P}(W_{ab} = 1) \in [0, 1]$. This encodes the desired local, separately specified, credal sets. The idea is that a and b select a particular configuration of Y_1 and Y_2, while c selects a particular extreme point of the corresponding local credal set (and then Z_{abc} carries the appropriate probability value). By repeating this structure to take into account any configuration of parents of X, we construct a joint credal set whose marginal is the strong extension of the original credal network (note that we may have to use additional variables with the same role as W_{ab}, in case we have more than two extreme points per credal set).

Given the generality of $\mathsf{Prop}(\wedge, \neg)$, we have that $\mathsf{INF_d}(\mathsf{Prop}(\wedge, \neg))$ is $\mathsf{NP}^{\mathsf{PP}}$-complete [10]. Now consider a more restricted language: denote by $\mathsf{Prop}(\wedge, (\neg))$ the language that uses only conjunction and *atomic* negation (defined as negation that can appear only before a proposition that is associated with a probabilistic assessment). Note that the credal network in Figure 1 belongs to $\mathcal{C}(\mathsf{Prop}(\wedge, (\neg)))$. We know that inference within $\mathsf{Prop}(\wedge, (\neg))$ for Bayesian networks is polynomial as long as evidence is "positive" [9]. Somewhat surprisingly, this result applies to credal networks:

Theorem 1 $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, (\neg)))$ *can be solved in polynomial time.*

Proof. Consider first a network with just conjunction, and consider a query $Q = \{X_Q = 1\}$. Note first that if a node X appears in Q or in $\mathbf{E}$, then its ascendants must all be set to true. So we first add to $\mathbf{E}$ all ascendants of nodes originally in $\mathbf{E}$; also, if a node has all parents set to true, then it must be true and can be added to $\mathbf{E}$, so we repeat this until no more nodes can be added to $\mathbf{E}$. Now if any descendant of X_Q is in the evidence, then X_Q is necessarily true, so we have $\underline{\mathbb{P}}(Q|\mathbf{E}) = \overline{\mathbb{P}}(Q|\mathbf{E}) = 1$. So, either we have evidence assigned to a descendant of X_Q, and then the solution is immediate, or all descendants are barren nodes that can be discarded. So, to proceed we suppose that X_Q has no descendants. Now continue by d-separation. Collect all nodes that are ascendants of X_Q; these are d-connected to X_Q. Now suppose one of these nodes, say W, points both to an ascendant of X_Q, and to a non-ascendant, say Y, of X_Q. Now if Y is not in $\mathbf{E}$, then it is a barren node that must be discarded. And if Y is in $\mathbf{E}$, then W itself must be in $\mathbf{E}$, hence Y is to be discarded. For instance, consider Figure 2, and suppose $\{Y = 1\}$ is the evidence. Then W, Z and W' are set to true, and we can discard them, as the path emanating from X_Q through W is blocked. Once we have discarded all nodes that are not required for our computation, we are left with an "inverted" tree whose root is X_Q, and where each leaf is either a node set to true, or a node associated with a probability interval. Denote by $X_1, \ldots, X_m$ the nodes that are not set to true in this tree; we can then write $X_Q \Leftrightarrow X_1 \wedge \cdots \wedge X_m$. So we have $\overline{\mathbb{P}}(X_Q|\mathbf{E}) = \prod_{i=1}^m \overline{\mathbb{P}}(X_i)$; in fact, we also have that $\underline{\mathbb{P}}(X_Q|\mathbf{E}) = \prod_{i=1}^m \underline{\mathbb{P}}(X_i)$. To complete the proof, suppose that atomic negation is allowed, so some variables appear negated. We can run the same procedure already described, with the novelty that we cannot have $X \wedge \neg X$ in the final expression (if that happens, the evidence is inconsistent). □

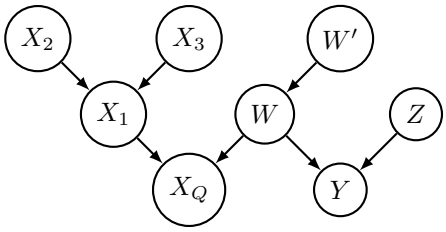


Figure 2: Network in Theorem 1.

It seems unlikely that polynomial-time inference can be obtained with other languages for Boolean credal networks, as several simple changes to $\mathsf{Prop}(\wedge, (\neg))$ move us into higher complexity.[1] Consider: even though $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, (\neg)))$ belongs to P, $\mathsf{INF_d}(\mathsf{Prop}(\wedge, (\neg)))$ does not (as it is PP-hard already when all probability intervals are singletons [9]). Also, if we move from $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, (\neg)))$ to $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, \neg))$, then clearly we obtain $\mathsf{NP}^{\mathsf{PP}}$-completeness. Finally, we might move to $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, \vee, (\neg)))$ by adding disjunction. In doing so, again we move away from polynomial-time behavior, as the following result shows.

Theorem 2 $\mathsf{INF_d}^+(\mathsf{Prop}(\wedge, \vee, (\neg)))$ *is* $\mathsf{NP}^{\mathsf{PP}}$*-complete.*

Proof. Consider an E-MAJSAT problem specified by (ϕ, k). We can code ϕ in CNF within $\mathsf{Prop}(\wedge, \vee, (\neg))$. For a given k, we can associate the first k variables X_i with assessments $\mathbb{P}(X_i) \in [0, 1]$, and the remaining variables X_j with assessments $\mathbb{P}(X_j) = 1/2$. We can then produce a network where each proposition is a root node, and all other nodes are either conjunctions or disjunctions of their parents. This network has size polynomial on the input. Denote by Q an assignment

[1] But note that if network topology is constrained to polytrees, then polynomial behavior is obtained by the 2U algorithm [13]. Hence, by suitably restricting the topology, we still get tractability.

for the leaf node that yields the final conjunction in the CNF. By deciding whether $\overline{\mathbb{P}}(Q) > 1/2$, we solve the E-MAJSAT problem. □

To close this section, we comment on an additional type of assessment that one might allow, namely, assessments where material implication is used instead of equivalence. For instance, suppose that in our previous example we change the logical assessment for X_5 to

$$X_5 \leftarrow X_3 \wedge X_4.$$

A sensible semantics here might be to consider every possible probability measure compatible with this logical constraint; that is, the assessment should mean $\mathbb{P}(A_5|A_3 \wedge A_4) = 1$ and $\mathbb{P}(A_5|\neg(A_3 \wedge A_4)) \in [0,1]$. This suggests that if we are willing to contemplate assessments based on material implication, we should be willing to entertain interval probabilities from the outset. We leave such a discussion for the future, noting here that existing languages such as Poole's Independent Choice Logic (ICL) [23] do have material implication in the syntax, but often adopt special semantics to guarantee sharp probabilities.

4 Relational Credal Networks

Many phenomena in real life depict repetitive patterns. For instance, social networks involve many individuals, several of which may share common characteristics. Epidemiological events may also bring together similar individuals; temporal sequences modeled by hidden Markov models often display similarities across time steps. There are indeed several formalisms that capture repetition in Bayesian network fragments [15, 16, 24, 25]. The simplest strategy is to allow random variables to be parameterized; for instance, we might extend the specification in the previous section as follows:

$$\mathbb{P}(X_1(\mathfrak{x}) = 1) \geq 1/2, \quad (1)$$

$$\mathbb{P}(X_2(\mathfrak{x}) = 1) \in [1/4, 1/3], \quad (2)$$

$$\mathbb{P}\Big(X_3(\mathfrak{x}, y) = 1\Big) = 1/5, \quad (3)$$

$$X_4(\mathfrak{x}) \Leftrightarrow X_1(\mathfrak{x}) \wedge X_2(\mathfrak{x}), \quad (4)$$

$$X_5(\mathfrak{x}, y) \Leftrightarrow X_3(\mathfrak{x}, y) \wedge X_4(\mathfrak{x}). \quad (5)$$

At this point we can simply refer to $\mathfrak{x}, y, \ldots$ as *logical variables*, and to $X_1(\mathfrak{x}), X_2(\mathfrak{x}), X_3(\mathfrak{x}, y)$ as *relations*. Again we write sentences that mix variables and Boolean operators. We say that $X(\mathfrak{x}_1, \ldots, \mathfrak{x}_k)$, where each $\mathfrak{x}_i$ is either a logical variable or an individual, is an *atom*. An atom with no logical variable is a *ground atom*.

We can then extend our specification framework as follows.

We assume that a directed acyclic graph is given, where each node is a relation, and that every k-ary relation X is associated with either:

- a logical assessment

  $$X(\mathfrak{x}_1, \ldots, \mathfrak{x}_k) \Leftrightarrow F(\mathfrak{x}_1, \ldots, \mathfrak{x}_k, Y_1, \ldots, Y_m),$$

  where F is a formula with free logical variables $\mathfrak{x}_1, \ldots, \mathfrak{x}_k$, and possibly with other logical variables that are bound, and where each Y_i is either a relation or an individual; or

- a probabilistic assessment

  $$\mathbb{P}(X(\mathfrak{x}_1, \ldots, \mathfrak{x}_k) = 1) \in [\alpha, \beta],$$

  where α and β are nonnegative rationals in $[0, 1]$.

We assume that our languages consist of subsets of function-free first-order logic (referred to as FFFO). Hence we allow existential and universal quantifiers in our syntax.

Concerning the semantics, as often happens when one moves from sharp to interval probabilities, there is more than one way to interpret assessments. In our setting, there are two sensible semantics for well-formed specifications, as we now discuss.

We assume that we have a set $\mathcal{D}$, the *domain*. In this paper every domain is finite, with N elements. Every individual refers to an element of the domain. We will always adopt the *rigidity* assumption that is common in probabilistic logic [3]; that is, we will always assume that the interpretation of individuals is constant across interpretations for a fixed domain. That is, the individual Ann is always mapped to the same element of $\mathcal{D}$, whatever the interpretation of relations. Hence our individuals can be identified with elements of the domain, and given labels such as $1, 2, \ldots, N$.

For instance, suppose we take assessments (1)–(5), and a domain with two individuals, say Ann and Bob, respectively denoted by a and b. We have several ground atoms: $X_1(a)$, $X_1(b)$, $X_2(a)$, $X_2(b)$, $X_3(a, a)$, $X_3(a, b)$, and so on. Consider a graph where each ground atom is a node, and where an edge is inserted between two nodes if an edge was present between the relations. In our example, we obtain the graph in Figure 3. Note that grounding produced two disjoint graphs in this case. However, suppose we keep assessments (1)–(4), but we turn X_5 into a unary relation such that:

$$X_5(\mathfrak{x}) \Leftrightarrow \forall y : X_3(\mathfrak{x}, y) \wedge X_4(y). \quad (6)$$

Then grounding takes us to Figure 4.

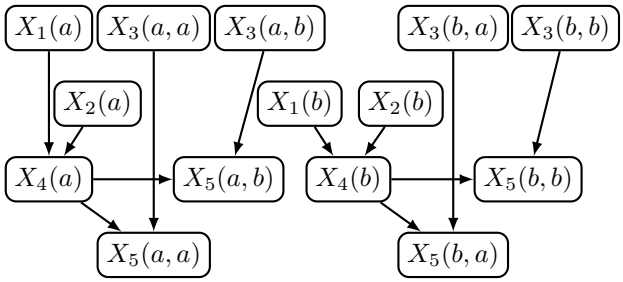


Figure 3: Grounding assessments (1)–(5) with respect to domain $\mathcal{D} = \{a, b\}$.

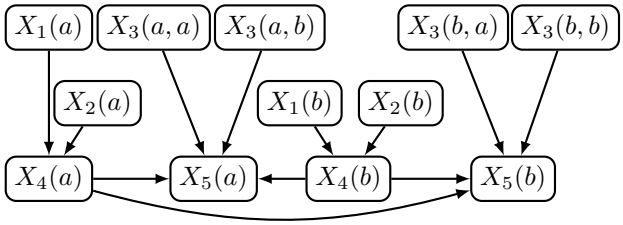


Figure 4: Grounding assessments (1)–(4) and (6) with respect to domain $\mathcal{D} = \{a, b\}$.

So far, our procedure to produce a single grounded graph out of the logical assessments and a fixed domain seems uncontroversial. Now consider the probability assessments; for instance, take

$$\mathbb{P}(X_1(\mathfrak{x})) \in [1/2, 1].$$

What does it mean? Does it mean that

- for each $\gamma \in [1/2, 1]$,

  $$\forall \mathfrak{x} \in \mathcal{D} : \mathbb{P}(X_1(\mathfrak{x})) = \gamma$$

  is a possible assessment, or that
- for each $\mathfrak{x} \in \mathcal{D}$,

  $$\mathbb{P}(X_1(\mathfrak{x})) = \gamma$$

  is a possible assessment for each $\gamma \in [1/2, 1]$?

The difference between these two interpretations is substantial, even though both share the same grounded graph (for given N). In the first interpretation the assessments are viewed as a set of relational Bayesian networks. That is, each selection of probability values defines a relational Bayesian network, that itself can be grounded into a Bayesian network given a domain. For assessments (1)–(5), we have 4 extreme Bayesian networks that are generated given $\mathcal{D} = \{a, b\}$; we have for instance an extreme Bayesian network where $\mathbb{P}(X_1(a)) = \mathbb{P}(X_1(b)) = 1/2$, and also we have a Bayesian network where $\mathbb{P}(X_1(a)) = \mathbb{P}(X_1(b)) = 1$ (but we do not have $\mathbb{P}(X_1(a)) = 1/2$ and $\mathbb{P}(X_1(b)) = 1$). In the second interpretation the assessments directly yield a credal network with separately specified local credal sets. In our example, the latter semantics yields grounded assessments

$$\mathbb{P}(X_1(a) = 1) \geq 1/2, \mathbb{P}(X_2(b) = 1) \geq 1/2,$$

$$\mathbb{P}(X_2(a) = 1) \in [1/4, 1/3], \ldots, \mathbb{P}(X_3(b,b) = 1) = 1/5,$$

and there are 16 extreme Bayesian networks given $\mathcal{D} = \{a, b\}$.

We will refer to a set of well-formed assessments as a *relational credal network*. When the first semantics is adopted, we say that the relational credal network has *coupled parameters*; when the latter semantics is used, we say the relational network has *decoupled parameters*. To simplify the language, we often refer to *coupled relational credal networks* and *decoupled relational credal networks*.

5 The Complexity of Relational Languages

We can now consider inference problems for selected relational languages $\mathcal{L}$. The input to our inference problems is a relational credal network, evidence, and the size of the domain, denoted by N. We assume that the arity of all relations is bounded.

Domain size N can be given either in binary or unary encoding. In computational terms, binary encoding for N implies that almost every calculation requires exponential effort (as there may be exponentially long numbers in the output) [9]. For this reason, it makes sense to assume that N is specified in unary notation. So, we denote by $\mathsf{INF_d}(\mathcal{L})$ and by $\mathsf{INF_d}^+(\mathcal{L})$ respectively the decision problems for language $\mathcal{L}$, as before, for unary N, where the query Q is an assignment to a grounded atom, and evidence $\mathbf{E}$ is a set of assignments for grounded atoms (evidence is understood as the conjunction of those assignments). Recall that all relations have bounded arity (and the bound is known). Note that for relatively simple languages we already have $\mathsf{NP}^{\mathsf{PP}}$-complete inference, from the results for propositional languages (Section 3).

Consider then function-free first-order logic (we refer to it by FFFO). The following result is not surprising:

Theorem 3 $\mathsf{INF_d}^+(\mathsf{FFFO})$ *is* $\mathsf{NP}^{\mathsf{PP}}$*-complete both for decoupled and for coupled relational credal networks.*

Proof. For pertinence, ground the relational credal network into a credal network specified using $\mathsf{Prop}(\wedge, \neg)$. Inference in the grounded credal network is a $\mathsf{NP}^{\mathsf{PP}}$-complete problem. For hardness, note that a domain

with a single individual can already define an arbitrarily complex credal network. □

To obtain more insightful results concerning complexity, we have previously proposed an analysis with respect to *data complexity* and to *domain complexity* [9]. We have started such an analysis for relational Bayesian networks, and we now present results for relational credal networks.

We refer to the complexity of computing a conditional probability, given a relational credal network, evidence (a set of assignments), and an integer N (the size of the domain in unary notation), as the *combined complexity*. Theorem 3 deals with combined complexity. We refer to the complexity of computing a conditional probability, for a fixed relational network, when evidence and N are inputs, as the *data complexity*. And we refer to the complexity of computing a conditional probability, for a fixed relational network and fixed evidence, when N is the input, as the *domain complexity*.[2]

When we focus on relational Bayesian networks, the data complexity of FFFO is PP-complete [9]. So the combined and data complexities are identical for relational Bayesian networks as far as first-order logic is concerned. For relational credal networks the data complexity depends on the semantics, as we now show: as often happens when one moves from sharp to indeterminate probabilities, concepts that collapse in the former case do not collapse in the latter case, and we must deal with more nuanced scenarios.

We use $\mathsf{DINF_d}(\mathcal{L})$ to indicate the data complexity of relational credal networks specified through language $\mathcal{L}$. We can state our main results:

Theorem 4 $\mathsf{DINF_d}(\mathsf{FFFO})$ *is* $\mathsf{NP}^{\mathsf{PP}}$*-complete for decoupled relational credal networks.*

Proof. For decoupled relational credal networks, pertinence to $\mathsf{NP}^{\mathsf{PP}}$ is easy (even the combined complexity is in $\mathsf{NP}^{\mathsf{PP}}$ by Theorem 3). To prove hardness, we adapt the proof for a similar result for Bayesian networks [9]. Take an E-MAJSAT problem with pair (ϕ, k), where ϕ is in CNF with m clauses, each one of them with three literals (for each clause, we refer to the "left" literal, the "middle" literal, and the "right" literal). Suppose propositions are $A_1, \ldots, A_n$. If the number of clauses m is smaller than n, then add trivial clauses such as $A_1 \vee A_1 \vee \neg A_1$ until $m = n$. These clauses do not change the output of MAJSAT. If instead $n < m$, then add fresh propositions $A_{n+1}, \ldots, A_m$. These propositions do not change the output of MAJSAT. Introduce unary relations $\mathsf{sat}(\mathfrak{x})$ and $\mathsf{choice}(\mathfrak{x})$; impose $\mathbb{P}(\mathsf{sat}(\mathfrak{x})) = 1/2$, $\mathbb{P}(\mathsf{choice}(\mathfrak{x})) \in [0,1]$. The idea is that $\mathsf{sat}(\mathfrak{x})$ refers to proposition $A_{\mathfrak{x}}$ for $\mathfrak{x} \in \{k+1, \ldots, n\}$, while $\mathsf{choice}(\mathfrak{x})$ refers to proposition $A_{\mathfrak{x}}$ for $\mathfrak{x} \in \{1, \ldots, k\}$. Introduce binary relations $\mathsf{aux}^{\mathsf{sat}}_{ij}(\mathfrak{x}, y)$ and $\mathsf{aux}^{\mathsf{choice}}_{ij}(\mathfrak{x}, y)$, where i can be either left, middle, and right, while j can be either $+$ or $-$. Adopt $\mathbb{P}\Big(\mathsf{aux}^{\mathsf{sat}}_{ij}(\mathfrak{x}, y)\Big) = \mathbb{P}\Big(\mathsf{aux}^{\mathsf{choice}}_{ij}(\mathfrak{x}, y)\Big) = \alpha$ for some $\alpha \in (0,1)$; the specific value of α will not matter. To be concrete, adopt $\alpha = 1/2$. Also, introduce auxiliary relations $\mathsf{literal}_i(\mathfrak{x})$ where i can be left, middle, right. Impose

$$\begin{aligned}\mathsf{literal}_i(\mathfrak{x}) \Leftrightarrow\ & (\exists y : \mathsf{aux}^{\mathsf{sat}}_{i+}(\mathfrak{x}, y) \wedge \mathsf{sat}(y)) \\ \vee\ & (\exists y : \mathsf{aux}^{\mathsf{sat}}_{i-}(\mathfrak{x}, y) \wedge \neg\mathsf{sat}(y)) \\ \vee\ & (\exists y : \mathsf{aux}^{\mathsf{choice}}_{i+}(\mathfrak{x}, y) \wedge \mathsf{choice}(y)) \\ \vee\ & (\exists y : \mathsf{aux}^{\mathsf{choice}}_{i-}(\mathfrak{x}, y) \wedge \neg\mathsf{choice}(y)).\end{aligned}$$

Introduce unary relation $\mathsf{clause}(\mathfrak{x})$ and impose

$$\mathsf{clause}(\mathfrak{x}) \Leftrightarrow \mathsf{literal}_{\mathsf{left}}(\mathfrak{x}) \vee \mathsf{literal}_{\mathsf{middle}}(\mathfrak{x}) \vee \mathsf{literal}_{\mathsf{right}}(\mathfrak{x}).$$

Finally, introduce query and

$$\mathsf{query} \Leftrightarrow \forall \mathfrak{x} : \mathsf{clause}(\mathfrak{x}).$$

Take $N = n$; given our previous discussion we have $n = m$. Individuals are referred as $\{1, \ldots, N\}$ and have a dual purpose, indexing both propositions and clauses.

Take evidence $\mathbf{E}$ as follows. For the ith clause, suppose the left literal is A_j. If $j > k$, set $\mathsf{aux}^{\mathsf{sat}}_{\mathsf{left}+}(i, j)$ to true, and all other $\mathsf{aux}^{\mathsf{sat}}_{\mathsf{left}+}(i, y)$ to false; also set all $\mathsf{aux}^{\mathsf{sat}}_{\mathsf{left}-}(i, y)$ to false, all $\mathsf{aux}^{\mathsf{choice}}_{\mathsf{left}+}(i, y)$ to false, and all $\mathsf{aux}^{\mathsf{choice}}_{\mathsf{left}-}(i, y)$ to false. If instead $i \leq k$, set $\mathsf{aux}^{\mathsf{choice}}_{\mathsf{left}+}(i, j)$ to true, and all other $\mathsf{aux}^{\mathsf{choice}}_{\mathsf{left}+}(i, y)$ to false; also set all $\mathsf{aux}^{\mathsf{choice}}_{\mathsf{left}-}(i, y)$ to false, all $\mathsf{aux}^{\mathsf{sat}}_{\mathsf{left}+}(i, y)$ to false, and all $\mathsf{aux}^{\mathsf{sat}}_{\mathsf{left}-}(i, y)$ to false.

Suppose instead that for the ith clause the left literal is $\neg A_j$; follow the previous paragraph, but exchange $+$ and $-$. Repeat similarly for middle and right literals, but using middle and right as appropriate. Finally, decide whether $\overline{\mathbb{P}}(\mathsf{query}(1) = 0) > 1/2$. If YES, the E-MAJSAT problem is accepted, if NO, it is not accepted. Hence we have the desired reduction. □

Theorem 5 $\mathsf{DINF_d}(\mathsf{FFFO})$ *is* PP*-complete for coupled relational credal networks.*

Proof. For coupled relational credal networks, PP-hardness is obtained by encoding any E-MAJSAT problem with $k = 0$ in the previous proof, and noting that MAJSAT is a PP-complete problem [27]. To prove pertinence to PP, we will use the fact that PP is

[2] Data and domain complexity are respectively related to the existing notions of lqe-liftability and liftability [17, 28]; lqe-liftability means that data complexity is polynomial, and liftability means that domain complexity is polynomial.

closed under union, a celebrated result in complexity theory [4]. Take a fixed relational credal network and note that there is a fixed number (possibly large) of relational Bayesian networks that can be generated by selecting each one of the possible endpoints of probability intervals. Each one of these M relational Bayesian networks specifies a set of strings, consisting of those strings containing N and associated evidence, such that the inference problem yields YES if a string is accepted. That is, we have M sets of accepted strings, and our problem is: given a string with N and evidence, accept it if any one of those M sets of strings contains it. But note that each set of strings defines a PP decision problem (the problem of accepting the strings), as each relational Bayesian network can be grounded into a polynomially larger Bayesian network, and inference can be conducted in the latter network. So our main problem is to consider a set of strings that is the union of the M set of strings; because PP is closed under union, the main problem is in PP as well. □

As noted previously, PP is the class of problems that can be solved by "majority" Turing machines with a polynomial bound; they are usually related to counting problems [20]. And $\mathsf{NP}^{\mathsf{PP}}$ is the class of problems that can be solved by a nondeterministic Turing machine with a polynomial bound, with the "help" of an oracle that returns the solution PP given problems. Intuitively, we should expect the latter problems to be significantly more taxing than the former (but current literature does not seem to have a result on whether they are different or not).

6 Conclusion

We have explored the balance of expressivity and complexity in Boolean credal networks. We have recently proposed a framework for such an analysis, geared to Bayesian networks [9]; this paper is a first step in extending the framework to credal networks.

We have discussed both propositional and relational languages, and for relational languages we have studied combined and data complexities. Theorem 1 reveals a class of credal networks that admits polynomial inference, a property shared by few other classes [10]; the result is surprising in that it reproduces the polynomial character of Bayesian networks under the same language. And in the opposite direction, Theorems 4 and 5 show distinctions between Bayesian and credal networks, as in the latter there is more than one reasonable semantics to choose from, and the choice does have an impact on complexity. Surprisingly, for coupled relational credal networks the data complexity is identical to the data complexity of relational Bayesian networks.

Perhaps the most compelling aspect of our framework is the number of questions it raises. Consider a simple fact. It is usually assumed that one can arbitrarily choose between computing an upper or a lower probability, as they are directly related by $\overline{\mathbb{P}}(A|B) = 1 - \underline{\mathbb{P}}(A^c|B)$ [29]. But if a language does not have negation, it may not be possible to formulate $\underline{\mathbb{P}}(A^c|B)$ as a query, and it may then be harder to produce a lower probability than an upper probability. This sort of phenomena can only be explored when we pay attention to languages. In fact, the key difference between Bayesian and credal networks is the language that is used to express assessments.

There are many languages to explore concerning the complexity of credal networks. There are several fragments of function-free first-order logic that are widely used, such as monadic logic [5]; there are guarded fragments and description logics such as DL-Lite, $\mathcal{EL}$, $\mathcal{ALC}$ [2]; there are also languages based on second-order logic and various modal logics. For all these logical languages, one can ask combined and data complexity, not only for inference, but also for other problems of common interest such as maximum a posteriori configurations (MAP). All such problems await detailed investigation.

Acknowledgements

The first author is partially supported by CNPq (grant 305395/2010-6) and the second author is supported by FAPESP (grant 2013/23197-4).

References

[1] Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to Imprecise Probabilities*. Wiley, 2014.

[2] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider. *Description Logic Handbook*. Cambridge University Press, 2002.

[3] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach*. MIT Press, Cambridge, 1990.

[4] Richard Beigel, Nick Reingold, and Daniel Spielman. PP is closed under intersection. *Journal of Computer and System Sciences*, 50(2):191–202, 1995.

[5] Egon Börger, Erich Grädel, and Yuri Gurevich. *The Classical Decision Problem*, Springer, 1997.

[6] Fabio G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

[7] Fabio G. Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186(2):577–600, 2012.

[8] Fabio G. Cozman. Independence for full conditional probabilities: Structure, factorization, non-uniqueness, and Bayesian networks. *International Journal of Approximate Reasoning*, 54:1261–1278, 2013.

[9] Fabio G. Cozman and Denis Deratani Mauá. Bayesian networks specified using propositional and relational constructs: Combined, data, and domain complexity. In *AAAI Conference on Artificial Intelligence*, pages 3519–3525, 2015.

[10] Cassio Polpo de Campos and Fabio G. Cozman. The inferential complexity of Bayesian and credal networks. In *International Joint Conference on Artificial Intelligence*, pages 1313–1318, Edinburgh, United Kingdom, 2005.

[11] Gert de Cooman and Enrique Miranda. Independent natural extension for sets of desirable gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012.

[12] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. *Artificial Intelligence*, 175:1911–1950, 2011.

[13] E. Fagiuoli and Marco Zaffalon. 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[14] Lance Fortnow. Counting complexity. In L. Hemaspaandra and A. Selman, editors, *Complexity Theory Retrospective II*, pages 81–107, Springer, 1997.

[15] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[16] Manfred Jaeger. Relational Bayesian networks. *Conference on Uncertainty in Artificial Intelligence*, pages 266–273, 1997.

[17] Manfred Jaeger and Guy Van Den Broeck. Liftability of probabilistic inference: Upper and lower bounds. In *2nd Statistical Relational AI (StaRAI-12) Workshop*, 2012.

[18] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[19] Michael L. Littman, Judy Goldsmith, and Martin Mundhenk. The computational complexity of probabilistic planning. *Journal of Artificial Intelligence Research*, 9:1–36, 1998.

[20] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley Publishing, 1994.

[21] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, United Kingdom, 2000.

[22] David Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64:81–129, 1993.

[23] David Poole. The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, 94(1/2):7–56, 1997.

[24] David Poole. First-order probabilistic inference. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 985–991, 2003.

[25] Luc De Raedt. *Logical and Relational Learning*. Springer, 2008.

[26] Taisuke Sato and Yoshitaka Kameya. Parameter learning of logic programs for symbolic-statistical modeling. *Journal of Artificial Intelligence Research*, 15:391–454, 2001.

[27] Leslie G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing* 8(3):410–421, 1979.

[28] Guy van den Broeck. On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *Neural Processing Information Systems*, 2011.

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

[30] Kurt Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. International Journal of Approximate Reasoning 24(2-3):149–170, 2000.

A Pointwise Ergodic Theorem for Imprecise Markov Chains

Gert de Cooman and **Jasper De Bock** and **Stavros Lopatatzidis**
Ghent University, SYSTeMS Research Group
{gert.decooman,jasper.debock,stavros.lopatatzidis}@UGent.be

Abstract

We prove a game-theoretic version of the strong law of large numbers for submartingale differences, and use this to derive a pointwise ergodic theorem for discrete-time Markov chains with finite state sets, when the transition probabilities are imprecise, in the sense that they are only known to belong to some convex closed set of probability measures.

Keywords. Imprecise probabilities, lower expectation, pointwise ergodic theorem, imprecise Markov chain, game-theoretic probability

1 Introduction

In Ref. [2], de Cooman and Hermans made a first attempt at laying the foundations for a theory of discrete-event (and discrete-time) stochastic processes that are governed by sets of, rather than single, probability measures. They showed how this could be done by connecting Walley's [1991] theory of coherent lower previsions with ideas and results from Shafer and Vovk's [2001] game-theoretic approach to probability theory. In later papers, de Cooman et al. [5] applied these ideas to finite-state discrete-time Markov chains, inspired by the work of Hartfiel [6]. They showed how to do efficient inferences in, and proved a Perron–Frobenius-like theorem for, so-called imprecise Markov chains, which are finite-state discrete-time Markov chains whose transition probabilities are imprecise, in the sense that they are only known to belong to a convex closed set of probability measures—typically due to partial assessments involving probabilistic inequalities. This work was later refined and extended by Hermans and de Cooman [7] and Škulj and Hable [15].

The Perron–Frobenius-like theorems in these papers give equivalent necessary and sufficient conditions for the uncertainty model—a set of probabilities—about the state X_n to converge, for $n \to +\infty$, to an uncertainty model that is independent of the uncertainty model for the initial state X_1.

In Markov chains with 'precise' transition probabilities, this convergence behaviour is sufficient for a pointwise ergodic theorem to hold, namely that:

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = E_\infty(f) \text{ almost surely}$$

for all real functions f on the finite state set $\mathscr{X}$, where E_∞ is the limit expectation operator that the expectation operators E_n for the state X_n at time n converge to pointwise, independently of the initial model E_1 for X_1, according to the classical Perron–Frobenius Theorem.[1]

The aim of the present paper is to extend this result to a version for imprecise Markov chains; see Theorem 11.

How do we mean to go about this? In Section 2, we explain what we mean by imprecise probability models: we extend the notion of an expectation operator to so-called lower (and upper) expectation operators, and explain how these can be associated with (convex and closed) sets of expectation operators.

In Section 3, we explain how these generalised uncertainty models can be combined with event trees to form so-called imprecise probability trees, to produce a simple theory of discrete-time stochastic processes. We show in particular how to combine local uncertainty models associated with the nodes in the tree into global uncertainty models (global conditional lower expectations) about the paths in the tree, and how this procedure is related to sub- and supermartingales. We also indicate how it extends and subsumes the (precise-)probabilistic approach.

In Section 4 we prove a very general strong law of large numbers for submartingale differences in our imprecise probability trees. Our pointwise ergodic theorem will turn out to be a consequence of this in the particular context of imprecise Markov chains. We briefly explain what imprecise Markov chains are in Section 5: how they are special

[1] Actually, much more general results can be proved, for functions f that do not depend on a single state only, but on the entire sequence of states; see for instance Ref. [8, Chapter 20]. In this paper, we will focus on the simpler version.

cases of imprecise probability trees, how to do efficient inference for them, and how to define Perron–Frobenius-like behaviour. We also explore the influence of time shifts on the global (conditional) lower expectations, and discuss stationarity and its relation with Perron–Frobenius-like behaviour.

In Section 6 we show that there is an interesting identity between the time averages that appear in our strong law of large numbers, and the ones that appear in the pointwise ergodic theorem. The discussion in Section 7 first focusses on a number of terms in this identity, and investigates the convergence of these terms for Perron–Frobenius-like imprecise Markov chains. This allows us to use the identity to prove our version of the pointwise ergodic theorem, whose significance we discuss briefly in Section 8.

2 Basic Notions from Imprecise Probabilities

Let us begin with a brief sketch of a few basic definitions and results about imprecise probabilities. For more details, we refer to Walley's [16] seminal book, as well as more recent textbooks [1, 13].

Suppose a subject is uncertain about the value that a variable Y assumes in a non-empty set of possible values $\mathscr{Y}$. He is therefore also uncertain about the value $f(Y)$ a so-called *gamble*—a bounded real-valued function—$f\colon \mathscr{Y} \to \mathbb{R}$ on the set $\mathscr{Y}$ assumes in $\mathbb{R}$. We will also call such an f a gamble *on* Y when we want to make explicit what variable Y the gamble f is intended to depend on. The subject's uncertainty is modelled by a *lower expectation*[2] $\underline{E}$, which is a real functional defined on the set $\mathscr{G}(\mathscr{Y})$ of all gambles on the set $\mathscr{Y}$, satisfying the following basic so-called *coherence axioms*:

LE1. $\underline{E}(f) \geq \inf f$ for all $f \in \mathscr{G}(\mathscr{Y})$;

LE2. $\underline{E}(f+g) \geq \underline{E}(f)+\underline{E}(g)$ for all $f,g \in \mathscr{G}(\mathscr{Y})$;

LE3. $\underline{E}(\lambda f) = \lambda \underline{E}(f)$ for all $f \in \mathscr{G}(\mathscr{Y})$ and real $\lambda \geq 0$.

One—but by no means the only[3]—way to interpret $\underline{E}(f)$ is as a lower bound on the expectation $E(f)$ of the gamble $f(Y)$. The corresponding upper bounds are given by the *conjugate upper expectation* $\overline{E}$, defined by $\overline{E}(f) := -\underline{E}(-f)$ for all $f \in \mathscr{G}(\mathscr{Y})$. It follows from the coherence axioms LE1–LE3 that

LE4. $\inf f \leq \underline{E}(f) \leq \overline{E}(f) \leq \sup f$ for all $f \in \mathscr{G}(\mathscr{Y})$;

LE5. $\underline{E}(f) \leq \underline{E}(g)$ and $\overline{E}(f) \leq \overline{E}(g)$ for all $f,g \in \mathscr{G}(\mathscr{Y})$ with $f \leq g$;

LE6. $\underline{E}(f+\mu) = \underline{E}(f)+\mu$ and $\overline{E}(f+\mu) = \overline{E}(f)+\mu$ for all $f \in \mathscr{G}(\mathscr{Y})$ and real μ.

Lower and upper expectations will be the basic uncertainty models we consider in this paper.

The *indicator* $\mathbb{I}_A$ of an *event* A—a subset of $\mathscr{Y}$—is the gamble on Y that assumes the value 1 on A and 0 outside A. It allows us to introduce the *lower* and *upper probabilities* of A as $\underline{P}(A) := \underline{E}(\mathbb{I}_A)$ and $\overline{P}(A) := \overline{E}(\mathbb{I}_A)$, respectively. They can be seen as lower and upper bounds on the probability $P(A)$ of A, and satisfy the conjugacy relation $\overline{P}(A) = 1-\underline{P}(\mathscr{Y} \setminus A)$.

When the lower bound $\underline{E}$ coincides with the upper bound $\overline{E}$, the resulting functional $E := \underline{E} = \overline{E}$ satisfies the defining axioms of an *expectation*:

E1. $E(f) \geq \inf f$ for all $f \in \mathscr{G}(\mathscr{Y})$;

E2. $E(f+g) = E(f)+E(g)$ for all $f,g \in \mathscr{G}(\mathscr{Y})$;

E3. $E(\lambda f) = \lambda E(f)$ for all $f \in \mathscr{G}(\mathscr{Y})$ and real λ.

When $\mathscr{Y}$ is finite, E is trivially the expectation associated with a (probability) mass function p defined by $p(y) := \underline{P}(\{y\}) = \overline{P}(\{y\})$ for all $y \in \mathscr{Y}$, because it follows from the expectation axioms that then $E(f) = \sum_{y\in\mathscr{Y}} f(y)p(y)$; see for instance also the detailed discussion in Ref. [13].

With any lower expectation $\underline{E}$, we can always associate the following convex and closed[4] set of *compatible* expectation operators:

$$\mathfrak{M}(\underline{E}) := \{E\colon (\forall f \in \mathscr{G}(\mathscr{Y}))\underline{E}(f) \leq E(f) \leq \overline{E}(f)\}, \quad (1)$$

and the properties LE1–LE3 then guarantee that

$$\left.\begin{array}{l}\underline{E}(f) = \min\{E(f)\colon E \in \mathfrak{M}(\underline{E})\}\\ \overline{E}(f) = \max\{E(f)\colon E \in \mathfrak{M}(\underline{E})\}\end{array}\right\} \text{for all } f \in \mathscr{G}(\mathscr{Y}). \quad (2)$$

In this sense, an imprecise probability model $\underline{E}$ can always be identified with a closed convex set $\mathfrak{M}(\underline{E})$ of compatible 'precise' probability models E.

3 Discrete-Time Finite-State Imprecise Stochastic Processes

We consider a discrete-time process as a sequence of variables, henceforth called *states*, $X_1, X_2, \ldots, X_n, \ldots$, where each state X_k is assumed to take values in a non-empty *finite* set $\mathscr{X}_k$.

[2] In the literature [16, 1, 13], other names, such as coherent lower expectation, or coherent lower prevision, have also been given to this concept.

[3] See Refs. [16, 10, 13] for other interpretations.

[4] The 'closedness' is associated with the weak* topology of pointwise convergence [16, Section 3.6].

3.1 Event Trees, Situations, Paths and Cuts

We will use, for any natural $k \leq \ell$, the notation $X_{k:\ell}$ for the tuple $(X_k, \ldots, X_\ell)$, which can be seen as a variable assumed to take values in the product set $\mathscr{X}_{k:\ell} := \times_{r=k}^{\ell} \mathscr{X}_r$. We denote the set of all natural numbers (without 0) by $\mathbb{N}$, and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

We call any $x_{1:n} \in \mathscr{X}_{1:n}$ for $n \in \mathbb{N}_0$ a *situation* and we denote the set of all situations by $\Omega^\Diamond$. So any situation is a finite string of possible values for the consecutive states, and if we denote the empty string by $\Box$, then in particular, $\mathscr{X}_{1:0} = \{\Box\}$. $\Box$ is called the *initial situation*. We also use the generic notations s, t or u for situations.

An infinite sequence of state values is called a *path*, and we denote the set of all paths—also called the *sample space*—by Ω. Hence

$$\Omega^\Diamond := \bigcup_{n \in \mathbb{N}_0} \mathscr{X}_{1:n} \text{ and } \Omega := \times_{r=1}^{\infty} \mathscr{X}_k.$$

We will denote generic paths by ω. For any path $\omega \in \Omega$, the initial sequence that consists of its first n elements is a situation in $\mathscr{X}_{1:n}$ that is denoted by ω^n. Its n-th element belongs to $\mathscr{X}_n$ and is denoted by ω_n. As a convention, we let its 0-th element be the initial situation $\omega^0 = \omega_0 = \Box$. The possible realisations ω of a process can be represented graphically as paths in a so-called *event tree*, where each node is a situation; see Figure 1.

We write that $s \sqsubseteq t$, and say that s *precedes* t or that t *follows* s, when every path that goes through t also goes through s. The binary relation $\sqsubseteq$ is a partial order, and we write $s \sqsubset t$ whenever $s \sqsubseteq t$ but not $s = t$. We say that s and t are *incomparable* when neither $s \sqsubseteq t$ nor $t \sqsubseteq s$.

A (partial) *cut* U is a collection of mutually incomparable situations, and represents a stopping time. For any two cuts U and V, we define the following sets of situations:

$$\begin{aligned}
[U,V] &:= \{s \in \Omega^\Diamond \colon (\exists u \in U)(\exists v \in V) u \sqsubseteq s \sqsubseteq v\} \\
[U,V) &:= \{s \in \Omega^\Diamond \colon (\exists u \in U)(\exists v \in V) u \sqsubseteq s \sqsubset v\} \\
(U,V] &:= \{s \in \Omega^\Diamond \colon (\exists u \in U)(\exists v \in V) u \sqsubset s \sqsubseteq v\} \\
(U,V) &:= \{s \in \Omega^\Diamond \colon (\exists u \in U)(\exists v \in V) u \sqsubset s \sqsubset v\}.
\end{aligned}$$

When a cut U consists of a single element u, then we will identify $U = \{u\}$ and u. This slight abuse of notation will for instance allow us to write $[u,v] = \{s \in \Omega^\Diamond \colon u \sqsubseteq s \sqsubseteq v\}$ and also $(U,v) = \{s \in \Omega^\Diamond \colon (\exists u \in U) u \sqsubset s \sqsubset v\}$. We also write $U \sqsubset V$ if $(\forall v \in V)(\exists u \in U) u \sqsubset v$. Observe that in that case $U \cap V = \emptyset$. In particular, $s \sqsupset U$ when there is some $u \in U$ such that $s \sqsupset u$, or in other words if $[U,s) \neq \emptyset$.

A *process* $\mathscr{F}$ is a map defined on $\Omega^\Diamond$. A *real process* is a real-valued process: it associates a real number $\mathscr{F}(x_{1:n}) \in \mathbb{R}$ with any situation $x_{1:n}$. It is called *bounded below* if there is some real B such that $\mathscr{F}(s) \geq B$ for all situations $s \in \Omega^\Diamond$.

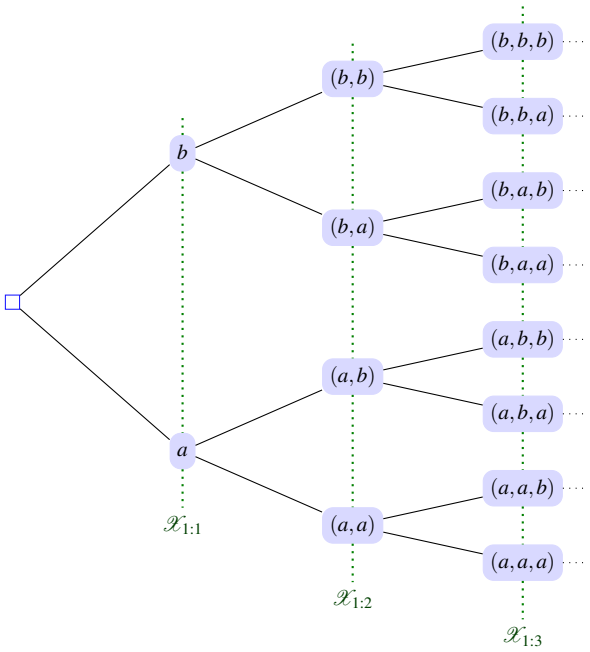


Figure 1: The (initial part of the) event tree for a process whose states can assume two values, a and b, and can change at time instants $n = 1, 2, 3, \ldots$ Each node in the tree corresponds to a situation. Also depicted are the respective sets of situations (cuts) $\mathscr{X}_{1:1}$, $\mathscr{X}_{1:2}$ and $\mathscr{X}_{1:3}$ where the states at times 1, 2 and 3 are revealed.

A *gamble process* $\mathscr{D}$ is a process that associates with any situation $x_{1:n}$ a gamble $\mathscr{D}(x_{1:n}) \in \mathscr{G}(\mathscr{X}_{n+1})$ on X_{n+1}. It is called *uniformly bounded* if there is some real B such that $|\mathscr{D}(s)| \leq B$ for all situations $s \in \Omega^\Diamond$. With any real process $\mathscr{F}$, we can always associate a gamble process $\Delta\mathscr{F}$, called the *process difference*. For every situation $x_{1:n}$, the gamble $\Delta\mathscr{F}(x_{1:n}) \in \mathscr{G}(\mathscr{X}_{n+1})$ is defined by[5]

$$\Delta\mathscr{F}(x_{1:n})(x_{n+1}) := \mathscr{F}(x_{1:n+1}) - \mathscr{F}(x_{1:n}) \quad \text{for all } x_{n+1} \in \mathscr{X}_{n+1}.$$

We will denote this more succinctly by $\Delta\mathscr{F}(x_{1:n}) = \mathscr{F}(x_{1:n}\cdot) - \mathscr{F}(x_{1:n})$, where the '$\cdot$' represents the generic value of the next state X_{n+1}.

Conversely, with a gamble process $\mathscr{D}$, we can associate a real process $\mathscr{I}^{\mathscr{D}}$, defined by

$$\mathscr{I}^{\mathscr{D}}(x_{1:n}) := \sum_{k=0}^{n-1} \mathscr{D}(x_{1:k})(x_{k+1}) \quad \text{for all } n \in \mathbb{N}_0 \text{ and } x_{1:n} \in \mathscr{X}_{1:n}.$$

Clearly, $\Delta\mathscr{I}^{\mathscr{D}} = \mathscr{D}$ and $\mathscr{F} = \mathscr{F}(\Box) + \mathscr{I}^{\Delta\mathscr{F}}$.

[5] Our assumption that $\mathscr{X}_{n+1}$ is finite is crucial here because it guarantees that $\Delta\mathscr{F}(x_{1:n})$ is bounded, which in turn implies that it is indeed a gamble.

Also, with any real process $\mathscr{F}$ we can associate the *path-averaged process* $\langle\mathscr{F}\rangle$, which is the real process defined by:

$$\langle\mathscr{F}\rangle(x_{1:n}) := \begin{cases} 0 & \text{if } n=0 \\ \frac{1}{n}\mathscr{F}(x_{1:n}) & \text{if } n>0 \end{cases}$$
$$\text{for all } n\in\mathbb{N}_0 \text{ and } x_{1:n}\in\mathscr{X}_{1:n}.$$

3.2 Imprecise Probability Trees, Submartingales and Supermartingales

The standard way to turn an event tree into a *probability tree* is to attach to each of its nodes, or situations $x_{1:n}$, a *local* probability model $Q(\cdot|x_{1:n})$ for what will happen immediately afterwards, i.e. for the value that the next state X_{n+1} will assume in $\mathscr{X}_{n+1}$. This local model $Q(\cdot|x_{1:n})$ is then an expectation operator on the set $\mathscr{G}(\mathscr{X}_{n+1})$ of all gambles $g(X_{n+1})$ on the next state X_{n+1}, conditional on observing $X_{1:n}=x_{1:n}$.

In a completely similar way, we can turn an event tree into an *imprecise probability tree* by attaching to each of its situations $x_{1:n}$ a local *imprecise* probability model $\underline{Q}(\cdot|x_{1:n})$ for what will happen immediately afterwards, i.e. for the value that the next state X_{n+1} will assume in $\mathscr{X}_{n+1}$. This local model $\underline{Q}(\cdot|x_{1:n})$ is then a *lower* expectation operator on the set $\mathscr{G}(\mathscr{X}_{n+1})$ of all gambles $g(X_{n+1})$ on the next state X_{n+1}, conditional on observing $X_{1:n}=x_{1:n}$. This is represented graphically in Figure 2.

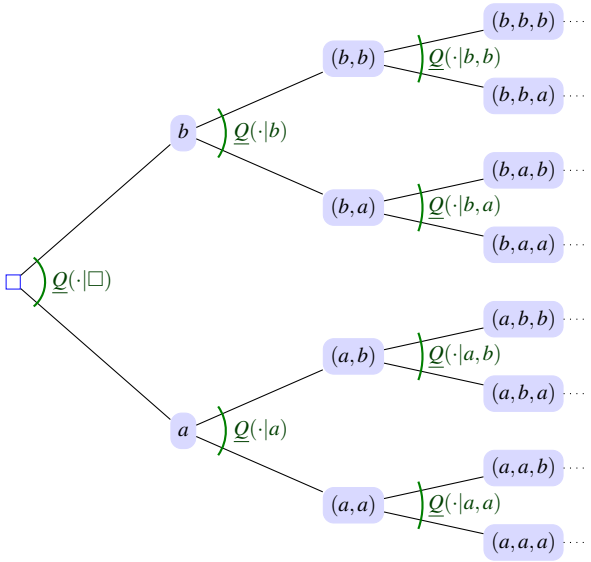


Figure 2: The (initial part of the) imprecise probability tree for a process whose states can assume two values, a and b, and can change at time instants $n=1,2,3,\ldots$

In a given imprecise probability tree, a *submartingale* $\mathscr{M}$ is a real process such that $\underline{Q}(\Delta\mathscr{M}(x_{1:n})|x_{1:n})\geq 0$ for all $n\in\mathbb{N}_0$ and $x_{1:n}\in\mathscr{X}_{1:n}$: all submartingale differences have non-negative lower expectation. A real process $\mathscr{M}$ is a *supermartingale* if $-\mathscr{M}$ is a submartingale, meaning that $\overline{Q}(\Delta\mathscr{M}(x_{1:n})|x_{1:n})\leq 0$ for all $n\in\mathbb{N}_0$ and $x_{1:n}\in\mathscr{X}_{1:n}$: all supermartingale differences have non-positive upper expectation. We denote the set of all submartingales for a given imprecise probability tree by $\underline{\mathbb{M}}$—whether a real process is a submartingale depends of course on the local uncertainty models. Similarly, the set $\overline{\mathbb{M}}:=-\underline{\mathbb{M}}$ is the set of all supermartingales.

In the present context of probability trees, we will also call *variable* any function defined on the so-called *sample space*—the set Ω of all paths. When this variable is real-valued and bounded, we will also call it a *gamble* on Ω. When it is extended real-valued, meaning that it assumes values in the set $\mathbb{R}^*:=\mathbb{R}\cup\{-\infty,+\infty\}$, we call in an *extended real variable*. An *event* A in this context is a subset of Ω, and its indicator $\mathbb{I}_A$ is a gamble on Ω assuming the value 1 on A and 0 elsewhere. With any situation $x_{1:n}$, we can associated the so-called *exact event* $\Gamma(x_{1:n})$ that $X_{1:n}=x_{1:n}$, which is the set of all paths $\omega\in\Omega$ that go through $x_{1:n}$:

$$\Gamma(x_{1:n}) := \{\omega\in\Omega\colon \omega^n = x_{1:n}\}.$$

For a given $n\in\mathbb{N}_0$, we call a variable ξ *n-measurable* if it is constant on the exact events $\Gamma(x_{1:n})$ for all $x_{1:n}\in\mathscr{X}_{1:n}$, or in other words, if it only depends on the values of the first n states $X_{1:n}$. We then use the obvious notation $\xi(x_{1:n})$ for its constant value $\xi(\omega)$ on all paths ω in $\Gamma(x_{1:n})$.

With a real process $\mathscr{F}$, we can associate in particular the following extended real variables $\liminf\mathscr{F}$ and $\limsup\mathscr{F}$, defined for all $\omega\in\Omega$ by $\liminf\mathscr{F}(\omega):=\liminf_{n\to\infty}\mathscr{F}(\omega^n)$ and $\limsup\mathscr{F}(\omega):=\limsup_{n\to\infty}\mathscr{F}(\omega^n)$. If $\liminf\mathscr{F}(\omega)=\limsup\mathscr{F}(\omega)$ on some path ω, then we also denote the common value there by $\lim\mathscr{F}(\omega)=\lim_{n\to\infty}\mathscr{F}(\omega^n)$.

3.3 Going from Local to Global Belief Models

So far, we have associated local uncertainty models with an imprecise probability tree. These represent, in any situation $x_{1:n}$, beliefs about what will happen immediately afterwards, or in other words about the step from $x_{1:n}$ to $x_{1:n}X_{n+1}$.

We now want to turn these local models into global ones: uncertainty models about which entire path ω is taken in the event tree, rather than which local steps are taken from one situation to the next. We use the following expression for the global lower expectation conditional on the situation s:

$$\underline{E}(f|s) := \sup\{\mathscr{M}(s)\colon \mathscr{M}\in\underline{\mathbb{M}}, \limsup\mathscr{M}\leq f \text{ on } \Gamma(s)\}, \tag{3}$$

and for the conjugate global upper expectation conditional on the situation s:

$$\overline{E}(f|s) := \inf\{\mathscr{M}(s)\colon \mathscr{M}\in\overline{\mathbb{M}}, \liminf\mathscr{M}\geq f \text{ on } \Gamma(s)\}, \tag{4}$$

where f is any extented real variable on Ω, and $s \in \Omega^\Diamond$ any situation. We use the simplified notations $\underline{E} = \underline{E}(\cdot|\Box)$ and $\overline{E} = \overline{E}(\cdot|\Box)$ for the (unconditional) global models, associated with the initial situation $\Box$.

Our reasons for using these so-called *Shafer–Vovk–Ville formulae*[6] are fourfold.

First of all, they are formally very closely related to the expressions for lower and upper prices in Shafer and Vovk's game-theoretic approach to probabilities, see for instance Refs. [11, Chapter 8.3], [12, Section 2] and [14, Section 6.3]. This allows us to import and adapt, with the necessary care, quite a number of powerful convergence results from that theory, as we shall see in Section 4. Moreover, Shafer and Vovk (see for instance Refs. [11, Proposition 8.8] and [14, Section 6.3]) have shown that they—or rather their restrictions to gambles—satisfy our defining properties for lower and upper expectations in Section 2, which is why we are calling them lower and upper expectations.

Secondly, we gather from Proposition 1 and Corollary 2 that the expressions (3) and (4) coincide for n-measurable gambles on Ω with the formulae derived in Ref. [2] as the most conservative[7] global lower and upper expectations that extend the local models—see Corollary 3.[8]

Proposition 1. *For any situation $x_{1:m} \in \Omega^\Diamond$ and any n-measurable extended real variable f, with $n,m \in \mathbb{N}_0$ such that $n \geq m$:*

$$\underline{E}(f|x_{1:m}) = \sup\{\mathscr{M}(x_{1:m})\colon \mathscr{M} \in \underline{\mathbb{M}} \text{ and } (\forall x_{m+1:n} \in \mathscr{X}_{m+1:n})\mathscr{M}(x_{1:n}) \leq f(x_{1:n})\}$$
$$\overline{E}(f|x_{1:m}) = \inf\{\mathscr{M}(x_{1:m})\colon \mathscr{M} \in \overline{\mathbb{M}} \text{ and } (\forall x_{m+1:n} \in \mathscr{X}_{m+1:n})\mathscr{M}(x_{1:n}) \geq f(x_{1:n})\}.$$

Corollary 2. *For any situation $x_{1:m} \in \Omega^\Diamond$ and any n-measurable extended real variable f, with $n,m \in \mathbb{N}_0$ such that $n \geq m$:*

$$\underline{E}(f|x_{1:m}) = \sup\{\underline{E}(g|x_{1:m})\colon g \in \mathscr{G}(\mathscr{X}_{1:n}) \text{ and } (\forall x_{m+1:n} \in \mathscr{X}_{m+1:n})g(x_{1:n}) \leq f(x_{1:n})\}$$
$$\overline{E}(f|x_{1:m}) = \inf\{\overline{E}(g|x_{1:m})\colon g \in \mathscr{G}(\mathscr{X}_{1:n}) \text{ and } (\forall x_{m+1:n} \in \mathscr{X}_{m+1:n})g(x_{1:n}) \geq f(x_{1:n})\}.$$

Corollary 3. *Consider $n \in \mathbb{N}_0$ and $x_{1:n} \in \Omega^\Diamond$. Then for any $(n+1)$-measurable gamble g on Ω: $\underline{E}(g|x_{1:n}) = \underline{Q}(g(x_{1:n}\cdot)|x_{1:n})$ and $\overline{E}(g|x_{1:n}) = \overline{Q}(g(x_{1:n}\cdot)|x_{1:n})$. Also, for any $(n+1)$-measurable extended real variable f:*

$$\underline{E}(f|x_{1:n}) = \sup\{\underline{Q}(h|x_{1:n})\colon h \in \mathscr{G}(\mathscr{X}) \text{ and } h \leq f(x_{1:n}\cdot)\}$$
$$\overline{E}(f|x_{1:n}) = \inf\{\overline{Q}(h|x_{1:n})\colon h \in \mathscr{G}(\mathscr{X}) \text{ and } h \geq f(x_{1:n}\cdot)\}.$$

Thirdly, it is (essentially) the expressions in Proposition 1 that we have used in Refs. [5, 7, 15] for our studies of imprecise Markov chains, which we report in Section 5. The main result of the present paper, Theorem 11 in Section 7, will build on the ergodicity results proved in those papers.

Fourthly, it was also shown in Ref. [2] that the expressions in Proposition 1 have an interesting interpretation in terms of (precise) probability trees. Indeed, we can associate with an imprecise probability tree a (usually infinite) collection of (so-called *compatible*) precise probability trees with the same event tree, by associating with each situation s in the event tree some arbitrarily chosen precise local expectation $Q(\cdot|s)$ that belongs to the convex closed set $\mathfrak{M}(\underline{Q}(\cdot|s))$ of expectations that are compatible with the local lower expectation $\underline{Q}(\cdot|s)$. For any n-measurable gamble f on Ω, the global precise expectations in the compatible precise probability trees will then range over a closed interval whose lower and upper bounds are given by the expressions in Proposition 1.

And finally, Shafer and Vovk have shown [11, Chapter 8] that when the local models are precise probability models, these formulae (3) and (4) lead to global models that coincide with the ones found in measure-theoretic probability theory. *This implies that the results we shall prove below, subsume, as special cases, the classical results of measure-theoretic probability theory.*

4 A Strong Law of Large Numbers for Submartingale Differences

We now discuss and prove two powerful convergence results for the processes we have defined in the previous section.

We call an event A *null* if $\overline{P}(A) = \overline{E}(\mathbb{I}_A) = 0$, and *strictly null* if there is some test supermartingale $\mathscr{T}$ that converges to $+\infty$ on A, meaning that:

$$\lim \mathscr{T}(\omega) = +\infty \quad \text{for all } \omega \in A.$$

Here, a *test supermartingale* is a supermartingale with $\mathscr{T}(\Box) = 1$ that is moreover non-negative in the sense that $\mathscr{T}(s) \geq 0$ for all situations $s \in \Omega^\Diamond$. Any strictly null event is null, but null events need not be strictly null [14].

Proposition 4. *Any strictly null event is null, but not vice versa.*[9]

[6] We give this name to these formulae because Glenn Shafer and Vladimir Vovk first suggested them, based on the ideas of Jean Ville; see the discussion of Ville's Theorem in Ref. [11, Appendix 8.5].

[7] By more conservative, we mean associated with a larger set of precise models, so pointwise smaller for lower expectations, and pointwise larger for upper expectations.

[8] We have also shown in recent, still unpublished work that in a more general context—where X_k takes values in a possibly *infinite* set $\mathscr{X}_k$—for arbitrary gambles on Ω they are the most conservative global models that extend the local ones and satisfy additional conglomerability and continuity properties.

[9] For the null and strictly null events to be the same, it is necessary to consider supermartingales that may assume extended real values, as is done in Refs. [14, 12]. We see no need for doing so here.

In this paper, we shall use the 'strict' approach, and prove that events are strictly null—and therefore also null—by actually showing that there is a test supermartingale that converges to $+\infty$ there.

As usual, an inequality or equality between two variables is said to hold *(strictly) almost surely* when the event that it does not hold is (strictly) null. Shafer and Vovk [11, 14] have proved the following interesting result, which we shall have occasion to use a few times further on. It can be seen as a generalisation of Doob's supermartingale convergence theorem [19, Sections 11.5–7] to imprecise probability trees.

Theorem 5 ([14, Section 6.5] Supermartingale convergence theorem)**.** *Let $\mathscr{M}$ be a supermartingale that is bounded below. Then $\mathscr{M}$ converges strictly almost surely to a real variable.*

We now turn to a very general version of the strong law of large numbers. Weak (as well as less general) versions of this law were proven by one of us in Refs. [3, 2]. It is this law that will, in Section 7, be used to derive our version of the pointwise ergodic theorem. Its proof is based on a tried-and-tested method for constructing test supermartingales that goes back to an idea in Ref. [11, Lemma 3.3].

Theorem 6 (Strong law of large numbers for submartingale differences)**.** *Let $\mathscr{M}$ be a submartingale such that $\Delta\mathscr{M}$ is uniformly bounded. Then $\liminf\langle\mathscr{M}\rangle \geq 0$ strictly almost surely.*

5 Imprecise Markov Chains

We are now ready to apply what we have learned in the previous sections to the special case of (time-homogeneous) imprecise Markov chains. These are imprecise probability trees where (i) all states X_k assume values in the same *finite* set $\mathscr{X}_k = \mathscr{X}$, called the *state space*, and (ii) all local uncertainty models satisfy the so-called *Markov condition*:

$$\underline{Q}(\cdot|x_{1:n}) = \underline{Q}(\cdot|x_n) \text{ for all situations } x_{1:n} \in \Omega^{\Diamond}, \qquad (5)$$

meaning that these local models only depend on the last observed state; see Figure 3.

We refer to Refs. [5, 7, 15] for detailed studies of the behaviour of these processes. We restrict ourselves here to a summary of the existing material that is relevant for the present discussion of ergodicity.

From now on, we start using a convenient notational device often encountered in texts on stochastic processes: when we want to indicate which states a process or variable depends on, we indicate them explicitly in the notation. Thus, we use for instance the notation $\mathscr{F}(X_{1:n})$ to indicate the 'uncertain' value of the process $\mathscr{F}$ after the first n time steps, and write $f(X_n)$ for a gamble that only depends on the value of the n-th state.

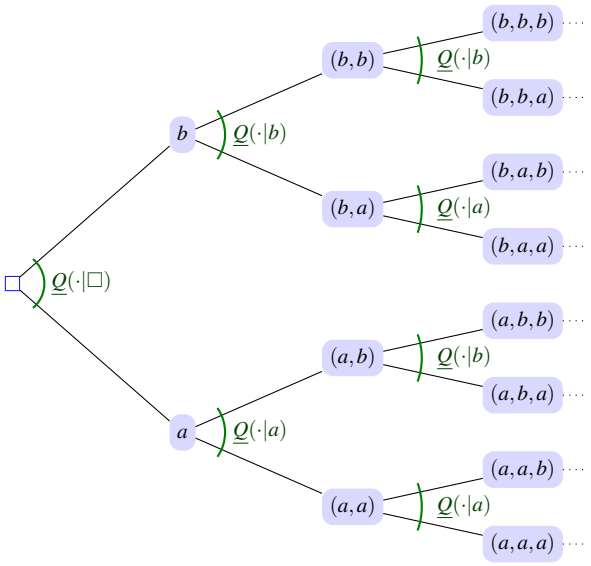


Figure 3: The (initial part of the) imprecise probability tree for an imprecise Markov process whose states can assume two values, a and b, and can change at time instants $n = 1,2,3,\ldots$

We can use the local uncertainty models to introduce a (generally non-linear) transformation $\underline{\mathrm{T}}$ of the set $\mathscr{G}(\mathscr{X})$ of all gambles on the state space $\mathscr{X}$. The so-called *lower transition operator* of the imprecise Markov chain is given by:

$$\underline{\mathrm{T}} : \mathscr{G}(\mathscr{X}) \to \mathscr{G}(\mathscr{X}) : f \mapsto \underline{\mathrm{T}} f,$$

where $\underline{\mathrm{T}} f$ is the gamble on $\mathscr{X}$ defined by

$$\underline{\mathrm{T}} f(x) := \underline{Q}(f|x) \text{ for all } x \in \mathscr{X}.$$

The conjugate *upper transition operator* $\overline{\mathrm{T}}$ is defined by $\overline{\mathrm{T}} f := -\underline{\mathrm{T}}(-f)$ for all $f \in \mathscr{G}(\mathscr{X})$. In particular, $\underline{\mathrm{T}}\mathbb{I}_{\{y\}}(x)$ is the lower probability to go from state value x to state value y in one time step, and $\overline{\mathrm{T}}\mathbb{I}_{\{y\}}(x)$ the conjugate upper probability. This seems to suggest that the lower/upper transition operators $\underline{\mathrm{T}}$ are generalisations of the concept of a Markov transition matrix for ordinary Markov chains. This is confirmed by the following result, proved in Ref. [5, Corollary 3.3] as a special case of the so-called Law of Iterated (Lower) Expectations [2, 11]. If, for any $n \in \mathbb{N}$, we denote by $\underline{E}_n(f)$ the value of the (global) lower expectation $\underline{E}(f(X_n))$ of a gamble $f(X_n)$ on the state X_n at time n, then

$$\underline{E}_n(f) = \underline{E}_1(\underline{\mathrm{T}}^{n-1} f), \text{ with } \underline{\mathrm{T}}^{n-1} f := \underbrace{\underline{\mathrm{T}}\underline{\mathrm{T}}\ldots\underline{\mathrm{T}}}_{n-1 \text{ times}} f,$$

and where, of course, $\underline{E}_1 = \underline{Q}(\cdot|\Box)$ is the marginal local model for the state X_1 at time 1. In a similar vein, for any $n \in \mathbb{N}_0$, $\underline{\mathrm{T}}^n\mathbb{I}_{\{y\}}(x)$ is the lower probability to go from state value x to state value y in n time steps, and $\overline{\mathrm{T}}^n\mathbb{I}_{\{y\}}(x)$ the conjugate upper probability.

We can formally call *lower transition operator* any transformation $\underline{\mathrm{T}}$ of $\mathscr{G}(\mathscr{X})$ such that for any $x \in \mathscr{X}$, the real functional $\underline{\mathrm{T}}_x$ on $\mathscr{G}(\mathscr{X})$, defined by $\underline{\mathrm{T}}_x(f) := \underline{\mathrm{T}} f(x)$ for all $f \in \mathscr{G}(\mathscr{X})$, is a lower expectation—satisfies the coherence axioms LE1–LE3. The composition of any two lower transition operators is again a lower transition operator. See Ref. [5] for more details on the definition and properties of such lower transition operators, and Ref. [4] for a mathematical discussion of the general role of these operators in imprecise probabilities.

We call an imprecise Markov chain with lower transition operator $\underline{\mathrm{T}}$ *Perron–Frobenius-like* if for all $f \in \mathscr{G}(\mathscr{X})$, the sequence of gambles $\underline{\mathrm{T}}^n f$ converges pointwise to a constant real number, which we shall then denote by $\underline{E}_{\mathrm{PF}}(f)$.

The following result was proved in Ref. [5, Theorem 5.1], together with a simple sufficient (and quite weak) condition on $\underline{\mathrm{T}}$ for a Markov chain to be Perron–Frobenius-like: there is some $n \in \mathbb{N}$ such that $\min \overline{\mathrm{T}}^n \mathbb{I}_{\{y\}} > 0$ for all $y \in \mathscr{X}$, or in other words, all state values can be reached from any state value with positive upper probability in (precisely) n time steps. More involved necessary *and* sufficient conditions were given later in Refs. [7, 15]; see also Theorem 8(iv) further on.

Proposition 7 ([5])**.** *The imprecise Markov chain with lower transition operator $\underline{\mathrm{T}}$ is Perron–Frobenius-like if and only if there is some real functional $\underline{E}_\infty$ on $\mathscr{G}(\mathscr{X})$ such that for any initial model $\underline{E}_1$ and any $f \in \mathscr{G}(\mathscr{X})$, it holds that $\underline{E}_n(f) = \underline{E}_1(\underline{\mathrm{T}}^{n-1} f) \to \underline{E}_\infty(f)$. Moreover, in that case the functional $\underline{E}_\infty$ is a lower expectation on $\mathscr{G}(\mathscr{X})$, called the* stationary lower expectation*, it coincides with $\underline{E}_{\mathrm{PF}}$, and it is the only lower expectation that is $\underline{\mathrm{T}}$-invariant in the sense that $\underline{E}_\infty \circ \underline{\mathrm{T}} = \underline{E}_\infty$.*

6 An Interesting Equality in Imprecise Markov Chains

We now prove an interesting equality for imprecise Markov chains, which will be instrumental in proving our pointwise ergodic theorem in the next section.

Consider, for any $f \in \mathscr{G}(\mathscr{X})$, the corresponding *gain* process $\mathscr{W}[f]$, defined by, for any $n \in \mathbb{N}$:

$$\mathscr{W}[f](X_{1:n}) := [f(X_1) - \underline{E}_1(f)] + \sum_{k=2}^{n} [f(X_k) - \underline{\mathrm{T}} f(X_{k-1})], \quad (6)$$

the corresponding *average gain* process $\langle \mathscr{W} \rangle[f]$, defined by:

$$\langle \mathscr{W} \rangle[f](X_{1:n}) := \frac{1}{n}\left[[f(X_1) - \underline{E}_1(f)] + \sum_{k=2}^{n} [f(X_k) - \underline{\mathrm{T}} f(X_{k-1})] \right], \quad (7)$$

and the *ergodic average* process $\mathscr{A}[f]$, defined by:

$$\mathscr{A}[f](X_{1:n}) := \frac{1}{n} \sum_{k=1}^{n} [f(X_k) - \underline{E}_k(f)]. \quad (8)$$

We can let these processes be 0 in the initial situation $\square$—the choice is immaterial. Now observe that, for any $n \in \mathbb{N}$ and any $f \in \mathscr{G}(\mathscr{X})$:

$$\begin{aligned} \sum_{\ell=0}^{n-1} \langle \mathscr{W} \rangle[\underline{\mathrm{T}}^\ell f](X_{1:n}) &= \frac{1}{n} \sum_{\ell=0}^{n-1} \left[\underline{\mathrm{T}}^\ell f(X_1) - \underline{E}_1(\underline{\mathrm{T}}^\ell f) \right] \\ &\quad + \frac{1}{n} \sum_{\ell=0}^{n-1} \sum_{k=2}^{n} \left[\underline{\mathrm{T}}^\ell f(X_k) - \underline{\mathrm{T}}^{\ell+1} f(X_{k-1}) \right], \end{aligned} \quad (9)$$

and moreover

$$\begin{aligned} &\sum_{\ell=0}^{n-1} \sum_{k=2}^{n} \left[\underline{\mathrm{T}}^\ell f(X_k) - \underline{\mathrm{T}}^{\ell+1} f(X_{k-1}) \right] \\ &= \sum_{\ell=0}^{n-1} \sum_{k=2}^{n} \underline{\mathrm{T}}^\ell f(X_k) - \sum_{\ell=0}^{n-1} \sum_{k=2}^{n} \underline{\mathrm{T}}^{\ell+1} f(X_{k-1}) \\ &= \sum_{\ell=0}^{n-1} \sum_{k=2}^{n} \underline{\mathrm{T}}^\ell f(X_k) - \sum_{\ell=1}^{n} \sum_{k=1}^{n-1} \underline{\mathrm{T}}^\ell f(X_k) \\ &= \sum_{k=2}^{n} f(X_k) + \sum_{\ell=1}^{n-1} \left(\underline{\mathrm{T}}^\ell f(X_n) + \sum_{k=2}^{n-1} \underline{\mathrm{T}}^\ell f(X_k) \right) \\ &\qquad - \sum_{k=1}^{n-1} \underline{\mathrm{T}}^n f(X_k) - \sum_{\ell=1}^{n-1} \left(\underline{\mathrm{T}}^\ell f(X_1) + \sum_{k=2}^{n-1} \underline{\mathrm{T}}^\ell f(X_k) \right) \\ &= \sum_{k=2}^{n} f(X_k) + \sum_{\ell=1}^{n-1} \underline{\mathrm{T}}^\ell f(X_n) - \sum_{k=1}^{n-1} \underline{\mathrm{T}}^n f(X_k) - \sum_{\ell=1}^{n-1} \underline{\mathrm{T}}^\ell f(X_1) \\ &= \sum_{k=1}^{n} f(X_k) + \sum_{\ell=1}^{n} \underline{\mathrm{T}}^\ell f(X_n) - \sum_{k=1}^{n} \underline{\mathrm{T}}^n f(X_k) - \sum_{\ell=0}^{n-1} \underline{\mathrm{T}}^\ell f(X_1), \end{aligned}$$

and if we substitute this back into Equation (9), we find, after getting rid of the cancelling terms, recalling that $\underline{E}_1(\underline{\mathrm{T}}^\ell f) = \underline{E}_{\ell+1}(f)$, and reorganising a bit, that:

$$\mathscr{A}[f](X_{1:n}) = \sum_{\ell=0}^{n-1} \langle \mathscr{W} \rangle[\underline{\mathrm{T}}^\ell f](X_{1:n}) + \frac{1}{n} \sum_{k=1}^{n} \underline{\mathrm{T}}^n f(X_k) - \frac{1}{n} \sum_{\ell=1}^{n} \underline{\mathrm{T}}^\ell f(X_n). \quad (10)$$

This is an important relationship between the ergodic average and the average gain. We now intend to show that under certain conditions the remaining terms on the right-hand side essentially cancel out for large enough n.

7 Consequences of the Perron–Frobenius-like Character

Let us associate with a lower transition operator $\underline{\mathrm{T}}$ the following (weak) *coefficient of ergodicity* [15, 7]:

$$\rho(\underline{\mathrm{T}}) := \max_{x,y\in\mathscr{X}} \max_{h\in\mathscr{G}_1(\mathscr{X})} |\underline{\mathrm{T}}h(x) - \underline{\mathrm{T}}h(y)| = \max_{h\in\mathscr{G}_1(\mathscr{X})} \|\underline{\mathrm{T}}h\|_{\mathrm{v}},$$

where $\mathscr{G}_1(\mathscr{X}) := \{h \in \mathscr{G}(\mathscr{X}) \colon 0 \leq h \leq 1\}$, and where for any $h \in \mathscr{G}(\mathscr{X})$, its *variation (semi)norm* is given by $\|h\|_{\mathrm{v}} := \max h - \min h$. If we define the following distance between two lower expectation operators $\underline{E}$ and $\underline{F}$ [15]:

$$d(\underline{E},\underline{F}) = \max_{h\in\mathscr{G}_1(\mathscr{X})} |\underline{E}(h) - \underline{F}(h)|,$$

then it is not difficult to see [using LE3, LE4 and LE6] that $0 \leq d(\underline{E},\underline{F}) \leq 1$, and that for any $f \in \mathscr{G}(\mathscr{X})$:

$$|\underline{E}(f) - \underline{F}(f)| \leq d(\underline{E},\underline{F})\|f\|_{\mathrm{v}}. \tag{11}$$

Škulj and Hable [15] have proved the following results, which will turn out to be crucial to our argument.

Theorem 8 ([15])**.** *Consider lower transition operators $\underline{\mathrm{S}}$ and $\underline{\mathrm{T}}$, and two lower expectations $\underline{E}_a$ and $\underline{E}_b$ on $\mathscr{G}(\mathscr{X})$. Then the following statements hold:*

(i) $0 \leq \rho(\underline{\mathrm{T}}) \leq 1$.

(ii) $\rho(\underline{\mathrm{S}}\underline{\mathrm{T}}) \leq \rho(\underline{\mathrm{S}})\rho(\underline{\mathrm{T}})$ *and therefore* $\rho(\underline{\mathrm{T}}^n) \leq \rho(\underline{\mathrm{T}})^n$ *for all* $n \in \mathbb{N}$.

(iii) $d(\underline{E}_a\underline{\mathrm{T}},\underline{E}_b\underline{\mathrm{T}}) \leq d(\underline{E}_a,\underline{E}_b)\rho(\underline{\mathrm{T}})$.

(iv) *The lower transition operator $\underline{\mathrm{T}}$ is Perron–Frobenius-like if and only if there is some $r \in \mathbb{N}$ such that $\rho(\underline{\mathrm{T}}^r) < 1$.*

Indeed, they allow us to derive useful bounds for the various terms on the right-hand side of Equation (10). For any non-negative real number a we denote by $\lfloor a \rfloor = \max\{n \in \mathbb{N}_0 \colon n \leq a\}$ the largest natural number that it still dominates—its integer part.

Proposition 9. *Let $\underline{\mathrm{T}}$ be a Perron–Frobenius-like lower transition operator, with invariant lower expectation $\underline{E}_\infty$, and let r be the smallest natural number such that $\rho := \rho(\underline{\mathrm{T}}^r) < 1$. Let $\underline{E}_a$ and $\underline{E}_b$ be any two lower expectations on $\mathscr{G}(\mathscr{X})$. Then for all $f \in \mathscr{G}(\mathscr{X})$, $\ell_1, \ell_2 \in \mathbb{N}_0$:*

$$\left|\underline{E}_a(\underline{\mathrm{T}}^{\ell_1} f) - \underline{E}_b(\underline{\mathrm{T}}^{\ell_2} f)\right| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\min\{\ell_1,\ell_2\}}{r} \rfloor}. \tag{12}$$

As a consequence, for all $f \in \mathscr{G}(\mathscr{X})$, $\ell, \ell_1, \ell_2 \in \mathbb{N}_0$ and $k, k_1, k_2 \in \mathbb{N}$:

$$\left|\underline{\mathrm{T}}^{\ell} f(X_k) - \underline{E}_\infty(f)\right| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\ell}{r} \rfloor}, \tag{13}$$

$$\left|\underline{E}_a(\underline{\mathrm{T}}^{\ell} f) - \underline{E}_\infty(f)\right| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\ell}{r} \rfloor}, \tag{14}$$

$$\left|\underline{\mathrm{T}}^{\ell} f(X_k) - \underline{E}_b(\underline{\mathrm{T}}^{\ell} f)\right| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\ell}{r} \rfloor}, \tag{15}$$

$$\left|\underline{\mathrm{T}}^{\ell_1} f(X_{k_1}) - \underline{\mathrm{T}}^{\ell_2} f(X_{k_2})\right| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\min\{\ell_1,\ell_2\}}{r} \rfloor}. \tag{16}$$

Proposition 10. *Consider an imprecise Markov chain with initial—or marginal—model $\underline{E}_1$ and lower transition operator $\underline{\mathrm{T}}$. Assume that $\underline{\mathrm{T}}$ is Perron–Frobenius-like, with invariant lower expectation $\underline{E}_\infty$, and let r be the smallest natural number such that $\rho := \rho(\underline{\mathrm{T}}^r) < 1$. Then the following statements hold for all $f \in \mathscr{G}(\mathscr{X})$, $\ell \in \mathbb{N}_0$ and $n \in \mathbb{N}$:*

(i) $|\langle \mathscr{W} \rangle [\underline{\mathrm{T}}^{\ell} f](X_{1:n})| \leq \|f\|_{\mathrm{v}} \rho^{\lfloor \frac{\ell}{r} \rfloor}$.

(ii) $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \underline{\mathrm{T}}^n f(X_k) = \underline{E}_\infty(f)$.

(iii) $\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} \underline{\mathrm{T}}^{\ell} f(X_n) = \underline{E}_\infty(f)$.

(iv) $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \underline{E}_k(f) = \underline{E}_\infty(f)$.

We can now state our main result.

Theorem 11 (Pointwise ergodic theorem)**.** *Consider an imprecise Markov chain with initial—or marginal—model $\underline{E}_1$ and lower transition operator $\underline{\mathrm{T}}$. Assume that $\underline{\mathrm{T}}$ is Perron–Frobenius-like, with invariant lower expectation $\underline{E}_\infty$. Then for all $f \in \mathscr{G}(\mathscr{X})$:*

$$\liminf \mathscr{A}[f] \geq 0 \text{ strictly almost surely},$$

and consequently,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) \geq \underline{E}_\infty(f) \text{ strictly almost surely}.$$

8 Conclusions and Discussion

We have proved a version of the pointwise ergodic theorem for imprecise Markov chains involving functions of a single state. It does not seem very difficult to extend this result to involve functions of a finite number of states, but it is still a subject of current research whether it can be extended to gambles that depend on the entire state trajectory, and not just on a finite number of states.

Our version subsumes the one for (precise) Markov chains, because there $\underline{E}_\infty(f) = \overline{E}_\infty(f) = E_\infty(f)$ and therefore

$$\begin{aligned} E_\infty(f) = \overline{E}_\infty(f) &\geq \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) \\ &\geq \liminf_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) \\ &\geq \underline{E}_\infty(f) = E_\infty(f) \\ &\qquad \text{strictly almost surely,} \end{aligned}$$

implying that $\frac{1}{n}\sum_{k=1}^{n} f(X_k)$ converges to $E_\infty(f)$ (strictly) almost surely. In our more general case, however, we cannot generally prove that there is almost sure *convergence*, and we retain only almost sure inequalities involving limits inferior and superior, as is also the case for our strong law of large numbers for submartingale differences. Indeed,

that such convergence should not really be expected for imprecise probability models was already argued by Walley and Fine [17].

Ergodicity results for Markov chains are quite relevant for applications in queuing theory, where they are for instance used to prove Little's Law [18], or ASTA (Arrivals See Time Averages) properties [9]. We believe the discussion in this paper could be instrumental in deriving similar properties for queues where the probability models for arrivals and departures are imprecise.

Acknowledgements

Research by Gert de Cooman and Stavros Lopatatzidis was funded through project number G012512N of the Research Foundation Flanders (FWO). Jasper De Bock is a PhD Fellow of the FWO and wishes to acknowledge its financial support. The authors would like to express their gratitude to three anonymous referees for their comments, and to Tom Ward and Volodya Vovk for taking the time to discuss some of the ideas behind this paper.

References

[1] Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. John Wiley & Sons, 2014.

[2] Gert de Cooman and Filip Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008. doi: 10.1016/j.artint.2008.03.001.

[3] Gert de Cooman and Enrique Miranda. Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138(8):2409–2432, 2008. doi: 10.1016/j.jspi.2007.10.020.

[4] Gert de Cooman and Enrique Miranda. Lower previsions induced by filter maps. *Journal of Mathematical Analysis and Applications*, 410(1):101–116, 2014.

[5] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, January 2009. doi: 10.1017/S0269964809990039. arXiv:0801.0980.

[6] Darald J. Hartfiel. *Markov Set-Chains*. Number 1695 in Lecture Notes in Mathematics. Springer, Berlin, 1998.

[7] Filip Hermans and Gert de Cooman. Characterisation of ergodic upper transition operators. *International Journal of Approximate Reasoning*, 53(4):573–583, 2012. doi: 10.1016/j.ijar.2011.12.008.

[8] Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, second edition, 2002.

[9] Armand Makowski, Benjamin Melamed, and Ward Whitt. On averages seen by arrivals in discrete time. In *28th IEEE Conference on Decision and Control*, pages 1084–1086. IEEE, 1989.

[10] Enrique Miranda and Gert de Cooman. Lower previsions. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*. John Wiley & Sons, 2014.

[11] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[12] Glenn Shafer, Vladimir Vovk, and Akimichi Takemura. Lévy's zero–one law in game-theoretic probability. *Journal of Theoretical Probability*, 25: 1–24, 2012.

[13] Matthias C. M. Troffaes and Gert de Cooman. *Lower Previsions*. Wiley, 2014.

[14] Vladimir Vovk and Glenn Shafer. Game-theoretic probability. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*. John Wiley & Sons, 2014.

[15] Damjan Škulj and Robert Hable. Coefficients of ergodicity for Markov chains with uncertain parameters. *Metrika*, 76(1):107–133, 2013. doi: 10.1007/s00184-011-0378-0.

[16] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[17] Peter Walley and Terrence L. Fine. Towards a frequentist theory of upper and lower probability. *Annals of Statistics*, 10:741–761, 1982.

[18] Ward Whitt. A review of $L = \lambda W$ and extensions. *Queueing Systems*, 9(3):235–268, 1991.

[19] David Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, UK, 1991.

Fully Conglomerable Coherent Upper Conditional Prevision Defined by the Choquet Integral with respect to its Associated Hausdorff Outer Measure

Serena Doria
Department of Engineering and Geology
University G.d'Annunzio, Chieti-Pescara, Italy
s.doria@dst.unich.it

Abstract

Let (Ω, d) be a metric space where Ω is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension and let $\mathbf{B}$ be a partition of Ω. The coherent upper conditional prevision defined as the Choquet integral with respect to its associated Hausdorff outer measure is proven to satisfy the disintegration property and the conglomerative principle on every partition.

Keywords. Coherent upper conditional previsions, Hausdorff outer measures, Choquet integral, disintegration property, conglomerability principle.

1 Introduction

In Walley [21, 6.8] full conglomerability is required as a rationality axiom for a coherent upper prevision since it assures that it can be coherently extended to coherent conditional previsions for any partition $\mathbf{B}$ of Ω. If the partition $\mathbf{B}$ represents an experiment that could be performed it is necessary to update the unconditional upper prevision after observing a set B of $\mathbf{B}$. Coherent upper conditional prevision is coherent with the unconditional prevision if the following *conglomerability principle* is satisfied: if a random variable X is B-desirable, i.e. we have a disposition to accept X for every set B in the partition $\mathbf{B}$, then X is desirable. If there is no coherent way of updating the initial prevision after learning the outcome of the experiment the upper prevision, which represents our knowledge, is unreasonable.

For linear unconditional prevision full conglomerability is equivalent to the disintegration property introduced by Dubins [10] which is a generalization to the class of all bounded random variables of the conglomerative principle, introduced by de Finetti [2, p.99], [3] for probabilities.

Coherent upper conditional previsions are functionals on a linear space of bounded random variables satisfying the axioms of separate coherence.

Coherent upper conditional previsions cannot always be defined as an extension of conditional expectation of measurable random variables defined by the Radon-Nikodym derivative, according to the axiomatic definition. It occurs because one of the defining properties of the Radon-Nikodym derivative, that is to be measurable with respect to the σ-field of the conditioning events, contradicts a necessary condition for coherence (Doria [8, Theorem 1]). So the necessity to find a new mathematical tool in order to define coherent upper conditional previsions arises. Since conditional expectation defined by the Radon-Nikodym derivative may fail to be coherent, it is important to prove that the price of coherence is not to lose disintegrability that is a property satisfied by conditional expectation in the axiomatic definition.

The relation between conglomerability and countable additivity has been investigated in Walley [21, section 6.9] and Schervish, Seidenfeld and Kadane [18]. In [18] it has been proven that when an additive probability P is defined at least on a σ-field and it assumes infinitely many different values then it is fully conglomerable if and only if it is countably additive on every partition of Ω. It means that we can find examples of merely additive probabilities defined on a field, that is not a σ-field, that assume only finitely many values and that are conglomerable with respect to a given partition (see Scozzafava [19, Example 5.5.], and Walley [21, Example 6.6.4]). But since every merely finitely additive probability defined on a field can be extended to a σ-field and to the power set, we have that every extension of this kind of probability to a σ-field is not fully conglomerable, since it fails conglomerability with respect to some countable partitions. In Kadane, Schervish and Seidenfeld [12, Example 6.1] it is proven that for non-countable partitions countable additivity of the unconditional probability is not a sufficient condition to assure that it is coherent with the conditional probability.

Examples of non-conglomerable linear previsions are

given in Walley [21, 6.6.6, 6.6.7].

Consequences of failure of conglomerability are investigated in decision making where non-conglomerability of finitely additive probabilities leads to a violation of the decision-theoretic principle of admissibility as proven in Kadane, Schervish and Seidenfeld [12]. Moreover failure of conglomerability has consequence in sequential decision problems (Kadane, Schervish and Seidenfeld [13]).

In the paper of Miranda, Zaffalon and de Cooman [14] it is shown that the natural extension of assessment after imposing conglomerability, does not yield in general the conglomerable natural extension.

In Doria [8], [7], [5] a new model of coherent upper conditional previsions defined by Hausdorff outer measures is proposed in a metric space. Coherent upper and lower conditional probabilities are obtained when only 0-1 valued random variables are considered.

Let (Ω, d) be a metric space and let $\mathbf{B}$ be a partition of Ω. For each $B \in \mathbf{B}$ denote by s the Hausdorff dimension of B and let h^s be the Hausdorff s-dimensional outer measure, which is called Hausdorff outer measure *associated* with the coherent upper conditional prevision $\overline{P}(X|B)$. For every bounded random variable X a coherent upper conditional prevision $\overline{P}(X|B)$ is defined ([8], [7]) by the Choquet integral with respect to its associated Hausdorff outer measure if the conditioning event has positive and finite Hausdorff outer measure in its Hausdorff dimension. Otherwise if the conditioning event has Hausdorff outer measure in its Hausdorff dimension equal to zero or infinity it is defined by a 0-1 valued finitely, but not countably, additive probability.

In this paper coherent upper conditional and unconditional previsions are proven to satisfy the disintegration property and the conglomerative principle on every partition $\mathbf{B}$ of Ω if Ω is set with positive and finite Hausdorff outer measure in its Hausdorff dimension t. It occurs because Hausdorff outer measures are submodular and every random variable and every constant are comonotonic so that the Choquet integral with respect to the t-dimensional Hausdorff outer measure is equal to the Choquet integral with respect to an additive measure, which agrees with the t-dimensional Hausdorff outer measure on the class of the h^t- measurable sets [4, Proposition 10.1].

The paper is organized as follows. In Section 2 the model of coherent upper conditional previsions defined with respect to Hausdorff outer measure and its properties are recalled. Moreover a characterization of measurable sets is given in terms of natural extensions and every set B belonging to a partition of Ω is proven to be measurable with respect to the coherent upper conditional probabilities $\overline{P}(\cdot|B)$ and $\overline{P}(\cdot|\Omega)$.

Let $\overline{P}(X|\mathbf{B})$ be the random variable equal to $\overline{P}(X|B)$ if $\omega \in B$. In Section 3 the given coherent upper conditional prevision $\overline{P}(X|\mathbf{B})$ is proven to satisfy the disintegration property on every partition $\mathbf{B}$ of Ω if Ω is a set with positive and finite outer measure in its Hausdorff dimension and X is a monotone random variable. The random variables X and $\overline{P}(X|\mathbf{B})$ are proven to be comonotonic so that the Choquet integral of $X + \overline{P}(X|\mathbf{B})$ is additive.

In Section 4 the given upper coherent conditional prevision $\overline{P}(X|\mathbf{B})$ is proven to satisfy the disintegration property and the conglomerative principle on every partition $\mathbf{B}$ of Ω if Ω is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension. A sufficient condition is given such that the Choquet integral of $X + \overline{P}(X|\mathbf{B})$ is additive.

2 Coherent Upper Conditional Previsions Defined by the Choquet Integral with respect to Hausdorff Outer Measure

Let (Ω, d) be a metric space and let $\mathbf{F}$ be the Borel σ-field, which is the σ-field generated by the open sets of the *metric topology*, induced by the metric d. Let $\mathbf{B}$ be a partition of Ω.

A bounded random variable is a function $X : \Omega \to \Re$ and $L(\Omega)$ is the class of all bounded random variables defined on Ω; for every $B \in \mathbf{B}$ denote by $X|B$ the restriction of X to B and by $\sup(X|B)$ the supremum value that X assumes on B. Let $L(B)$ be the class of all bounded random variables $X|B$ and let I_B the indicator function of the set B, that is $I_B(\omega) = 1$ if $\omega \in B$ and $I_B(\omega) = 0$ if $\omega \notin B$.

For every $B \in \mathbf{B}$ coherent upper conditional previsions $\overline{P}(\cdot|B)$ are functionals, defined on $L(B)$, satisfying the axioms of separate coherence [21].

Definition 1. *Separately coherent upper conditional previsions are functionals $\overline{P}(\cdot|B)$ defined on $L(B)$, such that the following conditions hold for every X and Y in $L(B)$ and every strictly positive constant λ:*

1) $\overline{P}(X|B) \leq \sup(X|B)$;

2) $\overline{P}(\lambda X|B) = \lambda \overline{P}(X|B)$ *(positive homogeneity);*

3) $\overline{P}(X + Y)|B) \leq \overline{P}(X|B) + \overline{P}(Y|B)$ *(subadditivity);*

4) $\overline{P}(I_B|B) = 1$.

Definition 2. *Given a partition $\boldsymbol{B}$ and a random variable $X \in L(\Omega)$ a coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$ is a random variable on Ω equal to $\overline{P}(X|B)$ if $\omega \in B$. The random variable $\overline{P}(X|\boldsymbol{B})$ is separately coherent if all the $\overline{P}(X|B)$ are separately coherent.*

Suppose that $\overline{P}(X|B)$ is a coherent upper conditional prevision on $L(B)$ then its conjugate coherent lower conditional prevision is defined by $\underline{P}(X|B) = -\overline{P}(-X|B)$. Let K be a linear space contained in $L(B)$; if for every X belonging to K we have $P(X|B) = \underline{P}(X|B) = \overline{P}(X|B)$ then $P(X|B)$ is called a coherent *linear* conditional prevision (de Finetti [2]) and it is a linear, positive functional on K.

Moreover $P(X|B)$ is dominated on K by the subadditive, positively homogeneous functional $\overline{P}(X|B)$ and for the Hahn-Banach Theorem (see Rudin [17, Theorem 3.2]) it can be extended to a linear functional on $L(B)$ dominated by $\overline{P}(X|B)$. The following extension theorem holds (see also Regazzini [15]):

Theorem 1. *Let $\overline{P}$ be a coherent upper prevision on $L(B)$ and let P be a coherent linear prevision on a linear space $K \subseteq L(B)$ such that $P(X) \leq \overline{P}(X)$ $\forall X \in K$. Then there exists a linear extension P^* of P to $L(B)$ such that $P^*(X) = P(X)$ $\forall X \in K$ and $P^*(X) \leq \overline{P}(X)$ $\forall X \in L(B)$.*

The unconditional coherent upper prevision $\overline{P} = \overline{P}(\cdot|\Omega)$ is obtained as a particular case when the conditioning event is Ω.

An upper prevision is a real-valued function defined on some class of bounded random variables $K \subseteq L(B)$. A necessary and sufficient condition for an upper prevision $\overline{P}$ to be coherent is to be the *upper envelope* of linear previsions defined on $L(B)$, i.e. there is a class M of linear previsions on $L(B)$ such that [21, 3.3.3]

$$\overline{P} = \sup\{P(X) : P \in M; X \in K\} .$$

The supremum is actually attained by some dominated linear prevision.

Let $\overline{P}$ be an upper prevision on an arbitrary domain K such that the class $M(\overline{P})$ of all linear previsions defined on $L(\Omega)$ and dominated by $\overline{P}$ on K, is nonempty. The maximal extension of $\overline{P}$ to $L(B)$, denoted by $\overline{E}$, is called [21, 3.1.1] the *natural extension* of $\overline{P}$. Moreover $\overline{P}$ is coherent on K if and only if its natural extension $\overline{E}$ agrees with $\overline{P}$ on K.

Coherent upper conditional probabilities are obtained when only 0-1 valued random variables are considered.

If P is a countably additive probability defined on a σ-field $S \subset \wp(\Omega)$ its natural extensions, defined on all subsets of Ω, are the inner and outer measures generated by it [21, Theorem 3.1.5], that is

$$\overline{E}(A) = \inf\{P(B) : B \supset A; B \in S\}, A \in \wp(\Omega)$$

$$\underline{E}(A) = \sup\{P(B) : B \subset A; B \in S\}, A \in \wp(\Omega).$$

Definition 3. *A subset A of Ω is called measurable with respect to a coherent upper conditional probability $\overline{P}(\cdot|B)$ defined on $\wp(B)$ if it decomposes every subset of B additively, that is if*

$$\overline{P}(E|B) = \overline{P}((A \cap E)|B) + \overline{P}((A^c \cap E)|B)$$

for all sets $E \subseteq B$.

The class of all measurable sets of Ω is a field and $\overline{P}(\cdot|B)$ is additive on it [4, Proposition 2.5].

If $\overline{P}(\cdot|B)$ is subadditive and continuous from below then the class of all measurable sets of Ω is a σ-field and $\overline{P}(\cdot|B)$ is countably additive on it [4, Proposition 2.6].

Let $P(\cdot|B)$ be an additive coherent conditional probability on a field $S \subset \wp(B)$, then the class of all measurable sets with respect to $P(\cdot|B)$ coincides with the class of sets such that the outer and inner measure are equal. [4, Proposition 2.9].

A characterization of measurable sets can be given in terms of natural extensions.

Proposition 1. *Let $\overline{P}(\cdot|B)$ be a coherent upper conditional probability such that its restriction $P(\cdot|B)$ to a σ-field S is a countably additive coherent conditional probability. A subset A of Ω is measurable with respect to $\overline{P}(\cdot|B)$ if and only if $\overline{E}(A|B) = \underline{E}(A|B)$.*

A functional $\Gamma : L(B) \to \Re$ can be represented as Choquet integral with respect to a coherent upper conditional probability μ on $\wp(B)$ if $\Gamma(X) = \int X d\mu$ $\forall X \in L(B)$. Then $\Gamma(I_A) = \mu(A)$. For every $x \in \Re$ let $\{X|B > x\} = \{\omega \in B : X(\omega) > x\}$.

Since X is a bounded random variable thus there exist a constant k such that $\widetilde{X} = X + k \geq 0$ and the *decreasing distribution function* of $\widetilde{X}$ with respect to μ is $G_{\mu,\widetilde{X}}(x) = G_{\mu,X}(x-k) = \mu\{X|B > x-k\}$ for every real number x [4, Proposition 4.1].

The *Choquet integral* [4] of a bounded random variable X with respect to μ is defined by

$$\int X d\mu = \int_0^{+\infty} G_{\mu,\widetilde{X}}(x) dx.$$

Let S be a class properly contained in $\wp(\Omega)$ and μ a coherent upper conditional probability on S. Denoted by μ^* and μ_* respectively the outer and inner set functions generated by μ, a random variable X is called *upper-μ-measurable* [4] if $G_{\mu^*,X}(x) = G_{\mu_*,X}(x)$

except on a μ-null set, that is equivalent to require that all the upper level sets $]x, +\infty[$ are μ^*-measurable.

X is called upper S-measurable if it is upper μ-measurable for any monotone set function on S; moreover if the sets $\{\omega \in \Omega : X(\omega) > x\}$ belong to S for every $x \in \Re$ then X is S-measurable.

If Ω is finite and μ defined on a field S, denote by $A_1, ..., A_n$ the atoms of S, which are the minimal elements of $S \ - \oslash$. If the atoms A_i are enumerated so that $x_i = X(A_i)$ are in in descending order, i.e. $x_1 \geq x_2 \geq ... \geq x_n$ and $x_{n+1} = 0$ the Choquet integral with respect to μ is given by

$$\int X d\mu = \sum_{i=1}^{n} (x_i - x_{i+1}) \mu(S_i)$$

where $S_i = A_1 \cup A_2 ... \cup A_i$, and $x_{n+1} = 0$.

Definition 4. *A coherent upper conditional probability μ is submodular or 2-alternating if for every $A, E \in \wp(B)$*

$$\mu((A \cup E)|B) + \mu((A \cap E)|B) \leq \mu(A|B) + \mu(E|B).$$

In Doria [5], [8] a new model of coherent upper conditional probability based on Hausdorff outer measures (see Rogers [16] and Falconer [11]) is introduced.

Let $\delta > 0$ and let s be a non-negative number. The *diameter* of a non empty set U of Ω is defined as $|U| = \sup \{d(x, y) : x, y \in U\}$ and if a subset A of Ω is such that $A \subseteq \bigcup_i U_i$ and $0 < |U_i| \leq \delta$ for each i, the class $\{U_i\}$ is called a δ-cover of A.

The *Hausdorff s-dimensional outer measure* of A, denoted by $h^s(A)$, is defined on $\wp(\Omega)$, the class of all subsets of Ω, as

$$h^s(A) = \lim_{\delta \to 0} \inf \sum_{i=1}^{+\infty} |U_i|^s$$

where the infimum is over all δ-covers $\{U_i\}$.

The *Hausdorff dimension* of a set A, $dim_H(A)$, is defined as the unique value, such that

$$h^s(A) = +\infty \text{ if } 0 \leq s < dim_H(A),$$

$$h^s(A) = 0 \text{ if } dim_H(A) < s < +\infty.$$

We can observe that if $0 < h^s(A) < +\infty$ then $dim_H(A) = s$ (the converse is not true). In any metric space a finite non-empty subset A of Ω has positive and finite counting measure h^0 so the Hausdorff dimension of a finite set is 0.

Hausdorff s-dimensional outer measures are submodular, continuous from below and their restriction on the Borel σ-field is countably additive.

Theorem 2. *[8, Theorem 2] Let m be a 0-1 valued finitely additive, but not countably additive, probability on $\wp(B)$ such that a different m is chosen for each B. Then for each $B \in \boldsymbol{B}$ the functionals $\overline{P}(X|B)$ defined on $L(B)$ by*

$$\overline{P}(X|B) = \frac{1}{h^s(B)} \int_B X dh^s \text{ if } 0 < h^s(B) < +\infty$$

and by

$$\overline{P}(X|B) = m(XB) \text{ if } h^s(B) = 0, +\infty$$

are separately coherent upper conditional previsions.

Coherent upper conditional probabilities are obtained when only indicator functions of events are considered.

Theorem 3. *[8, Theorem 3] Let m be a 0-1 valued finitely additive, but not countably additive, probability on $\wp(B)$ such that a different m is chosen for each B. Thus, for each $B \in \boldsymbol{B}$, the function defined on $\wp(B)$ by*

$$\overline{P}(A|B) = \frac{h^s(AB)}{h^s(B)} \text{ if } 0 < h^s(B) < +\infty$$

and by

$$\overline{P}(A|B) = m(AB) \text{ if } h^s(B) = 0, +\infty$$

is a coherent upper conditional probability.

A *fuzzy measure* (also called a *capacity*) μ on $\wp(B)$ is a set function such that $\mu(B) = 1$, $\mu(\oslash) = 0$, $\mu(A) \leq \mu(E)$ if $A \subseteq E$, i.e. a fuzzy measure is a monotone set function such that $\mu(B) = 1$, $\mu(\oslash) = 0$.

If $B \in \mathbf{B}$ is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension s, the fuzzy measure μ_B^* defined for every $A \in \wp(B)$ by $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$ is a coherent upper conditional probability, which is submodular, continuous from below and such that its restriction to the Borel σ-field is a Borel regular countably additive probability. Moreover the coherent upper conditional probability $\mu_B^* = \frac{h^s(AB)}{h^s(B)}$ is translation invariant.

The coherent upper unconditional probability $\overline{P} = \mu_\Omega^*$ defined on $\wp(\Omega)$ is obtained for B equal to Ω.

Theorem 4. *Let $\boldsymbol{B}$ be a partition of Ω. Every $B \in \boldsymbol{B}$, is a measurable set with respect to $\overline{P}(\cdot|B)$.*

Proof. Let $P(\cdot|B)$ the restriction, to the σ-field of the h^s- measurable sets, of the coherent upper conditional probability $\overline{P}(\cdot|B)$ defined in Theorem 3. For every $B \in \mathbf{B}$, with positive and finite Hausdorff outer measure in its Hausdorff dimension, $\overline{P}(\cdot|B)$ is the natural extension of the countably additive probability $P(\cdot|B)$. Then by the conjugacy property we have

$$\underline{P}(B|B) = \overline{P}(\Omega|B) - \overline{P}(B^c|B) = \overline{P}(B|B).$$

So by Proposition 1, the set B is measurable with respect to $\overline{P}(\cdot|B)$.

For every $B \in \mathbf{B}$ with Hausdorff outer measure equal to zero or infinity in its Hausdorff dimension $\overline{P}(\cdot|B)$ is a $0-1$ valued additive probability so by Definition 3 B is measurable with respect to $\overline{P}(\cdot|B)$. $\diamond$

Theorem 5. *Let Ω be a set with positive and finite Hausdorff measure in its Hausdorff dimension t. Then for every partition $\boldsymbol{B}$ there is at most a countable subclass $\boldsymbol{B}^*$ of $\boldsymbol{B}$ of sets B with positive upper coherent probability μ^*_Ω.*

Proof Since Ω is a set positive and finite Hausdorff outer measure in its Hausdorff dimension t, we have that the restriction $\mu_\Omega(\cdot) = \frac{h^t(\cdot)}{h^t(\Omega)}$ to the σ-field of h^t-measurable sets, of the upper conditional probability defined in Theorem 3, is a countably additive probability. Moreover since Hausdorff outer measure are regular for each $B \in \mathbf{B}$ there is a h^t-measurable set B' such that $B \subset B'$ and $h^t(B) = h^t(B')$ so for every partition $\mathbf{B}$ there is at most a countable subclass $\mathbf{B}^*$ of $\mathbf{B}$ of sets B with positive upper coherent probability μ^*_Ω. $\diamond$

Theorem 6. *Let Ω be a set with positive and finite Hausdorff measure in its Hausdorff dimension t and let $\boldsymbol{B}$ be a Borel countable partition of Ω. Then the random variable $\overline{P}(X|\boldsymbol{B})$ is h^t-measurable.*

Proof The random variable $\overline{P}(X|\mathbf{B})$ is h^t-measurable if the sets $\left\{\omega \in \Omega : \overline{P}(X|\mathbf{B}) \geq x\right\}$ are h^t-measurable for every $x \in \Re$. Since the random variable $\overline{P}(X|\mathbf{B})$ is $\mathbf{B}$-measurable, i.e. constant on the sets B, and $\mathbf{B}$ is a Borel countable partition of Ω, for every $x \in \Re$ the sets $\left\{\omega \in \Omega : \overline{P}(X|\mathbf{B}) \geq x\right\}$ are countable unions of h^t-measurable sets B , so they are h^t-measurable. $\diamond$

3 Conglomerability and Disintegration Property of Coherent Upper Conditional Prevision Defined by Hausdorff Outer Measure

In this section coherent upper conditional previsions defined as in Theorem 2, are proven to satisfy the conglomerability axiom and the disintegration property on every partition and for every monotone random variable.

Walley [21, 6.3] discusses when an unconditional lower prevision $\underline{P}$ is coherent with the lower conditional prevision $\underline{P}(\cdot|\mathbf{B})$.

Definition 5. *$\underline{P}$ and $\underline{P}(\cdot|\boldsymbol{B})$ defined on $\boldsymbol{L}(\Omega)$ are called coherent if and only if the following conditions hold for every X in $\boldsymbol{L}(\Omega)$ and $B \in \boldsymbol{B}$:*

$$\underline{P}(\textstyle\sum_{B\in\boldsymbol{B}} I_B(X - \underline{P}(X|B)) \geq 0$$

(Conglomerative axiom)

and

$$\underline{P}(I_B(X - \underline{P}(X|B)) = 0$$

(Generalized Bayes Rule).

In some special cases coherence of $\underline{P}$ and $\underline{P}(\cdot|\mathbf{B})$ can be characterized by simpler conditions. In particular in Walley [21, section 6.5.3 and section 6.5.7] it has been proven that if P and $P(\cdot|\mathbf{B})$ are respectively linear unconditional and conditional previsions on the class of all bounded random variables and $P(\cdot|\mathbf{B})$ are separately coherent, then P and $P(\cdot|\mathbf{B})$ are coherent if and only if the following *conglomerative property* is satisfied $P(X) = P(P(X|\mathbf{B}))$.

The notion of disintegrability given by Dubins [10] can be extended to coherent upper conditional previsions.

Definition 6. *A coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$ is disintegrable with respect to a partition $\boldsymbol{B}$ if the following equality is satisfied for every bounded variable $X \in L(\Omega)$*

$$\overline{P}(X) = \overline{P}(\overline{P}(X|\boldsymbol{B})).$$

Definition 7. *A coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$ is defined to be conglomerative with respect to a partition $\boldsymbol{B}$ of Ω if the following condition is satisfied: for every bounded variable $X \in L(\Omega)$*

$$\overline{P}(X|\boldsymbol{B}) \geq 0 \text{ implies } \overline{P}(X) \geq 0.$$

Definition 8. *Two random variables X and $Y \in L(\Omega)$ are comonotonic on Ω if and only if $\forall\omega_1, \omega_2 \in \Omega$*

$$(X(\omega_1) - X(\omega_2))(Y(\omega_1) - Y(\omega_2)) \geq 0.$$

A class $\mathbf{C}$ of random variables is *comonotonic* if and only if each pair of functions in $\mathbf{C}$ is comonotonic.

Let μ be a coherent upper probability which is is submodular and defined on $\wp(\Omega)$ and let $\mathbf{C}$ be a comonotonic class of random variables. By Proposition 10.1 of [4] for any random variable $X \in \mathbf{C}$ there exists an additive set function α on $\wp(\Omega)$, which agree with μ on the σ-field of μ-measurable sets, such that

$$\textstyle\int_\Omega X d\mu = \int_\Omega X d\alpha$$

Example 1. *Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and let P_1 and P_2 be two finitely additive probabilities defined by $P_1(\omega_i) = \mu^*_\Omega(\omega_i) = \frac{1}{4}$ for i=1,...,4 and $P_2(\omega_1) = P_2(\omega_2) = \frac{1}{8}$, $P_2(\omega_3) = \frac{1}{2}$, $P_2(\omega_4) = \frac{1}{4}$. Let $\overline{\mu}$ be the submodular coherent upper probability defined on*

$\wp(\Omega)$ by the upper envelope of P_1 and P_2, i.e. $\overline{\mu}(A) = \max_{j=1,2} P_j(A)$ for $A \in \wp(\Omega)$ and let $\underline{\mu}$ be the coherent lower probability defined by $\underline{\mu}(A) = \min_{j=1,2} P_j(A)$ for $A \in \wp(\Omega)$. Let consider the comonotonic random variables X and Y defined by

$X(\omega_1) = 0, X(\omega_2) = 1, X(\omega_3) = 2, X(\omega_4) = 3$ *and*

$Y(\omega_1) = 0, Y(\omega_2) = 1, Y(\omega_3) = 3, Y(\omega_4) = 4.$

We have that

$\int X d\overline{\mu} = \int X dP_2 = \frac{15}{8}, \int Y d\overline{\mu} = \int Y dP_2 = \frac{21}{8}.$

Moreover, by Proposition 10.1 of [4], for any other increasing random variable $Z \in L(\Omega)$ we have that

$\int Z d\overline{\mu} = \int Z dP_2.$

By coherence of the lower probability $\underline{\mu}$ we have

$\int X d\underline{\mu} = \int X dP_1 = \frac{3}{2}$ *and* $\int Y d\underline{\mu} = \int Y dP_1 = 2$

and by the asymmetry of the Choquet integral for every increasing random variable $Z \in L(\Omega)$ we have that

$\int (-Z) d\underline{\mu} = -\int Z d\overline{\mu}.$

Let $I(\Omega)$ be the class of all increasing random variables on Ω.

Theorem 7. *Let $X \in L(\Omega)$ be a monotone random variable, then X and $\overline{P}(X|\boldsymbol{B})$ are comonotonic.*

Proof Let consider $X \in I(\Omega)$. Let $Y(\omega) = \overline{P}(X|\mathbf{B})$. We have to prove that $\forall \omega_1, \omega_2 \in \Omega$

$$(X(\omega_1) - X(\omega_2))(Y(\omega_1) - Y(\omega_2)) \geq 0.$$

If $\omega_1, \omega_2 \in B$ then X and $\overline{P}(X|\mathbf{B})$ are comonotonic since $\overline{P}(X|\mathbf{B})$ is constant on the atoms of the partition $\mathbf{B}$ so that

$$Y(\omega_1) - Y(\omega_2) = \overline{P}(X|B) - \overline{P}(X|B) = 0.$$

If $\omega_1 < \omega_2$ and $\omega_1 \in B_1$ and $\omega_2 \in B_2$ since X is increasing and $\overline{P}(X|\mathbf{B})$ is separately coherent we have $\inf_{B_1} X \leq \overline{P}(X|B_1) \leq \sup_{B_1} X \leq \inf_{B_2} X \leq \overline{P}(X|B_2)$.

So $\forall \omega_1, \omega_2 \in \Omega$ with $\omega_1 < \omega_2$

$$X(\omega_1) \leq X(\omega_2) \text{ implies } \overline{P}(X|B_1) \leq \overline{P}(X|B_2).$$

So that X and $\overline{P}(X|\mathbf{B})$ are comonotonic.$\diamond$

In the next theorem sufficient conditions are given such that the coherent upper conditional prevision defined in Theorem 2 satisfies the disintegration property on every partition and for every monotone random variable.

Theorem 8. *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t and let $\boldsymbol{B}$ be a partition of Ω. Then for every monotone $X \in L(\Omega)$ the coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$, defined as in Theorem 2, satisfies the disintegration property, i.e.*

$$\overline{P}(X) = \overline{P}(\overline{P}(X|\boldsymbol{B})).$$

Proof We prove the theorem for $X \in I(\Omega)$. Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t then the coherent upper unconditional prevision is defined as the Choquet integral with respect to $\mu^*_\Omega(\cdot) = \frac{h^t(\cdot)}{h^t(\Omega)}$. Let $I(\Omega)$ be the class of all increasing random variables on $\wp(\Omega)$. Since h^t is submodular and defined on $\wp(\Omega)$ by [4, Proposition 10.1] for any random variable $X \in I(\Omega)$ there exists an additive set function α on $\wp(\Omega)$, which agrees with h^t on the σ-field of h^t-measurable sets, such that

$$\int_\Omega X dh^t = \int_\Omega X d\alpha.$$

By Theorem 5 there is at most a countable subclass $\mathbf{B}^*$ of $\mathbf{B}$ of sets B with positive upper coherent probability μ^*_Ω. By Theorem 7 X and $\overline{P}(X|\mathbf{B})$ are comonotonic so the disintegration property is satisfied for every partition $\mathbf{B}$ since the following equalities hold:

$$\begin{aligned}
\overline{P}(\overline{P}(X|\mathbf{B})) &= \frac{1}{h^t(\Omega)} \int_\Omega \overline{P}(X|\mathbf{B}) dh^t \\
&= \frac{1}{h^t(\Omega)} \int_\Omega \overline{P}(X|\mathbf{B}) d\alpha \\
&= \sum_{B \in \mathbf{B}^*} \left(\frac{1}{h^t(B)} \int_B X dh^t \right) \frac{h^t(B)}{h^t(\Omega)} \\
&= \frac{1}{h^t(\Omega)} \sum_{B \in \mathbf{B}^*} \int_B X dh^t \\
&= \frac{1}{h^t(\Omega)} \sum_{B \in \mathbf{B}^*} \int_B X d\alpha \\
&= \frac{1}{h^t(\Omega)} \int_\Omega X d\alpha \\
&= \frac{1}{h^t(\Omega)} \int_\Omega X dh^t = \overline{P}(X). \diamond
\end{aligned}$$

Theorem 9. *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t and let $\boldsymbol{B}$ be a partition of Ω. Then for every monotone $X \in L(\Omega)$ we have*

$$\int_\Omega (X + \overline{P}(X|\boldsymbol{B})) dh^t = \int_\Omega X dh^t + \int_\Omega \overline{P}(X|\boldsymbol{B}) dh^t.$$

Proof By Theorem 7 we have that X and $\overline{P}(X|\mathbf{B})$ are comonotonic so that the Choquet integral of their sum in additive.

$$\int_\Omega (X + \overline{P}(X|\mathbf{B})) dh^t = \int_\Omega X dh^t + \int_\Omega \overline{P}(X|\mathbf{B}) dh^t. \diamond$$

4 Full Conglomerability

In this section the coherent upper conditional prevision $\overline{P}(X|\mathbf{B})$ is proven to satisfy the disintegration property with respect to every partition $\mathbf{B}$ if Ω is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension.

Theorem 10. *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t. Thus the coherent conditional prevision $\overline{P}(X|\boldsymbol{B})$ satisfies the disintegration property on every partition $\boldsymbol{B}$ of Ω.*

Proof. Ω is a set with with positive and finite Hausdorff outer measure in its Hausdorff dimension t so that the restriction $\mu_\Omega(\cdot) = \frac{h^t(\cdot)}{h^t(\Omega)} = \frac{\alpha(\cdot)}{h^t(\Omega)}$ to the σ-field of the h^t-measurable sets, of the upper unconditional probability defined in Theorem 3, is a countably additive probability. Moreover by Theorem 5 for every partition $\mathbf{B}$, there is at most a countable subclass $\mathbf{B}^*$ of $\mathbf{B}$ of sets B with positive upper coherent probability μ^*_Ω.

Since every random variable X and every constant c in $L(\Omega)$ are comonotonic, we consider the two comonotonic classes $\mathbf{C} = \left\{\overline{P}(X|\mathbf{B}), c\right\}$ and $\mathbf{C}_1 = \{X, c\}$ so that by Proposition 10.1 of [4] there exist two additive set functions α, and α' on $\wp(\Omega)$, which agree with h^t on the σ-field of h^t-measurable sets, such that

$$\int_\Omega \overline{P}(X|\mathbf{B})dh^t = \int_\Omega \overline{P}(X|\mathbf{B})d\alpha'$$

and

$$\int_B X dh^t = \int_\Omega I_B X dh^t = \int_\Omega I_B X d\alpha = \int_B X d\alpha.$$

Then for every random variable $X \in L(\Omega)$ the disintegration property is satisfied for every partition $\mathbf{B}$ since the following equalities hold:

$$\begin{aligned}
\overline{P}(\overline{P}(X|\mathbf{B})) &= \frac{1}{h^t(\Omega)} \int_\Omega \overline{P}(X|\mathbf{B})dh^t \\
&= \frac{1}{h^t(\Omega)} \int_\Omega \overline{P}(X|\mathbf{B})d\alpha' \\
&= \sum_{B\in\mathbf{B}^*} \left(\frac{1}{h^t(B)} \int_B X d\alpha\right) \frac{h^t(B)}{h^t(\Omega)} \\
&= \frac{1}{h^t(\Omega)} \sum_{B\in\mathbf{B}^*} \int_B X d\alpha \\
&= \frac{1}{h^t(\Omega)} \int_\Omega X dh^t = \overline{P}(X). \diamond
\end{aligned}$$

Remark. *If $X \in L(\Omega)$ is not monotone then X and $\overline{P}(X|\boldsymbol{B})$ are not comonotonic so that the additivity of the integral $\int_\Omega (X + \overline{P}(X|\boldsymbol{B}))dh^t$ does not hold. Since h^t is submodular by the Subadditive Theorem we have*

$$\int_\Omega (X + \overline{P}(X|\boldsymbol{B}))dh^t \leq \int_\Omega X dh^t + \int_\Omega \overline{P}(X|\boldsymbol{B})dh^t$$

In the following theorem a sufficient condition for the additivity of the Choquet integral with respect to h^t of the random variable $X + \overline{P}(X|\mathbf{B})$ is given.

Theorem 11. *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t and let $\boldsymbol{B}$ be a Borel countable partition of Ω. Thus*

$$\int_\Omega (X + \overline{P}(X|\boldsymbol{B}))dh^t = \int_\Omega X dh^t + \int_\Omega \overline{P}(X|\boldsymbol{B})dh^t$$

Proof. Since $\mathbf{B}$ is a countable partition, by Theorem 6, we have that the random variable $\overline{P}(X|\mathbf{B})$ is h^t-measurable. So [4, Corollary 10.2] we have

$$\int_\Omega (X + \overline{P}(X|\mathbf{B}))dh^t = \int_\Omega X dh^t + \int_\Omega \overline{P}(X|\mathbf{B})dh^t. \diamond$$

In the next theorem we prove that the coherent upper unconditional prevision defined as in Theorem 2 satisfies the conglomerability principle.

Theorem 12. *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t and let $\boldsymbol{B}$ be a partition of Ω. Then for every $X \in L(\Omega)$ the coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$, satisfies the conglomerability principle, i.e.*

$$\overline{P}(X|\boldsymbol{B}) \geq 0 \text{ implies } \overline{P}(X) \geq 0.$$

Proof By the coherence of the unconditional upper prevision $\overline{P}$ defined in Theorem 2 we have that if

$$\overline{P}(X|\mathbf{B}) \geq 0 \Rightarrow \overline{P}(\overline{P}(X|\mathbf{B})) \geq 0.$$

Moreover from Theorem 10 the disintegration property is satisfied, that is $\overline{P}(X) = \overline{P}(\overline{P}(X|\mathbf{B}))$.

So we have $\overline{P}(X|\mathbf{B}) \geq 0$ implies $\overline{P}(\overline{P}(X|\mathbf{B})) = \overline{P}(X) \geq 0. \diamond$

In the following example [10] unconditional and conditional probabilities are given such that they are not coherent. It occurs because they satisfy the Generalized Bayes Rule but not the conglomerative axiom. The previous results can be applied to show that the coherent unconditional and conditional previsions defined with respect to Hausdorff outer measures are coherent.

Example 2. *Let $\Omega = [0,1]^2$ and let E be a subset of Ω such that $P(E) = P(E^c) = \frac{1}{2}$. Let $\boldsymbol{B}$ be a countable partition of Ω such that for each $B_n \in \boldsymbol{B}$*

$P(EB_n) = \frac{1}{2^{n+1}}$ and $P(E^cB_n) = \frac{\epsilon}{2^{n+1}}$ with $\epsilon > 0$ so that $P(B_n) = \frac{1+\epsilon}{2^{n+1}}$.

The Generalized Bayes Rule holds since for each $B_n \in \boldsymbol{B}$ $P(E|B_n) = \frac{P(EB_n)}{P(B_n)} = \frac{1}{1+\epsilon}$ while the conglomerative axiom does not hold since

$$\begin{aligned}
P(P(E|B_n)) &= \sum_{B_n\in\boldsymbol{B}} \frac{1}{1+\epsilon} P(B_n) \\
&= \frac{1}{1+\epsilon} \neq \frac{1}{2} = P(E).
\end{aligned}$$

If coherent unconditional and conditional probabilities are defined as in Theorem 3, they are proven to be coherent.

Let $P(\cdot) = \mu_\Omega(\cdot) = \frac{h^2(\cdot)}{h^2(\Omega)}$ *the unconditional countably additive probability defined on the* σ*-field of the* h^2*-measurable subsets. Let* E *be a* h^2*-measurable subset of* Ω *such that* $P(E) = P(E^c) = \frac{1}{2}$ *and let* $\boldsymbol{B}$ *be a countable partition of* Ω *such that for each* $B_n \in \boldsymbol{B}$ *we have* $P(B_n) = \frac{h^2(B_n)}{h^2(\Omega)} > 0$.

The Generalized Bayes Rule and the conglomerative axiom hold since for each $B_n \in \boldsymbol{B}$

$$P(E|B_n) = \frac{P(EB_n)}{P(B_n)} = \frac{h^2(EB_n)}{h^2(B_n)}$$

and

$$P(P(E|B_n)) = \sum_{B_n \in \boldsymbol{B}} \frac{h^2(EB_n)}{h^2(B_n)} P(B_n)$$
$$= \frac{h^2(E)}{h^2(\Omega)} = P(E|\Omega) = P(E).$$

The last equalities hold since the Hausdorff 2-dimensional measure h^2 *is countably additive.*

Let Ω be an uncountable set with positive and finite Hausdorff outer measure in its Hausdorff dimension.

In Walley [21, Example 6.9.6] it is proven that when Ω is uncountable a countably additive probability P defined on a σ-field of subsets of Ω can extended to a fully conglomerable lower prevision taking the natural extension of P.

Definition 9. *A coherent lower prevision* $\underline{P}$ *on* $\boldsymbol{L}(\Omega)$ *is called* $\boldsymbol{B}$*-conglomerable when it satisfies the axiom: if* $X \in \boldsymbol{L}(\Omega)$ *and* $B_1, B_2, ...$ *are distinct sets in* $\boldsymbol{B}$ *such that* $\underline{P}(B_n) > 0$ *and* $\underline{P}(B_n X) \geq 0$ *for all* $n \geq 1$ *then* $\underline{P}(\sum_{n=1}^{\infty} B_n X) \geq 0$.

Definition 10. *A coherent lower prevision on* $\boldsymbol{L}(\Omega)$ *is called fully conglomerable if it is* $\boldsymbol{B}$*-conglomerable on every countable partition* $\boldsymbol{B}$ *of* Ω*. This holds if and only if [21, 6.8.1]* $\underline{P}$ *satisfies the axiom:*

if $X \in \boldsymbol{L}(\Omega)$ *and* $\boldsymbol{B}$ *is a countable partition of* Ω *such that* $\underline{P}(B) > 0$ *and* $\underline{P}(BX) \geq 0$ *for all* $B \in \boldsymbol{B}$ *then* $\underline{P}(X) \geq 0$

In the next theorem the unconditional upper prevision defined as in Theorem 2 is proven to be fully conglomerable if Ω is an uncountable set with positive and finite Hausdorff outer measure in its Hausdorff dimension.

Theorem 13. *Let* Ω *be an uncountable set with positive and finite Hausdorff outer measure in its dimension* s*. Let* $P(\cdot|\Omega)$ *be the restriction of the upper conditional probability defined in Theorem 3 to the Borel* σ*-field* $\boldsymbol{F}$ *of subsets of* Ω*. Then the upper conditional prevision* $\overline{P}(\cdot|\Omega)$ *defined on* $\boldsymbol{L}(\Omega)$ *as in Theorem 2 is fully conglomerable.*

Proof. Since every Hausdorff s-dimensional outer measure is countably additive on the Borel σ-field $\mathbf{F}$ of Ω and Ω is a set with positive and finite Hausdorff outer measure in its dimension s thus $P(A|\Omega) = \frac{h^s(A)}{h^s(\Omega)}$ is a countably additive probability on $\mathbf{F}$. The lower conditional prevision $\underline{P}(\cdot|\Omega)$ defined as in Theorem 2 is the natural extension of P to $\mathbf{L}(\Omega)$ where Ω is an uncountable set thus [21, 6.9.6] $\underline{P}(\cdot|\Omega)$ is fully conglomerable. From the conjugacy property $\overline{P}(X|\Omega) = -\underline{P}(-X|\Omega)$ we have that the upper conditional prevision is fully conglomerable. $\diamond$

If Ω has Hausdorff outer measure in its Hausdorff dimension equal to zero or infinity then the upper conditional previsions defined as in Theorem 3 do not satisfy the disintegration property as shown by the following example.

Example 3. *Let* $\Omega = N$*,* $A = \{2n, n \in N\}$ *and let* $\boldsymbol{B}$ *be the partition whose elements are the sets* $B_n = \{2n-1, 2n\}$ *for* $n \in N$*. If upper conditional previsions are defined as in Theorem 3 and* X *is the indicator function of* A *we have that*

$P(X|B_n) = \frac{1}{h^0(B_n)} \int_{B_n} X dh^0 = \frac{1}{2}$;

$P(P(X|B_n)) = \int_\Omega \frac{1}{2} dm = \frac{1}{2}$

$P(X|\Omega) = m(\cdot|\Omega) = m(A)$.

Since m *is a 0-1 valued finite probability measure the disintegration property is not satisfied because* $m(A) \neq \frac{1}{2}$.

In [9] the notions of equivalent and indifferent random variables given B are proposed.

Definition 11. *Two random variables* X *and* $Y \in L(B)$ *are equivalent given* B *if* $\overline{P}(X|B) = \overline{P}(Y|B)$.

A weak order on $L(B)$ is a complete reflexive and transitive binary relation on $L(B)$. Let X and Y be two bounded random variables belonging to $L(B)$.

Definition 12. *We say that* X *is preferable to* Y *given* B *, i.e.* $X \succ Y$ *given* B *if and only if*

$$\overline{P}((X - Y)|B) > 0$$

and X *and* Y *are indifferent given* B*, i.e.* $X \approx Y$ *given* B *if and only if*

$$\overline{P}((X - Y)|B) = \overline{P}((Y - X)|B) = 0.$$

By Theorem 8 a bounded random variable X is equivalent to $\overline{P}(X|\mathbf{B})$, moreover, by Theorem 9, X and $\overline{P}(X|\mathbf{B})$ are indifferent with respect to the ordering

represented by the coherent upper prevision $\overline{P}(\cdot|\Omega)$ if X is monotone or, by Theorem 11, if $\mathbf{B}$ is a Borel countable partition.

5 Conclusions

In this paper a coherent upper conditional prevision, defined as the Choquet integral with respect to its associated Hausdorff outer measure, is proven to satisfy the disintegration property and the conglomerative principle on every partition $\mathbf{B}$ of a metric space (Ω, d) where Ω is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension. This result is due to the fact that Hausdorff outer measures are submodular and continuous from below and a random variable and a constant are always comonotonic. Submodularity of Hausdorff outer measures implies that the Choquet integral with respect to Hausdorff outer measure of every random variable is equal to the integral with respect to an additive measure, which agrees with it on the σ-field of measurable sets. By the given results a random variable X is equivalent to the random variable $\overline{P}(X|\mathbf{B})$ and X and $\overline{P}(X|\mathbf{B})$ are indifferent given Ω with respect to the ordering represented by $\overline{P}(\cdot|\Omega)$ if X is monotone or if $\mathbf{B}$ is a Borel countable partition. A future aim of this research is to investigate the consequences in decision theory of the results proven in this paper.

Acknowledgments

The author is grateful to the reviewers for their useful comments.

References

[1] P. Billingsley, Probability and measure, New York, Wiley, (1986)

[2] B. de Finetti, Probability, Induction and Statistics, Wiley, New York, (1972)

[3] B. de Finetti, Theory of Probability, Wiley, London, (1974)

[4] D. Denneberg, Non-additive measure and integral, Kluwer Academic Publishers, (1994)

[5] S. Doria, Probabilistic independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures, International Journal of Approximate Reasoning, 46, 617-635, (2007)

[6] S. Doria, Coherent upper conditional previsions and their integral representation with respect to Hausdorff outer measures, In Combining Soft Computing and Statistical Methods in Data Analysis (C. Borgelt et al. editors), Advances in Intelligent and Soft Computing 77, 209-216, Springer, (2010)

[7] S. Doria, Coherent Upper and Lower Conditional Previsions Defined by Hausdorff Outer Measures, Modeling, Designs and Simulation of Systems with Uncertainties, Eds A. Rauh and E. Auer, Springer, 175-195, (2011)

[8] S. Doria, Characterization of a coherent upper conditional prevision as the Choquet integral with respect to its associated Hausdorff outer measure, Annals of Operations Research, 33-48, (2012)

[9] S. Doria, Symmetric coherent upper conditional prevision defined by the Choquet integral with respect to by Hausdorff outer measure, Annals of Operations Research, DOI 10.1007/s10479-014-1752-x, (2014)

[10] L.E. Dubins, Finitely additive conditional probabilities, conglomerability and disintegrations, The Annals of Probability, Vol. 3, 89-99, (1975)

[11] K.J. Falconer, The geometry of fractals sets, Cambridge University Press (1986)

[12] J.B. Kadane, M. J. Schervish, T. Seidenfeld, Statistical implications of finitely additive probability. In Bayesian Inference and Decision Techniques With Applications (P.Goel and A. Zellner, eds.), North-Holland, Amsterdam, 59-76, (1986)

[13] J. B. Kadane, M. J. Schervish, T. Seidenfeld, Is Ignorance Bliss? The Journal of Philosophy 105, (1), 5-36, (2008)

[14] E. Miranda, M. Zaffalon, G. de Cooman, Conglomerable natural extension, International Journal of Approximate Reasoning, Vol 53, Issue 8, 1200-1227, (2012)

[15] E. Regazzini, De Finetti's coherence and statistical inference, The Annals of Statistics, Vol 15, No. 2, 845-864, (1987)

[16] C.A. Rogers, Hausdorff measures, Cambridge University Press Mc Graw-Hill, Science/Engineering (1970)

[17] W. Rudin, Functional Analysis, c Graw-Hill, Science/Engineering/Math, (1991)

[18] M.J. Schervish, T. Seidenfeld, and J.B. Kadane, The extent of non-conglomerability of finitely additive probabilities. Z. Warsch.Verw.Gebiete 66, 205-226, (1984)

[19] R. Scozzafava, Probabilita' σ-additive a non, Bollettino U.M.I., (6) 1-A, 1-33, (1986)

[20] T. Seidenfeld, M.J. Schervish, and 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)

[21] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, (1991)

Coherent Conditional Measures of Risk Defined by the Choquet Integral with respect to Hausdorff Outer Measures and Dependent Risks

Serena Doria
Department of Engineering and Geology,
University G.d'Annunzio, Chieti-Pescara, Italy
s.doria@dst.unich.it

Abstract

Let (Ω, d) be a metric space and let $\mathbf{B}$ be a partition of Ω. For every set B of $\mathbf{B}$ with positive and finite Hausdorff outer measure in its Hausdorff dimension, a coherent conditional measure of risk is defined as the Choquet integral with respect to Hausdorff outer measure. Two risks are defined to be s-independent if the atoms of the classes generated by their weak upper level sets are s-independent. The given notion permits to capture dependence between risks that are stochastically independent according to the axiomatic definition. Two risks which are surjective and injective are proven to be s-dependent and a sufficient condition is given such that s-independent simple risks satisfy the factorization property of their joint coherent measures of risk.

Keywords. Coherent conditional measures of risk, Hausdorff outer measures, Choquet integral, stochastic dependence.

1 Introduction

Partial knowledge is a natural interpretation of conditional probability. This interpretation can be formalized in a different way in the axiomatic approach (see Billingsley [2]) and in the subjective approach (de Finetti [3], [4], Regazzini [19], Walley [21]), where conditional probability is respectively defined by the Radon-Nikodym derivative or by the axioms of coherence. In both cases conditional probability is obtained as the restriction of conditional expectation or conditional prevision to the class of indicator functions of events. For a comparison between the two different approaches see Doria [6]. In the axiomatic approach conditional expectation is defined with respect to a σ-field $\mathbf{G}$ of conditioning events by the Radon-Nikodym derivative while in the subjective approach proposed by Walley conditional prevision is defined with respect to a partition $\mathbf{B}$; the definitions of conditional expectation and coherent linear conditional prevision can be compared when the σ-field $\mathbf{G}$ is generated by the partition $\mathbf{B}$. In particular, given a probability space $(\Omega, \mathbf{F}, P)$, let $\mathbf{G}$ be equal or contained in the σ-field generated by a countable class S of subsets of $\mathbf{F}$ and let $\mathbf{B}$ be the partition of the atoms generated by the class S. Denote $\Omega' = \mathbf{B}$, $P(A|\mathbf{B})$ the class of all $P(A|B)$ with $B \in \mathbf{B}$ and φ_B the function from Ω to Ω' that associates to every $\omega \in \Omega$ the atom B of the partition $\mathbf{B}$ that contains ω. Then we have that $P(X|\mathbf{G}) = P(X|\mathbf{B}) \circ \varphi_B$ for every random variable $\in L(B)$ [15, p.262].

Let $\mathbf{F}$ and $\mathbf{G}$ be two σ-fields of subsets of Ω with $\mathbf{G}$ contained in $\mathbf{F}$ and let X be an integrable random variable on $(\Omega, \mathbf{F}, P)$. Let P be a probability measure on $\mathbf{F}$; define a measure ν on $\mathbf{G}$ by $\nu(G) = \int_G X dP$. This measure is finite and absolutely continuous with respect to P. So there exists a function, the Radon-Nikodym derivative denoted by $E[X|\mathbf{G}]$, defined on Ω, $\mathbf{G}$-measurable, integrable and satisfying the functional equation

$$\int_G E[X|\mathbf{G}]dP = \int_G X dP \text{ with } G \text{ in } \mathbf{G}.$$

This function is unique up to a set of P-measure zero and it is a version of the conditional expected value.

If X is the indicator function of any event A belonging to $\mathbf{F}$ then $E[X|\mathbf{G}] = E[A|\mathbf{G}] = P[A|\mathbf{G}]$ is a version of the conditional probability.

In Doria [8], [10], [11], [12] it has been proven that conditional expectation, defined by the Radon-Nikodym derivative may fail to be coherent and a new model of coherent conditional previsions, based on Hausdorff outer measures, has been introduced.

In [2, Example 33.11] it is shown that the interpretation of conditional probability in terms of partial knowledge breaks down in certain cases. A probability space $(\Omega, \mathbf{F}, P)$ can be used to represent a random phenomenon or an experiment whose outcome is drawn

according to the probability given by P. Partial information about the experiment can be represented by a sub σ-field $\mathbf{G}$ of $\mathbf{F}$ in the following way: an observer does not know which ω has been drawn but he knows for each $H \in \mathbf{G}$, if ω belongs to H or if ω belongs to H^c. A sub σ-field $\mathbf{G}$ of $\mathbf{F}$ can be identified as partial information about the random experiment, and, fixed A in $\mathbf{F}$, conditional probability can be used to represent partial knowledge about A given the information on $\mathbf{G}$.

A concept related to the definition of conditional probability is stochastic independence for events and for random variables based on the factorization property ([2, p.48]). In particular two random variables are stochastically independent, in the axiomatic approach, if the σ-fields generated by them are independent. As a consequence we obtain that for independent random variables the joint distribution is equal to the product of the marginal distributions.

In a probability space (Ω, F, P), if partial information is represented by a sub σ-field G and conditional probability is defined by the Radon-Nykodim derivative, denoted by $P[A|\mathbf{G}]$, by the standard definition [2, p.52] we have that an event A is independent from the σ-field $\mathbf{G}$ if it is independent from each $H \in \mathbf{G}$, that is $P[A|\mathbf{G}] = P(A)$ with probability 1.

If $\mathbf{G} = \{\Omega, \oslash\}$ then $P[A|\mathbf{G}](\omega) = P(A)$ for every $A \in \mathbf{F}$ and for every $\omega \in \Omega$.

Example 1 *Let Ω = [0,1], let $\boldsymbol{F}$ be the Borel σ-field of [0,1] and let P be the Lebesgue measure on $\boldsymbol{F}$. Let $\boldsymbol{G}$ be the sub σ-field of sets that are either countable or co-countable. Then $P(A)$ is a version of the conditional probability $P[A|\boldsymbol{G}]$ defined by the Radon-Nikodym derivative because $P(G)$ is either 0 or 1 for evey $G \in \boldsymbol{G}$. So an event A is independent from the information represented by $\boldsymbol{G}$ and this is a contradiction according to the fact that the information represented by $\boldsymbol{G}$ is complete since $\boldsymbol{G}$ contains all singletons of Ω.*

In the subjective approach the concept of epistemic independence with respect to upper and lower probabilities has proposed by Walley [21]. It is based on the notion of irrelevance; given two events A and B, we say that B is irrelevant to A when $\overline{P}(A|B) = \overline{P}(A|B^c) = \overline{P}(A)$ and $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$.

The events A and B are epistemically independent when B is irrelevant to A and A is irrelevant to B. As a consequence of this definition we can obtain that the factorization property $P(AB) = P(A)P(B)$, which constitutes the standard definition of independence for events, holds both for $P = \overline{P}$ and $P = \underline{P}$. In a continuous probabilistic space (Ω, F, P), where Ω is equal to $[0,1]^n$ and the probability is usually assumed equal to the Lebesgue measure on Ω, we have that the finite, countable and fractal sets (i.e. the sets with Hausdorff dimension non- integer) have probability equal to zero. For these sets the standard definition of independence, given by the factorization property, is always satisfied since both members of the equality are zero. In Theorem 6 of [9] we prove that an event B is always irrelevant, according to the definition of Walley, to an event A if $dim_H(A) < dim_H(B) < dim_H(\Omega)$ and A and B have positive and finite Hausdorff outer measures in their dimensions; moreover if A and B are disjoint then they are epistemically independent. Thus logical independence is not a necessary condition for epistemic independence.

To avoid these problems the notions of s-irrelevance and s-independence with respect to upper and lower conditional probabilities assigned by a class of Hausdorff outer and inner measures are proposed to test independence ([7], [8], [9]). The definitions of s-independence and s-irrelevance are based on the fact that epistemic independence and irrelevance must be tested for events A and B, such that they and their intersection AB, have the same Hausdorff dimension.

In this paper coherent conditional measures of risk are defined equal to coherent upper conditional previsions defined by Hausdorff measures when the conditioning event has positive and finite Hausdorff outer measure in its Hausdorff dimension. The notion of *s*-irrelevance and *s*-independence for risks, represented by bounded random variables, are proposed to capture dependence.

2 Coherent Conditional Measures of Risk Defined by the Choquet Integral with respect to Hausdorff Outer Measures

Let (Ω, d) be a metric space and let $\mathbf{B}$ be a partition of Ω. For every $B \in \mathbf{B}$ with positive and finite Hausdorff outer measure in its Hausdorff dimension s a coherent conditional measure of risk is defined by the Choquet integral with respect to Hausdorff outer measure h^s.

2.1 Coherent Upper Conditional Previsions Defined by the Choquet Integral with respect to Hausdorff Outer Measures

A *risk* or *bounded random variable* is a function $X : \Omega \to \Re$ and $L(\Omega)$ is the class of all bounded random variables defined on Ω. For every $B \in \mathbf{B}$ denote by $X|B$ the restriction of X to B and by $\sup(X|B)$ the supremum value that X assumes on B. Let $L(B)$ be the class of all bounded random variables $X|B$. Denote by I_A the indicator function of any event $A \in \wp(B)$, i.e. $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ if $\omega \in A^c$.

Let $\delta > 0$ and let s be a non-negative number. The *diameter* of a non empty set U of Ω is defined as $|U| = \sup\{d(x,y) : x, y \in U\}$ and if a subset A of Ω is such that $A \subseteq \bigcup_i U_i$ and $0 < |U_i| \leq \delta$ for each i, the class $\{U_i\}$ is called a δ-cover of A.

The *Hausdorff s-dimensional outer measure* of A, denoted by $h^s(A)$, is defined on $\wp(\Omega)$, the class of all subsets of Ω, as

$$h^s(A) = \lim_{\delta \to 0} \inf \sum_{i=1}^{+\infty} |U_i|^s.$$

where the infimum is over all δ-covers $\{U_i\}$ of the set A.

For the definition of Hausdorff outer measure and its basic properties see Rogers [20] and Falconer [14].

The *Hausdorff dimension* of a set A, $dim_H(A)$, is defined as the unique value, such that

$$h^s(A) = +\infty \text{ if } 0 \leq s < dim_H(A),$$
$$h^s(A) = 0 \text{ if } dim_H(A) < s < +\infty.$$

If $0 < h^s(A) < +\infty$ then $dim_H(A) = s$ (the converse is not true). In any metric space a finite non-empty subset A of Ω has positive and finite counting measure h^0 so the Hausdorff dimension of a finite set is 0.

We assume that the Hausdorff dimension of the empty-set $\oslash$ is $-\infty$ so that no set has Hausdorff dimension equal to the Hausdorff dimension of the empty-set.

Hausdorff s-dimensional outer measures are submodular, continuous from below and their restrictions to the Borel σ-field which is the σ-field generated by open sets of the *metric topology*, induced by the metric d, are countably additive.

If $B \in \mathbf{B}$ is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension s the monotone set function μ^*_B is defined for every $A \in \wp(B)$ by $\mu^*_B(A) = \frac{h^s(AB)}{h^s(B)}$.

If X is a bounded random variable thus there exists a constant k such that $\widetilde{X} = X + k \geq 0$ and the *decreasing distribution function* of $\widetilde{X}$ with respect to μ^*_B is $G_{\mu^*_B,\widetilde{X}}(x) = G_{\mu^*_B,X}(x-k) = \mu^*_B\{X|B > x-k\}$ for every real number x [5, Proposition 4.1].

The Choquet integral [5] of a bounded random variable X with respect to μ^*_B is defined by

$$\int X d\mu^*_B = \int_0^{+\infty} \mu^*_B\left\{\omega \in B : \widetilde{X}(\omega) \geq x\right\} dx$$
$$= \frac{1}{h^s(B)} \int_0^{+\infty} h^s\{\omega \in B : X(\omega) \geq x\}\, dx.$$

If B is finite then μ^*_B is the counting measure defined on the field $\wp(B)$. If the atoms A_i are enumerated so that $x_i = X(A_i)$ are in in descending order, i.e. $x_1 \geq x_2 \geq ... \geq x_n$ and $x_{n+1} = 0$ the Choquet integral with respect to μ is given by

$$\int X d\mu = \sum_{i=1}^{n} (x_i - x_{i+1}) \mu^*_\Omega(S_i)$$

where $S_i = A_1 \cup A_2 ... \cup A_i$, and $x_{n+1} = 0$.

In [12] a model of coherent upper conditional prevision based on Hausdorff outer measure has been defined.

Theorem 1 *Let m be a 0-1 valued finitely additive, but not countably additive, probability on $\wp(B)$ such that a different m is chosen for each B. Then for each $B \in \boldsymbol{B}$ denote by s the Hausdorff dimension of B and by h^s the Hausdorff s-dimensional outer measure. The functionals $\overline{P}(X|B)$ defined on $L(B)$ by*

$$\overline{P}(X|B) = \frac{1}{h^s(B)} \int_B X dh^s \text{ if } 0 < h^s(B) < +\infty$$

and by

$$\overline{P}(X|B) = m(XB) \text{ if } h^s(B) = 0, +\infty$$

are separately coherent upper conditional previsions.

Definition 1 *Given a partition $\boldsymbol{B}$ of Ω and a random variable $X \in L(\Omega)$ a coherent upper conditional prevision $\overline{P}(X|\boldsymbol{B})$ is a random variable on Ω equal to $\overline{P}(X|B)$ if $\omega \in B$. The random variable $\overline{P}(X|\boldsymbol{B})$ is separately coherent if all the $\overline{P}(X|B)$ are separately coherent.*

Definition 2 *Given a partition $\boldsymbol{B}$ of Ω and a random variable $X \in L(\Omega)$, X is $\boldsymbol{B}$-measurable if it is constant on the sets of the partition $\boldsymbol{B}$.*

By Definition 1 the random variable $\overline{P}(X|\mathbf{B})$ is $\mathbf{B}$-measurable since it is constant on the atoms of the partition $\mathbf{B}$.

2.2 Coherent Conditional Measures of Risk

Coherent measures of risk are introduced in Artzener et al. [1] to manage risks.

Correspondence between the concepts of upper prevision and coherent measure of risk has been underlined by Pelessoni and Vicig [17] and Maaß [16].

In Pelessoni and Vicig [18] risk measures have been interpreted as coherent conditional previsions.

In this section, given a set $B \in \mathbf{B}$ with positive and finite Hausdorff outer measure in its Hausdorff dimension s, a coherent conditional measure of risk $\rho(\cdot|B)$

is defined as the Choquet integral with respect to Hausdorff outer measure h^s.

The advantage to define coherent measures of risk by Hausdorff outer measures is that they can be represented as Choquet integral since Hausdorff outer measures are submodular (see Doria [13, Proposition 1]); moreover coherent measures of risk defined with respect to Hausdorff outer measures are comonotonically additive and continuous from below (see Theorem 2).

Definition 3 *A coherent conditional measure of risk $\rho(\cdot|B)$ is a functional on $L(B)$ such that the following axioms are satisfied for every $X, Y \in L(B)$ and strictly positive λ:*

(i) monotonicity $X \leq Y$ implies $\rho(X|B) \leq \rho(Y|B)$;

(ii) translation invariance $\rho(X + h|B) = \rho(X|B) + h$;

(iii) subadditivity $\rho(X + Y|B) \leq \rho(X|B) + \rho(Y|B)$

(iv) positive homogeneity $\rho(\lambda X|B) = \lambda\rho(X|B)$

Two risks X and $Y \in L(B)$ are *comonotonic* if,

$$(X(\omega_1) - X(\omega_2))(Y(\omega_1) - Y(\omega_2)) \geq 0 \quad \forall \omega_1, \omega_2 \in B.$$

Definition 4 *A coherent conditional measure of risk $\rho(\cdot|B)$ on $L(B)$ is*

(v) comonotonically additive if and $\rho(X + Y|B) = \rho(X|B) + \rho(Y|B)$ for every comonotonic risks X and Y;

(vi) continuous from below if $lim_{n\to\infty}\rho(X_n|B) = \rho(X|B)$ if X_n is an increasing sequence of risks $\in L(B)$ converging to X.

In [12] it has been proven that if B is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension, the coherent upper conditional prevision defined in Theorem 1 is comonotonically additive and continuous from below. So the following result holds:

Theorem 2 *Let $B \in \boldsymbol{B}$ be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension s, i.e. $0 < h^s(B) < +\infty$; the functional $\rho(\cdot|B)$ defined on $L(B)$ by*

$$\rho(X|B) = \overline{P}(X|B) = \frac{1}{h^s(B)} \int_B X dh^s$$

is a coherent conditional measure of risk, which is comonotonically additive and continuous from below.

3 S-Independence for Risks

In Doria [7], [8], [9] the notions of s-irrelevance and s-independence with respect to conditional probabilities assigned by a class of Hausdorff dimensional measures have been introduced.

Definition 5 *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension and let E and $F \in \wp(\Omega)$. E and F are s-independent if the following conditions hold*

(s1) $dim_H E = dim_H F = dim_H E \cap F$,

(s2) $\overline{P}(E|F) = \overline{P}(E|F^c) = \overline{P}(E)$,

$\underline{P}(E|F) = \underline{P}(E|F^c) = \underline{P}(E)$,

(s3) $\overline{P}(F|E) = \overline{P}(F|E^c) = \overline{P}(F)$,

$\underline{P}(F|E) = \underline{P}(F|E^c) = \underline{P}(F)$,

F is s-irrelevant to E if conditions s1) and s2) hold and E is s-irrelevant to F if the conditions s1) and s3) hold.

We can observe that the notion of s-irrelevance is not symmetric, that is E can be s-irrelevant to F because conditions (s1) and (s2) hold but F is not s-irrelevant to E since condition (s3) does not hold.

In the sequel we assume that Ω is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension.

In [8, Theorem 4] the following result is proven:

Corollary 1 *If E is s-irrelevant to F then $\overline{P}$ satisfies the factorization property, i.e. $\overline{P}(E \cap F) = \overline{P}(E)\overline{P}(F)$*

Theorem 3 *Let $F \in \wp(\Omega)$ then F and Ω are s-independent if and only if $dim_H(F) = dim_H(\Omega)$.*

Proof. If E is equal to Ω conditions $(s1), (s2), (s3)$ become

$(s1)$ $dim_H(\Omega) = dim_H(F)$,

$(s2)$ $\overline{P}(\Omega|F) = \overline{P}(\Omega|F^c) = \overline{P}(\Omega)$,

$\underline{P}(\Omega|F) = \underline{P}(\Omega|F^c) = \underline{P}(\Omega)$,

$(s3)$ $\overline{P}(F|\Omega) = \overline{P}(F)$,

$\underline{P}(F|\Omega) = \underline{P}(F)$.

Since conditions $(s2), (s3)$ are always satisfied then Ω and F are s-independent if and only if $dim_H(\Omega) = dim_H(F)$. $\diamond$

Remark 1 *The previous result permits to put in evidence dependence between events. In fact, according to the axiomatic definition of stochastic independence, given a probability space* $(\Omega, \boldsymbol{F}, P)$ *and the* σ*-field* $\boldsymbol{G} = \{\oslash, \Omega\}$*, if conditional probability is defined by the Radon-Nikodym derivative we have that* $P(F|\boldsymbol{G}) = P(F)\ \forall F \in \boldsymbol{F}$*, that is any event* F *is independent from* $\boldsymbol{G}$*.*

If Ω is a finite set and F is a non-empty set $\in \wp(\Omega)$, Ω and F are s-independent since any non-empty finite set has Hausdorff dimension equal to 0.

We extend the notions of s-irrelevance and s-independence to risks. Let S be a subclass of $\wp(\Omega)$ closed under intersection. The *atoms* of S are the minimal sets with respect to inclusion in $S - \{\oslash\}$.

Definition 6 *Let* $S = \{C_i\}_{i \in N}$ *be a countable class, the constituents of* S *are the sets* $C = \bigcap_{i \in N} \widetilde{C}_i$ *where* $\widetilde{C}_i = C_i$ *or* $\widetilde{C}_i = C_i^c$*.*

Definition 7 *The partition* $\boldsymbol{C}(S)$ *generated by a countable class* $S \subseteq \wp(\Omega)$ *is the partition of its constituents.*

Definition 8 *The* σ*-field generated by a class* $S \subseteq \wp(\Omega)$ *is the smallest* σ*-field* $\sigma(S)$ *containing* S*.*

In [15, Proposition 4.30] the following result has been proved.

Proposition 1 *Let* S *be a countable class of subsets of* Ω *and* $\boldsymbol{G}$ *a* σ*-field such that the class of the constituents* $\boldsymbol{C}(S) \subseteq \boldsymbol{G} \subseteq \sigma(S)$*. The atoms of* $\boldsymbol{G}$ *are the constituents of* S*.*

Remark 2 *If* (Ω, d) *is the Euclidean metric space the* σ*-field generated by the class* $\{[x, +\infty); x \in \Re\}$ *is the Borel* σ*-field, which can be also generated by the countable class* $\{[x, +\infty); x \in Q\}$*. The class of the singletons* $\{x\}$ *in* $\Re$ *is contained in the Borel* σ*-field but it does not generate it. So the Borel* σ*-field is countably generated.*

Definition 9 *Let* $X \in L(\Omega)$*. The partition* $\boldsymbol{B}(X)$ *generated by the random variable* X *is the partition of the non-empty constituents generated by the countable class* $S = \left\{X^{-1}[x, +\infty); x \in Q\right\}$*. It is countably generated.*

Proposition 2 *We have that*

$$\boldsymbol{B}(X) = \left\{X^{-1}\{x\}; x \in \Re\right\} - \{\oslash\}.$$

Example 2 *Let* $X = I_A$ *be the indicator function of a set* $A \subseteq \Omega$*, then*

$$X^{-1}[x, +\infty) = \Omega \text{ if } x \leq 0,$$
$$X^{-1}[x, +\infty) = A \text{ if } 0 < x \leq 1,$$
$$X^{-1}[x, +\infty) = \oslash \text{ if } x > 1.$$

Thus $S = \left\{X^{-1}[x, +\infty); x \in \Re^+\right\} = \{\Omega, A, \oslash\}$ *and the class of atoms* $S(X)$ *generated by the random variable* X *is* $S(X) = \{A\}$*. The partition generated by the random variable* X *is is* $\boldsymbol{B}(X) = \boldsymbol{C}(S) - \{\oslash\} = \{A, A^c\}$

Definition 10 *Two classes of events* S_1 *and* S_2 *are s-independent if for every* $E \in S_1$ *and* $F \in S_2$ *the events* E *and* F *are s-independent.*

The class S_2 *is s- irrelevant to the class* S_1 *if for every* $E \in S_1$ *and every* $F \in \boldsymbol{S}_2$ F *is s-irrelevant to* E*.*

The class S_1 *is s-irrelevant to the class* S_2 *if for every* $E \in S_1$ *and every* $F \in S_2$ E *is s-irrelevant to* F*.*

Definition 11 *Given a risk* X *the class of the weak upper level sets of* X *is* $\left\{X^{-1}\{x\}; x \in \Re\right\}$*. Let* X *and* $Y \in L(\Omega)$ *be two risks and let* $S(X)$ *and* $S(Y)$ *be the classes of atoms generated respectively by the class of the weak upper level sets of the risks* X *and* Y*. Then*

- X *and* Y *are s-independent if* $S(X)$ *and* $S(Y)$ *are s-independent;*
- Y *is s-irrelevant to* X *if* $S(Y)$ *is s-irrelevant to* $S(X)$*;*
- X *is s-irrelevant to* Y *if* $S(X)$ *is s-irrelevant to* $S(Y)$*.*

Example 3 *Let* $\Omega = [0, 1]$*, let* $E = [0; \frac{1}{3}]$*,* $E_1 = [\frac{2}{3}; 1]$ $E_2 = [\frac{1}{3}; \frac{2}{3}]$ *and let* X *and* $Y \in L(\Omega)$ *be two risks defined by* $Y(\omega) = K$ *and* $X(\omega) = 1$ *if* $\omega \in E$*,* $X(\omega) = 2$ *if* $\omega \in E_1$ *and* $X(\omega) = 0$ *if* $\omega \in E_2$ *.*

Then

$$X^{-1}[x, +\infty) = \Omega \text{ if } x \leq 0,$$
$$X^{-1}[x, +\infty) = E \cup E_1 \text{ if } 0 < x \leq 1,$$
$$X^{-1}[x, +\infty) = E_1 \text{ if } 1 < x \leq 2,$$
$$X^{-1}[x, +\infty) = \oslash \text{ if } x > 2.$$

So $S(X) = \{E_1\}$ *and* $S(Y) = \{\Omega\}$*. By Theorem 3* E_1 *and* Ω *are s-independent since* $dim_H E_1 = dim_H \Omega = 1$ *so that* X *and* Y *are s-independent.*

Theorem 4 *Let* $X = I_A$ *and* $Y = I_E$ *be the indicator functions of two sets* $A, E \subseteq \Omega$*, then* Y *is s-irrelevant to* X *if and only if* E *is s-irrelevant to* A*.*

Proof. By Definition 11 Y is s-irrelevant to X if and only if $S(Y)$ is s-irrelevant to $S(X)$. Since $S(Y) = \{E\}$ and $S(X) = \{A\}$ then Y is s-irrelevant to X if and only if E is s-irrelevant to A.$\diamond$

Remark 3 *S-independence of the indicator functions of two events A and E does not imply s-independence of the indicator functions of their complements because the classes of atoms respectively generated by $X = I_A$ and $Y = I_E$ do not contain the sets A^c and E^c. It is the main difference with the standard definition of independence, according to which two random variables are independent if and only if the σ-field generated by them are independent. Since the σ-fields generated by I_A and by I_{A^c} are equal to $\sigma(X) = \{\Omega, A, A^c, \oslash\}$ and the σ-fields generated by I_E and I_{E^c} are equal to $\sigma(Y) = \{\Omega, E, E^c, \oslash\}$, independence of I_A and I_E implies that I_{A^c} and I_{E^c} are independent.*

In [9, Theorem 9] it has been proven that curves filling the space like Peano curve and Hilbert curve are s-independent, so that by Definition 5 and Theorem 4 the indicator functions of curves filling the space are s-independent.

Example 4 *Let $\Omega = [0,1]$ and let X and $Y \in L(\Omega)$ be two risks defined by $Y(\omega) = K$ and $X(\omega) = 1$ if $\omega \in Q \cap [0,1]$ and $X(\omega) = 0$ otherwise.*

Then

$$X^{-1}[x, +\infty) = \Omega \text{ if } x \leq 0,$$
$$X^{-1}[x, +\infty) = Q \cap [0,1] \text{ if } 0 < x \leq 1,$$
$$X^{-1}[x, +\infty) = \oslash \text{ if } x > 1.$$

We have that $S(X) = \{Q \cap [0,1]\}$ and $S(Y) = \{\Omega\}$. By Theorem 3 $Q \cap [0,1]$ and Ω are s-dependent since

$$dim_H Q \cap [0,1] = 0 \neq 1 = dim_H \Omega$$

and so X and Y are s-dependent.

According to the axiomatic definition of stochastic independence X and Y are independent since the σ-field $\mathbf{G}(Y)$ generated by Y is $\mathbf{G}(Y) = \{\Omega, \oslash\}$ so that $P[A|\mathbf{G}(Y)] = P(A)$ with probability 1 for every A belonging to the σ-field generated by X.

Theorem 5 *Let $X, Y \in L(\Omega)$ and let $\boldsymbol{B}(X)$ and $\boldsymbol{B}(Y)$ be the partitions generated by them. A sufficient condition for s-irrelevance of Y to X is that $\boldsymbol{B}(Y)$ is s-irrelevant to $\boldsymbol{B}(X)$.*

Proof. The class of atoms is contained in the partition generated by a random variable, that is $S(X) \subseteq \mathbf{B}(X)$ and $S(Y) \subseteq \mathbf{B}(Y)$ $\forall X, Y \in L(\Omega)$, so if $\mathbf{B}(Y)$ is s-irrelevant to $\mathbf{B}(X)$ it implies that $S(Y)$ is s-irrelevant to $S(X)$. $\diamond$

Theorem 6 *Let $X, Y \in L(\Omega)$. A necessary condition for s-irrelevance of Y to X is that, for every $E \in \boldsymbol{S}(X)$, for every $F \in \boldsymbol{S}(Y)$ and $\omega \in F$ the following equality holds*

$$\overline{P}(E|\boldsymbol{B}(Y))(\omega) = \overline{P}(E|F) = \overline{P}(E)$$

Proof. Let Y be s-irrelevant to X; then by conditions s2) and s3) of Definition 4 we have that $\overline{P}(E|F) = \overline{P}(E)$ for every $E \in \mathbf{S}(X)$ and $F \in \mathbf{S}(Y)$, that is for $\omega \in \Omega$ with $\omega \in F$

$$\overline{P}(E|\mathbf{B}(Y))(\omega) = \overline{P}(E|F) = \overline{P}(E). \quad \diamond$$

Lemma 1 *Let $X, Y \in L(\Omega)$ be two risks such that X is $\boldsymbol{B}(Y)$-measurable then $\forall E \in S(X)$ we have that I_E is $\boldsymbol{B}(Y)$-measurable.*

Proof. Since X is $\mathbf{B(Y)}$-measurable, the sets of the partition $\mathbf{B}(X) = \left\{X^{-1}(x); x \in \Re\right\}$ are union of sets belonging to $\mathbf{B(Y)}$ and for every $E \in \mathbf{B}(X)$ the indicator function I_E is $\mathbf{B}(Y)$-measurable. Moreover $S(X) \subseteq \mathbf{B}(X)$ so the Lemma is proven.$\diamond$

Theorem 7 *Let $X, Y \in L(\Omega)$ such that X is $\boldsymbol{B}(Y)$-measurable and $S(X) \neq \{\Omega\}$ and $S(Y) \neq \{\Omega\}$. Then Y is s-relevant to X.*

Proof. Since X is $\boldsymbol{B}(Y)$-measurable by the coherence of $\overline{P}$ [21, property (f), p. 292] we have $\overline{P}(X|\boldsymbol{B}(Y)) = X$.

Let $B \in \mathbf{B}(Y)$, for every $\omega \in \Omega$ with $\omega \in B$ we have

$$\overline{P}(X|\boldsymbol{B}(Y))(\omega) = \overline{P}(X|B) = X(\omega).$$

Since X is $\mathbf{B}(Y)$-measurable by Lemma 1, for every $E \in \mathbf{B}(X)$ the indicator function I_E is $\mathbf{B}(Y)$-measurable so that

$$\overline{P}(I_E|\boldsymbol{B}(Y)) = I_E$$

so, since $S(X) \subseteq \mathbf{B}(X)$ and $S(X) \neq \{\Omega\}$, the necessary condition for s-irrelevance of Y to X given in Theorem 6 is not satisfied and Y is s-relevant to X.$\diamond$

The previous theorem does not hold if $S(X) = \{\Omega\}$ or $S(Y) = \{\Omega\}$.

Example 5 *Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and let $X, Y \in L(\Omega)$ such that $X(\omega_1) = 1, X(\omega_2) = 2, X(\omega_3) = 3$ and $Y(\omega_i) = 1$ for $i = 1, ..., 3$. Then*

$$X^{-1}[x, +\infty) = \Omega \text{ if } x \leq 1,$$
$$X^{-1}[x, +\infty) = \{\omega_2, \omega_3\} \text{ if } 1 < x \leq 2,$$
$$X^{-1}[x, +\infty) = \{\omega_3\} \text{ if } 2 < x \leq 3,$$
$$X^{-1}[x, +\infty) = \oslash \text{ if } x > 3.$$

and

$$Y^{-1}[x,+\infty) = \Omega \text{ if } x \leq 1,$$
$$Y^{-1}[x,+\infty) = \oslash \text{ if } x > 1.$$

So we have that $S(X) = \{\omega_3\}$ *and* $S(Y) = \{\Omega\}$ *and by Definition 11 and Theorem 3* X *and* Y *are s-independent since* $dim_H\{\omega_3\} = 0 = dim_H\Omega$.

Proposition 3 *Let* $X, Y \in L(\Omega)$ *such that* $X(\omega) = k$ *and* $Y(\omega) = h$ *for every* $\omega \in \Omega$ *Then* X *and* Y *are s-independent.*

Proof. X and Y are constant thus $S(X) = S(Y) = \{\Omega\}$ and by Definition 11 X and Y are s-independent.$\diamond$

Example 6 *Let* $X, Y \in L(\Omega)$ *such that* $\boldsymbol{B}(Y)$ *is the partition of singletons of* Ω *and* $S(X) \neq \{\Omega\}$. *Then* Y *is s-relevant to* X *since* X *is* $\boldsymbol{B}(Y)$*-measurable and from Theorem 7 we have that* Y *is s-relevant to* X.

4 Dependent Risks

In this section sufficient conditions for s-dependence for risks are given. Surjective strictly monotone risks are proven to be s-dependent and every risk is proven to be s-dependent on a bijective risk.

Theorem 8 *Let* X *and* $Y \in L(\Omega)$ *be two risks and let* $S(X)$ *and* $S(Y)$ *be the classes of atoms generated respectively by the class of the weak upper level sets of the risks* X *and* Y. *If there exist* $A \in S(X)$ *and* $F \in S(Y)$ *such that* $dim_H A \neq dim_H F$ *then* X *and* Y *are s-dependent.*

Proof. Let $A \in S(X)$ and $F \in S(Y)$ such that $dim_H A \neq dim_H F$ so A and F are s-dependent since condition $s1$) of Definition 5 is not satisfied. Thus the classes $S(X)$ and $S(Y)$ are s-dependent and by Definition 10 X and Y are s-dependent. $\diamond$

Example 7 *Let* $([0,1], d)$ *be the Euclidean metric space and let* $X, Y \in L(\Omega)$ *be two risks defined by* $X(\omega) = 1$ *if* $0 < \omega < \frac{1}{2}, X(\omega) = 2$ *if* $\omega = \frac{1}{2}$ *and* $X(\omega) = 0$ *otherwise,* $Y(\omega) = 1$ *if* $0 < \omega < \frac{1}{2}$, $Y(\omega) = \frac{1}{2}$ *if* $\omega \geq \frac{1}{2}$;

So we have

$$X^{-1}[x,+\infty) = \Omega \text{ if } x \leq 0,$$
$$X^{-1}[x,+\infty) = [0,\frac{1}{2}] \text{ if } 0 < x \leq 1,$$
$$X^{-1}[x,+\infty) = \left\{\frac{1}{2}\right\} \text{ if } 1 < x \leq 2,$$
$$X^{-1}[x,+\infty) = \oslash \text{ if } x > 2,$$

and

$$Y^{-1}[x,+\infty) = \Omega \text{ if } x \leq \frac{1}{2},$$
$$Y^{-1}[x,+\infty) = [0,\frac{1}{2}] \text{ if } \frac{1}{2} < x \leq 1,$$
$$Y^{-1}[x,+\infty) = \oslash \text{ if } x > 1.$$

Thus $S(X) = \left\{\left\{\frac{1}{2}\right\}\right\}$ *and* $S(Y) = \left\{[0,\frac{1}{2}]\right\}$ *and by Theorem 8 the risks* X *and* Y *are s-dependent. We can observe that the events belonging to* $S(X)$ *and* $S(Y)$ *satisfy the factorization property with respect to* μ^*_Ω, *which is the Lebesgue measure* h^1 *since the Hausdorff dimension of* $\Omega = [0,1]$ *is 1. In fact the following equalities hold:*

$$h^1(\left\{\frac{1}{2}\right\})h^1([0,\frac{1}{2}]) = 0 = h^1(\left\{\frac{1}{2}\right\} \cap [0,\frac{1}{2}]).$$

Theorem 9 *Let* $([0,1], d)$ *be the Euclidean metric space and let* $X, Y \in L(\Omega)$ *be two risks such that* Y *is bijective. Then* X *ad* Y *are s-dependent.*

Proof. Since Y is bijective, the partition generated by Y is $\mathbf{B}(Y) = \left\{Y^{-1}\{x\}; x \in \Re\right\} - \{\oslash\}$, that is the partition of singletons of $[0,1]$. So for any risk X condition $(s1)$ of Definition 5 is not satisfied for every $F \in S(X)$, then we have that X and Y are s-dependent. $\diamond$

Corollary 2 *Let* $([0,1], d)$ *be the Euclidean metric space and let* $X, Y \in L(\Omega)$ *be two surjective and strictly monotone random variables. Then* X *and* Y *are s-dependent.*

Proof. Since X and Y are surjective and strictly monotone the partition generated by them is the partition of singletons of [0,1]; from the fact that strictly monotonicity implies injectivity by Theorem 9 X is s-relevant to Y and Y is s-relevant to X; so X and Y are s-dependent.$\diamond$

Theorem 10 *Let* $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ *and let* $X, Y \in L(\Omega)$ *be two risks such that* $Y : \{\omega_1, \omega_2, ..., \omega_n\} \to \{x_1, x_2, ..., x_n\}$ *is injective. Then* X *and* Y *are s-dependent.*

Proof. Since Y is surjective if and only if it is injective, so the partition generated by Y is $\mathbf{B}(Y) = \left\{Y^{-1}\{x\}; x \in \Re\right\} - \{\oslash\}$, that is the partition of singletons of Ω. So for any risk X condition $(s1)$ of Definition 5 is not satisfied for every $F \in S(X)$, then we have that X and Y are s-dependent. $\diamond$

5 Joint Coherent Conditional Measure of Risk

Let X and Y be two risks belonging to $L(\Omega)$. In this section the joint measure of risk $\rho(X,Y)$ of X and Y is defined and some properties are proven.

Definition 12 *Let X and Y be two risks belonging to $L(\Omega)$ and let $\boldsymbol{B}(X)$ be the partition generated by X. The coherent upper conditional prevision of Y given X, denoted by $\overline{P}(Y|X)(\omega)$ is the random variable on Ω defined by*

$$\overline{P}(Y|X)(\omega) = \overline{P}(Y|\boldsymbol{B}(X)) = \overline{P}(Y|B)$$

if $\omega \in B$ and $B \in \boldsymbol{B}(X)$.

Moreover if B has positive and finite Hausdorff outer measure in its Hausdorff dimension s the coherent conditional measure of risk $\rho(Y|X)$ of Y given X is defined by

$$\rho(Y|X) = \overline{P}(Y|\boldsymbol{B}(X)) = \overline{P}(Y|B) = \int_B Y d\mu_B^*.$$

If Y is $\mathbf{B}(X)$-measurable (i.e. it is constant on the sets of $\mathbf{B}(X)$) then $\overline{P}(Y|X) = Y$.

Given a risk X let $A = X^{-1}(A')$ be the inverse image of A' for every Borelian set A' of $\Re$.

Definition 13 *Given a risk $X \in L(\Omega)$ and denoted by t the Hausdorff dimension of Ω, the coherent upper probability $\overline{P}_X$ induced by X on $\Re$ is defined by*

$$\overline{P}_X(A') = \overline{P}(A) = \frac{h^t\{\omega \in \Omega : X(\omega) \in A'\}}{h^t(\Omega)}$$

for every Borelian set A' of $\Re$.

Definition 14 *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension t. Given two risks $X, Y \in L(\Omega)$ the joint coherent upper probability $\overline{P}_{(X,Y)}$ induced by the pair (X,Y) on $\Re \times \Re$ is defined by*

$$\begin{aligned}\overline{P}_{(X,Y)}(A' \times B') &= \overline{P}(A \cap B) \\ &= \frac{h^t\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A' \times B'\}}{h^t(\Omega)}\end{aligned}$$

for every pair of Borelian sets A' and B' of $\Re$.

Definition 15 *Given two risks $X, Y \in L(\Omega)$ such that the Hausdorff dimension of the inverse image $A = X^{-1}(A')$ is s and $0 < h^s(A) < +\infty$ the joint coherent upper probability $\overline{P}_{(X,Y)|X}$ given X is defined by*

$$\begin{aligned}\overline{P}_{(X,Y)|X}((A' \times B') \times A') &= \overline{P}((A \cap B)|A) \\ &= \frac{h^s\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A' \times B'\}}{h^s(A)}\end{aligned}$$

for every pair of Borelian sets A' and B' of $\Re$.

Definition 16 *Given two risks $X, Y : \Omega \longrightarrow \Re$ and a partition $\boldsymbol{B}$ of Ω let $B \in \boldsymbol{B}$ be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension s; the joint coherent conditional measure of risk $\rho((X,Y)|B)$ is defined by*

$$\begin{aligned}\rho((X,Y)|B) &= \int_B \overline{P}_{(X,Y)} d\mu_B^* \\ &= \int \frac{h^s\{\omega \in B : (X(\omega), Y(\omega)) \in A' \times B'\}}{h^s(B)} dx.\end{aligned}$$

6 Properties of s-Independent Risks

In this section the relation between s-independence and the factorization of the joint distribution $\overline{P}_{(X,Y)}$ into the product of the marginal distributions $\overline{P}_X$ and $\overline{P}_Y$ is investigated.

According to the axiomatic definition two random variables X and Y are independent if and only if $\sigma(X)$ and $\sigma(Y)$, the σ-fields generated by them, are independent. Since the σ-field generated by a random variable X is the smallest σ-field with respect to which X is measurable, it contains the inverse image of all Borelian sets of $\Re$. So if the random variables X and Y are independent then the joint distribution is equal to the product of the marginal distributions.

S-independence of risk X and Y does not imply that the joint distribution $\overline{P}_{(X,Y)}$ is equal to the product of the marginal $\overline{P}_X$ and $\overline{P}_Y$. It occurs because s-independence between risks implies that the factorization property holds only for the atoms (i.e. minimal ses with respect to the inclusion) of the classes generated by the weak upper level sets of X and Y (see Corollary 1) and not for all sets of the σ-fields generated by X and Y.

In the next theorem a sufficient condition is given to assure that the joint distribution of two simple risks is equal to the product of their marginal distributions.

Theorem 11 *Let Ω be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension and let X and $Y \in L(\Omega)$ be two simple risks such*

that the partition $\boldsymbol{B}(Y)$ *is s-irrelevant to the partition* $\boldsymbol{B}(X)$. *Then*

$$\rho(XY) = \overline{P}(XY) = \overline{P}(X)\overline{P}(Y) = \rho(X)\rho(Y).$$

Proof. Let X and Y be simple risks. Let $A_1, ..., A_n$ the atoms of $\mathbf{B}(X)$ and $B_1, ..., B_m$ the atoms of $\mathbf{B}(Y)$ where the atoms A_i and B_j are enumerated so that $x_i = X(A_i)$ for $i = 1, ..., n$ and $y_j = Y(B_j)$ for $j = 1, ...m$ are in descending order, i.e. $x_1 \geq x_2 \geq ... \geq x_n$ with $x_{n+1} = 0$ and $y_1 \geq y_2 \geq ... \geq y_m$ with $y_{m+1} = 0$. Thus $X = \sum_{i=1}^{n} x_i I_{A_i}$ and $Y = \sum_{j=1}^{m} x_j I_{B_j}$ and since $\mathbf{B}(Y)$ is s-irrelevant to $\mathbf{B}(X)$ by Corollary 1 we have that

$$\mu^*_\Omega(A_i \cap B_j) = \mu^*_\Omega(A_i)\mu^*_\Omega(B_j).$$

The coherent upper probability μ^*_Ω is submodular and since the random variable $Z = XY$ and any constant c are comonotonic we consider the class $\mathbf{C} = \{XY, c\}$ so that by Proposition 10.1 of [5] there exists an additive set function α on $\wp(\Omega)$ which agrees with μ^*_Ω on the class of μ^*_Ω-measurable sets (and so on the atoms of $\mathbf{B}(X)$ and on the atoms of $\mathbf{B}(Y)$) such that

$$\begin{aligned}\int_\Omega XY d\mu^*_\Omega &= \int_\Omega XY d\alpha \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \alpha(I_{A_i} I_{B_j}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mu^*_\Omega(I_{A_i} I_{B_j}).\end{aligned}$$

Thus the following equalities hold

$$\begin{aligned}\rho(XY) = \overline{P}(XY) &= \int XY d\mu^*_\Omega \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mu^*_\Omega(I_{A_i} I_{B_j}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mu^*_\Omega(A_i \cap B_j) \\ &= \sum_{i=1}^{n} x_i \mu^*_\Omega(A_i) \sum_{J=1}^{m} y_j \mu^*_\Omega(B_j) \\ &= \overline{P}(X)\overline{P}(Y) = \rho(X)\rho(Y). \quad \diamond\end{aligned}$$

Corollary 3 *Let* $B \in \boldsymbol{B}$ *be a set with positive and finite Hausdorff outer measure in its Hausdorff dimension and let* $X|B$ *and* $Y|B \in L(B)$ *be two simple risks such that* $\boldsymbol{B}(Y|B)$ *is s-irrelevant to* $\boldsymbol{B}(X|B)$. *Then*

$$\begin{aligned}\rho(XY|B) &= \overline{P}(XY|B) \\ &= \overline{P}(X|B) \cdot \overline{P}(Y|B) = \rho(X|B)\rho(Y|B).\end{aligned}$$

The next example shows that s-independence of two risks X and Y does not imply s-independence of $X|B$ and $Y|B$ where B is a set with positive and finite Hausdorff outer measure in its Hausdorff dimension belonging to a partition $\mathbf{B}$ of Ω.

Example 8 *Let* X *and* $Y \in L(\Omega)$ *be the two s-independent risks defined in Example 3 and let* $B = [0, \frac{2}{3}]$. *Since* $S(X|B) = \left\{\frac{2}{3}\right\}$ *and* $S(Y|B) = B$ *then by Theorem 8* $X|B$ *and* $Y|B$ *are s-dependent.*

7 Conclusions

Two risks X and Y are s-independent if the atoms of the classes generated by their weak upper level sets are s-independent. A crucial difference with respect to the axiomatic definition of independence for random variables is that s-independence of the indicator functions of two events does not imply s-independence of the indicator functions of their complements. Moreover s-independence for risks does not imply the factorization of the joint distribution into the product of the marginal distributions. In the model, different experiences of two decision makers are represented by different conditioning events with different Hausdorff dimension and this produces two different measures of risk for the same random variable. This allows us to mathematically represent the fact that different decision makers can retain certain actions more risky or less risky according to their own experiences or the information that they hold. That is to say what is for one a disadvantage for another person could be considered a convenient choice .

Acknowledgments

The author is grateful to the reviewers for their useful comments.

References

[1] P. Artzner, F. Delbaen, J.M. Eber, D. Heath, Coherent Measures of Risk, Math. Finance, 3, 203-228, (1999)

[2] P. Billingsley, Probability and measure, New York, Wiley, (1986)

[3] B. de Finetti, Probability, Induction and Statistics, Wiley, New York, (1972)

[4] B. de Finetti, Theory of Probability, Wiley, London, (1974)

[5] D. Denneberg, Non-additive measure and integral, Kluwer Academic Publishers, (1994)

[6] S. Doria, Conditional Upper Probabilities Assigned by a Class of Hausdorff Outer measures, Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications, Eds. G. de Cooman, T. Fine, T. Seidenfeld, Ithaca, NY, 147-151, (2001)

[7] S. Doria, Independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures, Proceedings of the 3th International Symposium on Imprecise Probabilities and their Applications, Eds. J.M. Bernard, T. Seidenfeld, M. Zaffalon, Lugano, Switzerland, 231-244, (2003)

[8] S. Doria, Probabilistic independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures, International Journal of Approximate Reasoning, 46, 617-635, (2007)

[9] S. Doria, Stochastic independence with respect to upper and lower conditional probabilities defined by Hausdorff outer and inner measures, in Stochastic Control, Editor C. Myers, Sciyo, 87-101, (2010)

[10] S. Doria, Coherent upper conditional previsions and their integral representation with respect to Hausdorff outer measures, Methods in Data Analysis. Eds. C. Borgelt et al., Advances in Intelligent and Soft Computing 77, 209-216, Springer, (2010)

[11] S. Doria, Coherent Upper and Lower Conditional Previsions Defined by Hausdorff Outer Measures, Modeling, Designs and Simulation of Systems with Uncertainties, Eds A. Rauh and E. Auer, Springer, 175-195, (2011)

[12] S. Doria, Characterization of a coherent upper conditional prevision as the Choquet integral with respect to its associated Hausdorff outer measure, Annals of Operations Research, 33-48, (2012)

[13] S. Doria, Symmetric coherent upper conditional prevision defined by the Choquet integral with respect to by Hausdorff outer measure, Annals of Operations Research, DOI 10.1007/s10479-014-1752-x, (2014)

[14] K.J. Falconer, The geometry of fractals sets, Cambridge University Press, (1986)

[15] G. Koch, La matematica del probabile, Aracne Editrice, (1997)

[16] S. Maaß, Coherent lower previsions as exact functionals and their (sigma-)core, Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA '01, Ed. by G.de Cooman, T. Fine, T. Seidenfeld, Cornell University, Ithaca, NY,USA, 230-236, (2001)

[17] R. Pelessoni, P.Vicig, Coherent risk measures and upper previsions, Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA '01, Ed. by G.de Cooman, T. Fine, T. Seidenfeld, Cornell University, Ithaca, NY,USA, 307-315, (2001)

[18] R. Pelessoni, P.Vicig, Envelope Theorems and Dilation with Convex Conditional Previsions, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, ISIPTA '05, Ed. by F. Cozman, R. Nau, T. Seidenfeld, Carnegie Mellon University, Pittsburgh, Pennsylvania,USA, 307-315, (2005)

[19] E. Regazzini, De Finetti's coherence and statistical inference, The Annals of Statistics, Vol 15, No. 2, 845-864, (1987)

[20] C.A. Rogers, Hausdorff measures, Cambridge University Press, (1970)

[21] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, (1991)

Imprecise Random Variables, Random Sets, and Monte Carlo Simulation

Thomas Fetz
University of Innsbruck, Austria
thomas.fetz@uibk.ac.at

Michael Oberguggenberger
University of Innsbruck, Austria
michael.oberguggenberger@uibk.ac.at

Abstract

The paper addresses the evaluation of upper and lower probabilities induced by functions of an imprecise random variable. Given a function g and a family X_λ of random variables, where the parameter λ ranges in an index set Λ, one may ask for the upper/lower probability that $g(X_\lambda)$ belongs to some Borel set B. Two interpretations are investigated. In the first case, the upper probability is computed as the supremum of the probabilities that $g(X_\lambda)$ lies in B. In the second case, one considers the random set generated by all $g(X_\lambda)$, $\lambda \in \Lambda$, e.g. by transforming X_λ to standard normal as a common probability space, and computes the corresponding upper probability. The two results are different, in general. We analyze this situation and highlight the implications for Monte Carlo simulation. Attention is given to efficient simulation procedures and an engineering application is presented.

Keywords. Upper and lower probabilities, imprecise random variables, random sets, propagation of uncertainty through a function, Monte Carlo simulation.

1 Introduction

Methods of imprecise probability have increasingly attracted interest in the engineering community, see e.g. the recent survey article [3]. In most industrial applications, engineering structures are described by black box input-output models, given by large computer programs. To evaluate probabilities of output quantities, Monte Carlo simulation is the method of choice. Computational effort is aggravated, if the input variables are imprecise random variables (or even random fields) with set-valued parameters.

The present paper is motivated by recent papers on simulation of random sets [16, 17] and on simulation of upper and lower probabilities in engineering reliability [2]. These papers raise the question what is the appropriate model of imprecision and what are cost-saving ways of simulating the output quantities. The key issue is to keep the number of required evaluations of the expensive input-output map as low as possible. Suppose the imprecision of the input is described in terms of a family of random variables. Two interpretations of resulting lower and upper probabilities are at hand: the first one by taking infima and suprema over the probabilities generated by the individual members of the family, the second one by first forming random sets based on the family and then evaluating the lower and upper probability as belief and plausibility (see e.g. [4]). In general, the results differ. The second question is the appropriate simulation method in the two cases.

In civil engineering, the use of random sets on continuous probability spaces goes back at least to [20], see also the text book [4] and the survey in [15]. Various alternative approaches to simulation of random sets have been proposed in recent years [1, 14, 22, 21]. The employed notion of imprecise probability is of course crucial. This paper focusses on the two limiting cases given by a family of random variables versus the random set generated by it. Once the random set is given, other ways of generating upper and lower probabilities are known: by means of all measurable selections of the random set, by means of measurable selections of the measures living on the focal elements, or by means of its core (the set of probabilities dominated by the upper probability of the random set). For these notions we refer to [5, 11, 12] as well as the exhaustive comparison in [6] (see also the remarks in Section 3 of the paper).

The paper is organized as follows. In Section 2, the set-up is explained and sufficient conditions are given making the formation of a random set possible. Section 3 is devoted to comparing the two versions of lower and upper probabilities and to stating conditions under which they coincide. Section 4 presents an engineering example, which is used in the subsequent sections for the purpose of illustration. Section 5 addresses issues of simulating random sets and the corresponding lower/upper probabilities, while Section 6 addresses the issue of simulation for lower/upper probabilities derived from families of random variables. Section 7 contains a summary and conclusions.

2 The Set-Up

The issue of the paper is how to evaluate and to simulate a function of an imprecise random variable. To fix the notation, let Λ be an index set. Consider a family of random variables $\{X_\lambda : \lambda \in \Lambda\}$, defined on some probability space $(\Omega, \Sigma, \mathsf{m})$. At fixed $\lambda \in \Lambda$, the random variable X_λ may be univariate or multivariate, with values in some $\mathbb{R}^n$. Further, let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuous function.

We wish to investigate lower and upper probabilities induced by the family of random variables $g(X_\lambda)$, $\lambda \in \Lambda$. From a probabilistic point of view, the function g could be suppressed, because $Y_\lambda = g(X_\lambda)$ is again a family of random variables. However, in the applications we have in mind, the function g will be an input-output map of a complex system, for which we want to compute the output distribution. Using Monte Carlo simulation, large samples of $g(X_\lambda)$ should be computed. While generating large samples of the input is computationally inexpensive, the evaluation of the function g might be very expensive. As mentioned, a focus of the paper will be on devising simulation methods of low computational cost. For this reason, the function g will play a role in Sections 4 to 6. One could also consider more generally parametrized functions of the form $g(\lambda, X_\lambda)$. For the initial probabilistic analysis, we drop mentioning the function g and consider families of real valued random variables.

Let $(\Omega, \Sigma, \mathsf{m})$ be the basic probability space, which we assume to be complete. Let $\{X_\lambda\}_{\lambda\in\Lambda}$ be a family of random variables $X_\lambda : \Omega \to \mathbb{R} : \omega \to X_\lambda(\omega)$. Then the probability of a Borel set $B \in \mathcal{B}(\mathbb{R})$ for a fixed random variable X_λ is

$$P(X_\lambda \in B) = \int_\Omega \mathbb{1}_{X_\lambda(\omega)\in B} \, \mathsf{dm}(\omega) \tag{1}$$

where

$$\mathbb{1}_E = \begin{cases} 1 & \text{if event } E \text{ occurs,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

denotes the indicator function of an event E. Modeling uncertainty by a family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables, it is natural to define lower and upper probabilities:

Definition 1 *For a family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables, the lower probability $\underline{P}$ and the upper probability $\overline{P}$ are given by*

$$\underline{P}(B) = \inf_{\lambda\in\Lambda} P(X_\lambda \in B) = \inf_{\lambda\in\Lambda} \int_\Omega \mathbb{1}_{X_\lambda(\omega)\in B} \, \mathsf{dm}(\omega), \tag{3}$$

$$\overline{P}(B) = \sup_{\lambda\in\Lambda} P(X_\lambda \in B) = \sup_{\lambda\in\Lambda} \int_\Omega \mathbb{1}_{X_\lambda(\omega)\in B} \, \mathsf{dm}(\omega), \tag{4}$$

where $B \subset \mathbb{R}$ is a Borel set.

On the other hand, it is also natural to construct a random set based on the family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables. To this end, we define the set-valued map $\mathcal{X} : \Omega \to \mathbb{R}$ by

$$\mathcal{X}(\omega) = \{X_\lambda(\omega) : \lambda \in \Lambda\}. \tag{5}$$

By definition, this set-valued map $\mathcal{X}$ is a random set, provided the upper inverses

$$\mathcal{X}^-(B) = \{\omega \in \Omega : \mathcal{X}(\omega) \cap B \neq \varnothing\} \tag{6}$$

are measurable subsets of Ω for every Borel set B. Then the same holds also for the lower inverses

$$\mathcal{X}_-(B) = \{\omega \in \Omega : \mathcal{X}(\omega) \subset B\}. \tag{7}$$

These lower and upper inverses lead to lower and upper probabilities based on the random set $\mathcal{X}$:

Definition 2 *For a random set $\mathcal{X}$, the lower probability $\underset{\sim}{P}$ and the upper probability $\widetilde{P}$ is defined by*

$$\underset{\sim}{P}(B) = \mathsf{m}(\mathcal{X}_-(B)) = \int_\Omega \mathbb{1}_{\mathcal{X}(\omega)\subseteq B} \, \mathsf{dm}(\omega), \tag{8}$$

$$\widetilde{P}(B) = \mathsf{m}(\mathcal{X}^-(B)) = \int_\Omega \mathbb{1}_{\mathcal{X}(\omega)\cap B\neq\varnothing} \, \mathsf{dm}(\omega). \tag{9}$$

We will show in the next section that

$$\underset{\sim}{P} \leq \underline{P} \leq \overline{P} \leq \widetilde{P} \tag{10}$$

with equality only in special cases, but first we give sufficient conditions so that $\mathcal{X}$ generated by the family $\{X_\lambda\}_{\lambda\in\Lambda}$ is a random set.

Theorem 1 *Let a family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables be given. Assume that Λ is a compact subset of a metric space and that the maps $\lambda \to X_\lambda(\omega)$ are continuous for each fixed $\omega \in \Omega$. Then*

(a) *the set-valued map $\mathcal{X}$ defined by*

$$\mathcal{X}(\omega) = \{X_\lambda(\omega) : \lambda \in \Lambda\} \tag{11}$$

is a compact random set;

(b) *the map $\overline{\mathbb{1}}_B : \omega \to \sup_{\lambda\in\Lambda} \mathbb{1}_{X_\lambda(\omega)\in B}$ is measurable;*

(c) *the map $\underline{\mathbb{1}}_B : \omega \to \inf_{\lambda\in\Lambda} \mathbb{1}_{X_\lambda(\omega)\in B}$ is measurable;*

where B is a Borel set.

Proof. (a) As a compact subset of a metric space, Λ is separable. Take a countable dense subset of parameter values λ_k in Λ. From the continuity assumption, it follows that the sequence $\{X_{\lambda_k}(\omega) : k = 1, 2, 3, \ldots\}$ is dense in $\mathcal{X}(\omega)$ for

every fixed ω. (That is, it is a Castaing representation.) Let B be an open set. Then the set

$$\begin{aligned}\mathcal{X}^-(B) &= \{\omega\in\Omega : \mathcal{X}(\omega)\cap B\neq\varnothing\} \\ &= \{\omega\in\Omega : \text{ there is } k \text{ such that } X_{\lambda_k}(\omega)\in B\} \\ &= \bigcup_k \{\omega\in\Omega : X_{\lambda_k}(\omega)\in B\}\end{aligned} \tag{12}$$

is measurable as a countable union of measurable sets. (The individual sets are measurable, because each X_λ is a random variable.) The fundamental measurability theorem (see e.g. [13]) implies that $\mathcal{X}^-(B)$ is measurable for every Borel set B. In addition, $\mathcal{X}(\omega)$ is the continuous image of a compact set. Thus $\mathcal{X}$ is a compact random set.

(b) Considering $\overline{\mathbb{1}}_B$, we observe the equivalence

$$\begin{aligned}\overline{\mathbb{1}}_B(\omega) &= \sup_{\lambda\in\Lambda} \mathbb{1}_{X_\lambda(\omega)\in B} = 1 \\ &\iff \exists\lambda\in\Lambda : X_\lambda(\omega)\in B \iff \mathcal{X}(\omega)\cap B\neq\varnothing.\end{aligned} \tag{13}$$

Thus $\mathcal{X}^-(B) = \overline{\mathbb{1}}_B^{-1}(\{1\})$. Since $\overline{\mathbb{1}}_B$ takes only the two values 0 and 1, this proves that $\overline{\mathbb{1}}_B$ is measurable.

(c) For $\underline{\mathbb{1}}_B$ we have

$$\begin{aligned}\underline{\mathbb{1}}_B(\omega) &= \inf_{\lambda\in\Lambda} \mathbb{1}_{X_\lambda(\omega)\in B} = 1 \\ &\iff \forall\lambda\in\Lambda : X_\lambda(\omega)\in B \iff \mathcal{X}(\omega)\subseteq B\end{aligned} \tag{14}$$

which leads to $\mathcal{X}_-(B) = \underline{\mathbb{1}}_B^{-1}(\{1\})$. The same arguments as in (b) yield the measurability of $\underline{\mathbb{1}}_B$. □

3 Comparison Results

The purpose of this section is to prove the chain of inequalities formulated in Eq. (10), exhibit some circumstances when they are equal and illustrate the behavior by means of simple examples.

Theorem 2 *Let $\{X_\lambda\}_{\lambda\in\Lambda}$ be a family of random variables according to the assumptions of Theorem 1 and let $\mathcal{X}$ be the random set induced by this family together with the map $\mathcal{X}(\omega) = \{X_\lambda(\omega) : \lambda\in\Lambda\}$. Then it holds*

$$\underset{\sim}{P} \leq \underline{P} \leq \overline{P} \leq \widetilde{P} \tag{15}$$

for the lower and upper probabilities in Def. 1 and 2.

Proof. For the upper probabilities of a Borel set B we have

$$\begin{aligned}\overline{P}(B) &= \sup_{\lambda\in\Lambda} P(X_\lambda\in B) = \sup_{\lambda\in\Lambda}\int_\Omega \mathbb{1}_{X_\lambda(\omega)\in B}\,\mathrm{dm}(\omega) \\ &\leq \int_\Omega \sup_{\lambda\in\Lambda}\mathbb{1}_{X_\lambda(\omega)\in B}\,\mathrm{dm}(\omega) = \int_\Omega \overline{\mathbb{1}}_B\,\mathrm{dm}(\omega) \\ &= \int_\Omega \mathbb{1}_{\mathcal{X}(\omega)\cap B\neq\varnothing}\,\mathrm{dm}(\omega) = \widetilde{P}(B)\end{aligned} \tag{16}$$

using Eq. (13). Together with Eq. (14) we get

$$\begin{aligned}\underline{P}(B) &= \inf_{\lambda\in\Lambda} P(X_\lambda\in B) = \inf_{\lambda\in\Lambda}\int_\Omega \mathbb{1}_{X_\lambda(\omega)\in B}\,\mathrm{dm}(\omega) \\ &\geq \int_\Omega \inf_{\lambda\in\Lambda}\mathbb{1}_{X_\lambda(\omega)\in B}\,\mathrm{dm}(\omega) = \int_\Omega \underline{\mathbb{1}}_B\,\mathrm{dm}(\omega) \\ &= \int_\Omega \mathbb{1}_{\mathcal{X}(\omega)\subseteq B}\,\mathrm{dm}(\omega) = \underset{\sim}{P}(B)\end{aligned} \tag{17}$$

for the lower probabilities. □

Remark. Let

$$\mathcal{M}(\widetilde{P}) = \{P : P(A)\leq\widetilde{P}(A), A\in\Sigma\} \tag{18}$$

be the set of all probability measures dominated by the upper probability $\widetilde{P}$ induced by the random set $\mathcal{X}$. Further let

$$P(\mathcal{X}) = \{P_X : X\in S(\mathcal{X})\} \tag{19}$$

be the set of all probability measures generated by the measurable selections

$$S(\mathcal{X}) = \{X : \Omega\to\mathbb{R} \text{ measurable} : X(\omega)\in\mathcal{X}(\omega)\} \tag{20}$$

of the random set $\mathcal{X}$. In [5, 11, 12] these two sets $P(\mathcal{X})$ and $\mathcal{M}(\widetilde{P})$ are investigated and it is proven that the relation $P(\mathcal{X})\subseteq\mathcal{M}(\widetilde{P})$ holds and that we have $P(\mathcal{X}) = \mathcal{M}(\widetilde{P})$ under certain conditions.

In our case the random variables $X\in\{X_\lambda\}_{\lambda\in\Lambda}$ are measurable selections of the random set $\mathcal{X}$ but in general not all of the selections in $S(\mathcal{X})$. That means we have $\{X_\lambda\}_{\lambda\in\Lambda}\subseteq S(\mathcal{X})$ and the following relations

$$\begin{aligned}\overline{P}(B) &= \underbrace{\sup_{\lambda\in\Lambda} P(X_\lambda\in B) = \sup_{X\in\{X_\lambda\}_{\lambda\in\Lambda}} P_X(B)}_{\text{family of random variables}} \leq \\ &\leq \underbrace{\sup_{X\in S(\mathcal{X})} P_X(B)}_{\text{all measurable selections}} \leq \underbrace{\sup_{P\in\mathcal{M}(\widetilde{P})} P(B)}_{\text{dominated probabilities}} = \widetilde{P}(B)\end{aligned}$$

for the upper probabilities and vice versa for the lower probabilities.

Example 1 Let $(\Omega,\Sigma,\mathsf{m}) = (\mathbb{R},\mathcal{B}(\mathbb{R}),\mathsf{m})$ be the probability space with probability measure

$$\mathsf{m}(B) = \int_{\mathbb{R}} \mathbb{1}_{\omega\in B}\,\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\omega^2/2}\,\mathrm{d}\omega, \quad B\in\mathcal{B}(\mathbb{R}). \tag{21}$$

On the one hand, the family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables is given by

$$\{X_{(\mu,\sigma)}\}_{(\mu,\sigma)\in\Lambda}, \quad \Lambda = [\underline{\mu},\overline{\mu}]\times[\underline{\sigma},\overline{\sigma}], \quad \underline{\sigma}>0 \tag{22}$$

where

$$X_{(\mu,\sigma)}(\omega) = \sigma\omega+\mu. \tag{23}$$

This means that $X_{(\mu,\sigma)} \sim \mathcal{N}(\mu,\sigma^2)$ is a Gaussian random variable parameterized by (μ,σ). In particular, we have $X_{(0,1)} \sim \mathcal{N}(0,1)$ and $X_{(0,1)}(\omega) = \omega$.

On the other hand, the random set $\mathcal{X}$ is generated by the set-valued map

$$\mathcal{X}(\omega) = \{X_\lambda(\omega) : \lambda \in \Lambda\}. \tag{24}$$

In this case, $\mathcal{X}(\omega)$ is an interval $[\underline{\mathcal{X}}(\omega), \overline{\mathcal{X}}(\omega)]$ with lower bound

$$\underline{\mathcal{X}}(\omega) = \inf_{\substack{\mu\in[\underline{\mu},\overline{\mu}]\\ \sigma\in[\underline{\sigma},\overline{\sigma}]}} X_{(\mu,\sigma)}(\omega) = \begin{cases} \overline{\sigma}\omega + \underline{\mu} & \omega < 0, \\ \underline{\sigma}\omega + \underline{\mu} & \omega \geq 0, \end{cases} \tag{25}$$

and upper bound

$$\overline{\mathcal{X}}(\omega) = \sup_{\substack{\mu\in[\underline{\mu},\overline{\mu}]\\ \sigma\in[\underline{\sigma},\overline{\sigma}]}} X_{(\mu,\sigma)}(\omega) = \begin{cases} \underline{\sigma}\omega + \overline{\mu} & \omega < 0, \\ \overline{\sigma}\omega + \overline{\mu} & \omega \geq 0. \end{cases} \tag{26}$$

As a specific example, we determine the lower and upper probabilities $\underline{P}(B)$ and $\overline{P}(B)$ given $\Lambda = [-0.5,2] \times [1,2]$ and $B = [1,2.5]$. For the family $\{X_{(\mu,\sigma)}\}_{(\mu,\sigma)\in\Lambda}$ of random variables we get

$$\begin{aligned} \underline{P}(B) &= \inf_{(\mu,\sigma)\in\Lambda} P(X_{(\mu,\sigma)} \in B) = P(X_{(-0.5,1)} \in B) \\ &= 0.065457, \\ \overline{P}(B) &= \sup_{(\mu,\sigma)\in\Lambda} P(X_{(\mu,\sigma)} \in B) = P(X_{(1.75,1)} \in B) \\ &= 0.546745, \end{aligned} \tag{27}$$

cf. Fig. 1 where the probability $P(X_{(\mu,\sigma)} \in [1,2])$ on Λ is visualized as a contour plot. The maximum probability is achieved at parameter values $(\mu,\sigma) = (1.75,1)$ ($\triangle$) and the minimum probability at $(-0.5,1)$ ($\triangledown$).

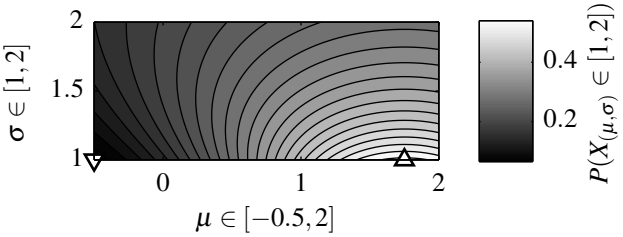


Figure 1: Set $\Lambda = [-0.5,2] \times [1,2]$ and probabilities $P(X_{(\mu,\sigma)} \in [1,2])$ visualized as a contour plot.

In Fig. 2 the random set $\mathcal{X}$ (gray area) corresponding to the above family of random variables, the bounds $\underline{\mathcal{X}}$, $\overline{\mathcal{X}}$, a single random variable $X_{(1.5,1.3)} \in \{X_{(\mu,\sigma)}\}_{(\mu,\sigma)\in\Lambda}$ and the focal set $\mathcal{X}(\omega)$ at $\omega = 1$ are depicted.

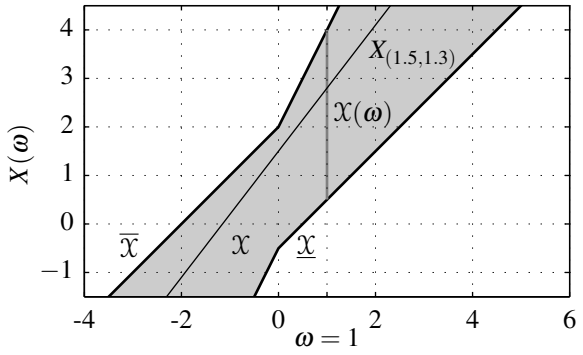


Figure 2: Random set $\mathcal{X}$, bounds $\underline{\mathcal{X}}$ and $\overline{\mathcal{X}}$, a single random variable $X_{(1.5,1.3)}$ and a focal set $\mathcal{X}(\omega)$.

To compute the lower and upper probabilities $\underset{\sim}{P}(B)$ and $\widetilde{P}(B)$ we need the lower and upper inverses. The lower inverse $\mathcal{X}_-(B)$ is the empty set, because there are no focal sets $\mathcal{X}(\omega)$ which are subsets of $B = [1,2.5]$. The upper inverse $\mathcal{X}^-(B)$ is the interval $[-1,3]$, cf. Fig. 3 where the random set $\mathcal{X}$ and the set B are depicted. In addition, some of the focal sets $\mathcal{X}(\omega)$, $\omega \in \mathcal{X}^-(B)$, with non-empty intersection with B are visualized as vertical lines. For comparison, the random variables $X_{(-0.5,1)}$ and $X_{(1.75,1)}$ resulting in $\underline{P}(B)$ and $\overline{P}(B)$ are plotted as well.

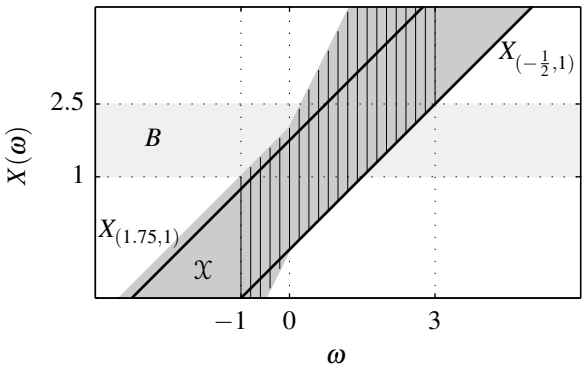


Figure 3: Set $B = [1,2.5]$, random set $\mathcal{X}$, random variables $X_{(1.75,1)}$ and $X_{(-\frac{1}{2},1)}$ resulting in $\underline{P}(B)$ and $\overline{P}(B)$. The vertical lines are focal sets with non-empty intersection with B leading to $\widetilde{P}(B)$.

Then the lower and upper probabilities are easily obtained:

$$\begin{aligned} \underset{\sim}{P}(B) &= \mathsf{m}(\mathcal{X}_-(B)) = \mathsf{m}(\varnothing) = 0 \\ &\leq 0.065457 = \underline{P}(B), \end{aligned} \tag{28}$$

$$\begin{aligned} \widetilde{P}(B) &= \mathsf{m}(\mathcal{X}^-(B)) = \mathsf{m}([-1,3]) = \Phi(3) - \Phi(-1) \\ &= 0.839994 \geq 0.546745 = \overline{P}(B) \end{aligned} \tag{29}$$

where $\Phi(\omega) = \mathsf{m}((-\infty,\omega])$ denotes the Gaussian cumulative distribution function.

In the following theorem we present special cases where $\underset{\sim}{P}(B) = \underline{P}(B)$ and/or $\widetilde{P}(B) = \overline{P}(B)$ holds for a fixed Borel set B. Since we are interested in probabilities $P(g(X) \le 0)$, we focus on sets B of type $(-\infty, b]$.

Theorem 3 *Let $\{X_\lambda\}_{\lambda\in\Lambda}$ be a family of random variables, Λ a compact subset of a metric space and assume that the maps $\lambda \to X_\lambda(\omega)$ are continuous for each fixed $\omega \in \Omega$. Further let $\mathcal{X}(\omega) = \{X_\lambda(\omega) : \lambda \in \Lambda\}$ and denote the lower and upper bounds by*

$$\underline{\mathcal{X}}(\omega) = \min \mathcal{X}(\omega) \quad \textit{and} \quad \overline{\mathcal{X}}(\omega) = \max \mathcal{X}(\omega). \tag{30}$$

(a) *If there is a $\lambda \in \Lambda$ such that $\mathbb{1}_{X_\lambda(\omega)\in B} = \underline{\mathbb{1}}_B(\omega)$ m-almost everywhere, then we have $\underline{P}(B) = \underset{\sim}{P}(B)$.*

If there is a $\lambda \in \Lambda$ such that $\mathbb{1}_{X_\lambda(\omega)\in B} = \overline{\mathbb{1}}_B$ m-almost everywhere, then we have $\overline{P}(B) = \widetilde{P}(B)$.

(b) *Let $B = (-\infty, b]$. If there is a $\lambda^* \in \Lambda$ such that $X_{\lambda^*}^{-1}(B) = \overline{\mathcal{X}}^{-1}(B)$, then we have $\underline{P}(B) = \underset{\sim}{P}(B)$.*

If there is a $\lambda_* \in \Lambda$ such that $X_{\lambda_*}^{-1}(B) = \underline{\mathcal{X}}^{-1}(B)$, then we have $\overline{P}(B) = \widetilde{P}(B)$.

(c) *Let $B = (-\infty, b]$. If there is a $\lambda^* \in \Lambda$ such that $X_{\lambda^*} = \overline{\mathcal{X}}$, then we have $\underline{P}(B) = \underset{\sim}{P}(B)$.*

If there is a $\lambda_* \in \Lambda$ such that $X_{\lambda_*} = \underline{\mathcal{X}}$, then we have $\overline{P}(B) = \widetilde{P}(B)$.

(d) *Let $B = (-\infty, b]$, $(\Omega, \Sigma, \mathsf{m}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathsf{m})$ and $\Lambda \times \mathbb{R} \to \mathbb{R} : (\lambda, \omega) \to X_\lambda(\omega)$ continuous and strictly monotonically increasing (decreasing) in the ω-direction, then $\underline{P}(B) = \underset{\sim}{P}(B)$ and $\overline{P}(B) = \widetilde{P}(B)$.*

Proof. (a) It follows directly from the Eqs. (16) and (17) in the proof of Theorem 2.

For the further proof we use that $\mathcal{X}_-(B) = \overline{\mathcal{X}}^{-1}(B)$ and $\mathcal{X}^-(B) = \underline{\mathcal{X}}^{-1}(B)$ holds for a set of the form $B = (\infty, b]$.

(b) Let $B = (-\infty, b]$. Then

$$\begin{aligned} \underset{\sim}{P}(B) &= \mathsf{m}(\mathcal{X}_-(B)) = \mathsf{m}(\overline{\mathcal{X}}^{-1}(B)) = \mathsf{m}(X_{\lambda^*}^{-1}(B)) \\ &= \int_\Omega \mathbb{1}_{X_{\lambda^*}(\omega)\in B}\, \mathsf{dm}(\omega) \ge \underline{P}(B) \end{aligned} \tag{31}$$

and

$$\begin{aligned} \widetilde{P}(B) &= \mathsf{m}(\mathcal{X}^-(B)) = \mathsf{m}(\underline{\mathcal{X}}^{-1}(B)) = \mathsf{m}(X_{\lambda_*}^{-1}(B)) \\ &= \int_\Omega \mathbb{1}_{X_{\lambda_*}(\omega)\in B}\, \mathsf{dm}(\omega) \le \overline{P}(B). \end{aligned} \tag{32}$$

(c) It is a special case of (b). E.g., if there is a $\lambda_* \in \Lambda$ such that $X_{\lambda_*} = \underline{\mathcal{X}}$, then it follows that $\underline{\mathcal{X}}^{-1}(B) = X_{\lambda_*}^{-1}(B)$.

(d) Let $B = (-\infty, b]$ and $(\lambda, \omega) \to X_\lambda(\omega)$ continuous and strictly monotonically increasing in the ω-direction. In this case the bound $\underline{\mathcal{X}}$ is continuous and strictly monotonically increasing in ω. (Assume $\underline{\mathcal{X}}(\omega_1) \ge \underline{\mathcal{X}}(\omega_2)$ for $\omega_1 < \omega_2$. Then it follows from Eq. (30) that there is a λ such that $\underline{\mathcal{X}}(\omega_2) = X_\lambda(\omega_2)$. This leads to the contradiction $X_\lambda(\omega_1) \ge \underline{\mathcal{X}}(\omega_1) \ge X_\lambda(\omega_2)$.)

There are three cases:

(i) Case $b < \underline{\mathcal{X}}$. Then $\underline{\mathcal{X}}^{-1}(B) = \varnothing$, $\widetilde{P}(B) = \mathsf{m}(\varnothing) = 0$ and $\overline{P}(B) = 0$.

(ii) Case $b > \underline{\mathcal{X}}$. Then $\underline{\mathcal{X}}^{-1}(B) = \mathbb{R}$, $\widetilde{P}(B) = \mathsf{m}(\mathbb{R}) = 1$. Further, $\overline{P}(B) = 1$. Indeed, take any $\overline{\omega} \in \mathbb{R}$. Then there is $\lambda \in \Lambda$ such that $\underline{\mathcal{X}}(\overline{\omega}) \le X_\lambda(\overline{\omega}) < b$ and by strict monotonicity $\underline{\mathcal{X}}(\omega) \le X_\lambda(\omega) < b$ for all $\omega \le \overline{\omega}$. This implies $\overline{P}(B) \ge \mathsf{m}((\infty, \overline{\omega}])$. Since $\overline{\omega}$ is arbitrary, we get $\overline{P}(B) = 1$ for $\overline{\omega} \to \infty$.

(iii) Case $b \in \underline{\mathcal{X}}(\mathbb{R})$. Then there is an $\omega^* \in \mathbb{R}$ such that $\underline{\mathcal{X}}(\omega^*) = b$.

In case (iii) we have on the one hand

$$\underline{\mathcal{X}}((-\infty, \omega^*]) \subseteq (-\infty, b] \tag{33}$$

because of the strict monotonicity of $\underline{\mathcal{X}}$ and on the other hand for the complement of $(-\infty, \omega^*]$

$$\underline{\mathcal{X}}((\omega^*, \infty)) \cap (-\infty, b] = \varnothing \tag{34}$$

which means that $\underline{\mathcal{X}}^{-1}((-\infty, b]) = (-\infty, \omega^*]$.

Further there is a $\lambda_* \in \Lambda$ such that $\underline{\mathcal{X}}(\omega^*) = b = X_{\lambda_*}(\omega^*)$ because of Eq. (30). This and the monotonicity of $X_{\lambda_*} \ge \underline{\mathcal{X}}$ leads to

$$\begin{aligned} &X_{\lambda_*}((-\infty, \omega^*]) \subseteq (-\infty, b], \\ &X_{\lambda_*}((\omega^*, \infty)) \cap (-\infty, b] = \varnothing, \\ &X_{\lambda_*}^{-1}((-\infty, b]) = (-\infty, \omega^*] \end{aligned} \tag{35}$$

and this in turn leads to the assumption

$$X_{\lambda_*}^{-1}((-\infty, b]) = \underline{\mathcal{X}}^{-1}((-\infty, b]) \tag{36}$$

of case (b). In particular, we get $\omega^* = \underline{\mathcal{X}}^{-1}(b) = X_{\lambda_*}^{-1}(b)$ and $\overline{P}(B) = \widetilde{P}(B) = \mathsf{m}((-\infty, \omega^*])$.

By the same arguments one can prove for the lower probabilities that $\underline{P}(B) = \underset{\sim}{P}(B) = \mathsf{m}((-\infty, \omega_*])$ with $\omega_* = \overline{\mathcal{X}}^{-1}(b) = X_{\lambda^*}^{-1}(b)$ and the results for decreasing functions $(\lambda, \omega) \to X_\lambda(\omega)$ as well. □

Remark on Theorem 3c. If there is a $\lambda_* \in \Lambda$ such that $X_{\lambda_*} = \underline{\mathcal{X}}$, then we have $F_{X_{\lambda_*}} = F_{\underline{\mathcal{X}}}$ for the cumulative distribution functions and therefore

$$\widetilde{P}((-\infty, b])) = F_{\underline{\mathcal{X}}}(b) = F_{X_{\lambda_*}}(b) = \overline{P}((-\infty, b]). \tag{37}$$

The relation $F_{X_{\lambda_*}} = F_{\underline{\mathcal{X}}}$ means in the notion of p-boxes [7], that one of the distribution functions F_{X_λ}, $\lambda \in \Lambda$, coincides with the upper envelope $\overline{F} = F_{\underline{\mathcal{X}}}$, cf. [7], of the p-box.

Example 2 We continue with Example 1. Here we want to compute the lower and upper probabilities for the set $B = (-\infty, b]$.

Since $(\mu, \sigma, \omega) \to X_{(\mu,\sigma)}(\omega) = \sigma\omega + \mu$, $\sigma > 0$, is continuous and a strictly monotonically increasing function in ω, we can apply Theorem 3 (d).

First we determine the inverses of $\underline{\mathcal{X}}$ and $\overline{\mathcal{X}}$ (cf. Eqs. (25) and (26)):

$$\underline{\mathcal{X}}^{-1}(x) = \begin{cases} (x-\underline{\mu})/\overline{\sigma} = X^{-1}_{(\underline{\mu},\overline{\sigma})}(x) & x < \underline{\mu}, \\ (x-\underline{\mu})/\underline{\sigma} = X^{-1}_{(\underline{\mu},\underline{\sigma})}(x) & x \geq \underline{\mu}, \end{cases} \tag{38}$$

and

$$\overline{\mathcal{X}}^{-1}(x) = \begin{cases} (x-\overline{\mu})/\underline{\sigma} = X^{-1}_{(\overline{\mu},\underline{\sigma})}(x) & x < \overline{\mu}, \\ (x-\overline{\mu})/\overline{\sigma} = X^{-1}_{(\overline{\mu},\overline{\sigma})}(x) & x \geq \overline{\mu}. \end{cases} \tag{39}$$

Now let again $\Lambda = [\underline{\mu}, \overline{\mu}] \times [\underline{\sigma}, \overline{\sigma}] = [-0.5, 2] \times [1, 2]$ and $B = (-\infty, b] = (-\infty, 2.5]$. Then $2.5 \geq \overline{\mu} \geq \underline{\mu}$ which means that we have to take the second parts of Eqs. (38) and (39) to determine the lower and upper probabilities:

$$\begin{aligned} \underline{P}((-\infty, b]) &= \underline{P}((-\infty, b]) = \mathsf{m}((-\infty, \overline{\mathcal{X}}^{-1}(b)]) \qquad (40) \\ &= \mathsf{m}((-\infty, \omega_*]) = \Phi(X^{-1}_{(2,2)}(2.5)) = \Phi(0.25) \\ &= 0.598706, \\ \widetilde{P}((-\infty, b]) &= \overline{P}((-\infty, b]) = \mathsf{m}((-\infty, \underline{\mathcal{X}}^{-1}(b)]) \qquad (41) \\ &= \mathsf{m}((-\infty, \omega^*]) = \Phi(X^{-1}_{(-0.5,1)}(2.5)) = \Phi(3) \\ &= 0.998650 \end{aligned}$$

with $b = 2.5$, $\omega_* = \overline{\mathcal{X}}^{-1}(b) = 0.25$ and $\omega^* = \underline{\mathcal{X}}^{-1}(b) = 3$, see also Fig. 4.

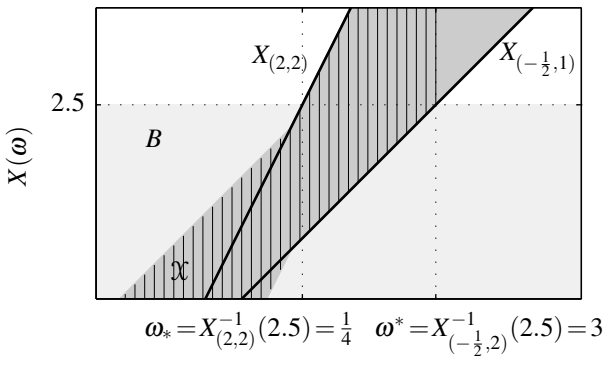


Figure 4: Set $B = (-\infty, b] = (-\infty, 2.5]$, random set $\mathcal{X}$, ω_*, ω^*, random variables $X_{(2,2)}$ and $X_{(-\frac{1}{2},1)}$ resulting in $\underline{P}(B)$ and $\overline{P}(B)$. The vertical lines are focal sets with non-empty intersection with B leading to $\widetilde{P}(B) = \overline{P}(B)$.

4 Numerical Example

As a simple engineering example we consider a beam of length $L = 3$ m supported on both ends and additionally bedded on a spring, cf. Fig. 5. The values of the beam rigidity $EI = 1\,\text{kNm}^2$, of the elastic limit moment $M_{\text{yield}} = 21$ kNm and of the load $f(\xi) = q = 100$ kN/m are deterministic, but the value of the spring constant x (in our notation for the variables of the function g) is assumed to be uncertain.

The beam will fail in the case where the value of limit state function g depending on the spring constant x is less or equal to 0. This means that we are interested in the failure probability $P(g(X) \leq 0)$ where the random variable X describes the uncertainty of the spring constant.

In this example the limit state function g is given as

$$\begin{aligned} \mathrm{g}(x) &= M_{\text{yield}} - \max_{\xi \in [0,3]} |M(\xi, x)| \qquad (42) \\ &= M_{\text{yield}} - \frac{qL^2}{4} \max\left(\frac{(1-c(x))^2}{2}, c(x) - \frac{1}{2} \right) \end{aligned}$$

with $c(x) = 5x/(384EI/L^3 + 8x)$, see Fig. 5 and [9]. Obviously this function is cheap to evaluate. Nevertheless we will apply the strategies for handling time consuming functions g as described in the following sections.

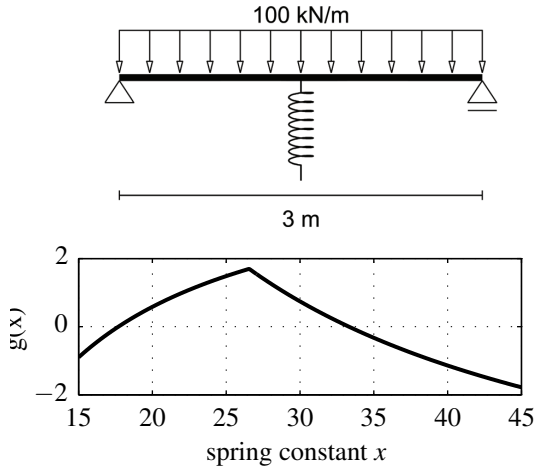


Figure 5: Beam bedded on a spring and deterministic limit state function g depending on the spring constant.

We model the uncertainty of the spring constant by a family $\{X_{(\mu,\sigma)}\}_{(\mu,\sigma)\in\Lambda}$ of random variables $X_{(\mu,\sigma)} \sim \mathcal{N}(\mu, \sigma^2)$ and, alternatively, by the induced random set $\mathcal{X}$. For this purpose, we simply continue with the set-up of Example 1. However, the set Λ is now given by

$$\Lambda = [\underline{\mu}, \overline{\mu}] \times [\underline{\sigma}, \overline{\sigma}] = [20, 30] \times [0.5, 3]. \tag{43}$$

In the following sections the function g, the family of random variables as defined above and the corresponding random set will be used to exemplify the simulation techniques presented.

5 Simulation of Random Sets

First, let us recall how the propagation of random set data through a map is accomplished. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and $\mathcal{X}(\omega)$, $\omega \in \Omega$ be a random set. The image set $\mathcal{G}(\omega) = g(\mathcal{X}(\omega))$, $\omega \in \Omega$, where $g(\mathcal{X}(\omega))$ is the set of values the function $y = g(x)$ attains when x ranges in $\mathcal{X}(\omega)$ is a random set. Computational aspects of random sets have been discussed at many places, see e.g. the references in [3]. In our context, the principles of Monte Carlo simulation of random sets of the form $g(\mathcal{X}(\omega)) = g(\{X_\lambda(\omega) : \lambda \in \Lambda\})$ have to be explained.

We assume that Λ is a compact subset of a metric space and that the maps $\lambda \to X_\lambda(\omega)$ are continuous for each fixed $\omega \in \Omega$. Then $\mathcal{G}(\omega) = g(\mathcal{X}(\omega))$ is contained in a random interval $[\underline{\mathcal{G}}(\omega), \overline{\mathcal{G}}(\omega)]$, where $\underline{\mathcal{G}}(\omega) = \min g(\mathcal{X}(\omega))$ and $\overline{\mathcal{G}}(\omega) = \max g(\mathcal{X}(\omega))$. Suppose we wish to compute its upper and lower distribution functions $\overline{F}(y)$ and $\underline{F}(y)$. Note that

$$\overline{F}(y) = P\big((-\infty, y] \cap [\underline{\mathcal{G}}, \overline{\mathcal{G}}] \neq \varnothing\big) = P\big(\underline{\mathcal{G}} \leq y\big) = F_{\underline{\mathcal{G}}}(y),$$

the cumulative distribution function of the random variable $\underline{\mathcal{G}}$. Similarly,

$$\underline{F}(y) = P\big([\underline{\mathcal{G}}, \overline{\mathcal{G}}] \subset (-\infty, y]\big) = P\big(\overline{\mathcal{G}} \leq y\big) = F_{\overline{\mathcal{G}}}(y).$$

Recall that $g(\mathcal{X}(\omega)) = \{g(X_\lambda(\omega))) : \lambda \in \Lambda\}$. Thus, in order to compute $\underline{\mathcal{G}}(\omega)$ and $\overline{\mathcal{G}}(\omega)$, an optimization problem has to be solved that determines the smallest and the largest value of the set $\{g(X_\lambda(\omega))) : \lambda \in \Lambda\}$. This leads to the following algorithm for computing the upper distribution function $\overline{F}(x)$.

- Generate a sample $\omega_1, \ldots, \omega_{N_{\text{samp}}}$ random elements of Ω, distributed according to m .
- For each ω_n, estimate $\underline{\mathcal{G}}(\omega_n) = \min g(X_\lambda(\omega_n))$ by minimization with respect to $\lambda \in \Lambda$.
- The empirical distribution function of the sample $\{\underline{\mathcal{G}}(\omega_n) : n = 1, \ldots, N_{\text{samp}}\}$ is an approximation to $\overline{F}(y)$.

In order to compute the respective minima and maxima, the parameter set should be discretized into $\lambda_1, \ldots, \lambda_{N_{\text{grid}}}$. The algorithm requires $N_{\text{grid}} \cdot N_{\text{samp}}$ evaluations of the function g. Generally, this is too expensive for large scale applications. Computational cost can be saved by suitably approximating the input-output function g by a surrogate model. For such an approximation, two levels are at hand:

$$\Omega \xrightarrow{X_\lambda} \mathbb{R}^n \xrightarrow{g} \mathbb{R}.$$

There are two possibilities to construct a surrogate model: either by a surrogate model $\widetilde{g}$ of the map $g : \mathbb{R}^n \to \mathbb{R}$ or by a family of stochastic surrogate models of the maps $\Omega \to g \circ X_\lambda$. Both approaches start with a set of collocation points x_j, $j = 1, \ldots, N_{\text{coll}}$ in $\mathbb{R}^n$, together with the corresponding function values $y_j = g(x_j)$. This requires N_{coll} evaluations of the input-output function g. In the first approach, the functions $\widetilde{g} \circ X_\lambda$ have to be simulated for λ belonging to a grid of representative parameter values λ_i, $i = 1, \ldots, N_{\text{grid}}$. Each X_{λ_i} has a different probability distribution. If a Monte Carlo sample of size N_{samp} is desired, this still requires $N_{\text{grid}} \cdot N_{\text{samp}}$ evaluations of the function $\widetilde{g}$. We will show below that reweighting technique can reduce the computational cost to N_{samp} evaluations.

We first discuss the second approach. Here the collocation points are pulled back to Ω as follows: For each λ_i and x_j, define a collocation point in Ω by $\omega_{ij} = X_{\lambda_i}^{-1}(x_j)$. Clearly, $y_j = g(X_{\lambda_i}(\omega_{ij}))$ for every i. Fitting a surrogate model $\widetilde{g}_i$ for each i, based on the data (ω_{ij}, g_j), $j = 1, \ldots, N_{\text{coll}}$ is computationally inexpensive. Typically, when $\Omega = \mathbb{R}^m$ and the measure $\mathsf{dm}(\omega)$ is absolutely continuous with respect to Lebesgue measure, one may use orthogonal polynomials with respect to the measure $\mathsf{dm}(\omega)$ (Hermite expansion in the Gaussian case) and then compute the coefficients by weighted regression through the data.

At fixed ω, the lower bound $\underline{\mathcal{G}}(\omega)$ of the focal set $\mathcal{G}(\omega)$ can simply be estimated by the smallest value among the $\widetilde{g}_i(\omega)$, $i = 1, \ldots, N_{\text{grid}}$. Repeating this procedure for a sample $\omega_1, \ldots, \omega_{N_{\text{samp}}}$ produces a Monte Carlo sample $\{\underline{\mathcal{G}}(\omega_n) : n = 1, \ldots, N_{\text{samp}}\}$ which can be used to estimate the desired upper distribution function $\overline{F}(y)$, and similarly for the lower distribution function $\underline{F}(y)$.

This approach requires N_{coll} evaluations of the expensive full model and N_{samp} evaluations of the inexpensive surrogate model. Details of the procedure can be found in [16]; further information on the construction of stochastic surrogate models is in [10].

Remarks. (a) In case X_λ are Gaussian variables with $\lambda = (\mu, \sigma)$ and $\Omega = \mathbb{R}$ with the standard Gaussian density, we simply have $X_\lambda(\omega) = \mu + \sigma\omega$ and $X_{\lambda_i}^{-1}(x) = (x - \mu)/\sigma$. The same procedure can be applied to a non-Gaussian family X_λ by transforming it to standard Gaussian space, i.e., $X_\lambda(\omega) = F_\lambda^{-1}(\Phi(\omega))$ where F_λ and Φ denote the cumulative distribution functions of X_λ and of a standard normal variable, respectively.

(b) Some indications on multivariate families X_λ are in order. We consider the case of an n-dimensional Gaussian variable $X_\lambda \sim \mathcal{N}\big(\mu(\lambda), \mathbf{S}(\lambda)\big)$ with mean $\mu(\lambda)$ and covariance $\mathbf{S}(\lambda)$, both assumed to depend continuously on a possibly multidimensional parameter λ. Performing the Cholesky factorization $\mathbf{S}(\lambda) = \mathbf{C}(\lambda)\mathbf{C}(\lambda)^\mathsf{T}$, the random variables X_λ can be realized as $X_\lambda = \mathbf{C}(\lambda)Y$ where Y is an n-dimensional standard Gaussian variable. The procedure outlined above can be applied in the same way, employing n-dimensional Hermite polynomials.

In this framework, finite dimensional discretizations of Gaussian random fields can be accommodated as well, either using the Cholesky factorization or – equivalently – a truncated Karhunen-Loève expansion.

Example 3 We continue with our engineering example. Due to its small size, we may use the full model without constructing a surrogate model.

Let the grid points (μ_i, σ_j) with

$$\mu_i = 20, 21, \ldots, 30 \quad \text{and} \quad \sigma_j = 0.5, 1, 1.5, \ldots, 3 \tag{44}$$

define a grid on $\Lambda = [\underline{\mu}, \overline{\mu}] \times [\underline{\sigma}, \overline{\sigma}] = [20, 30] \times [0.5, 3]$. Then a focal set $[\underline{\mathcal{G}}(\omega), \overline{\underline{\mathcal{G}}}(\omega)]$ of the random set $\mathcal{G}$ at ω is approximated by

$$\begin{aligned} \underline{\mathcal{G}}(\omega) &\approx \min_{i,j} g \circ X_{(\mu_i,\sigma_j)}(\omega), \\ \overline{\mathcal{G}}(\omega) &\approx \max_{i,j} g \circ X_{(\mu_i,\sigma_j)}(\omega). \end{aligned} \tag{45}$$

We approximate the upper probability of failure of the beam by means of Monte Carlo simulation:

$$\begin{aligned} \widetilde{P}(g \le 0) &= \int_{\mathbb{R}} \mathbb{1}_{\mathcal{G}(\omega) \cap (-\infty, 0] \neq \varnothing} \, \mathsf{dm}(\omega) \\ &= \int_{\mathbb{R}} \mathbb{1}_{\underline{\mathcal{G}}(\omega) \le 0} \, \mathsf{dm}(\omega) \\ &\approx \sum_{k=1}^{N_{\text{samp}}} \mathbb{1}_{\underline{\mathcal{G}}(\omega_k) \le 0} \cdot \frac{1}{N_{\text{samp}}} = 0.358. \end{aligned} \tag{46}$$

where $\omega_1, \ldots, \omega_{N_{\text{samp}}}$ is a standard normally distributed sample.

In Fig. 6 the random set $\mathcal{G}$ and one of the functions which are generating $\mathcal{G}$, namely $g \circ X_{(24,2)}$, are depicted. Further in Fig. 7 a sample of focal sets $\mathcal{G}(\omega_k)$, $k = 1, \ldots, N_{\text{samp}}$, is visualized. Counting the focal sets with non-empty intersection with $(-\infty, 0]$ leads to the upper probability $\widetilde{P}(g \le 0)$.

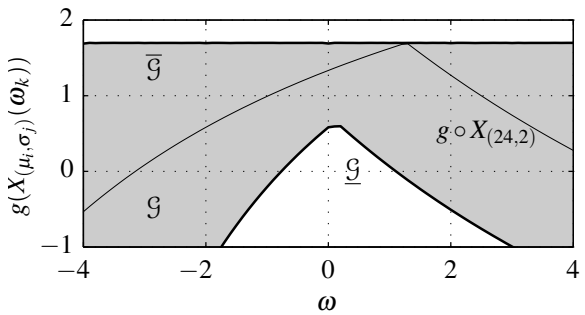


Figure 6: Random set $\mathcal{G}$, $g \circ X_{(24,2)}$ (one of the transformations of g), lower and upper bounds $\underline{\mathcal{G}}$ and $\overline{\mathcal{G}}$.

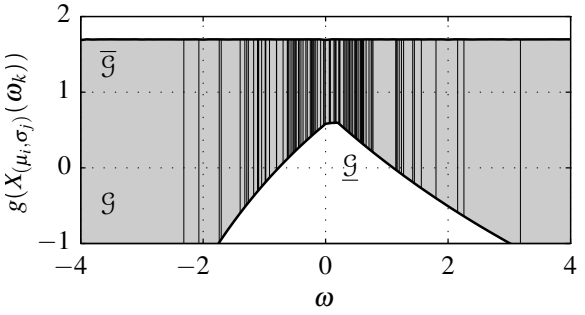


Figure 7: Random set $\mathcal{G}$ and focal sets $\mathcal{G}(\omega_k)$ for sample $\omega_1, \ldots, \omega_{N_{\text{samp}}}$ visualized as vertical lines.

6 Simulation of Families of Random Variables

In this section we show how to simulate families of random variables and how to save computational cost applying sample reweighting techniques.

We shortly present the ideas of resampling: Let v be a measurable function (later on an indicator function $\mathbb{1}_{g(x) \le y}$), $f_X > 0$ the strictly positive density function of a random variable X and $f_Y > 0$ the strictly positive density of a random variable Y. Then, for the random variable X we have the approximation

$$\int_{\mathbb{R}} v(x) f_X(x) \, \mathrm{d}x \approx \sum_{k=1}^{N_{\text{samp}}} v(x_k) \cdot \frac{1}{N_{\text{samp}}} \tag{47}$$

of the integral where $x_1, \ldots, x_{N_{\text{samp}}}$ is a sample distributed according to X together with weights $1/N_{\text{samp}}$. Similarly, for Y we get

$$\int_{\mathbb{R}} v(x) f_Y(x) \, \mathrm{d}x \approx \sum_{k=1}^{N_{\text{samp}}} v(y_k) \cdot \frac{1}{N_{\text{samp}}} \tag{48}$$

where now $y_1, \ldots, y_{N_{\text{samp}}}$ is a sample distributed as Y with weights $1/N_{\text{samp}}$. In this version, replacing the density function f_X by f_Y needs new samples and new function evaluations $v(y_k)$.

Alternatively, applying sample reweighting, we have

$$\begin{aligned} \int_{\mathbb{R}} v(x) f_Y(x) \, \mathrm{d}x &= \int_{\mathbb{R}} v(x) \cdot \frac{f_Y(x)}{f_X(x)} \cdot f_X(x) \, \mathrm{d}x \\ &\approx \sum_{k=1}^{N_{\text{samp}}} v(x_k) \cdot \frac{f_Y(x_k)}{f_X(x_k)} \frac{1}{N_{\text{samp}}} \end{aligned} \tag{49}$$

using the original sample $x_1, \ldots, x_{N_{\text{samp}}}$, but now with new weights

$$w_k := \frac{f_Y(x_k)}{f_X(x_k)} \frac{1}{N_{\text{samp}}} \tag{50}$$

instead of the uniform weights $1/N_{\text{samp}}$.

Our goal is to approximate probabilities $P(g(X_\lambda) \leq y)$ by means of Monte Carlo simulation using only one sample for all X_λ, $\lambda \in \Lambda$.

As a first step, we start with the generation of a sample $x_1, \ldots, x_{N_{\text{samp}}}$. This sample may be distributed as one of our random variables X_λ, $\lambda \in \Lambda$. But a better choice is a "basic" distribution covering a greater range than a distribution of a single X_λ, $\lambda \in \Lambda$, does. In our example with the family $\{X_{(\mu,\sigma)}\}_{(\mu,\sigma)\in\Lambda}$ of Gaussian random variables, this can be achieved by using an appropriate high variance $\sigma_*^2 > \overline{\sigma}^2$ and $\mu_* = (\underline{\mu} + \overline{\mu})/2$ for generating a Gaussian sample $x_1, \ldots, x_{N_{\text{samp}}} \sim \mathcal{N}(\mu_*, \sigma_*^2)$. In general we say that this "basic" sample is distributed as a random variable X_*.

As a second step, we compute all function values $g(x_k)$, $k = 1, \ldots, N_{\text{samp}}$, of the limit state function, either directly evaluating g or using a surrogate model $\tilde{g}$.

As a third step, we have to perform a reweighting of the sample generated above, since we need samples distributed according to certain random variables X_λ. These weights for a given X_λ are obtained by

$$w_k(\lambda) = \frac{f_{X_\lambda}(x_k)}{f_{X_*}(x_k)} \frac{1}{N_{\text{samp}}}. \tag{51}$$

Now we are able to approximately compute probabilities $P(g(X_\lambda) \leq y)$ for different random variables X_λ without additional function evaluations of g:

$$P(g(X_\lambda) \leq y) \approx \sum_{k=1}^{N_{\text{samp}}} \mathbb{1}_{g(x_k)\leq y} \cdot w_k(\lambda). \tag{52}$$

For the computation of the upper/lower probabilities we use a grid of representative parameter values λ_i as mentioned in the previous section, estimate the probabilities $P(g(X_{\lambda_i}) \leq y)$ at the grid points λ_i by means of Monte Carlo simulation as in Eq. (52) and take the maximum/minimum value as an approximation:

$$\begin{aligned} \overline{P}(g \leq y) &= \sup_{\lambda\in\Lambda} P(g(X_\lambda) \leq y) \approx \max_{i=1,\ldots,N_{\text{grid}}} P(g(X_{\lambda_i}) \leq y) \\ &\approx \max_{i=1,\ldots,N_{\text{grid}}} \sum_{k=1}^{N} \mathbb{1}_{g(x_k)\leq y} \cdot w_k(\lambda_i), \end{aligned} \tag{53}$$

$$\underline{P}(g \leq y) \approx \min_{i=1,\ldots,N_{\text{grid}}} \sum_{k=1}^{N} \mathbb{1}_{g(x_k)\leq y} \cdot w_k(\lambda_i). \tag{54}$$

Example 4 Again, we continue with the engineering example.

We approximately compute the failure probability $P(g(X_{(\mu,\sigma)}) \leq 0)$ of the beam for a fixed pair $(\mu,\sigma) \in \Lambda$ using either Monte Carlo simulation in the space of the variables of the limit state function g, Eq. (57), or in the standard normal space, Eq. (56):

$$P(g(X_{(\mu,\sigma)}) \leq 0) = \int_{\mathbb{R}} \mathbb{1}_{g(X_{(\mu,\sigma)}(\omega))\leq 0} \, \mathrm{dm}(\omega) \tag{55}$$

$$\approx \sum_{k=1}^{N_{\text{samp}}} \mathbb{1}_{g(X_{(\mu,\sigma)}(\omega_k))\leq 0} \cdot w_k(\mu,\sigma) \tag{56}$$

$$\approx \sum_{k=1}^{N_{\text{samp}}} \mathbb{1}_{g(x_k)\leq 0} \cdot w_k(\mu,\sigma) \tag{57}$$

where

$$X_{(\mu,\sigma)}(\omega_k) = \sigma\omega_k + \mu = x_k \tag{58}$$

and in the reverse direction

$$\omega_k = X_{(\mu,\sigma)}^{-1}(x_k) = (x_k - \mu)/\sigma. \tag{59}$$

The weights are given by

$$w_k(\mu,\sigma) = \frac{f_{X_{(\mu,\sigma)}}(x_k)}{f_{X_*}(x_k)} \frac{1}{N_{\text{samp}}} \tag{60}$$

where the sample $x_1, \ldots, x_{N_{\text{samp}}}$ is distributed according $X_* \sim \mathcal{N}(25, 6^2)$ for $N_{\text{samp}} = 100\,000$.

In the next step we have to compute

$$\overline{P}(g \leq 0) = \sup_{(\mu,\sigma)\in\Lambda} P(g(X_{(\mu,\sigma)}) \leq 0) \tag{61}$$

which is approximated using grid points (μ_i, σ_j) with

$$\mu_i = 20, 21, \ldots, 30 \quad \text{and} \quad \sigma_j = 0.5, 1, 1.5, \ldots, 3. \tag{62}$$

The probabilities $P(g(X_{(\mu_i,\sigma_j)}) \leq 0)$ at these grid points are computed as in Eq. (52) and depicted in Fig. 8. Then we simply take the maximum of all these probabilities similar to Eq. (53) and obtain the result

$$\overline{P}(g \leq 0) \approx \max_{i,j} P(g(X_{(\mu_i,\sigma_j)}) \leq 0) \approx 0.221 \tag{63}$$

which is the upper probability of failure of the beam considered.

Comparing this result with the result in Section 5 it holds that $0.221 \approx \overline{P}(g \leq 0) \leq \widetilde{P}(g \leq 0) \approx 0.358$.

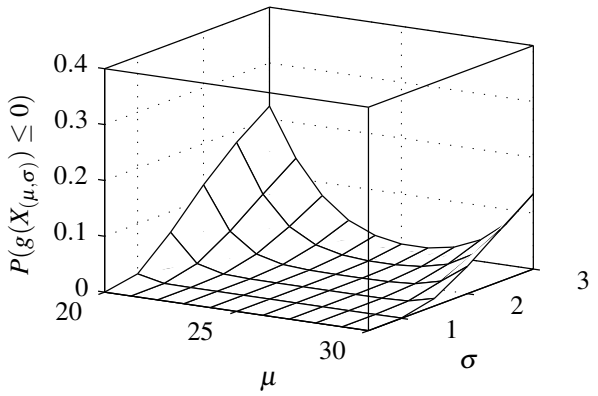


Figure 8: Failure probability $P(g(x_{(\mu,\sigma)}) \leq 0)$ computed at grid points (μ_i, σ_j).

7 Summary and Conclusions

We have discussed two interpretations of upper and lower probabilities, given a family of random variables. The random set approach is supported by the availability of a rich theory as well as various recent applications, e.g. [16, 17, 18, 19]. The approach based directly on the family of random variables has been favored e.g. in [2]. We have shown here that the latter approach gives tighter bounds, i.e., smaller probability intervals, in general. For both approaches, cost saving simulation methods have been presented. We hope that the paper stimulates further research into the computational aspects of imprecise probability.

References

[1] D. A. Alvarez. A Monte Carlo-based method for the estimation of lower and upper probabilities of events using infinite random sets of indexable type. *Fuzzy Sets and Systems*, 160:384–401, 2009.

[2] M. de Angelis, E. Patelli and M. Beer. Line Sampling for Assessing Structural Reliability with Imprecise Failure Probabilities. In: M. Beer, S.-K. Au, J.W. Hall (Eds.). *Vulnerability, Uncertainty, and Risk Quantification: Mitigation and Management*. CD-ROM Proceedings of the 2nd International Conference on Vulnerability and Risk Analysis and Management, and 6th International Symposium on Uncertainty Modelling and Analysis. American Society of Civil Engineers, Reston VA 2014, pp. 915–924.

[3] M. Beer, S. Ferson and V. Kreinovich. Imprecise probabilities in engineering analysis. *Mechanical Systems and Signal Processing*, 37:4–29, 2013.

[4] A. Bernardini and F. Tonon. *Bounding Uncertainty in Civil Engineering - Theoretical Background*. Berlin: Springer-Verlag, 2010.

[5] A. Castaldo and M. Marinacci. Random correspondences as bundles of random variables. In: G. de Cooman, T. Fine, T. Seidenfeld (Eds.), *ISIPTA'01, Proceedings of the Second Symposium on Imprecise Probabilities and Their Applications*, Shaker Publ. BV, Maastricht 2001, pp. 77–82.

[6] S. Destercke, D. Dubois and E. Chojnacki. Relating practical representations of imprecise probabilities. In: G.de Cooman, J. Vejnarova and M. Zaffalon (Eds.), *ISIPTA '07, Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications.* Action M Agency, SIPTA, Prague 2007, pp. 155–164.

[7] S. Ferson, V. Kreinovich, 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.

[8] 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.

[9] T. Fetz. Modelling uncertainties in limit state functions. *Int. J. of Approximate Reasoning*, 53(1):1–23, 2012.

[10] O. Le Maître and O. Knio. *Spectral methods for uncertainty quantification. With applications to computational fluid dynamics*. New York: Springer, 2010.

[11] E. Miranda, I. Couso and P. Gil. Random intervals as a model for imprecise information. *Fuzzy Sets and Systems*, 154(3):386–412, 2005.

[12] E. Miranda, I. Couso and P. Gil. Random sets as imprecise random variables. *Journal of Mathematical Analysis and Applications*, 307(1):32–47, 2005.

[13] I. Molchanov. *Theory of Random Sets*. Berlin: Springer-Verlag, 2005.

[14] R. I. Mullen and R. I. Muhanna. Probability bounds for the system response of non-linear structures. In G. Deodatis, B.R. Ellingwood and D.M. Frangopol (Eds.), *Safety, Reliability, Risk and Life-Cycle Performance of Structures & Infrastructures*. Taylor & Francis Group, London, 2013, pp. 499–502.

[15] M. Oberguggenberger. Engineering. In: T. Augustin, F. Coolen, G. de Cooman, M. Troffaes (Eds.), *Introduction to Imprecise Probabilities.* John Wiley & Sons Ltd, Chichester 2014, pp. 291–304.

[16] M. Oberguggenberger. Stochastic Response Surfaces, Interval Analysis, and the Reliability of Structures. In: M. Beer, S.-K. Au, J.W. Hall (Eds.). *Vulnerability, Uncertainty, and Risk Quantification: Mitigation and Management*. CD-ROM Proceedings of the 2nd International Conference on Vulnerability and Risk Analysis and Management, and 6th International Symposium on Uncertainty Modelling and Analysis. American Society of Civil Engineers, Reston VA 2014, pp. 864–875.

[17] M. Oberguggenberger. Analysis and computation with hybrid random set stochastic models. *Structural Safety*, 52:233–243, 2015.

[18] B. Schmelzer. On solutions to stochastic differential equations with parameters modeled by random sets. *International Journal of Approximate Reasoning*, 51:1159–1171, 2010.

[19] B. Schmelzer. Set-valued assessments of solutions to stochastic differential equations with random set parameters. *Journal of Mathematical Analysis and Applications*, 400:425–438, 2013.

[20] F. Tonon, A. Bernardini and A. Mammino. Determination of parameters range in rock engineering by means of random set theory. *Reliability Engineering and System Safety*, 70:241–261, 2000.

[21] L. Utkin and S. Destercke. Computing expectations with continuous p-boxes: Univariate case. *Int. J. of Approximate Reasoning*, 50:778–798, 2009.

[22] R. Zhang. Structural reliability analysis with imprecise probability using interval importance sampling method. In G. Deodatis, B.R. Ellingwood and D.M. Frangopol (Eds.), *Safety, Reliability, Risk and Life-Cycle Performance of Structures & Infrastructures*. Taylor & Francis Group, London, 2013, pp. 563–570.

Robust Parameter Estimation of Density Functions under Fuzzy Interval Observations

Romain Guillaume
IRIT - Université de Toulouse, France
guillaum@irit.fr

Didier Dubois
IRIT, CNRS & Université de Toulouse, France
dubois@irit.fr

Abstract

This paper deals with the derivation of a probabilistic parametric model from interval or fuzzy data using the maximum likelihood principle. In contrast with classical techniques such as the EM algorithm, that define a precise likelihood function by averaging inside each imprecise observations, our approach presupposes that each imprecise observation underlies a precise one, and that the uncertainty that pervades its observation is epistemic, rather than representing noise. We define an interval-valued likelihood function and apply robust optimisation methods to find a safe plausible estimate of the statistical parameters. The resulting density has a standard deviation that is large enough to cover the imprecision of the observations, making a pessimistic assumption on dispersion. This approach is extended to fuzzy data by optimizing the average of lower likelihoods over a collection of data sets obtained from cuts of the fuzzy intervals, as a trade off between optimistic and pessimistic interpretations of fuzzy data. The principles of this method are compared with those of other existing approaches to handle incompleteness of observations, especially the EM technique.

Keywords. Possibility theory, fuzzy intervals, maximum likelihood, robust optimisation, epistemic uncertainty

1 Introduction

Interval observations, and more generally, set-valued ones, do not always refer to the same situation [1]. Intervals may either represent exact observations of items taking the form of intervals (for instance, the daily min-max temperature ranges across one year), or, on the contrary, imprecise observations of precise quantities. In the first situation, interval data are a special kind of functional data where observations lie in a space of characteristic functions equipped with the suitable metric structure, enabling precise statistical parameters to be derived [18]. In this paper we are interested in the statistical analysis of data when observations are imprecise, more specifically, when we only know that the precise values of observations are restricted by intervals or fuzzy intervals. This kind of fuzzy interval is an epistemic set [1] which attaches to each value the possibility that it is the true observed value (unreachable for the observer). Under the epistemic approach, the expected value and the variance of a set of fuzzy intervals are fuzzy intervals [2].

This paper presents a general iterative approach to compute estimates of the parameters of a density function under imprecise observations, where the lack of precision is an epistemic rather than an aleatory phenomenon. To estimate the quality of parameters of the underlying precise random process, we use the maximum likelihood principle. Nevertheless, under imprecise observations, the likelihood function itself becomes imprecisely appraised too and is thus interval-valued. In this paper we adopt a pessimistic point of view and maximize the lower bound of the likelihood function, with a view to obtain a robust probability density whose standard variation accounts for potentially extreme variability across imprecision intervals.

The paper is organized as follows. In Section 2, we propose an algorithm that evaluates minimal and maximal bounds for the likelihood function. Then, we formulate the estimation problem for interval data as a robust optimization problem, which consists in maximizing the minimal expected likelihood. We study the cases of unimodal and Gaussian distributions. In Section 3, we define an extension of this approach to fuzzy interval data. Especially we discuss how to define a likelihood function for fuzzy interval data. In the literature, a classical approach to handling incomplete data in estimation is the famous EM algorithm [3]. It considers that the likelihood function is a precise function, even if observations are imprecise. In Section 4 we briefly discuss the difference between the two approaches, as well as the optimistic counterpart of ours.

2 Interval Uncertainty

Before solving the problem with fuzzy intervals, we focus on the problem with classical intervals. Firstly, we present a general framework for handing uncertainty on observations whatever the parametrized family of distributions. Secondly, we present an algorithm to solve the problem for unimodal density distributions. Finally, we study the case of normal density distributions.

2.1 General Framework

Let $\{x_i : i \in N\}$ with $|N| = n$, be a set of precise observations. To evaluate the quality of the parameters of the distribution that represents these observations, the usual approach is to define a likelihood measure $f(x_i|\theta)$ for each piece of data. Note that $f(x_i|\theta)$ can be understood as the possibility that the generation process for x_i is based on the parameter value θ [4]. The density function with vector of parameters θ and independent observations $\{x_i : i \in N\}$ takes the form of a product of likelihood functions:

$$L = \prod_{i \in N} f(x_i|\theta) \tag{1}$$

A standard criterion to define the parameters of the density function is the maximization of this likelihood function

$$\max_{\theta} \prod_{i \in N} f(x_i|\theta) \tag{2}$$

Under uncertainty, observations are of limited precision, and take the form of intervals $x_i \in [\underline{x_i}, \overline{x_i}], \forall i \in N$. Let $\Gamma = [\underline{x_1}, \overline{x_1}] \times ... \times [\underline{x_n}, \overline{x_n}]$ be the set of possible n-tuples of observations, we call *selections*. Namely, the selection $X \in \Gamma$ with $X = (x_1, ..., x_n)$ is a possible realization of the imprecise observation Γ. Fixing the parameter θ, one may argue that, in the spirit of [1], if observations are imprecise, the likelihood evaluation should become imprecise too, that is, $L(\theta) \in [\underline{L}; \overline{L}]$ with $\underline{L}$ and $\overline{L}$ respectively defined by:

- Lower likelihood

$$\underline{L}(\theta) = \min_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta). \tag{3}$$

- Upper likelihood

$$\overline{L}(\theta) = \max_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta). \tag{4}$$

To find robust solutions that cover potential variability, we can determine the parameter value (denoted by θ^{Rob}) which maximizes the *lower* likelihood. It can be formulated as a robust optimization problem:

$$\max_{\theta} \min_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta) \tag{5}$$

This is equivalent to the log-likelihood problem:

$$\max_{\theta} \min_{X \in \Gamma} \sum_{i \in N} ln(f(x_i|\theta)) \tag{6}$$

2.2 Resolution Method

In this section, we propose an algorithm that evaluates the lower and upper bounds of the likelihood function for given parameters θ for a density function under the form (1). For a given data vector $X^* = (x_1^*, ..., x_n^*)$, the log-likelihood function $\sum_{i \in N} ln(f(x_i^*|\theta))$ is supposed to be convex with θ and to have a derivative.

Assumption 1 *$\exists x^m \in \mathbb{R}$ such that $f(x_i^*|\theta)$ is an increasing function on $]-\infty, x^m]$ and decreasing on $[x^m, +\infty[$.*

If the distribution is unimodal, x^m is the mode of the distribution.

2.2.1 Determining Upper and Lower Likelihood Functions

Note that the upper and lower likelihoods are of the form $f(X|\theta)$ for some $X \in \Gamma$. Moreover, from Property 1, we know that for a given parameter value θ, the minimum of function $f(x_i|\theta)$, where $x_i \in [\underline{x_i}, \overline{x_i}]$, is attained at the boundary of the domain ($x_i = \underline{x_i}$ or $x_i = \overline{x_i}$). It is called a *worst case selection*. This is not true for the *best case selection* obtained from the upper likelihood. Since the observations are assumed to be independent, the solution of problems (3) (worst case X^w for $\underline{L}(\theta)$) and (4) (best case X^b for $\overline{L}(\theta)$) can be computed using the following rules:

$$\text{if } x^m \in]-\infty; \underline{x_i}[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \tag{7}$$

$$\text{if } x^m \in]\overline{x_i}; \infty[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \tag{8}$$

$$\text{if } x^m \in [\underline{x_i}; \overline{x_i}] \text{ then } \begin{cases} x_i^b = x^m \\ x_i^w = \underline{x_i} \text{ if } f(\underline{x_i}|\theta) > f(\overline{x_i}|\theta), \\ \overline{x_i} \text{ otherwise.} \end{cases} \tag{9}$$

2.2.2 Computing Robust Parameters

The worst case selection $X^w(\theta) \in \Gamma$ is the one that minimizes the lower likelihood ($\underline{L}$) with parameter θ. If the density function is unimodal, it follows that the maximum likelihood problem comes down to discrete optimisation, that is we can restrict the selections of observations to extreme selections $X^w \in \Gamma_{\text{dis}}$, with $\Gamma_{\text{dis}} = \{\underline{x_1}, \overline{x_1}\} \times ... \times \{\underline{x_n}, \overline{x_n}\}$, the set of extreme assignments of x_i. Using Lagrange relaxation, problem (5) can be transformed into the following problem:

$$h^* = \max_{\theta} \sum_{X \in \Gamma_{\text{dis}}} \lambda_X \times (\sum_{i \in N} ln(f(x_i|\theta))) \tag{10}$$

where the Lagrange coefficients λ_X respect the conditions

$$\forall X \in \Gamma_{\text{dis}}, \lambda_X = \begin{cases} 1 \text{ if } h^{min} = \sum\limits_{i \in N} ln(f(x_i|\theta)) \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

h^{min} being the minimal value of the log-likelihood over the selections:

$$h^{min} = \min_{X \in \Gamma_{\text{dis}}} \sum_{i \in N} ln(f(x_i|\theta)) \tag{12}$$

Proposition 1 *Expression (10) gives the optimal solution of the problem (5) if the Lagrange coefficients $\lambda_X, \forall X \in \Gamma_{dis}$ satisfy the conditions (11).*

Proof: Note that if the Lagrange coefficients respect the conditions (11), the expression (10) is equivalent to $h^* = \max\limits_{\theta} k \times (\min\limits_{X \in \Gamma_{\text{dis}}} \sum\limits_{i \in N} ln(f(x_i|\theta)))$ where k is the number of functions $\sum\limits_{i \in N} ln(f(x_i|\theta))$ that intersect at the maximum; it is the number of Lagrange coefficients $\lambda_X = 1$. Hence, the optimal solution of the previous expression (10) is the same as the optimal solution of problem $h^* = \max\limits_{\theta} \min\limits_{X \in \Gamma_{\text{dis}}} \sum\limits_{i \in N} ln(f(x_i|\theta)))$ which is equivalent to the problem (5) □

To solve problem (10), we use an iterative algorithm (Algorithm 1), which is an adaptation of the Uzawa method [5] to our problem.

Nevertheless the number of extreme selections is equal to 2^n. So we construct an iterative algorithm for solving problem (5) based on iterative relaxation scheme for min-max problems proposed in [6] and developed for min-max regret linear programming problems with an interval objective function [7, 8] coupled with Uzawa method.

Let RX-ROB be the problem (10) with a given set of assignments $\Gamma^*_{\text{dis}} \subseteq \Gamma_{\text{dis}}$. Obviously, the maximal cost h^* of problem RX-ROB over the discrete assignment set Γ^*_{dis} is an upper bound on the maximal cost of problem (6). Our algorithm (Algorithm 2) starts with zero upper bound $UB = 0$ and initial parameters θ^* (for instance the optimal parameter for the assignment of the mid-points of intervals) and empty discrete scenario set, $\Gamma^*_{\text{dis}} = \emptyset$. At each iteration, a worst case assignment X^w for θ^* is computed using rules (7, 8) and (9). Clearly, $\underline{L}(\theta^*)$ is an upper bound of $\underline{L}(\theta^{Rob})$. If a termination criterion is fulfilled (usually $\underline{L}(\theta^*) \leq UB - \epsilon$, $\epsilon > 0$ is a given constant) then the algorithm stops with an optimal robust parameter θ^*. Otherwise the worst case selection X^w is added to Γ^*_{dis}. Next the updated problem (RX-ROB) is solved to obtain a better candidate θ^* for an optimal solution to (5) and a new upper bound $UB = h^{min}$, based on Γ^*_{dis}. Since set Γ^*_{dis} is updated during the course of the algorithm, the computed values are upper bounds that form a nonincreasing sequence of values. Then, a new iteration is started.

Algorithm 1: A robust solution under a set of discrete scenarios

Input: Initial parameters $k = 0$, λ^0_X, the set of selections Γ_{dis}, and a convergence tolerance parameter $\rho > 0$.
Output: An optimal solution θ^{Rob},h^{Rob}
Step 1. Compute θ^k the optimal solution of problem (10) using $\lambda^k_X, X \in \Gamma_{\text{dis}}$
Step 2. If $\forall X \in \Gamma_{\text{dis}}$ the condition (11) is satisfied, then output θ^k, h^{min} and STOP.
Step 3. Compute the λ^{k+1}_X:
if $h^{min} = \sum\limits_{i \in N} ln(f(x_i|\theta^{k+1})$ then $\lambda^{k+1}_X = 1$ else
decrease the Lagrange parameter using
$\lambda^{k+1}_X = \max(0, \lambda^k_X - \rho \times (\sum\limits_{i \in N} ln(f(x_i|\theta^{k+1}) - h^{min}))$
Step 4. $k := k + 1$, and go to Step 1.

Algorithm 2: Finding optimal robust parameters.

Input: Observations $x_i = [\underline{x_i}; \overline{x_i}], \forall i \in N$, initial parameters θ^*, a convergence tolerance parameter $\epsilon > 0$.
Output: An optimal robust parameter θ^{Rob}
Step 0. $k := 0$, $UB := 0$, $\Gamma^*_{\text{dis}} := \emptyset$.
Step 1. $\theta^k := \theta^*$.
Step 2. Compute a worst case selection X^w for θ^k by solving problem (3) using rules (7), (8), (9). Then let $h = \sum\limits_{i \in N} ln(f(x^w_i|\theta))$
Step 3. If $(h \leq UB - \epsilon)$ then output θ^k and STOP.
Step 4. $k := k + 1$.
Step 5. $X^k := X^w$, $\Gamma^*_{\text{dis}} := \Gamma^*_{\text{dis}} \cup \{X^k\}$
Step 6. Compute an optimal solution θ^* by Algorithm 1, using Γ^*_{dis}; then set $UB = h^{min}$ and go to Step 1.

2.3 The Case of Normal Distributions

We suppose that the random variable follows a normal distribution:

$$f(x|(\mu,\sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (13)$$

Upper and lower likelihoods can be reformulated into

$$\underline{L}(\mu,\sigma) = \min_{X\in\Gamma} \prod_{i\in N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (14)$$

$$\overline{L}(\mu,\sigma) = \max_{X\in\Gamma} \prod_{i\in N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (15)$$

2.3.1 Determining the Upper and Lower Likelihoods

The lower log-likelihood in the case of normal distributions becomes:

$$ln(\underline{L}(\mu,\sigma)) = -(\tfrac{Nln(\sigma^2)}{2} + \frac{1}{\sigma^2} \max_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2)$$

Likewise, the upper log-likelihood in the case of normal distributions becomes:

$$ln(\overline{L}(\mu,\sigma)) = -(\tfrac{Nln(\sigma^2)}{2} + \frac{1}{\sigma^2} \min_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2)$$

In this case the mode $x^m = \mu^*$ is the mean, and since the normal distribution is symmetric, the general equations (7, 8) and (9) that compute the worst and the best case selections become respectively:

$$\text{if } \mu^* \in]-\infty; \underline{x_i}[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \quad (16)$$

$$\text{if } \mu^* \in]\overline{x_i}; \infty[\text{ then } \begin{cases} x_i^b = \overline{x_i} \\ x_i^w = \underline{x_i} \end{cases} \quad (17)$$

$$\text{if } \mu^* \in [\underline{x_i}; \overline{x_i}] \text{ then}$$

$$\begin{cases} x_i^b = \mu^* \\ x_i^w = \underline{x_i} \text{ if } (\underline{x_i}-\mu^*)^2 > (\overline{x_i}-\mu^*)^2, \\ \overline{x_i} \text{ otherwise} \end{cases} \quad (18)$$

It follows that the complexity for evaluating the lower and the upper likelihoods is $O(n)$.

2.3.2 Computing Robust Parameters

We can further decompose the problem of finding robust parameters into a sequence of two problems:

- first find the robust optimal μ^{rob}, solving the problem

$$ROB_{N,\mu}: \quad \min_{\mu} \max_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2 \quad (19)$$

- and then compute the robust optimal σ^{rob}. We get the variance around μ^{rob} using the optimal selection X^w obtained at the previous step:

$$ROB_{N,\sigma}: \quad \sigma^{rob} = \sqrt{\frac{\sum_{i\in N} (x_i^w-\mu^{rob})^2}{n}} \quad (20)$$

Let us now focus on the problem $ROB_{N,\mu}$. Let $\underline{\mu} = \frac{1}{n}\sum_{i\in N} \underline{x_i}$ and $\overline{\mu} = \frac{1}{n}\sum_{i\in N} \overline{x_i}$.

Proposition 2 *The optimal solution μ^{rob} of the problem ROB_N lies in $[\underline{\mu}, \overline{\mu}]$.*

Proof Suppose $\exists \mu^{rob} < \underline{\mu}$. We have two cases. The first one is: the selection X^w associated to μ^{rob} is the same as the one for $\underline{\mu}$. We also know that, if $\mu^{rob} < \underline{\mu}$ then $\sum_{i\in N} (x_i-\mu^{rob})^2 > \sum_{i\in N} (x_i-\underline{\mu})^2$, since $\forall X \in \Gamma$, the optimal value $\mu^{op} \in [\underline{\mu}, \overline{\mu}]$, that contradicts the assumption that μ^{rob} is the optimal robust solution.

The second case is $X^w = X^w_{\underline{\mu}} + \delta$ where $X^w_{\underline{\mu}}$ is the worst case selection induced by $\underline{\mu}$ and δ is a vector of non-negative values. So $\sum_{i\in N} (x_i^w-\mu^{rob})^2 > \sum_{i\in N} (y_i-\mu^{rob})^2$ and $y_i = (X^w_{\underline{\mu}})_i$. We know that if $\mu^{rob} < \underline{\mu}$ then $\sum_{i\in N} (x_i-\mu^{rob})^2 > \sum_{i\in N} (x_i-\underline{\mu})^2$. Hence, $\sum_{i\in N} (x_i^w-\mu^{rob})^2 > \sum_{i\in N} (y_i-\underline{\mu})^2$, which contradicts the assumption that μ^{rob} is the optimal robust solution. The proof for the upper bound is similar. □

In the following Algorithm 3, we use the derivative

$$\frac{d(\sum_{i\in N} (x_i-\mu)^2)}{d\mu} = 2n\mu - 2\sum_{i\in N} x_i$$

.

Theorem 1 *Algorithm 3 finds the optimal robust parameter μ^{rob}.*

Proposition 3 *The complexity of computing the optimal robust solution μ^{rob} and σ^{rob} is $O(n.ln(|\mu|)$*

Algorithm 3: Finding optimal robust parameters for normal distribution.

Input: Observations $[\underline{x_i}; \overline{x_i}], \forall i \in N$, a convergence tolerance parameter $\epsilon > 0$.
Output: An optimal robust parameter μ^{Rob}
Step 0. $k := 0$, $a = \underline{\mu}$, $b = \overline{\mu}$.
Step 1. Compute a worst case selection X_c^w for the value $c = \frac{1}{2}(a + b)$.
Step 2. Compute the value $D = 2n\mu - 2\sum_{i \in N} x_i$ for the worst case selection X_c^w
Step 3. If $D < 0$ then $a := c$, else $b := c$.
Step 4. If $a - b > \epsilon$ then go to Step 1 else return $\frac{1}{2}(a + b)$ and STOP.

Proof: The major part of one iteration of the dichotomy algorithm is spent computing the worst case selection, which is $O(n)$. Since the dichotomy algorithm is $O(ln(|\mu|))$ where $|\mu|$ depends on the width of the interval $[\underline{\mu}, \overline{\mu}]$ and the precision parameter ϵ, the complexity of Algorithm 3 is $O(n.ln(|\mu|))$. And since σ^{rob} is directly computed from μ^{rob}, the complexity of computing the optimal robust solution μ^{rob} and σ^{rob} is $O(n.ln(|\mu|))$. □

2.3.3 Robust Solution vs. Maximal Variance

The robust solution can be understood as the parameter μ that minimizes the maximal possible variance under uncertainty (across all scenarios compatible with the interval data). Note that the problem $ROB_{N,\mu}$ is a relaxation of the problem of maximization of the variance of interval data [9]:

$$\max_{X \in \Gamma} \sum_{i \in N} (x_i - \sum_{i \in N} x_i/n)^2 \tag{21}$$

since, in the latter, $\mu = \sum_{i \in N} x_i/n$, while in problem $ROB_{N,\mu}$, μ is an independent variable. Let $(\sigma^{\max})^2$ be the maximal variance in problem (21). An imprecise probability solution to the estimation problem could be the set of normal distributions with $\mu \in [\underline{\mu}, \overline{\mu}]$ and $\sigma = \sigma^{\max}$. However the robust solution has the following property:

Proposition 4 $\sigma^{rob} \geq \sigma^{\max}$

Proof: It is enough to notice that

$$\begin{aligned} n(\sigma^{\max})^2 &= \max_{X \in \Gamma} \min_{\mu} \sum_{i \in N} (x_i - \mu)^2 \\ &\leq \min_{\mu} \max_{X \in \Gamma} \sum_{i \in N} (x_i - \mu)^2 = n(\sigma^{rob})^2 \quad \square \end{aligned}$$

Assume there is a single worst case solution X^w in problem $ROB_{N,\mu}$. In that case, the minimum is attained for the mean value $\mu^{rob} = \sum_{i \in N} x_i^w/n$, hence $\sigma^{rob} = \sigma^{\max}$. The maximal variance solution is then robust. However if there are several worst case solutions $X_j^w, j = 1, \ldots k$ in problem $ROB_{N,\mu}$, μ^{rob} is the intersection point of k parabolas

$$f_j(\mu) = \sum_{i \in N} (x_{ji}^w - \mu)^2,$$

while the maximal variance corresponds to the maximal ordinate of the minima of each parabola whose abscissa is

$$\mu^j = \sum_{i \in N} x_{ji}^w/n.$$

So the robust solution is a kind of compromise between extreme data selections, and is more pessimistic than the maximal variance solution.

3 Fuzzy Interval Uncertainty

We now use a more refined modeling of uncertainty pervading the observations. They are modeled by fuzzy intervals $\widetilde{x}_i, \forall i \in N$.

3.1 Selected Notions of Possibility Theory

A *fuzzy interval* $\tilde{A}$ is a fuzzy set in $\mathbb{R}$ whose membership function $\mu_{\tilde{A}}$ is normal, quasi concave and upper semicontinuous. Usually, it is assumed that the support of a fuzzy interval is compact. The main property of a fuzzy interval is the fact that all its α-cuts, that is, the sets $\tilde{A}^{[\alpha]} = \{x : \mu_{\tilde{A}}(x) \geq \alpha\}$, $\alpha \in (0, 1]$, are closed intervals. We will assume that $\tilde{A}^{[0]}$ is the smallest closed set containing the support of $\tilde{A}$. So, every fuzzy interval $\tilde{A}$ can be represented as a family of closed intervals $\tilde{A}^{[\alpha]} = [\underline{a}^{[\alpha]}, \overline{a}^{[\alpha]}]$, parametrized by the value of $\alpha \in [0, 1]$.

Let us now recall the possibilistic interpretation of fuzzy intervals. *Possibility theory* [10] is an approach to handle incomplete information and it relies on two dual measures: *possibility* and *necessity*, which express plausibility and certainty of events. Both measures are built from a *possibility distribution*. Let a fuzzy interval $\widetilde{A}$ be attached with a single-valued variable a (an uncertain real quantity). The membership function $\mu_{\widetilde{A}}$ is understood as a possibility distribution, $\pi_a = \mu_{\widetilde{A}}$, which describes the set of more or less plausible, mutually exclusive values of the variable a. It can encode a family of probability functions [11]. In particular, a degree of possibility can be viewed as the upper bound of a degree of probability [11]. The value of $\pi_a(v)$ represents the possibility degree of the assignment $a = v$, i.e. $\Pi(a = v) = \pi_a(v) = \mu_{\widetilde{A}}(v)$, where $\Pi(a = v)$ is the possibility of the event that a will take the value of v. In particular, $\pi_a(v) = 0$ means

that $a = v$ is impossible and $\pi_a(v) = 1$ means that $a = v$ is fully plausible. Equivalently, it means that the value of a belongs to an α-cut $\widetilde{A}^{[\alpha]}$ with confidence (or degree of necessity) $1 - \alpha$. It can be viewed as a random set defined by a multi-mapping from the unit interval equipped with Lebesgue measure to intervals consisting of cuts $\tilde{A}^{[\alpha]}$ [14]. Discrete approximations of π can also be viewed as random sets $(m, F)_\pi$, with nested focal sets E_i and masses $m(E_i)$, such that:

$$\begin{cases} E_i = \{x \in \mathbb{R} | \pi(x) \geq \alpha_i\} \\ m(E_i) = \alpha_i - \alpha_{i-1} \end{cases} \tag{22}$$

The possibility distribution is then approximated by: $\pi'(x) = \sum_{x \in E_i} m(E_i)$ [12].

3.2 Fuzzy Interval Datasets

A fuzzy interval data set is a collection of fuzzy intervals $\widetilde{x}_i, i = 1 \ldots, N$ whose membership functions are regarded as possibility distributions π_i restricting the values of the x_i's. The x_i's are stochastically independent but their uncertainties are non-interactive. We have thus extended the scenario set Γ from intervals (see Section 2) to the fuzzy case and now $\widetilde{\Gamma}$ is a fuzzy set of scenarios with membership function $\mu_{\widetilde{\Gamma}}(X) = \pi(X)$, $X \in \mathbb{R}^n$. The value of $\pi(X)$ stands for the possibility of the event that scenario $X \in \mathbb{R}^n$ has occurred. Hence, the possibility distributions associated with the observations x_i, forming the vector X, induce the following possibility distribution over all assignments in $X \in \mathbb{R}^n$ (see [13]):

$$\pi(X) \quad = \quad \min_{i=1,\ldots,n} \pi_i(x_i). \tag{23}$$

We see at once that the α-cuts of $\widetilde{\Gamma}$ for every $\alpha \in [0, 1]$ are such that: $\widetilde{\Gamma}^{[\alpha]} = \{X : \pi(X) \geq \alpha\} = [x_1^{-[\alpha]}, x_1^{+[\alpha]}] \times \cdots \times [x_T^{-[\alpha]}, x_T^{+[\alpha]}]$, from (23) and the definition of α-cut. Notice that $\widetilde{\Gamma}^\alpha$, $\alpha \in [0, 1]$, is the Cartesian product containing all selections (scenarios) whose possibility of occurrence is not less than α.

3.3 Formulations of Likelihood under Fuzzy Observations

In this section, we extend the definition of interval likelihood to the case of fuzzy intervals. There are several ideas that can be implemented to bring the fuzzy interval maximal likelihood problem back to a standard interval problem:

1. The simplest one is to turn fuzzy intervals into intervals by taking the interval mean [14], the Aumann integral $I(\widetilde{x}_i) = \int_0^1 [x_i^{-[\alpha]}, x_i^{+[\alpha]}] d\alpha$. However, one may then wonder why to start with fuzzy intervals in the first place.

2. Alternatively, we can solve the interval maximum likelihood problem for each α-cut, which would provide a set of possible solutions. If we remember that the fuzzy interval can also be interpreted in terms of subjective uncertainty, whereby $1 - \alpha$ is the degree of certainty that $[x^{-[\alpha]}, x^{+[\alpha]}]$ contains the actual observation x, the optimal parameter θ^*_α obtained from applying the interval approach to the α-cuts $\{[x_i^{-[\alpha]}, x_i^{+[\alpha]}] : i = 1, \ldots, n\}$ can be interpreted as the robust value of the model parameter corresponding to certainty $1-\alpha$, which can be viewed as a degree of pessimism of the solution. Indeed, if $\alpha = 1$ we take an optimistic view on the precision of the data, while if $\alpha = 0$, we assume the data is very imprecise and we try to be robust in the face of large interval uncertainty.

3. Yet another approach consists in considering all cuts of all fuzzy data $\widetilde{x}_i$ namely,

   $$\{[x_i^{-[\alpha]}, x_i^{+[\alpha]}] : i = 1, \ldots, n, \alpha \in [0, 1]\}$$

   as an equivalent set of interval data. In practice, this data set can be approximated using a finite set of cuts using equation (22). This approach considers the set of fuzzy data as a convex set of probabilities, induced by a random set in the spirit of Couso and Sanchez [15]. Indeed, the fuzzy data set is viewed as equivalent to a set of intervals generated as follows: Picking i at random in $\{1, \ldots, n\}$ and picking an α-cut at random ($[0, 1]$ is equipped with Lebesgue measure), obtaining the interval $[x^{-[\alpha]}, x^{+[\alpha]}]$.

All above approaches are amenable to a solution via the above proposed algorithms. These methods can be considered as somewhat extreme, as the first one does away with gradual membership, the second is difficult to use in practice (how to choose the best cut), and the third one considers two cuts of the same fuzzy observations as equivalent to two cuts each from a different observation, or in other words, fuzzy observations are the result of grouping together nested interval observations. Our next approach is a kind of trade-off between these views. Here we rely not on the mean interval of each fuzzy interval separately, but on the average of interval likelihoods obtained from all data sets $\widetilde{\Gamma}^\alpha$, $\alpha \in [0, 1]$.

3.4 The Average Robust Estimation Problem

We define a mean interval likelihood as follows:

Definition 1 *The mean interval likelihood under fuzzy observations is* $[\int_{\alpha\in[0,1]} \underline{L}^{\alpha} d\alpha, \int_{\alpha\in[0,1]} \overline{L}^{\alpha} d\alpha]$.

It can be approximated using a finite set of cuts using equation (22) and the average likelihood can then be expressed as:

$$[\sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha})\underline{L}^{\alpha_j}, \sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha})\overline{L}^{\alpha_j}] \quad (24)$$

The estimation of minimal and maximal likelihood under fuzzy observations can then be computed using the formulae (7, 8) and (9) $\forall i \in N, \forall j = 1, .., k$.

This average interval likelihood approach can be viewed as a balanced solution between working with the cores of the fuzzy intervals and their supports, while not letting cuts of a single fuzzy interval play the same role as cuts of different fuzzy intervals.

In the context of fuzzy information, the average robust problem can be formulated as follows:

$$\underline{L}^{rob} = \max_{\theta} \sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha}) \min_{X\in\widetilde{\Gamma}_j^{\alpha}} \sum_{i\in N} ln(f(x_i|\theta)) \quad (25)$$

The reader may object to this formulation, as it seems that we give up our pessimistic point of view on interval data. However, it is not straightforward to define pessimism in a simple way in the face of fuzzy intervals. Indeed, fuzzy intervals carry two dimensions of pessimism, horizontal and vertical. On the one hand, the vertical dimension pertains to the choice of a cut of a fuzzy interval. Taking a cut at level 1, is optimistic in the sense that it is a narrow plausible range. Taking the support is safe but perhaps yields too imprecise an interval. On the other hand, the horizontal dimension (which end of the cut to choose ?) is the one at work in our approach to interval data, leading to take a pessimistic view on variance in the presence of imprecision. The approach proposed here achieves a global trade-off between vertical optimism and pessimism, and retains a pessimistic horizontal view.

3.4.1 The General Case

For simplicity we assume the discretisation of the membership set is such that $\forall j = 1, .., k, m(\widetilde{\Gamma}_j^{\alpha}) = 1/k$ (equidistant cuts).

Proposition 5 *The problem (25) can be reformulated as follows,*

$$h^* = \max_{\theta} \sum_{j=1}^{k} \sum_{X\in\widetilde{\Gamma}_{dis}^{\alpha_j}} \lambda_X^{\alpha_j} \times (\sum_{i\in N} ln(f(x_i|\theta))) \quad (26)$$

under the conditions: $\forall X \in \Gamma_{dis}$,

$$\lambda_X^{\alpha_j} = \begin{cases} \frac{1}{n_{\alpha_j}} \text{ if } h^* = \sum\limits_{i\in N} ln(f(x_i|\theta)), \\ 0, \text{ otherwise,} \end{cases} \quad (27)$$

where n_{α_j} *is the number of non-zero coefficients* $\lambda_X^{\alpha_j}$.

In fact, n_{α_j} is the number of functions $\sum\limits_{i\in N} ln(f(x_i^{[\alpha]}|\theta))$ that intersect at the maximum.

Proof: Note that if the Lagrange coefficients respect the conditions (27), the expression (26) is equivalent to $h^* = \max\limits_{\theta} \sum\limits_{j=1}^{k} \frac{n_{\alpha_j}}{n_{\alpha_j}} \times (\min\limits_{X\in\Gamma_{\text{dis}}} \sum\limits_{i\in N} ln(f(x_i|\theta)))$. Hence, the optimal solution of the previous expression (26) is the same as the optimal solution of problem (25). □

We can also generalize Algorithms 1 and 2 to the case of fuzzy observations by modifying Step 3 of Algorithm 1. This step becomes, for all $X \in \widetilde{\Gamma}_{\text{dis}}^{\alpha_j}$:

- if $h_j^{min} = \sum\limits_{i\in N} ln(f(x_i|\theta))$, then $\lambda_X^{k+1,\alpha_j} = \frac{1}{n_{\alpha_j}}$
- else $\lambda_X^{k+1,\alpha_j} = \max(0, \min(\lambda_X^{k,\alpha_j}, \frac{1}{n_{\alpha_j}}) - \rho(\sum\limits_{i\in N} ln(f(x_i|\theta^{k+1}) - h_j^{min}))$.

And Step 2 of Algorithm 2 must be used to find the worst case selection for each $j = 1, .., k$.

3.4.2 The Case of Normal Distributions

In the case of fuzzy observations, the optimal mean μ^* belongs to the set of means μ for scenarios with $\alpha = 0$.

Proposition 6 *The optimal value of the mean* μ *of the problem* ROB_N *is* $\mu^{rob} \in [\underline{\mu}^{[0]}, \overline{\mu}^{[0]}]$ *with* $\underline{\mu}^{[0]} = \frac{1}{n}\sum\limits_{i\in N} \underline{x_i}^{[0]}$ *and* $\overline{\mu}^{[0]} = \frac{1}{n}\sum\limits_{i\in N} \overline{x_i}^{[0]}$

To generalize Algorithm 3 to fuzzy observations, Step 1 becomes: Compute the worst case selection $X_j^w, \forall j = 1, .., k$ for the value $c = \frac{1}{2}(a + b)$. And the derivative of the likelihood function becomes $2nk\mu - 2\sum\limits_{j=1}^{k}\sum\limits_{i\in N} x_i^k$

Proposition 7 *The complexity of computing the optimal robust solution* μ^{rob} *and* σ^{rob} *is* $O(n.k.ln(|\mu^{[0]}|))$.

It is the same complexity as in the interval case but increased by a factor k (the number of cuts of the fuzzy intervals used in the algorithm).

4 Related Works

In the literature, a definition of likelihood under incomplete observations have been proposed by Dempster et al. [3]. This is the basis of the classical EM algorithm. Applied to our interval observations, it comes down to

Definition 2 *The likelihood of θ under one imprecise observation ($x \in [\underline{x};\overline{x}]$) is $L(\theta,[\underline{x};\overline{x}]) = P([\underline{x};\overline{x}]|\theta) = \int_{x\in[\underline{x};\overline{x}]} f(x|\theta)\,\mathrm{d}x$.*

The problem (5) is replaced by

$$\max_\theta P(\Gamma|\theta) = \max_\theta \prod_{i\in N} P_i([\underline{x_i},\overline{x_i}]|\theta) \qquad (28)$$

The EM method relies on the choice of an initial probability density, then compute averages over the intervals $[\underline{x_i},\overline{x_i}]$, which provides a precise dataset from which another density is obtained via maximal likelihood, and the process starts again until convergence.

However, this approach is often presented as handling latent unobserved variables, or missing values, not especially interval-valued observations. Namely, the authors using the EM algorithm rather present the framework as one with two kinds of variables: $\mathcal{X}$, a set of observed variables with precise realizations $\mathbf{x}$ and $\mathcal{Z}$ a set of non-observed variables, while here we consider a set of incomplete observations of the same variable. In the setting of the EM algorithm, an incomplete observation is thus of the form $\mathbf{x}\times Dom(\mathcal{Z})$, where Dom is short for domain. In other words, observations are set-valued, but form a partition of $Dom(\mathcal{X}\cup\mathcal{Z})$ into disjoint classes. So, moving from $\mathcal{X}$ to $\mathcal{X}\cup\mathcal{Z}$ corresponds to a change in granularity, whereby the second space is finer and the first can be viewed as a partition of the second. In such a situation, it sounds natural to consider that the likelihood $f(\mathbf{x}|\theta)$ is equated to the integral $\int_{Dom(\mathcal{Z})} f(\mathbf{x}\times\mathbf{z}|\theta)d\mathbf{z}$ (because the data $\mathcal{Z}$ is supposed to be missing at random, i.e., $f(\mathbf{x}|\theta) = f(\mathbf{x}|\mathbf{z},\theta)$). Insofar as one is only interested in events in the algebra formed by the coarse partition, the formulation of the likelihood function as in the EM approach is justified.

Recently, the EM algorithm has been generalized by Denoeux [16] to the case of data taking the form of mass functions of belief functions, and he also uses a scalar likelihood function defined as a weighted average of the EM likelihood:

Definition 3 *The likelihood of of a single imprecise observation x_i described by a belief function with mass function m_i bearing on intervals $[\underline{x}_j;\overline{x}_j]$ is $L(\theta,m) = \sum_j m([\underline{x}_j;\overline{x}_j])L(\theta,[\underline{x}_j;\overline{x}_j])$.*

Note that this definition also applies to fuzzy interval data viewed as consonant belief functions, and thus leads to yet another extension of maximum likelihood estimation to fuzzy data, studied by Denoeux [17].

In the case of interval-valued observations viewed as imprecise data (dealt with in our paper), the formulation of likelihood after the EM algorithm looks questionable.

On the one hand, there seems to be no point averaging the probabilities $f(x_i|\theta)$ over the interval $[\underline{x_i};\overline{x_i}]$, as if computing the frequency of this event from a sample space. Indeed, each observation $[\underline{x_i};\overline{x}_i]$ is a disjunctive set, one value of which is the real (unique) realization x_i. This defect in observing the x_i's leads to possibilistic uncertainty about the actual probability of their realizations (knowing θ), better expressed by the interval likelihood function $[\underline{L};\overline{L}]$. It contrasts with the problem of computing a frequency $P([\underline{x};\overline{x}]|\theta)$ based on a collection of precise observations. In the latter case, all observations inside $[\underline{x};\overline{x}]$ have been observed (it is a conjunctive set). In contrast, each interval $[\underline{x}_i,\overline{x}_i]$ is the incomplete description of a single precise observation x_i. The EM approach seems to interpret the equal possibility of all values in $[\underline{x};\overline{x}]$ as being an equal probability, or at least, it seems to admit the existence of a random process generating precise values inside this interval. In the case of latent or unobserved variables, it makes sense if they are indeed driven by an unobserved random process. But this is not our assumption in the case of interval uncertainty.

On the other hand, the overlapping nature of the interval valued data makes it hard to assume the existence of auxiliary random processes inside each interval, while if the incomplete observations partition the sample space, this assumption may look more natural.

Besides, Definition 2 does not generalize the definition of likelihood in the context of perfect observations, since under this definition, the likelihood of precise values is 0.

Here we view the intervals as describing epistemic uncertainty bearing on the observations that stem from a random process generating precise (even if grossly observed) data. Another option could be to assume that there is a second random process generating the intervals surrounding the outcomes of the first one. This purely aleatory view of imprecise observations is also at work in the trend on fuzzy random variables after Puri and Ralescu where they are interpreted as standard random variables whose images are functions [18]. However in this case, there would be three random variables, say $x, u > 0, v > 0$, such that the observed intervals are realizations of the form $[x-u, x+v]$. Then

we could rightfully define the likelihood function:

$$P([\underline{x}_i, \overline{x}_i]|\theta) = \int_{\underline{x}_i = x-u, \overline{x}_i = x+v} P(x, u, v|\theta) dx du dv$$

and apply the EM algorithm. However, here we consider the uncertainty pervading the observations x_i is not aleatory at all, it is just sheer lack of information due to the coarseness of the observation tool, and the width of the interval surrounding the x_i's is supposedly not generated by a random process.

More recently, Hüllermeier [19] proposed an approach similar to ours for simultaneously optimizing a model and disambiguating the (interval-valued) data. He presents the approach in terms of minimizing a loss function, and applies it to regression and classification problems. However, his idea comes down to maximizing the product of upper likelihoods in our setting, while our proposal, more in the spirit of robust optimisation, takes a pessimistic view on imprecise observations. Note that our approach also leads to disambiguating the data, albeit taking an opposite view, covering potential dispersion of the actual data, as testified by the use of extreme selections induced by rules (7, 8, 9).

Let us compare the two approaches on a simple case with two interval observations $x_1 \in [20, 30]$ and $x_2 \in [30, 40]$. The result of minimizing the loss function (maximizing the upper likelihood) is $x_1 = 30, x_2 = 30$ so $\mu = 30$ and $\sigma = 0$. In contrast, the robust approach applied to normal distributions will select the values $x_1 = 20, x_2 = 40$ so $\mu = 30$ and $\sigma = 10$. The Hüllermeier approach can be understood as the fusion of information items $x \in [20, 30]$ and $x \in [30, 40]$, privileging the common parts, which is optimistic, while our approach tends to assume the information could be very dissonant, with a variance equal to 100. See [20] for more comments along this line on the optimistic approach. Note that the EM approach on this case would maximise the product

$$L_{EM}(\theta) = P([20, 30]|\theta) \cdot P([30, 40]|\theta).$$

Using normal distributions, it would lead to the optimistic solution of Hüllermeier (a Dirac measure at 30) to ensure $L_{EM} = 1$. Note that the algorithm, assuming the initial distribution is fixed through the choice of θ_0, will compute the expectations $\hat{x}_1$ and $\hat{x}_2$ inside their intervals, and perform a likelihood maximization using these precise expectations as new data, based on the product of densities, getting a new value θ_1, and so on. This process will tend to shrink the expected interval $[\hat{x}_1, \hat{x}_2]$. However if the support of the current symmetric distribution lies inside $[20, 40]$, $L_{EM}(\theta)$ will remain constant (0.25) and jumps to 1 for the Dirac function on 30.

A natural issue is whether in the case of missing data, one may replace them by the whole range of the random variable, say an interval $[a, b]$, or not. This is immaterial for the EM algorithm applied to our setting as $P([a, b]|\theta) = 1$ in any case. So, in our setting, the EM algorithm would just neglect missing observations, whether unsuccessful experiments are carried out or not. On the contrary, in our approach, it makes a difference, as can be seen in the next example.

Consider the case when the range of x is [0, 200], 10 precise observations at 100 were made and the result of one experiment could not be properly observed, so its value lies in [0,200]. The mean value is $\mu = 100$ for the three approaches including the robust one, but the resulting distribution is the Dirac for the optimistic solution of Hüllermeier and the EM approach while the σ value of the robust approach is around 30 (not excluding a maximal deviation from 100 in the failed experiment). If now we have 10 precise observations at 100 and 10 completely imprecise ones modelled by [0,200]. The solution returned by the optimistic approach of Hüllermeier and the EM approach will still be a Dirac measure at 100 while the σ value of the robust approach becomes close to 70. In our approach, the imprecision of observations directly impacts the variance of the identified density. So, unsuccessful observations are not treated as observations not yet carried out. Whether this distinction is meaningful or not in all situations is a matter of debate.

More generally, in the case (perhaps unlikely in practice) where the dataset consists of overlapping intervals, it is clear that any density function with support inside the intersection of the intervals will ensure that the EM likelihood function $L_{EM} = 1$ in (28) since each term has probability 1 in the product (the same remark applies if one maximizes the upper likelihood). However our method will give a density whose standard deviation reflects the width of the uncertainty intervals. In this case, though, using a possibility measure to represent the data may sound more appropriate than a density that turns incompleteness into variability.

5 Conclusion

In this paper we propose to propagate the epistemic interval uncertainty pervading a data set over to the estimation of the likelihood. Then we propose an iterative algorithm which finds parameter values that maximize the lower likelihood values among all data sets compatible with the interval observations, under not too restrictive conditions on the density function. We have studied the case of normal distribution and have shown that the computation of optimal mean and variance can be achieved efficiently. As perspectives,

first we plan to compute robust parameter estimations for other classical distributions. In particular, the algorithm that finds optimal solutions can be improved taking into account the specificities of density functions (as for the normal distribution in this paper). Another perspective is the study of robust linear regression under imprecise observations. Finally, an experimental validation step will be useful to compare our results to those obtained by optimizing upper likelihoods, and methods in the style of the EM algorithm. This approach will be applied to the determination of robust production plans under ill-known demand modelled by fuzzy intervals, in the production engineering environment.

Acknowledgements

The authors are grateful to referees for interesting thought-provoking comments that led us to improve the paper while confirming our intuitions.

References

[1] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reasoning*, 55(7):1502–1518, 2014.

[2] R. Kruse and K. Meyer. *Statistics with Vague Data*. D. Reidel, Dordrecht, 1987.

[3] A. Dempster, N. M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Stat. Soc., series B*, 39:1–38, 1977.

[4] D. Dubois, S. Moral, and H. Prade. A semantics for possibility theory based on likelihoods. *J. of Math. Anal. and Appl.*, 205:359–380, 1997.

[5] H. Uzawa. Iterative methods for concave programming. In K. J. Arrow, L. Hurwicz, and H. Uzawa, editors, *Studies in linear and nonlinear programming*. Stanford University Press, 1958.

[6] K. Shimizu and E. Aiyoshi. Necessary Conditions for Min-Max Problems and Algorithms by a Relaxation Procedure. *IEEE Trans. on Automatic Control*, 25:62–66, 1980.

[7] M. Inuiguchi and M. Sakawa. Minimax regret solution to linear programming problems with an interval objective function. *Eur. J. of Operational Research*, 86:526–536, 1995.

[8] H. E. Mausser and M. Laguna. A heuristic to minimax absolute regret for linear programs with interval objective function coefficients. *Eur. J. of Operational Research*, 117:157–174, 1999.

[9] V. Kreinovich, G. Xiang, and S. Ferson. Computing mean and variance under dempster-shafer uncertainty: Towards faster algorithms. *Int. J. Approx. Reasoning*, 42(3):212–227, 2006.

[10] D. Dubois and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, 1988.

[11] D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

[12] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M.M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, Amsterdam, 1982.

[13] D. Dubois, H. Fargier, and V. Galvagnon. On latest starting times and floats in activity networks with ill-known durations. *European Journal of Operational Research*, 147:266–280, 2003.

[14] D. Dubois and H. Prade. The mean value of a fuzzy number. *Fuzzy Sets and Systems*, 24:279–300, 1987.

[15] I. Couso and L. Sánchez. Upper and lower probabilities induced by a fuzzy random variable. *Fuzzy Sets and Systems*, 165(1):1–23, 2011.

[16] T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. Knowl. Data Eng.*, 25(1):119–130, 2013.

[17] T. Denoeux. Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy Sets and Systems*, 183(1):72–91, 2011.

[18] G. González-Rodríguez, A. Colubi, and M. Angeles Gil. Fuzzy data treated as functional data: A one-way ANOVA test approach. *Computational Statistics & Data Analysis*, 56(4):943–955, 2012.

[19] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reasoning*, 55(7):1519–1534, 2014.

[20] D. Dubois. On various ways of tackling incomplete information in statistics. *Int. J. Approx. Reasoning*, 55(7):1570–1574, 2014.

On Two Composition Operators in Dempster-Shafer Theory

Radim Jiroušek
Faculty of Management, Univ. of Economics
Jindřichuv Hradec, Czech Republic
radim@utia.cas.cz

Abstract

Efficient computations with probabilistic multidimensional models are made possible if the respective probability measure (distribution) is in the form of a decomposable model. Some of the advantageous properties of these models are based on the fact that factorization and conditional independence coincide. It means that a decomposable multidimensional model can be assembled (composed) from its low-dimensional marginals with the help of an *operator of composition*, which introduces conditional independence relations among the variables.

The problem arises when we also want to apply these ideas in Dempster-Shafer theory of evidence, because two different operators of composition have been introduced in literature. The present paper serves as a survey of results on these two operators, recollects their common properties and differences, and tries to find a proper role for each of them.

Keywords. Factorization, conditional independence, combination, composition, decomposable model, IPFP.

1 Introduction

In every textbook dealing with Bayesian network theory there inevitably appears a basic theorem saying that in probability theory, factorization and conditional independence coincide. To express this property more exactly (and simultaneously in its simplest version), consider a probability measure π defined on a finite three-dimensional Cartesian product space $\mathbb{X} \times \mathbb{Y} \times \mathbb{Z}$. Then

$$\pi(\mathbf{a}) \cdot \pi^{\downarrow\{Z\}}(\mathbf{a}^{\downarrow\{Z\}}) = \pi^{\downarrow\{X,Z\}}(\mathbf{a}^{\downarrow\{X,Z\}}) \cdot \pi^{\downarrow\{Y,Z\}}(\mathbf{a}^{\downarrow\{Y,Z\}}) \quad (1)$$

for all $\mathbf{a} \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Z}$, if and only if there exist two functions

$$\phi : \mathbb{X} \times \mathbb{Z} \longrightarrow \mathbb{R}^+,$$
$$\psi : \mathbb{Y} \times \mathbb{Z} \longrightarrow \mathbb{R}^+,$$

such that

$$\pi(\mathbf{a}) = \phi(\mathbf{a}^{\downarrow\{X,Z\}}) \cdot \psi(\mathbf{a}^{\downarrow\{Y,Z\}}) \quad (2)$$

for all $\mathbf{a} \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Z}$.

The equality (1) says that for probability measure π variables X and Y are conditionally independent given variable Z, and equality (2) expresses the fact that measure π factorizes with respect to cover $\{\mathbb{X} \times \mathbb{Z}, \mathbb{Y} \times \mathbb{Z}\}$.

The importance of the factorization stems from the fact that it describes formal conditions under which it is possible to represent a multidimensional Bayesian network with a reasonable number of parameters (conditional probabilities) and to design computationally tractable inference procedures. On the other hand, the concept of conditional independence is comprehensible to users. Thus, verification of the formal conditions for factorization is made possible by the very fact that these two concepts coincide in probability theory. Namely, expressing a probability distribution in a factorized form is equivalent to introducing a conditional independence relation among the variables. And to verify the model, the users should consider whether the introduced conditional independence relations are justifiable (or at least acceptable).

Trying to uncover a similar relationship between factorization and conditional independence in Dempster-Shafer theory of evidence, one easily comes to the conclusion that conditional independence coincides with the factorization of commonality functions, which are, unfortunately, completely illegible to users. This is one of the reasons why we will focus on factorization of basic probability assignments in this paper. We will show that the notions of conditional independence and factorization (of basic probability assignments) correspond to two composition operators studied previously,

in ISIPTA paper [8].

The present paper, which is in fact a synthesis of known results about the two operators of composition in D-S theory, is organized as follows. In the next section the necessary notions and notation are introduced. Section 3 is devoted to the properties of the composition operators and uncovers their pros and cons from the point of view of their computational complexity. In Section 4 we describe (using a simple example) how to make some computations with belief functions tractable. The basic idea is the same as the one used for computations with Bayesian networks. We will show how to represent multidimensional belief functions in the form of decomposable models, for which we will employ both the studied operators of composition. Section 5 explains the role of one of the operators in computation of conditionals.

2 Basic Notions and Notation

In this paper we consider a finite set of finite valued variables $N = \{X_1, X_2, \ldots, X_n\}$; $\mathbb{X}_i$ denotes the set of states of variable X_i. $\mathbb{X}_N = \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_n$ denotes a finite multidimensional space of states of variables N, and its subspaces[1] (for all $K \subseteq N$) are denoted by

$$\mathbb{X}_K = \bigtimes_{i \in K} \mathbb{X}_i.$$

For a state $x = (x_1, x_2, \ldots, x_n) \in \mathbb{X}_N$ its projection into subspace $\mathbb{X}_K$ is denoted by $x^{\downarrow K} = (x_{i, i \in K})$, and for $\mathbf{a} \subseteq \mathbb{X}_N$

$$\mathbf{a}^{\downarrow K} = \{y \in \mathbb{X}_K : \exists x \in \mathbf{a}, x^{\downarrow K} = y\}.$$

Symbol $2^{\mathbf{a}}$ denotes the set of all nonempty subsets of $\mathbf{a}$. By a *join* of two sets $\mathbf{a} \subseteq \mathbb{X}_K$ and $\mathbf{b} \subseteq \mathbb{X}_L$ we understand a set

$$\mathbf{a} \bowtie \mathbf{b} = \{x \in \mathbb{X}_{K \cup L} : x^{\downarrow K} \in \mathbf{a} \ \ \& \ \ x^{\downarrow L} \in \mathbf{b}\}.$$

Realize that if K and L are disjoint, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \times \mathbf{b}$, if $K = L$ then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \cap \mathbf{b}$, and, generally, for $\mathbf{c} \subseteq \mathbb{X}_{K \cup L}$, $\mathbf{c}$ is a subset of $\mathbf{c}^{\downarrow K} \bowtie \mathbf{c}^{\downarrow L}$, which may be proper. This is why we will often use the symbol

$$2^{\mathbb{X}_{K \bowtie L}} = \{\mathbf{c} \subseteq \mathbb{X}_{K \cup L} : \mathbf{c} \neq \emptyset \ \& \ \mathbf{c} = \mathbf{c}^{\downarrow K} \bowtie \mathbf{c}^{\downarrow L}\}.$$

Let us mention that sets $\mathbf{c} \in 2^{\mathbb{X}_{K \bowtie L}}$ are called *Z-layered rectangles* in [2].

In what follows it will be important to realize that cardinality of $2^{\mathbb{X}_{K \bowtie L}}$, though growing exponentially with $|\mathbb{X}_{K \cup L}|$, is much smaller than $|2^{\mathbb{X}_{K \cup L}}|$. For example, for binary variables

$$\begin{aligned} |\mathbb{X}_{\{1,2,3\}}| &= 8, \\ |2^{\mathbb{X}_{\{1,2,3\}}}| &= 255, \\ |2^{\mathbb{X}_{\{1,2\} \bowtie \{2,3\}}}| &= 99, \end{aligned}$$

and for ternary variables

$$\begin{aligned} |\mathbb{X}_{\{1,2,3\}}| &= 27, \\ |2^{\mathbb{X}_{\{1,2,3\}}}| &= 134\,217\,727, \\ |2^{\mathbb{X}_{\{1,2\} \bowtie \{2,3\}}}| &= 124\,999. \end{aligned}$$

2.1 Belief Functions

The role played in probability theory by probability measures (or probability distributions), is played by belief functions in Dempster-Shafer theory. It is well known [14] that these functions can be equivalently represented in several ways. In this paper we will use just basic probability assignments, and commonality functions.

Basic probability assignment (bpa) on $\mathbb{X}_K$ is a function

$$\mu : 2^{\mathbb{X}_K} \longrightarrow \mathbf{R}.$$

Though most authors require this function to be non-negative and normalized, in this paper we accept the more general approach of Shenoy [15], who does not restrict the class of considered functions and says that bpa is *proper* if this function is non-negative, and further says that it is *normal* if

$$\sum_{\mathbf{a} \subseteq \mathbb{X}_K} \mu(\mathbf{a}) = 1.$$

Nevertheless, not even in this paper will we consider all possible functions. When speaking about a bpa we will always assume that its corresponding *commonality function* (comf), which is a function on $\mathbb{X}_K$ defined for each nonempty $\mathbf{a} \subseteq \mathbb{X}_K$

$$\theta(\mathbf{a}) = \sum_{\mathbf{b} \supseteq \mathbf{a}} \mu(\mathbf{b}), \tag{3}$$

is strictly positive. Recall that this transformation of bpas into comfs is unique, and that a bpa can be reconstructed from its comf using the following formula (Möbius transform – see [14])

$$\mu(\mathbf{a}) = \sum_{\mathbf{b} \supseteq \mathbf{a}} (-1)^{|\mathbf{b} \setminus \mathbf{a}|} \theta(\mathbf{b}), \tag{4}$$

for all nonempty $\mathbf{a} \subseteq \mathbb{X}_K$. This enables us to call comf θ *proper* (*normal*) if the corresponding bpa is proper (normal).

$\mathbf{a} \in 2^{\mathbb{X}_K}$ is said to be a *focal element* of bpa μ if $\mu(\mathbf{a}) \neq 0$. Quite often, the list of focal elements and

[1] In our examples we will use a simplified notation. Instead of a correct notation for a subset of variables, say, $K = \{X_1, X_3, X_7\}$ we will use just $K = \{1, 3, 7\}$.

the respective values of bpa are used for belief function representation. However, as we will see later, it does not mean that a belief function may be represented by the list of values of the respective comf on the focal elements, because comfs are quite often positive also for non-focal elements. The exceptions are so-called *probabilistic* (some authors call them *Bayesian*) belief functions, for which all focal elements are *singletons* (the cardinality of each focal element is one), and, as it can be immediately seen from Formula (3), $\mu = \theta$.

The last notion introduced in this section is that of dominance. Consider two bpas μ_1 and μ_2 defined on the same $\mathbb{X}_K$. We say that μ_1 dominates μ_2 (and write $\mu_1 \gg \mu_2$) if all focal elements of μ_2 are also focal elements of μ_1, i.e., if

$$\mu_1(\mathbf{a}) = 0 \implies \mu_2(\mathbf{a}) = 0$$

for all $\mathbf{a} \in 2^{\mathbb{X}_K}$.

2.2 Combination

An important notion in D-S theory is the notion of combination $\oplus$, usually called *Dempster's rule of combination.*

Definition 1 *Consider two arbitrary bpas μ_1 on $\mathbb{X}_K$ and μ_2 on $\mathbb{X}_L$ ($K \neq \emptyset \neq L$). A* combination $\mu_1 \oplus \mu_2$ *is defined as follows:*
if $\mathbf{c} \in 2^{\mathbb{X}_{K \bowtie L}}$ then

$$(\mu_1 \oplus \mu_2)(\mathbf{c}) = \alpha^{-1} \sum_{\mathbf{a} \subseteq \mathbb{X}_K, \mathbf{b} \subseteq \mathbb{X}_L : \mathbf{a} \bowtie \mathbf{b} = \mathbf{c}} \mu_1(\mathbf{a}) \cdot \mu_2(\mathbf{b})$$

where α is a normalization constant

$$\alpha = \sum_{\mathbf{d} \in 2^{\mathbb{X}_{K \bowtie L}}} \sum_{\mathbf{a} \subseteq \mathbb{X}_K, \mathbf{b} \subseteq \mathbb{X}_L : \mathbf{a} \bowtie \mathbf{b} = \mathbf{d}} \mu_1(\mathbf{a}) \cdot \mu_2(\mathbf{b}),$$

and $(\mu_1 \oplus \mu_2)(\mathbf{c}) = 0$ for all $\mathbf{c} \in 2^{\mathbb{X}_{K \cup L}} \setminus 2^{\mathbb{X}_{K \bowtie L}}$.

It is well known (e.g., [14]) that Dempster's rule of combination can also be (even more elegantly) defined with the help of the respective comfs. Namely, if θ_1, θ_2 correspond to μ_1, μ_2, respectively, then the comf corresponding to $\mu_1 \oplus \mu_2$ is

$$(\theta_1 \oplus \theta_2)(\mathbf{c}) = \frac{\theta_1(\mathbf{c}^{\downarrow K}) \cdot \theta_2(\mathbf{c}^{\downarrow L})}{\sum_{\mathbf{d} \in 2^{\mathbb{X}_{K \cup L}}} (-1)^{|\mathbf{d}|+1} \ \theta_1(\mathbf{d}^{\downarrow K}) \cdot \theta_2(\mathbf{d}^{\downarrow L})}.$$

The introduced operator of combination models a belief update. So, it is not surprising that this operator is not idempotent, which means that, generally, $\mu \oplus \mu \neq \mu$. Indeed, when hearing the same piece of information from two independent sources we get more convinced about its currency.

The reader can easily verify the non-idempotence of Dempster's rule of combination on a simple bpa μ with two focal elements: $\mu(\mathbf{a}) = \frac{1}{3}, \mu(\mathbf{b}) = \frac{2}{3}$. Namely, the resulting bpa $\mu \oplus \mu$ has, naturally, the same two focal elements

$$(\mu \oplus \mu)(\mathbf{a}) = \frac{1}{5}, \qquad (\mu \oplus \mu)(\mathbf{b}) = \frac{4}{5}.$$

Contrary to the belief update, when assembling a global knowledge from its local pieces we need a tool that *is* idempotent. If the local pieces of knowledge are consistent, each of them should be preserved unchanged in its global representation. Using a mathematical terminology, we can also say that in this case we are looking for a *join extension* of the local pieces. And this is the goal for which the operator of composition introduced in the following section is designed.

2.3 Operators of Composition

In this paragraph we introduce two composition operators. Definition 2 is based on Dempster's rule of combination and its equivalence to (normalized) multiplication of the respective comfs. As we showed above, this operation is not idempotent; we thus have to avoid double counting of contributions on the overlapping subspace. In other words, we have to ensure that each piece of local information is considered only once (for details see [15] or [9]). Such a removal of information is performed by an operation inverse to Dempster's rule of combination (division of the respective comfs). Notice that Definition 2 is described with the help of comfs while Definition 3 makes use of bpas. Nonetheless, because of the one-to-one correspondence between bpas and comfs both these operators are naturally extended to both bpas and comfs. Note that Definition 2 is from [9], and Definition 3 first appeared in [10].

Definition 2 *Consider two arbitrary bpas, μ_1 on $\mathbb{X}_K$ and μ_2 on $\mathbb{X}_L$ ($K \neq \emptyset \neq L$) and assume that $\mu_2^{\downarrow K \cap L} \gg \mu_1^{\downarrow K \cap L}$. Let θ_1 and θ_2 be the respective comfs. A* composition $\theta_1 \overset{d}{\triangleright} \theta_2$ *is defined for each nonempty $\mathbf{c} \subseteq \mathbb{X}_{K \cup L}$ by the following formula:*

$$(\theta_1 \overset{d}{\triangleright} \theta_2)(\mathbf{c}) = \begin{cases} \alpha^{-1} \frac{\theta_1(\mathbf{c}^{\downarrow K}) \cdot \theta_2(\mathbf{c}^{\downarrow L})}{\theta_2^{\downarrow K \cap L}(\mathbf{c}^{\downarrow K \cap L})} & \text{if } \theta_2^{\downarrow K \cap L}(\mathbf{c}^{\downarrow K \cap L}) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where α is a normalization constant defined as

$$\alpha = \sum_{\mathbf{d} \in 2^{\mathbb{X}_{K \cup L}} : \theta_2^{\downarrow K \cap L}(\mathbf{d}^{\downarrow K \cap L}) > 0} (-1)^{|\mathbf{d}|+1} \ \frac{\theta_1(\mathbf{d}^{\downarrow K}) \cdot \theta_2(\mathbf{d}^{\downarrow L})}{\theta_2^{\downarrow K \cap L}(\mathbf{d}^{\downarrow K \cap L})}.$$

Definition 3 *Consider two normal bpas, μ_1 on $\mathbb{X}_K$ and μ_2 on $\mathbb{X}_L$ ($K \neq \emptyset \neq L$). A* composition $\mu_1 \,^{f}\!\triangleright\, \mu_2$ *is defined for each nonempty $\mathbf{c} \subseteq \mathbb{X}_{K\cup L}$ by one of the following expressions:*

(i) *if $\mu_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L}) > 0$ and $\mathbf{c} \in 2^{\mathbb{X}_{K\bowtie L}}$ then*

$$(\mu_1 \,^{f}\!\triangleright\, \mu_2)(\mathbf{c}) = \frac{\mu_1(\mathbf{c}^{\downarrow K}) \cdot \mu_2(\mathbf{c}^{\downarrow L})}{\mu_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L})};$$

(ii) *if $\mu_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L}) = 0$ and $\mathbf{c} = \mathbf{c}^{\downarrow K} \times \mathbb{X}_{L\setminus K}$ then*

$$(\mu_1 \,^{f}\!\triangleright\, \mu_2)(\mathbf{c}) = m_1(\mathbf{c}^{\downarrow K});$$

(iii) *in all other cases, $(\mu_1 \,^{f}\!\triangleright\, \mu_2)(\mathbf{c}) = 0$.*

Having two operators of composition, quite a natural question arises: why do we need both of them? The operators differ in many aspects. As we will discuss in the next section, each of them has different computational complexity and different semantics. Even so, the answer to the previous question is not straightforward; in fact, information found throughout this entire paper should help readers form their own opinions.

To simplify the following exposition, let us make two conventions. First, whenever we use symbol $\triangleright$ then the respective assertion holds for both the defined operators $^{d}\!\triangleright$ and $^{f}\!\triangleright$. Second, to avoid frequent repetition of the condition on dominance of arguments required in the definition of $^{d}\!\triangleright$, and the requirement of normality needed in the definition of $^{f}\!\triangleright$, whenever the operators are used in the following text, we will assume that the necessary conditions under which the respective operator is defined are fulfilled.

3 Properties of Operators of Composition

Both the operators of composition introduced in the preceding section comply with the properties expected from composition. These properties were originally proven for probability theory in [6], and later also for Shenoy's Valuation-Based systems (VBS) in [9], from which it follows that all of them hold for D-S theory.

Theorem 1 *Suppose μ_1, μ_2 and μ_3 are bpas on $\mathbb{X}_K$, $\mathbb{X}_L$, and $\mathbb{X}_M$, respectively. Then the following statements hold:*

1. (Domain): *$\mu_1 \triangleright \mu_2$ is a bpa on $\mathbb{X}_{K\cup L}$.*
2. (Composition preserves first marginal): *$(\mu_1 \triangleright \mu_2)^{\downarrow K} = \mu_1$.*
3. (Reduction:) *If $L \subseteq K$ then, $\mu_1 \triangleright \mu_2 = \mu_1$.*
4. (Non-commutativity): *In general, $\mu_1 \triangleright \mu_2 \neq \mu_2 \triangleright \mu_1$.*
5. (Commutativity under consistency): *If μ_1 and μ_2 have a common marginal on $\mathbb{X}_{K\cap L}$, i.e., $\mu_1^{\downarrow K\cap L} = \mu_2^{\downarrow K\cap L}$, then $\mu_1 \triangleright \mu_2 = \mu_2 \triangleright \mu_1$.*
6. (Non-associativity): *In general, $(\mu_1 \triangleright \mu_2) \triangleright \mu_3 \neq \mu_1 \triangleright (\mu_2 \triangleright \mu_3)$.*
7. (Associativity under special condition I): *If $K \supset (L \cap M)$ then, $(\mu_1 \triangleright \mu_2) \triangleright \mu_3 = \mu_1 \triangleright (\mu_2 \triangleright \mu_3)$.*
8. (Associativity under special condition II): *If $L \supset (K \cap M)$ then, $(\mu_1 \triangleright \mu_2) \triangleright \mu_3 = \mu_1 \triangleright (\mu_2 \triangleright \mu_3)$.*
9. (Stepwise combination): *If $(K \cap L) \subseteq M \subseteq L$ then, $(\mu_1 \oplus \mu_2^{\downarrow M}) \triangleright \mu_2 = \mu_1 \oplus \mu_2$.*
10. (Stepwise composition): *If $(K \cap L) \subseteq M \subseteq L$ then, $(\mu_1 \triangleright \mu_2^{\downarrow M}) \triangleright \mu_2 = \mu_1 \triangleright \mu_2$.*
11. (Exchangeability): *If $K \supset (L \cap M)$ then, $(\mu_1 \triangleright \mu_2) \triangleright \mu_3 = (\mu_1 \triangleright \mu_3) \triangleright \mu_2$.*
12. (Simple marginalization): *If $(K\cap L) \subseteq M \subseteq K\cup L$ then, $(\mu_1 \triangleright \mu_2)^{\downarrow M} = \mu_1^{\downarrow K\cap M} \triangleright \mu_2^{\downarrow L\cap M}$.*
13. (Irrelevant combination): *If $M \subseteq K \setminus L$ then, $\mu_1 \triangleright (\mu_2 \oplus \mu_3) = \mu_1 \triangleright \mu_2$.*

From the formal point of view, the main difference between $^{d}\!\triangleright$ and $^{f}\!\triangleright$ is in their computational complexity. We need not subject them to a detailed and precise complexity analysis, because the difference is visible at the first sight.

3.1 Computational Complexity

Let us start with a simpler task. Consider the formulae in Definition 3, which give direct instructions for how to compute the composition $^{f}\!\triangleright$. To compute $\mu_1 \,^{f}\!\triangleright\, \mu_2$, it is necessary to find out all of its focal elements $\mathbf{c} \in 2^{\mathbb{X}_{K\cup L}}$ and compute the respective value $(\mu_1 \,^{f}\!\triangleright\, \mu_2)(\mathbf{c})$. Therefore, the computational complexity of this process is linear in the number of those $\mathbf{c} \in 2^{\mathbb{X}_{K\cup L}}$ that must be checked to determine whether they are focal elements. There are two direct ways in which to do it. One is based on the fact that all focal elements of $\mu_1 \,^{f}\!\triangleright\, \mu_2$ are from $2^{\mathbb{X}_{K\bowtie L}}$. So, we can generate all the elements of $2^{\mathbb{X}_{K\bowtie L}}$, which is, as we said (and illustrated) in Section 2, substantially smaller than $2^{\mathbb{X}_{K\cup L}}$.

The other, quite often more efficient, possibility is based on the fact that for all focal elements $\mathbf{c} \in 2^{\mathbb{X}_{K\cup L}}$ of $\mu_1 \,^{f}\!\triangleright\, \mu_2$ the following two conditions must hold simultaneously

(1) $\mu_1(\mathbf{c}^{\downarrow K}) \neq 0$;

(2) either $\mu_2(\mathbf{c}^{\downarrow L}) \neq 0$, or $\mathbf{c}^{\downarrow L} = \mathbf{c}^{\downarrow K\cap L} \times \mathbb{X}_{L\setminus K}$.

This means that the number of potential focal elements of $\mu_1 \overset{f}{\triangleright} \mu_2$ cannot exceed the product of two numbers: the number of focal elements of μ_1 and the number of focal elements of μ_2. In the case when the considered belief functions are represented by the lists of their focal elements, these numbers are usually limited.

For operator $\overset{d}{\triangleright}$ the situation is different. Regardless of the fact that the number of focal elements is also limited, namely, by $|2^{\mathbb{X}_{K\bowtie L}}|$, one can hardly expect that it would be possible to find an algorithm whose computational complexity would be linear in the number of (potential) focal elements. This pessimistic statement holds for Dempster's rule of combination (each value of $\mu_1 \oplus \mu_2$ is computed as a summation), the more it holds for $\mu_1 \overset{d}{\triangleright} \mu_2$. To the best of our knowledge, up to now, no other way to compute $\mu_1 \overset{d}{\triangleright} \mu_2$ has been known than to convert the composed bpas into comfs, and afterward carry out the computations described in Definition 2. However, even for bpa μ_1 with a limited number of focal elements, the number of those $\mathbf{c} \in 2^{\mathbb{X}_K}$ for which $\theta_1(\mathbf{c}) > 0$ holds may be very high. Let us illustrate the situation on an extreme (but appearing in practice) situation.

3.2 Example

Consider bpas μ_1 and μ_2 on $\mathbb{X}_K$, $\mathbb{X}_L$, respectively, and assume they define lower probabilities on the respective subspaces, i.e., their focal elements are only singletons plus the whole $\mathbb{X}_K$ for μ_1, and singletons plus the whole $\mathbb{X}_L$ for μ_2. It means that μ_1 has no more than $|\mathbb{X}_K| + 1$ focal elements, and μ_2 has no more than $|\mathbb{X}_L| + 1$ focal elements.

However, one can immediately see from Formula (3) that any of the corresponding comfs θ_1, θ_2 may easily be positive for all the elements of $2^{\mathbb{X}_K}$, $2^{\mathbb{X}_L}$, respectively. It means that when computing $\theta_1 \overset{d}{\triangleright} \theta_2$ we have to compute the respective values for all $\mathbf{c} \in 2^{\mathbb{X}_{K\cup L}}$.

Computation of $\mu_1 \overset{f}{\triangleright} \mu_2$ is different in this example. If $\mu_2^{\downarrow K\cap L} \gg \mu_1^{\downarrow K\cap L}$ then the resulting bpa $\mu_1 \overset{f}{\triangleright} \mu_2$ has the same property as μ_1 and μ_2: its focal elements are only singletons plus the whole $\mathbb{X}_{K\cup L}$ (this is true because for the considered bpas case **(ii)** of Definition 3 can never assign a value different from 0). If $\mu_2^{\downarrow K\cap L} \not\gg \mu_1^{\downarrow K\cap L}$, i.e., for some $\mathbf{b} \in \mathbb{X}_{K\cap L}$, $\mu_2^{\downarrow K\cap L}(\mathbf{b}) = 0$ and $\mu_1^{\downarrow K\cap L}(\mathbf{b}) \neq 0$, then the number of focal elements $\mathbf{c} \in \mathbb{X}_{K\cup L}$, for which $\mathbf{c}^{\downarrow K\cap L} = \mathbf{b}$ does not exceed

$$|\{\mathbf{a} \in \mathbb{X}_K : \mathbf{a}^{\downarrow K\cap L} = \mathbf{b}\}|,$$

and therefore we see that the number of focal elements of $\mu_1 \overset{f}{\triangleright} \mu_2$ cannot exceed $|\mathbb{X}_{K\cup L}| + 1$.

3.3 Factorization

As a direct consequence of Properties 2 and 5 in Theorem 1, one can see that if two bpas μ_1 and μ_2 (assume again they are defined on $\mathbb{X}_K$ and $\mathbb{X}_L$, respectively) have a common marginal (i.e., $\mu_1^{\downarrow K\cap L} = \mu_2^{\downarrow K\cap L}$), then both $\mu_1 \overset{d}{\triangleright} \mu_2$ and $\mu_1 \overset{f}{\triangleright} \mu_2$ are common extensions of μ_1 and μ_2. Generally, these two extensions differ from each other: each of them has its own semantics.

From the point of view of this paper it is important to say that $\overset{d}{\triangleright}$ reflects the notion of conditional independence in the sense used by Shafer [14] and Shenoy [15]. More precisely, $\mu^{\downarrow K\cup L} = \mu^{\downarrow K} \overset{d}{\triangleright} \mu_2^{\downarrow L}$ holds if and only if variables $K \setminus L$ and $L \setminus K$ are for the considered belief function (bpa μ) *conditionally independent given variables* $K \cap L$.

The semantics of the fact that $\mu^{\downarrow K\cup L} = \mu^{\downarrow K} \overset{f}{\triangleright} \mu_2^{\downarrow L}$ is different. Namely, it is a direct consequence of Lemma 3 in [16] that $\mu^{\downarrow K\cup L} = \mu^{\downarrow K} \overset{f}{\triangleright} \mu^{\downarrow L}$ if and only if $\mu^{\downarrow K\cup L}$ factorizes with respect to a couple $\{K, L\}$ in the sense of the following definition.

Definition 4 *Consider bpa μ on $\mathbb{X}_M$, and two subsets of variables $K, L \subset M$. We say that μ* factorizes *with respect to $\{K, L\}$ if*

(a) $\mu^{\downarrow K\cup L}(\mathbf{c}) = 0$ *for all* $\mathbf{c} \in (2^{\mathbb{X}_{K\cup L}} \setminus 2^{\mathbb{X}_{K\bowtie L}})$, *and*

(b) there exist two functions

$$\phi : \mathbb{X}_K \longrightarrow \mathbb{R},$$
$$\psi : \mathbb{X}_L \longrightarrow \mathbb{R},$$

such that

$$\mu^{\downarrow K\cup L}(\boldsymbol{c}) = \phi(\boldsymbol{c}^{\downarrow X_K}) \cdot \psi(\boldsymbol{c}^{\downarrow X_L})$$

for all $\boldsymbol{c} \in 2^{\mathbb{X}_{K\bowtie L}}$.

As we already said above, $\mu_1 \overset{d}{\triangleright} \mu_2$ and $\mu_1 \overset{f}{\triangleright} \mu_2$ generally differ from each other. Nevertheless, there are special situations in which they coincide. For example, it is easy to show that these compositions coincide when $K \cap L = \emptyset$, or when the composed bpas are probabilistic. Nevertheless, note that specification of necessary and sufficient conditions under which the two operators coincide has remained an open problem for several years.

4 Decomposable Models

By decomposable probability distributions we understand the distributions whose conditional dependence structures can be well depicted with the help of so-called *decomposable graphs* [11]. The latter is an important class of graphs that were introduced in graph theory under several different names (triangulated graphs,

chordal graphs – see [13]). One of the characteristic properties of these graphs is that their cliques can be ordered to meet the so-called running intersection property (see below).

In general, by decomposable models we understand multidimensional measures/distributions/valuations that can be decomposed, and thereafter reconstructed from a system of its marginals without loss of information, and the structure of the system of marginals can be depicted with the help of a decomposable graph. The latter condition is equivalent to the requirement that the marginals can be ordered to meet the running intersection property. The purpose of such decomposition is twofold. One reason is to decrease the number of necessary parameters representing the multidimensional model. The other reason is to decrease the computational complexity of the procedures that process this model (e.g., when used for inference). As a rule, these two goals are mutually connected. Usually, the fewer parameters necessary to define a model, the more efficient the computational procedures will be.

Let us also apply this general idea to bpas in D-S theory. Consider a multidimensional bpa μ on $\mathbb{X}_N$. Let $\{K_1, K_2, \ldots, K_m\}$ be a cover of N (i.e., $\cup_{i=1}^m K_i = N$) meeting the *running intersection property* (RIP):

$$\forall i = 2, \ldots, m \;\; \exists j < i: \;\; (K_1 \cup \ldots \cup K_{i-1}) \cap K_i \subseteq K_j.$$

We say that μ is a *decomposable model with structure* $\{K_1, K_2, \ldots, K_m\}$ if

$$\mu = \mu^{\downarrow K_1} \triangleright \mu^{\downarrow K_2} \triangleright \ldots \triangleright \mu^{\downarrow K_m}. \tag{5}$$

Since the operator of composition is not associative, we must explain how to interpret the right hand side of Formula (5): whenever the order of operators is not specified by parentheses, they are performed from left to right, i.e.,

$$\begin{aligned} &\mu^{\downarrow K_1} \triangleright \mu^{\downarrow K_2} \triangleright \ldots \triangleright \mu^{\downarrow K_m} \\ &\quad = \quad \big(\ldots(\mu^{\downarrow K_1} \triangleright \mu^{\downarrow K_2}) \triangleright \ldots \triangleright \mu^{\downarrow K_{m-1}}\big) \triangleright \mu^{\downarrow K_m}. \end{aligned}$$

As we used the general symbol $\triangleright$, the reader certainly understands that, in principal, either of the operators ${}^{d}\!\triangleright$ and ${}^{f}\!\triangleright$ can be used. This is why we will use the notions of d-decomposability and f-decomposability when we need to stress which of the two operators of composition is being applied.

The following property of decomposable models in VBS framework was proven in [9], and therefore it also holds for the considered decomposable models in D-S theory: If $K_{j_1}, K_{j_2}, \ldots, K_{j_m}$ is a permutation of $K_1, K_2, \ldots, K_m$ such that it also meets RIP, then

$$\begin{aligned} &\mu^{\downarrow K_1} \triangleright \mu^{\downarrow K_2} \triangleright \ldots \triangleright \mu^{\downarrow K_m} \\ &\quad = \mu^{\downarrow K_{j_1}} \triangleright \mu^{\downarrow K_{j_2}} \triangleright \ldots \triangleright \mu^{\downarrow K_{j_m}}. \end{aligned}$$

Let us conclude this section by highlighting that in this paper we restrict our attention only to sequential models, i.e., the distributions that can be expressed in the form of Formula (5). For the properties of more general compositional models see [12].

4.1 Marginal Problem Example

Perhaps the best way to illustrate both advantages and problems connected with computations on decomposable models is to consider a real (maximally simplified) task. Let $\mu_1, \ldots, \mu_4$ be four bpas on $\mathbb{X}_{\{1,2\}}$, $\mathbb{X}_{\{2,3\}}$, $\mathbb{X}_{\{3,4\}}$, $\mathbb{X}_{\{1,4\}}$, respectively. The goal is to find a bpa μ^* on $\mathbb{X}_{\{1,2,3,4\}}$ such that all μ_i are its marginals.

To the best of our knowledge, there is no better way to solve this problem than to apply *Iterative Proportional Fitting Procedure* (IPFP) [4, 3], which proceeds as follows:

I Define bpa $\lambda(\mathbf{c}) := \frac{1}{|2^{\mathbb{X}_{\{1,2,3,4\}}}|}$ for all nonempty $\mathbf{c} \subseteq \mathbb{X}_{\{1,2,3,4\}}$

II Repeat the following cycle (four steps) until the procedure converges:

(i) $\lambda := \mu_1 \triangleright \lambda$,

(ii) $\lambda := \mu_2 \triangleright \lambda$,

(iii) $\lambda := \mu_3 \triangleright \lambda$,

(iv) $\lambda := \mu_4 \triangleright \lambda$.

We dealt with this procedure previously in ISIPTA contribution [8] where we showed that

(a) for both operators ${}^{d}\!\triangleright$ and ${}^{f}\!\triangleright$ it holds that if the procedure converges then all bpas $\mu_1, \ldots, \mu_4$ are marginals of the resulting bpa (more precisely of the limit bpa, to which the procedure converges);

(b) if there exists a bpa having all four bpas $\mu_1, \ldots, \mu_4$ for its marginals then the procedure with ${}^{f}\!\triangleright$ converges;

(c) it may happen that the procedure with ${}^{d}\!\triangleright$ does not converge even if there exists a bpa having all four bpas $\mu_1, \ldots, \mu_4$ for its marginals.

These results give a hint that using IPFP with ${}^{f}\!\triangleright$ should be preferred to ${}^{d}\!\triangleright$; in fact, ${}^{f}\!\triangleright$ is computationally less demanding and the convergence of the procedure is guaranteed by the existence of a single join extension of the given marginal.

Nevertheless, let us note that, regardless whether ${}^{d}\!\triangleright$ or ${}^{f}\!\triangleright$ is used, the procedure cannot be applied to multidimensional bpas because at each step we have to

compute all $|2^{\mathbb{X}_N}|$ values when computing bpa λ. Even in the considered simple four-dimensional case it means that we have to compute $|2^{\mathbb{X}_{\{1,2,3,4\}}}|$ values (which equals 65 535 for binary, and 43 046 720 for ternary variables), by which bpa λ (or the respective comf, in case ${}^{d}\triangleright$ is used) is defined. In analogy to computation within the probabilistic framework [5], a principal simplification can be achieved when representing the computed bpa λ in the form of a decomposable model.

When computing on decomposable models, the general approach starts with finding a RIP cover of N that is a coarsening of $(K_1, K_2, \ldots, K_m)$. In the considered example it means that we look for a RIP cover of $\{1,2,3,4\}$ that is a coarsening of $\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$. This property is met by $\{\{1,2,3\},\{1,3,4\}\}$ (the other possibility would be $\{\{1,2,4\},\{2,3,4\}\}$). Thus, we will consider a decomposable model

$$\lambda = \lambda_1 \triangleright \lambda_2,$$

where $\lambda_1 = \lambda^{\downarrow\{1,2,3\}}$ and $\lambda_2 = \lambda^{\downarrow\{1,3,4\}}$. This type of representation of the four-dimensional pba λ claims a formal change of step II of the above described IPFP algorithm; each step of the cycle is split into two simpler steps – see the modified algorithm below.

However, if we decide to decrease computational complexity of the presented algorithm by the decomposition of the computed bpa, we must be ready to face new problems that do not appear in the probabilistic framework. Namely, in D-S theory the "uniform" bpa (i.e., the initial bpa that is assigned in step I of the algorithm) is *not* decomposable. The reader can see it immediately from the fact that in the considered example, decomposable bpas have only focal elements from $2^{\mathbb{X}_{\{1,2,3\}\bowtie\{2,3,4\}}}$, which contains only 9 999 sets out of 65 535 from $2^{\mathbb{X}_{\{1,2,3,4\}}}$. Therefore, in the considered four-dimensional example, when applying IPFP to the considered decomposable models we also have to modify the initializing step I. So, for the considered example we are getting the following modified algorithm.

I Define bpas:
$$\lambda_1(\mathbf{c}) := \frac{1}{|2^{|\mathbb{X}_{\{1,2,3\}}|}} \text{ for all } \mathbf{c} \in 2^{\mathbb{X}_{\{1,2,3\}}},$$
$$\lambda_2(\mathbf{c}) := \frac{1}{|2^{|\mathbb{X}_{\{1,3,4\}}|}} \text{ for all } \mathbf{c} \in 2^{\mathbb{X}_{\{1,3,4\}}}.$$

II Repeat the following cycle (eight steps) until the procedure converges:

(i) $\lambda_1 := \mu_1 \triangleright \lambda_1$,
(ii) $\lambda_2 := \lambda_1^{\{1,3\}} \triangleright \lambda_2$,
(iii) $\lambda_1 := \mu_2 \triangleright \lambda_1$,
(iv) $\lambda_2 := \lambda_1^{\{1,3\}} \triangleright \lambda_2$,
(v) $\lambda_2 := \mu_3 \triangleright \lambda_2$,
(vi) $\lambda_1 := \lambda_2^{\{1,3\}} \triangleright \lambda_1$,
(vii) $\lambda_2 := \mu_4 \triangleright \lambda_2$,
(viii) $\lambda_1 := \lambda_2^{\{1,3\}} \triangleright \lambda_1$.

(Let us note that it does not matter that λ_1, λ_2 defined in the initializing step of the algorithm may be inconsistent. The only condition we have to guarantee is that $\lambda_1 \triangleright \lambda_2$ is positive for all $\mathbf{c} \in 2^{\mathbb{X}_{\{1,2,3\}\bowtie\{2,3,4\}}}$.)

The achieved simplification is obvious. Although we have to perform twice as many steps in one cycle when considering the simple decomposable model as we do in the general case, we only compute 255 numbers at each step for binary variables instead of 65 535. Naturally, the computational savings for ternary variables would be even more progressive. So, at this point, the question remains (it will be discussed in the next section) whether it is preferable to consider d-decomposable or f-decomposable models.

Before leaving the example let us set right one theoretical issue about this modified algorithm. Recalling the results from the last ISIPTA paper, we said (see point (b) above) that the algorithm with ${}^{f}\triangleright$ converges if there exists a bpa having all $\mu_1, \ldots, \mu_4$ for its marginals. This holds because the general IPFP algorithm is initialized with bpa λ, which is positive on $2^{\mathbb{X}_{\{1,2,3,4\}}}$, and therefore it dominates all bpas on $\mathbb{X}_{\{1,2,3,4\}}$. It is just a question of going through the proof of the assertion guaranteeing convergence of the IPFP procedure in [8], to show that, for the decomposable version of the IPFP algorithm, only a weaker assertion holds:

(b) if there exists a decomposable bpa (decomposable with structure $\{\{1,2,3\},\{1,3,4\}\}$) having all four bpas $\mu_1, \ldots, \mu_4$ for its marginals, then the procedure with ${}^{f}\triangleright$ converges;

5 Inference

The simplest inference scenario is based on computation of conditionals. Instructions for computing conditionals (and a clarification of what their properties are) can be found in [9]. In that paper it was shown that, for bpa μ on $\mathbb{X}_M$ and $X_j, X_k \in M$,

$$\mu(X_k|X_j = a) = \left(\nu_{X_j=a} \, {}^{d}\triangleright \, \mu\right)^{\downarrow X_k},$$

where $\nu_{X_j=a}$ is a one-dimensional bpa on $\mathbb{X}_j$ having just one focal element $\{a\} \subset \mathbb{X}_j$, for which $\nu_{X_j=a}(\{a\}) = 1$. (Note that $\nu_{X_j=a}$ is thus normal and proper.) This bpa expresses the fact that we are sure that variable X_j achieves value a.

Maybe it is worth showing that for computation of conditional bpas we *must* use the operator $\overset{d}{\triangleright}$ and not $\overset{f}{\triangleright}$.

5.1 Example

Consider two variables X_1, X_2 with $\mathbb{X}_1 = \mathbb{X}_2 = \{b, c, d, e\}$, and bpa μ with just one focal element

$$\mu(\{(b,b),(c,c),(d,d),(e,e)\}) = 1.$$

This bpa describes the situation when we do not know which of the values occurs, but we are sure that both variables X_1 and X_2 certainly have the same value.

To compute $\nu_{X_1=b} \overset{d}{\triangleright} \mu$, we proceed according to Definition 2. First notice that $\nu_{X_1=b}$ is a probabilistic bpa, and therefore (see Formula (3)) it is the same as the corresponding comf $\theta_{X_1=b} = \nu_{X_1=b}$. Denote by θ (with no index) the comf corresponding to μ. The marginal $\mu^{\downarrow X_1}$ is a vacuous bpa with just one focal element $\mathbb{X}_1 = \{b, c, d, e\}$. Therefore, the corresponding comf $\theta^{\downarrow X_1}$ equals 1 for all nonempty subsets of $\mathbb{X}_1$, and therefore the denominator in the formula appearing in Definition 2 equals 1. So, in this specific case we get

$$\theta_{X_1=b} \triangleright \theta = \theta_{X_1=b} \cdot \theta.$$

Due to Formula (3), θ equals 1 for all nonempty subsets of $\{(b,b),(c,c),(d,d),(e,e)\}$. Therefore $\theta_{X_1=b} \overset{d}{\triangleright} \theta$ equals one for all those subsets $\mathbf{a} \subseteq \{(b,b),(c,c),(d,d),(e,e)\}$ in which no other value than b appears at the first position, i.e., $\mathbf{a}^{\downarrow\{X_1\}} = \{b\}$, and it is only $\{(b,b)\}$. Therefore we get

$$(\theta_{X_1=b} \overset{d}{\triangleright} \theta)(\{(b,b)\}) = 1,$$

and $(\theta_{X_1=b} \overset{d}{\triangleright} \theta)(\mathbf{a}) = 0$ for all $\mathbf{a} \subseteq \mathbb{X}_1 \times \mathbb{X}_2$, for which $\mathbf{a} \neq \{(b,b)\}$.

This means that we get a probabilistic bpa equaling 1 for $\{(b,b)\}$, from which we get (after marginalization) that $\mu(X_2|X_1 = b)$ equals 1 if and only if $X_2 = b$.

However, if we computed $\mu(X_2|X_1 = b) = \nu_{X_1=b} \overset{f}{\triangleright} \mu$ according to Definition 3 we would get that $\nu_{X_1=b} \overset{f}{\triangleright} \mu$ has, again, only one focal element, but this time it would be $\{(b,b),(b,c),(b,d),(b,e)\}$. Therefore, marginalizing this bpa for variable X_2 we would get a vacuous bpa for which

$$(\nu_{X_1=b} \overset{f}{\triangleright} \mu)^{\downarrow X_2}(\{b,c,d,e\}) = 1.$$

This equality does not correspond to what we expected.

5.2 Conditioning in Decomposable Models

Consider decomposable bpa μ with the structure $\{K_1, K_2, \ldots, K_m\}$ ($\bigcup_{i=1}^m K_i = N$), and assume, first, that it is a d-decomposable model, i.e.,

$$\mu = \mu^{\downarrow K_1} \overset{d}{\triangleright} \mu^{\downarrow K_2} \overset{d}{\triangleright} \ldots \overset{d}{\triangleright} \mu^{\downarrow K_m}.$$

If we want to compute a conditional

$$\begin{aligned}&\mu(X_k|X_j = a)\\ &\quad = \left(\nu_{X_j=a} \overset{d}{\triangleright} (\mu^{\downarrow K_1} \overset{d}{\triangleright} \mu^{\downarrow K_2} \overset{d}{\triangleright} \ldots \overset{d}{\triangleright} \mu^{\downarrow K_m})\right)^{\downarrow X_k}\end{aligned}$$

we can be facing a computationally hard problem, unless we take into account the fact that $K_1, K_2, \ldots, K_m$ are ordered to meet RIP. This enables us to carry out the necessary computations *locally* in the way that was shown in [7]. This computationally tractable process takes advantage of the well-known fact (an immediate consequence of the existence of a join tree, see [1]) that if $K_1, K_2, \ldots, K_m$ can be ordered to meet RIP, then for each $\ell \in \{1, 2, \ldots, m\}$ there exists an ordering that meets RIP and in which K_ℓ is the first one. So consider any K_ℓ for which $X_j \in K_\ell$, and find the ordering that meets RIP and starts with K_ℓ. Without loss of generality, let it be $K_1, K_2, \ldots, K_m$ (so, in this case we assume that $X_j \in K_1$). This fact makes the application of Property 8 of Theorem 1 possible; applying it $(m-1)$ times we get

$$\begin{aligned}&\nu_{X_j=a} \overset{d}{\triangleright} (\mu^{\downarrow K_1} \overset{d}{\triangleright} \mu^{\downarrow K_2} \overset{d}{\triangleright} \ldots \overset{d}{\triangleright} \mu^{\downarrow K_m})\\ &\quad = \nu_{X_j=a} \overset{d}{\triangleright} (\mu^{\downarrow K_1} \overset{d}{\triangleright} \mu^{\downarrow K_2} \overset{d}{\triangleright} \ldots \overset{d}{\triangleright} \mu^{\downarrow K_{m-1}})\\ &\qquad\qquad \overset{d}{\triangleright} \mu^{\downarrow K_m} = \ldots\\ &\quad = \nu_{X_j=a} \overset{d}{\triangleright} \mu^{\downarrow K_1} \overset{d}{\triangleright} \mu^{\downarrow K_2} \overset{d}{\triangleright} \ldots \overset{d}{\triangleright} \mu^{\downarrow K_m}\end{aligned}$$

Notice that the whole process can be repeated several times in case one wants to compute a conditional like $\mu(X_k|X_1 = a, X_3 = b, \ldots)$ – the only additional problem is that before each reordering of marginals (i.e., after each conditioning by one variable) one has to ensure that the model is composed from marginals of one bpa (This can be done by a simple computational process described in Proposition 7 of [7]). So, we can conclude that the computation of conditionals in d-decomposable models can be carried out locally.

Another question is whether the same computationally local process can be also used if we consider f-decomposable model

$$\mu = \mu^{\downarrow K_1} \overset{f}{\triangleright} \mu^{\downarrow K_2} \overset{f}{\triangleright} \ldots \overset{f}{\triangleright} \mu^{\downarrow K_m}.$$

However, though we conjecture that the answer is positive, at this moment the question still remains open. We would need to prove the following assertion to confirm the validity of our expectation.

Conjecture Suppose μ_1, μ_2 and μ_3 are bpas on $\mathbb{X}_K$, $\mathbb{X}_L$, and $\mathbb{X}_M$, respectively. If $L \supset (K \cap M)$ then, $(\mu_1 \overset{d}{\triangleright} \mu_2) \overset{f}{\triangleright} \mu_3 = \mu_1 \overset{d}{\triangleright} (\mu_2 \overset{f}{\triangleright} \mu_3)$.

6 Summary and Conclusions

The primary goal of this paper is to convince the reader that introducing two operators of composition for be-

lief functions is not an end in itself. Each of them has its own raison d'être. $^{d}\triangleright$ is, in a way, a generalization of probabilistic composition, introducing a conditional independence among the variables, whereas $^{f}\triangleright$ generalizes probabilistic factorization. Since these two notions coincide in probability theory, it is sufficient to use just one operator of composition in probability theory.

The role of $^{d}\triangleright$ for computation of conditionals is irreplaceable. On the other hand, computational procedures for f-decomposable models are much more efficient than those for d-decomposable models. The only problem spoiling their mutually advantageous coexistence will disappear once the presented conjecture is proven. Nevertheless, even if the conjecture is disproved there will still be a chance to design computationally efficient procedures employing both studied operators; they just will not be as simple as the procedure described in the last section.

Acknowledgements

This work was partially supported by GAČR under Grant No. 15-00215S.

References

[1] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis. On the Desirability of Acyclic Database Schemes. *J. ACM*, 30, 3, pp. 479–513, 1983.

[2] B. Ben Yaghlane, Ph. Smets, and K. Mellouli. Belief functions independence: II. the conditional case. *Int. J. Approx. Reasoning*, 31, pp. 31–75, 2002.

[3] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* 3, pp 146–158, 1975.

[4] W. E. Deming and F. F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11, 427–444, 1940.

[5] R. Jiroušek. Solution of the Marginal Problem and Decomposable Distributions. *Kybernetika* 27, 5, pp. 403–412, 1991.

[6] R. Jiroušek. Foundations of compositional model theory. *Int. J. General Systems*, 40, 6, pp 623–678, 2011.

[7] R. Jiroušek. Local computations in Dempster-Shafer theory of evidence. *Int. J. Approx. Reasoning*, 53, 8, pp. 1155–1167, 2012.

[8] R. Jiroušek and V. Kratochvíl. On Open Problems Connected with Application of the Iterative Proportional Fitting Procedure to Belief Functions. In *Proc. of the 8th Symp. on Imprecise Probabilities and Their Applications,* F. Cozman, T. Denœux, S. Desterecke, T. Seidenfeld, Eds., Compiègne, France, pp. 149–158, 2013.

[9] R. Jiroušek and P. P. Shenoy. Compositional models in valuation-based systems. *Int. J. Approx. Reasoning*, 55, 1, pp. 277–293, 2014.

[10] R. Jiroušek, J. Vejnarová and M. Daniel. Compositional models of belief functions. In *Proc. of the 5th Symp. on Imprecise Probabilities and Their Applications,* G. de Cooman, J. Vejnarová, M. Zaffalon, Eds., Praha, pp. 243–252, 2007.

[11] S. L. Lauritzen, *Graphical models.* Oxford University Press, 1996.

[12] F. Malvestuto. Equivalence of compositional expressions and independence relations in compositional models. *Kybernetika*, 50, 3, pp. 322-362, 2014.

[13] R. Merris. *Graph theory.* John Wiley, New York, 2001.

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

[15] P. P. Shenoy. Conditional independence in valuation-based systems. *Int. J. Approx. Reasoning*, 10, 3, pp. 203–234, 1994.

[16] J. Vejnarová. On conditional independence in evidence theory. In *Proc. of the 6th Symposium on Imprecise Probability: Theories and Applications,* Durham, UK, pp. 431–440, 2009.

Common Knowledge, Ambiguity, and the Value of Information in Games

Hailin Liu
Department of Philosophy
Carnegie Mellon University, Pittsburgh
hailinl@andrew.cmu.edu

Abstract

This paper asks whether the salient result about non-negative value of cost-free information holds in the context of games. By reexamining Osborne's example where information may hurt, it argues that the failure of this result is mainly driven by the assumption of common knowledge in the traditional framework of incomplete information games, since it leads to act-state dependence in a sequential setting. This paper also shows that such a failure occurs when we extend the framework of incomplete information games to allow for a representation of uncertainty using sets of probabilities and the use of Γ-maximin. Nevertheless, the key to this negative result is that a phenomenon called dilation of sets of probabilities obtains in this generalized setting.

Keywords. Bayesian games, value of information, ambiguity, act-state dependence, dilation, Γ-maximin.

1 Introduction

It is well known that in an individual decision problem a Bayesian decision maker should not refuse to receive cost-free information. To understand this result intuitively, note that choosing the expected utility maximizer d^* in the absence of new information has the same expectation as the plan of always making decision d^* no matter what additional information is forthcoming. But it might be that the decision d^* does not always maximize expected utility upon on receiving new information. It follows that the expected utility of choosing d^* initially cannot be greater than the prior expectation of choosing the best decisions after having more information. This ensures that additional information cannot be harmful to one's prior expectations. Thus it is rational for a Bayesian decision maker to postpone her terminal decision in order to acquire cost-free information. Indeed, such an intuitive idea can be traced back to Ramsey (1990), and has been formalized by Good (1967). However, this satisfactory result about the non-negative value of cost-free information fails to hold in many cases. For instance, Kadane et al. (2008) have shown that certain modifications of standard expected utility theory may require a decision maker to strictly prefer less information to more, thereby implying a negative value of information to the individual.

For our purposes, two cases under which this positive result does not obtain are worth mentioning. The first involves the idea of *act-state dependence*, namely probabilistic dependence between act and state, which is precluded by conventional expected utility theory. For example, within a small geographic market, consumers' inquiry about the price of a certain good may cause its price to rise. In this case, because of act-state dependence, a potential consumer strictly prefers not to learn the (cost-free) information about the price. As a result, the value of information is negative to the consumer. Another relevant instance of such a violation occurs in the context of decision making under uncertainty, where uncertainty is assumed to be represented by sets of probabilities rather than a single probability distribution. When extending expected utility theory to accommodate uncertainty aversion, it may happen that the set of unconditional probabilities for an event is properly included in the set of probabilities conditional on every event of some partition, which is a phenomenon known as *dilation* in the literature[1]. Given the presence of dilation, it should come as no surprise that a rational decision maker may refuse a free offer to learn a piece of new information. Therefore, the result introduced at the beginning of the paper is not robust with respect to the introduction of act-state dependence, as well as the choice of the modeling of uncertainty in the setting of single-agent decision making.

Moving beyond individual decision making, it has also been demonstrated that, in the context of games, more information may hurt the player who possesses the information[2]. More specifically, the player may be worse off when she has more information than when she does not. The simple logic behind this negative result is the following. In a game where one player has certain information

[1] See Seidenfeld and Wasserman (1993), Herron et al. (1994), and Herron et al. (1997) for an extensive and systematic study of the phenomenon of dilation.

[2] See for instance Akerlof (1970) and Osborne (2004) for several prominent examples that illustrate this observation concerning the negative value of information in games.

about the situation, this fact being commonly known among the players has a strategic impact on the players' rational strategy choices, which results in an inferior equilibrium outcome than the one for a game with more information. Thus, if that specific player can first choose between these two games and then play the chosen one, the optimal strategy is to play the game with less information, implying that new information has a negative value for that player. In light of this, we can conclude that the familiar result about the non-negative expected worth of cost-free information is not robust with respect to the introduction of strategic interaction either.

These findings suggest a need for a reconsideration of the information value problem in games. The aim of this paper is to provide a better understanding of this issue in games based on previously known results in the literature of decision theory. As a starting point of our investigation, we review the example introduced by Osborne (2004) exhibiting the counterintuitive result that in the context of Bayesian games a rational player may strictly prefer less information to more. In addition, the example helps us clarify the central issue and highlight the sequential problem involved in the comparison between a game where one player has less information, and a different version of the game where the player has more information. It is within such a sequential setting that the issue of whether information has a positive value can be properly assessed.

Next we show that the non-negative value of information can be restored by a weakening of the common knowledge assumption. Our analysis of Osborne's example makes it clear that the assumption of common knowledge plays a crucial role in the result of the negative information value in games. This leads us to consider a variant of the original game in which one player has more information while the other players are not aware of this fact. By comparing it with the original game, we show that more information to a player does lead to a better equilibrium outcome and thus has a positive value for the player[3]. We can intuitively understand this result by recognizing that act-state dependence is the real factor that has driven the result of the negative information value in games. The lack of common knowledge merely provides a convenient way to abstract from the kind of probabilistic dependence between a player's choice of the games and her probability about the opponent's choices.

We then deal with the question of whether the finding about the negative information value in games still holds in the presence of ambiguity. First, we briefly discuss how the traditional framework of Bayesian games developed by Harsanyi (1967/68) can be naturally extended to accommodate the idea of employing sets of probabilities to model uncertainty. Such an extension allows the players' initial beliefs about the state to be represented by closed and convex sets of probability distributions, which is more normatively appropriate and empirically grounded than the modeling of uncertainty through a single precise prior[4]. Following the pioneer work of Kajii and Ui (2005), we introduce a solution concept called Γ-*maximin equilibrium* that generalizes the concept of Bayesian Nash equilibrium to incomplete information games under ambiguity. Roughly speaking, our solution concept requires that each player chooses the optimal strategy in the sense of maximizing her minimum expected utility for each realization of her private signals. Using this solution concept, we then show that the phenomenon of the negative value of information reappears in the context of incomplete information games under ambiguity. This finding is perhaps not that surprising given the discussion on Osborne's example.

Nevertheless, we further argue that there exists an important distinction between Osborne's observation and the finding presented here. Specifically, we demonstrate that, when ambiguity is present, the conclusion about the negative value of information in games is robust with respect to the relaxation of the common knowledge assumption. This stands in a sharp contrast to what happens in Bayesian games when the assumption of common knowledge is weakened. In order to account for this difference, we briefly describe the phenomenon of dilation of sets of probabilities, and suggest that the role dilation together with the use of Γ-maximin plays in our examples is fundamental to understanding the result about the negative value of information in games under ambiguity. This implies that our finding bears a closer relationship to the results reported in Seidenfeld (2004) and Kadane et al. (2008) concerning the effect of dilation on the value of information in the context of single-agent decision making.

The remainder of this paper is organized as follows. In Section 2 we examine the well-known result about the negative value of information in Bayesian games in detail, and also present a different approach for determining the value of information in games by relaxing the assumption of common knowledge. Section 3 shows by examples that a non-expected utility player may pay not to receive cost-free information in incomplete information games under ambiguity. Section 4 discusses whether our finding still holds true even if there is a lack of common knowledge in games, and then relates it to the phenomenon of dilation. Section 5 contains a few concluding remarks.

[3] Neyman (1991) similarly argues that a player cannot be worse off by having more information provided that other players are not aware of it. The author thanks an anonymous reviewer for pointing out this reference.

[4] See for instance Knight (1921), Ellsberg (1961), Levi (1974), and Walley (1991) for a number of compelling arguments that justify the idea of using imprecise probabilities to model uncertainty.

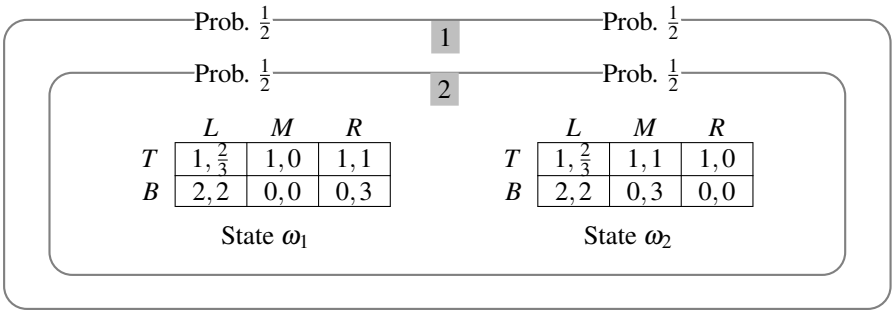


Figure 1: A Bayesian game with no information about the states

2 Common Knowledge and Negative Value of Information in Games

In this section we reconsider the example presented by Osborne (2004) where a player may assign a negative value to cost-free information in the context of Bayesian games. That is, the player becomes worse off in the version of the game where she has more information. It thus follows that the player would rationally pay to avoid learning this piece of new information, even though there is no cost associated with it. To keep things simple, we will return to this example later and argue that the assumption of common knowledge plays a critical role in Osborne's analysis, which throws considerable light on the nature of this negative result in games.

Example 2.1. *Consider the game structure shown in Figure 1. In this game, there are two players and two possible states* $\{\omega_1, \omega_2\}$. *The action set of player 1 is given by* $A_1 = \{T, B\}$ *and that of player 2 is given by* $A_2 = \{L, M, R\}$. *Players' payoffs under each strategy profiles are specified by two numbers in the corresponding box of the following table, with the first number being the payoff of player* 1. *Moreover, we assume that both players do not know the state and assign probability* $\frac{1}{2}$ *to each state. That is, both players believe that each state will occur with probability* $\frac{1}{2}$, *although neither of them knows which state they are actually in. And no further information concerning the state will be revealed to the players. Following normal convention, we further assume that everything about the game is common knowledge.*

Game theorists commonly regard the notion of Bayesian Nash equilibrium as the reasonable solution concept for solving Bayesian games. Informally speaking, a *Bayesian Nash equilibrium* consists of a collection of strategies such that each player's strategy is a *best reply* to the strategies the other players have chosen. It is easy to see by expectation that player 2's strategy L strictly dominates the other strategies. And player 1 has a unique best reply against L, namely, the action B. Hence, the strategy profile (B, L) is the unique Bayesian Nash equilibrium for the Bayesian game considered in this example. More importantly, note that this equilibrium gives rise to the outcome $(2,2)$ with the first component being the payoff to player 1.

It is well known in the decision-theoretic literature that in single-agent decision problems it is weakly better for an expected utility maximizer to wait for more information prior to make a terminal decision. In other words, every expected utility decision maker prefers more information to less. In order to test whether this is valid within the context of games, Osborne suggests to contrast Example 2.1 described above with the following case.

Example 2.2. *As before, we assume that both players do not know the state before receiving their private information and assign probability* $\frac{1}{2}$ *to each state. However, player 2 learns the state from her private information, whereas player 1 does not. In other words, after having received her private information, player 2 knows whether she is playing the strategic game on the left or the one on the right. By contrast, player 1 still does not know the state in this interim stage and thinks that with equal probability* $\frac{1}{2}$ *she is playing one of these two games. Nevertheless, player 1 knows the fact that player 2 is informed of the state. In a similar way, this strategic situation can be described by Figure 2 where the two frames labeled* 2 *enclosing each table indicate that player 2 can perfectly distinguish between these two tables.*

Notice that each type of player 2 has a strictly dominant action, namely, R and M respectively. This means that player 2 should choose to play the strategy (R, M), no matter what player 1 intends to do. Knowing that player 2 is informed of the state, player 1 can then anticipate the above reasoning on the behalf of player 2. Given this, player 1 should respond optimally by playing T, which is the unique best response to (R, M). Therefore, this game has a unique Bayesian Nash equilibrium $(T, (R, M))$, which gives rise to the outcome $(1, (1, 1))$.

Now suppose that in a sequential setting player 2 is first asked to choose either to play the Bayesian game in Example 2.1, or to play the Bayesian game in Example 2.2, and then enters the chosen game. Note that player 2 would obtain more information about the state in Example 2.2 than

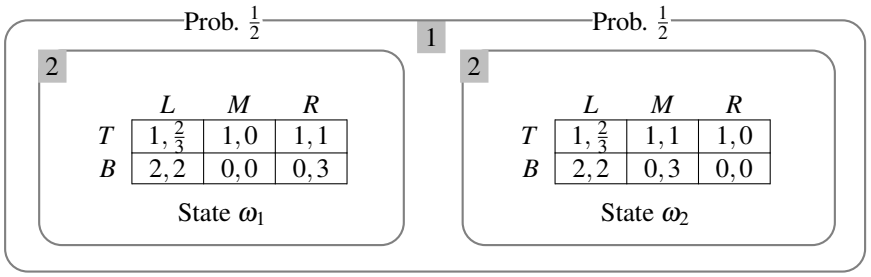


Figure 2: A Bayesian game in which player 2 knows the state

she would in Example 2.1. Contrary to our intuition, player 2 would rather choose to play the game in Example 2.1 instead of the one in Example 2.2, where she is actually informed of the state. It is important to remark, however, that such a choice of player 2 is perfectly rational in the exact sense of maximizing expected utility, since the unique Bayesian Nash equilibrium of the game in Example 2.1 yields player 2 a higher expected payoff than the unique Bayesian Nash equilibrium of the game in Example 2.2 does. This implies that player 2 strictly prefers to play the game with less information rather than more, which stands in stark contrast to the familiar result in decision theory about the non-negative value of information.

Generally, more information is valuable for making better decisions. In a strategic situation, however, knowing only that a certain kind of information becomes available to someone may alter one's behavior, even if she herself does not know what the information is. It is thus important to keep in mind that one should separate the real issue concerning the value of information in games, from the question whether the knowledge of others holding certain private information plays a role in determining its value. Next we suggest an instructive way to investigate whether more information is better for a player by restricting others from knowing that fact. Such an approach then avoids the kind of complication just described.

Example 2.3. *Consider a slightly modified version of the Bayesian game introduced in Example 2.2. As before, we assume that both players do not know the state before receiving their private information and assign probability $\frac{1}{2}$ to each state. Upon receipt of their private information, player 2 learns the state, whereas player 1 still does not know the state. Unlike in the previous game, in this case we assume that player 1 does not know the fact that player 2 is informed of the state. To be a bit more specific, player 1 still believes that player 2 assigns probability $\frac{1}{2}$ to each state like she does. Given that player 1 is not informed of player 2's updated belief about the state, it is obvious that the conventional assumption about common knowledge is no longer valid in the current case.*

The only difference between Example 2.2 and Example 2.3 lies in the fact that in the latter case we specifically make the information concerning player 2 knowing the state unavailable to player 1. Nevertheless, such a slight modification has greatly changed the strategic interaction between these two players. To see this, note that in Example 2.3 two players have rather different information about the situation: At the interim stage where players learn of their private information, player 1 does not know player 2's actual belief about the state, whereas player 2 has a complete knowledge of the strategic situation, including her own information being unknown to player 1. In this sense, player 1 holds a false belief about player 2's updated belief. In this setting, the usual strategic impact caused by new information goes away, since the information would only affect the person who has it. By restricting player 1 from knowing the fact that player 2 learns the state, we can then ask whether this piece of new information has any value to player 2 or not.

We first need to figure out how to resolve these kinds of strategic situations involving "imperfect" beliefs, as the notion of Bayesian Nash equilibrium is designed to solve only standard Bayesian games. As a minimal requirement of rationality, a reasonable solution should respect the information available to each player. Moreover, we want to follow the tradition of modeling the players as expected utility maximizers. Given these requirements, it seems reasonable to suggest that each player should choose a strategy that maximizes her expected payoff given the (possibly false) beliefs about the state and the strategies chosen by the other players, as long as these beliefs can be justified in terms of information available to her.

Now let us apply the above idea to the situation described in Example 2.3. First, we claim that player 1's optimal choice is to play B. To see this, note that in the interim stage player 1 still believes that player 2 does not know the state and assigns probability $\frac{1}{2}$ to each state. Given such a belief, player 1 would believe that player 2 will choose the action L, which strictly dominates the other two actions. Anticipating this conjecture, player 1 responds optimally by playing B. On the other hand, player 2 has perfect knowledge about the situation, even including the fact that player 1 holds incorrect belief about whether she knows the state. Since player 2 learns the state at the interim stage, player 2 will choose to play her strictly dominant actions in

each state, that is, R and M respectively. Although player 2 can anticipate player 1 to select B rather than her equilibrium strategy L, player 2 would keep her optimal choice of (R,M) unchanged, since it still maximizes her expected payoff given B. Hence, the recommended play in this case is the strategy choice $(B,(R,M))$.

It is important to note that player 2's payoff under the optimal play in Example 2.3 is 3, which is greater than her payoff of 2 associated with the equilibrium of the game in Example 2.1. Thus, if player 2 is faced with an initial choice between whether to play the game in Example 2.1 or instead to play the game in Example 2.3, the rational decision is to choose the latter one. It follows that player 2 strictly prefers to have more information rather than less, which is in accordance with the familiar result concerning the non-negative value of cost-free information. This contrasts sharply with the foregoing analysis.

Note that common knowledge plays a critical role in solving both games in terms of Bayesian Nash equilibrium. In both games, everything about the games is common knowledge, and thus both players can reason about each other's strategy choice on the behalf of her opponent. In Example 2.1, based on her prior belief about the state, player 2 select her best action L, which leads player 1 to choose B. Similarly, in Example 2.2, player 2 utilizes her information of the state and singles out the optimal strategy (R,M), which induces player 1 to choose T. So, when player 2 is faced with the sequential problem of choosing first between these two games and then playing the chosen one, there exists probabilistic dependence between her choice of the games and her probability about the opponent's strategy choice. In this sense, act-state dependence arises in such a sequential problem.

On the other hand, act-state dependence does not arise when player 2 is asked to choose first whether to play the game in Example 2.1 or to play the game in Example 2.3, and then to play the selected game. To see this, recall that in Example 2.3 player 1 does not know the fact that player 2 has more information. In the light of this assumption, player 1 is not able to arrive at the same conclusion as player 2 does. Instead, player 1 would obtain the same conjecture about player 2's strategy choice as the one in Example 2.1, namely, L. Since player 2 has perfect knowledge about the situation, she can deduce from her information player 1's strategy choice, which is identical to the one in Example 2.1. So, if player 2 is asked to first choose between the game in Example 2.1 and the one in Example 2.3 and then to play the chosen game, there is no act-state dependence, since both players' probabilities for how the other player chooses are unchanged. For this reason player 2 would assign a positive value to the information about the state.

It has already been shown (Kadane et al., 2008) that in individual decision problems the result about the non-negative value of cost-free information does not hold provided that there is act-state dependence in personal probabilities. Then it should come as no surprise that, in the context of games, players may have negative value for new information in the presence of act-state dependence. It is act-state dependence that has driven the counterintuitive result about the negative value of information discussed above[5]. Relaxing the assumption of common knowledge in Example 2.3 enables us to prevent the occurrence of act-state dependence in the sequential problem.

3 Negative Value of Information in Games under Ambiguity

This section is devoted to extending the analysis of the value of information in Bayesian games to accommodate ambiguity aversion by considering the multiple priors models developed by Gilboa and Schmeidler (1989). In order to explore whether the phenomenon of negative value of information is robust with respect to extensions of the expected utility theory, we need to investigate how to incorporate models of imprecise probabilities into Bayesian games. We will not attempt to present a formal model of incomplete information games under ambiguity here[6]. Instead, we shall introduce the basic ideas in an informal way.

Unlike in Harsanyi (1967/68), here we take the view that, due to limited information, a player may not be able to identify a unique prior to describe her belief about the states, which can be characterized as a set of probability distributions. More precisely, we assume that in an incomplete information game each player's perception of uncertainty about the states is modeled by a closed and convex set of probability measures, instead of a single common probability distribution. And we adopt the principle of Γ-maximin as the decision rule used by all the players. In the same spirit of Bayesian Nash equilibrium, we propose a new solution concept in which each player chooses the optimal action in the sense of maximizing the minimum expected utility for each realization of her private signal. In this sense, our model constitutes only a minor departure from the standard approach to Bayesian games.

In order to help better understand how our game model works, let us consider the following example through which we introduce the major components of an incomplete information game under ambiguity and the solution concept for this kind of games.

Example 3.1. *Consider the game structure shown in Figure 3 (Game 4). As in Bayesian games, nature moves first and chooses which state to occur. Moreover, assume that both players in this game are uncertain about nature's*

[5] The author thanks Teddy Seidenfeld for pointing this out. Seidenfeld (2009) first gives such an analysis of Osborne's example regarding negative value of information in games.

[6] See Kajii and Ui (2005) for a formal definition of such a game model.

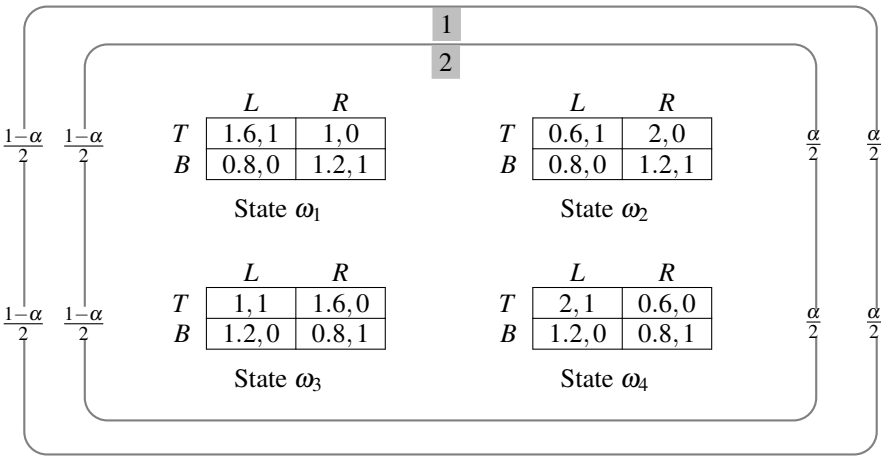


Figure 3: An incomplete information game under ambiguity

move in choosing the state. Nevertheless, suppose that both players' prior beliefs about the states are depicted by the following common set of priors over the states with $\alpha \in [0.1, 0.9]$.

$$\mathscr{P} = \{p \in \Delta(\Omega) : p(\omega_1) = p(\omega_3) = \frac{1-\alpha}{2}, \quad p(\omega_2) = p(\omega_4) = \frac{\alpha}{2}\}. \tag{1}$$

Like in the framework of Bayesian games, the solution concept proposed here requires that each player's conjecture regarding the opponents' choices is correct in the standard sense. Roughly speaking, our solution concept called Γ-*maximin equilibrium* is defined as a strategy profile such that each player's strategy is optimal in the sense of maximizing the minimum expected payoff, given the other players' strategy choices. Clearly, the concept of Γ-maximin equilibrium generalizes the notion of Bayesian Nash equilibrium to games under ambiguity.

Now let us apply the concept of Γ-maximin equilibrium to the game in Figure 3. We can transform the game into an ordinary game in strategic form by calculating each player's expected payoffs under different strategy profiles using each of her posterior probabilities defined over the state. Given a strategy profile, each player's payoff generally becomes an interval instead of a precise value. However, it turns out that in this case the players' payoffs are all determinate and independent of the variable α. By a simple calculation, it follows that this game can be turned into the following 2×2 game in strategic form.

	L	R
T	1.3, 1	1.3, 0
B	1, 0	1, 1

Figure 4: Game 4 in strategic form

It is obvious that this game can be easily solved by strict dominance, which leads to a unique solution, namely, the strategy profile (T, L). Hence, this is the unique Γ-maximin equilibrium for the incomplete information game under ambiguity in Figure 3, where its corresponding outcome is $(1.3, 1)$. One may notice that the principle of Γ-maximin does not play a role in solving this specific game. It is thus worthwhile to point out that this is not generally true for incomplete information games under ambiguity, since the payoffs would typically form intervals. As we shall see in the next example, the game is solved by explicitly applying the idea of Γ-maximin.

It has already been demonstrated that in single-agent decision problems a non-expected utility decision maker, especially a Γ-maximin decision maker, may prefer less information to more[7]. Thus one should expect that a similar phenomenon would arise in incomplete information games under ambiguity. In the following, we present a case where a Γ-maximin player would rationally pay not to receive cost-free information, which is exactly in the same spirit as the negative result presented in the previous section.

In order to draw the needed contrast with the game in Figure 3, we consider a situation in which player 2 has more information and player 1 knows that fact. In this sense, we follow exactly the same construction as Osborne has proposed for the Bayesian case.

Example 3.2. *Consider the game structure shown in Figure 5, which is similar to the game in Example 3.1. Likewise, assume that both players' prior beliefs about the states are represented by the same set of priors over the states given in Equation* (1)*. As opposed to the previous case, we assume in this game that player 1 learns more information about the state, whereas player 2 does not. To be more specific, player 1 may receive two signals; when player 1*

[7] See Wakker (1988), Seidenfeld (2004) and AI-Najjar and Weinstein (2009) for various examples that illustrate this point.

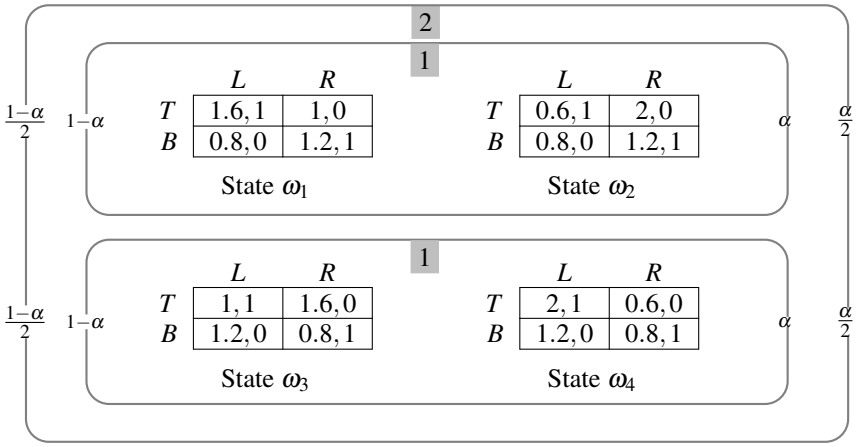


Figure 5: The second incomplete information game under ambiguity

gets one of the signals, she knows that the state is either ω_1 or ω_2; when she gets another signal, she knows that the state is either ω_3 or ω_4. Formally, player 1's signal function can be defined as follows $\tau_1(\omega_1) = \tau_1(\omega_2) = t_1$ and $\tau_1(\omega_3) = \tau_1(\omega_4) = t_1'$. By contrast, player 2's signal function is given as follows: $\tau_2(\omega_k) = t_2$ for $k = 1,2,3,4$. This indicates that player 2 receives a single signal in the interim stage.

Similarly, we want to solve this incomplete information game under ambiguity by considering the concept of Γ-maximin equilibrium. However, it is important to remark that, as opposed to the notion of Bayesian Nash equilibrium, this solution concept is sensitive to whether we solve the game from an *ex ante* or *interim* perspective. It is important to note that the ex ante and interim Γ-maximin equilibrium may lead to rather different solutions to the same incomplete information game under ambiguity, since the normal form and extensive form of a decision problem are not equivalent under Γ-maximin[8]. Here we shall only focus on the concept of interim Γ-maximin equilibrium, which can be regarded as a direct generalization of interim Bayesian Nash equilibrium.

Before introducing this interim solution, we need to specify how the players would update their beliefs upon receiving new information, which is critical for determining players' expected payoffs in the interim stage. The updating problem lies at the heart of the theory of incomplete information games. In the framework of Bayesian games, it is widely accepted that Bayes' rule provides a convenient and useful way of revising players' initial beliefs in the light of new information. On the contrary, there has been little agreement in the literature on how to update one's beliefs in the presence of ambiguity as new information is gathered[9]. Nevertheless, the so-called *full Bayesian updating rule* is often regarded as the straightforward generalization of Bayes' rule to the context of imprecise probabilities. Thus we will apply this rule to incomplete information games with the modeling of uncertainty through sets of probabilities.

Given player 1's private information, the set of posterior probabilities can be derived from the set of priors by applying the full Bayesian updating rule:

$$\mathscr{P}(\cdot\,|\,t_1) = \{p \in \Delta(\Omega) : p(\omega_1) = 1-\alpha, p(\omega_2) = \alpha\}$$
$$\mathscr{P}(\cdot\,|\,t_1') = \{p \in \Delta(\Omega) : p(\omega_3) = 1-\alpha, p(\omega_4) = \alpha\},$$

where $\alpha \in [0.1, 0.9]$. And it is clear that player 2's initial beliefs about the state would remain unchanged. Moreover, note that player 1's payoffs to some strategy profiles depends upon the value of α and thus become intervals. This stands in direct contrast with the game in Figure 3.

Similarly to the notion of interim Bayesian Nash equilibrium, the concept of interim Γ-maximin equilibrium requires that each type of each player chooses a strategy that maximizes her minimum *interim expected payoff* given the strategies chosen by the other types of every other player. We claim that this game has a unique interim Γ-maximin equilibrium in pure strategy, namely, the strategy profile $((B,B),R)$[10]. To see this, note that for each type of player 1 the action B always yields a higher minimum expectation than T does, no matter whether player 2 chooses L or R. Thus, player 1 should choose to play the strategy (B,B). Given such a conjecture, player 2 should respond optimally by playing R. Importantly, observe that this unique equilibrium yields type t_1 of player 1 an expected payoff of 1.2

[8] Seidenfeld (1988) explicitly shows that the rule of Γ-maximin distinguishes between sequential decisions in extensive form and the normal form one-stage decisions.

[9] See Gilboa and Schmeidler (1993) and Grove and Halper (1998) for more detailed discussions about various updating rules when using sets of probabilities to model ambiguity.

[10] In fact, this game has no other interim Γ-maximin equilibrium in mixed strategy. For the current purpose, however, the notion of interim Γ-maximin equilibrium in pure strategy is sufficient. Moreover, one can easily verify that this equilibrium cannot be justified a Bayesian Nash equilibrium using any element of the set of priors.

and type t_1' of player 1 an expected payoff of 0.8, both of which are less than the payoff in the unique Γ-maximin equilibrium of the game in Figure 3.

Now if player 1 has an initial choice between the game in Example 3.1 and the game in Example 3.2, player 1 strictly prefers to play the former one in which player 1 has no information about the state. In other words, player 1 assigns a negative value to the information that she may receive in the latter game. Unsurprisingly, then, this example demonstrates that the phenomenon of the negative value of information would arise in the context of incomplete information games under ambiguity. In this sense, such a phenomenon is indeed robust with respect to extensions of Bayesian games using certain classes of non-expected utility models.

In view of the analysis in Section 2, one may think that this is due to the same fact that act-state dependence obtains in this sequential problem. This is correct to some extent, since player 1's new information about the state does have a strategic impact on her conjecture about player 2's strategy choice. As we shall see in the next section, however, act-state dependence is not the key factor that accounts for this negative result in games under ambiguity. The use of imprecise probabilities to represent uncertainty in games actually introduces additional complexity to the value of information in games.

4 Negative Value of Information without Common Knowledge

In this section we consider a variant of the game in Example 3.2 where the assumption of common knowledge is slightly weakened, and demonstrate that the conclusion of the previous section still holds even if some player has more information that is not commonly known to the other players. Then, on the basis of previously known results from decision theory, we provide a more in-depth account of the nature of the negative value of information in incomplete information games with and without ambiguity.

As we announced, we reexamine the information value problem in games by comparing the game in Example 3.1 with a game where player 1 has more information but player 2 is not informed of this fact. Similarly, the construction of the latter game is deliberately designed to eliminate the strategic effects caused by common knowledge of new information.

Example 4.1. *Consider a slightly modified version of the incomplete information game under ambiguity introduced in Example 3.2. The modified game is very similar to the one in Figure 5 except that player 2's belief about player 1's information in the interim stage does not match up with the actual information possessed by player 1. More precisely, we assume here that player 1 obtains more information about the state, but that is not revealed to player 2. Instead, player 2 still thinks that player 1 holds the same belief as in the ex ante stage.*

It is worth emphasizing that the fundamental difference between Example 3.2 and the current example lies in the question of whether or not player 2 knows the fact that player 1 has more information about the state in the interim stage. As we have seen, such a modification has a profound impact on the solution to the games, which in turn substantially changes the analysis of the information value in games.

In a similar fashion, we propose to solve this game by demanding only that the strategy chosen by each player is optimal in the sense of maximizing the lower expectation on the basis of her information about the state and the other players' strategy choices. In contrast with Γ-maximin equilibrium, this notion does not impose the consistency requirement on each player's conjectures about the other players' strategy choices, since some player may not have all the information about the opponents' characteristics. For instance, in the current example, the fact that player 1 learns more information about the state is not available to player 2. For this reason it is not surprising to see that some player may hold a belief that is inconsistent with the opponents' actual behavior. In light of such a weakening of the common knowledge assumption, however, it seems quite reasonable to require each player to justify the strategy choice on the grounds of information that is available to her.

We first argue that player 2 should choose to play L. The argument goes like this. In this case, player 2 still believes that neither player obtains any further information about the state, and both of them employ the same set of priors $\mathscr{P}$ to represent their uncertainty about the state. Given such a belief, player 2 will then think that the game can be transformed into the strategic form game depicted in Figure 4, which is solvable using strict dominance. Player 2 would thus expect player 1 to select T and then choose to play her unique best response L.

Second, we claim that player 1's optimal strategy is (B,B). Unlike player 2, player 1 not only has more information about the state in the interim stage, but also knows that player 2 does not learn of this fact. Based on her private information, player 1 can reason as follows. As noted before, the action B always gives player 1 a higher minimum expected payoff than T does, regardless of the action chosen by player 2. Thus, the optimal choice for player 1 is to choose the strategy (B,B). We should also remark that, according to Γ-maximin, player 1's strategy (B,B) is optimal as well given player 2's choice of L. This means that player 1 does not want to alter her choice, even though she can expect that player 2 will choose to play L instead of the equilibrium strategy R. Importantly, note that under the specified optimal plays for both players, the payoffs to two types of player 1 are 0.8 and 1.2 respectively.

We now turn our attention to the question of whether player 1 prefers more information to less without the common knowledge assumption. Suppose that player is presented with the following sequential problem: First, decide whether to play the game in Example 3.1 or instead to play the game in Example 4.1, and then enter the selected game. What would be player 1's initial choice? Our foregoing analysis of these two games suggests that the initial choice for player 1 is to play the game in Figure 3 in which player 1 has no information about the state. That is, player 1 would rationally pay to avoid learning more information about the state in this case. Thus, the same phenomenon occurs again in the presence of ambiguity, even if player 2 is assumed to not know the fact that player 1 has more information. In contrast to the Bayesian case, one cannot explain away this counterintuitive result by weakening the assumption of common knowledge.

Let us reflect on this negative result obtained in games under ambiguity. First, it is important to note that the relaxation of the common knowledge assumption does block the kind of strategic effects caused by new information. For instance, player 2's strategy choice in the game of Example 4.1 is not affected by player 1's extra information about the state, since player 2 does not learn that fact. In this respect, the current case is quite similar to the Bayesian game. It thus follows that act-state dependence does disappear when we restrict player 2 from knowing that player 1 has more information. In the absence of act-state dependence, we still infer that player 1 strictly prefers less information to more when ambiguity is present. Taken these together, it suggests that there must be something other than act-state dependence, which leads to the undesirable result of the negative value of information in games under ambiguity.

A natural question then arises: What is the real reason behind this counterintuitive result? To address this question, we first point out a distinctive feature of the game introduced in Example 3.2 by comparing the two sets of probabilities that are used to represent player 1's prior and posterior beliefs about the state. At the beginning of the game, player 1's belief about the state is represented by the set $\mathscr{P}$ as depicted by Equation (1). By contrast, upon arrival of new information player 1 changes the probabilities by applying the full Bayesian updating rule, which gives rise to the sets $\mathscr{P}(\cdot\,|\,t_1)$ and $\mathscr{P}(\cdot\,|\,t_1')$. Observe that the prior probability interval for event $\{\omega_1,\omega_4\}$ is strictly contained within its conditional probability interval given by $\mathscr{P}(\cdot\,|\,t_1)$. Intuitively, this means that player 1's probability judgment about the event $\{\omega_1,\omega_4\}$ becomes more imprecise after she observes t_1. We can say the same thing about the probability interval for $\{\omega_2,\omega_3\}$. In a similar fashion, the same facts can be reported regarding player 1's probability estimates for these events in the case that t_1' is observed. We can thus conclude that player 1's probability intervals for nature's choice of the state become wider, regardless of the signal revealed to her. Such a phenomenon is called *dilation of sets of probabilities*.

Intuitively, one may expect that one's probability interval for some hypothesis should become narrower after learning the outcome of some experiment. Contrary to common sense, when dilation obtains, the probability interval actually expands, no matter what the outcome is. It is not surprising to find that a Γ-maximin decision maker may refuse to learn new information when dilation is present. Because the agent becomes more uncertain, no matter what the outcome of the experiment is. In single-agent decision problems, Γ-maximin requires that a decision maker using that decision rule should pay to not receive new information whenever dilation occurs.

In the context of games, it is reasonable to expect that this would also arise in the presence of ambiguity if dilation occurs. As explained above, this is exactly what happens in the example of incomplete information games under ambiguity. Hence, it is the phenomenon of dilation, together with the use of Γ-maximin, rather than act-state dependence that has really driven the result of the negative value of information in games under ambiguity. In addition, it is important to remark that in the presence of dilation the same result holds with or without the common knowledge assumption. This is why we arrive at the same conclusion in both comparisons that player 1 would rather choose the situation in which she does not have more information. As our examples have shown, when considering whether new information has any value in games, act-state dependence is very sensitive to strategic impact whereas dilation does not. In this sense, dilation is more robust than act-state dependence regarding the result about the negative value of information in the context of games under ambiguity.

5 Summary

The discussion here is concerned with the issue of whether the familiar result about the non-negative value of information is still valid under strategic situations. Osborne (2004) has shown that in Bayesian games more information to one player may make her worse off in terms of equilibrium payoffs. This finding raises the concern that a wide variety of game-theoretic models with important economic applications may fail to satisfy a seemingly reasonable requirement on rational choice. In this paper we have re-examined the information value problem in the context of incomplete information games by considering the strategic effect of common knowledge on players' strategies and also the introduction of ambiguity.

We draw two conclusions from this investigation. First, the role act-state dependence plays in the sequential setting is fundamental to understanding the result that a Bayesian player may rationally pay not to receive cost-free information in strategic interactions. We have argued that the

kind of probabilistic dependence between a player's initial choice of the games and her probability about the opponents' strategy choice leads the player to assign a negative value to the information. We have further shown that the non-negative value of information can be restored by weakening the assumption of common knowledge. This is mainly due to the fact that the lack of common knowledge isolates the strategic effect of information on equilibrium play, and thus removes act-state dependence. So this result lends an additional support for our account of the result of the negative value of information in Bayesian games.

Second, in the presence of ambiguity, more information may also damage the player who holds it, thereby implying a negative value of information. Yet, we should emphasize that the negative value of information can occur in the context of incomplete information games under ambiguity, even if we isolate the effect of act-state dependence by relaxing the common knowledge assumption. Unlike what happens in Bayesian games, in this case the result about the negative value of information still holds true mainly due to dilation of sets of probabilities. The upshot is that, within the generalized game-theoretic framework that allows for the modeling of uncertainty through sets of probabilities, one needs to pay attention to both the strategic effect of common knowledge and the role of ambiguity in order to respect the value of information.

Acknowledgements

I would like to thank Teddy Seidenfeld and Kevin Zollman who originally suggested this topic, and provided constant encouragement and valuable suggestions throughout the investigation. I am grateful to Martin Osborne for helping me with a technical problem when using the latex package sgame.sty created by him. I also want to thank the four anonymous reviewers for their insightful comments on an earlier version of the paper.

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84(3), 488-500.

Al-Najjar, N. I., and Weinstein J. (2009). The Ambiguity Aversion Literature: A Critical Assessment. *Economics and Philosophy* 25, 249-284.

Ellbsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* 75, 643-649.

Gilboa, I., and Schmeidler, D. (1989). Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics* 18, 141-153.

Gilboa, I., and Schmeidler, D. (1993). Updating Ambiguous Beliefs. *Journal of Economic Theory* 59, 33-49.

Good, I. J. (1967). On the Principle of Total Evidence. *British Journal of Philosophy of Science* 17, 319-321.

Grove, A. J., and Halpern, J. (1998). Updating Sets of Probabilities. In *Proceedings of the Fourteenth Conference on Uncertainty in AI*, 173-182.

Harsanyi, J. C. (1967/68). Games with Incomplete Information Played by 'Bayesian' Players. *Management Science* 14, 159–182, 320–334, and 486–502.

Herron, T., Seidenfeld, T., and Wasserman, L. (1994). The Extent of Dilation of Sets of Probabilities and the Asymptotics of Robust Bayesian Inference. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1, 250-259.

Herron, T., Seidenfeld, T., and Wasserman, L. (1997). Divisive Conditioning: Further Results on Dilation. *Philosophy of Science* 64, 411-444.

Kadane, J. B., Schervish, M., and Seidenfeld, T. (2008). Is Ignorance Bliss? *The Journal of Philosophy* 105, 5-36.

Kajii, A., and Ui, T. (2005). Incomplete Information Games with Multiple Priors. *Japanese Economic Review* 56, 332-351.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit.* Boston, MA: Houghton Mifflin Company.

Levi, I. (1974). On Indeterminate Probabilities. *Journal of Philosophy* 71, 391-418.

Neyman, A. (1991). The Positive Value of Information. *Games and Economic Behavior* 3, 350-355.

Osborne, M. J. (2004). *An Introduction to Game Theory*. New York, NY: Oxford University Press.

Ramsey, F. P. (1990). Weight or the Value of Knowledge. *British Journal for the Philosophy of Science* 41, 1-4.

Seidenfeld, T. (1988). Decision Theory without "Independence" or without "Ordering": What is the Difference? *Economics and Philosophy* 4, 267-290.

Seidenfeld, T. (2004). A Contrast between Two Decision Rules for Use with (Convex) Sets of Probabilities: Γ-maximin versus E-admissibility. *Synthese* 140, 69-88.

Seidenfeld, T. (2009). On the Value of Information in Games (Online). Available: www.hss.cmu.edu/philosophy/faculty-seidenfeld.php.

Seidenfeld, T., and Wasserman, L. (1993). Dilation for Sets of Probabilities. *Annals of Statistics* 21, 1139-1154.

Wakker, P. (1988). Non-Expected Utility as Aversion to Information. *Journal of Behav Decis Mak* 1, 169-175.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall.

Calculating Bounds on Expected Return and First Passage Times in Finite-State Imprecise Birth-Death Chains

Stavros Lopatatzidis and **Jasper De Bock** and **Gert de Cooman**
Ghent University, SYSTeMS Research Group
{Stavros.Lopatatzidis,Jasper.DeBock,Gert.deCooman}@UGent.be

Abstract

We provide simple methods for computing exact bounds on expected return and first passage times in finite-state birth-death chains, when the transition probabilities are imprecise, in the sense that they are only known to belong to convex closed sets of probability mass functions. These so-called imprecise birth-death chains are special types of time-homogeneous imprecise Markov chains. We also present numerical results and discuss the special case where the local models are linear-vacuous mixtures, for which our methods simplify even more.

Keywords. Birth-death chain, Markov chain, imprecise, return time, first passage time, credal set.

1 Introduction

A birth-death chain [11, Section 9.4] is a special type of time-homogeneous Markov chain that is used in various scientific fields, including evolutionary biology and queueing theory. We consider the generalised case of an imprecise birth-death chain, where the transition probabilities are imprecise, in the sense that they are only known to belong to convex closed sets of probability mass functions—credal sets. This may be the case because the transition probabilities are based on partial expert knowledge or limited data, or for the purposes of conducting a sensitivity analysis. Similar models were already considered in Reference [2], which presented results on limiting conditional distributions for imprecise birth-death chains with one absorbing state. Imprecise birth-death chains are themselves a special case of so-called (time-homogeneous) imprecise Markov chains, which were studied in—amongst others—References [5, 7, 9].

This paper focusses on—upward and downward—first passage times and return times.[1] For precise birth-death chains, these have been studied in, for example, Reference [8]. For the more general case of imprecise birth-death chains, we are not aware of any results. Our most important contribution are simple methods for computing lower and upper—exact bounds for—expected values of first passage times and return times in finite-state imprecise birth-death chains. We also present numerical results and discuss the special case where the local models are linear-vacuous mixtures, for which our methods simplify even more.

We start in section 2 by discussing the notion of a precise birth-death chain and then introduce our imprecise version of it in Section 3. Section 4 defines return and—upward and downward—first passage times and their lower and upper expected values. In Sections 5 and 6, we provide our methods for computing lower and upper expected upward and downward first passage times. We use these methods in section 7 to calculate lower and upper expected return times. Section 8 discusses the special case where the local models are linear-vacuous mixtures and Section 9 presents numerical results. We conclude the paper in Section 10.

2 Birth-Death Chains

Finite-state birth-death chains are special cases of time-homogeneous finite-state Markov chains. Their state space, denoted by $\mathcal{X}$, is finite and can be linearly ordered by an integer. Without loss of generality, we may assume that $\mathcal{X} = \{0, \dots, L\}$, with $L \in \mathbb{N}$.[2] At any time point $n \in \mathbb{N}$, the state of the chain is represented by a random variable, denoted by X_n, which takes values in the state space $\mathcal{X}$. For every $n \in \mathbb{N}$, the sequence of variables $X_1, \dots, X_n$ is denoted by $X_{1:n}$ and takes values $x_{1:n} := x_1, \dots, x_n$ in $\mathcal{X}^n$. Similarly, we use $X_{1:\infty}$ as a shorthand notation for the infinite sequence $X_1, \dots, X_n, \dots$ Also, for every $k \in \mathbb{N}$ such that $k \leq n$, we let $X_{k:n}$ and $X_{k:\infty}$ be the sequences of states from time point k to n or infinity, respectively.

[1] These are often called recurrence times as well.

[2] We do not consider zero to be a natural number.

Since finite-state birth-death chains are special cases of (time-homogeneous) Markov chains, they satisfy the Markov condition, which requires that

$$E_{n+1}(\cdot|x_{1:n}) = E_{n+1}(\cdot|x_n) \text{ for all } x_{1:n} \in \mathcal{X}^n, \quad (1)$$

where $E_{n+1}(\cdot|x_n)$ is the expectation operator that corresponds to the probability mass function $p(X_{n+1}|x_n)$ for X_{n+1}, conditional on $X_n = x_n$, and similarly for $E_{n+1}(\cdot|x_{1:n})$. If the Markov chain is furthermore time-homogeneous, then $p(X_{n+1}|x_n)$—and therefore also $E_{n+1}(\cdot|x_n)$—does not depend on n, which implies that all the transition probabilities can be summarised by means of a single stochastic matrix P of dimension $L+1$, by letting $P_{ij} := p(j|i)$ for all $i, j \in \mathcal{X}$. In the special case of a birth-death chain, this stochastic matrix is tridiagonal, which expresses that transitions are only possible between adjacent states. Hence, P is of the form

$$P = \begin{pmatrix} r_0 & p_0 & 0 & \cdots & \cdots & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & q_{L-1} & r_{L-1} & p_{L-1} \\ 0 & \cdots & \cdots & 0 & q_L & r_L \end{pmatrix}$$

where the elements of each row sum to 1. For any $i \in \mathcal{X}/\{0, L\}$, we will assume that the probabilities p_i, q_i and r_i are positive, and similarly for r_0, p_0, q_L, r_L. Figure 1 depicts a graphical representation of a finite-state birth-death chain.

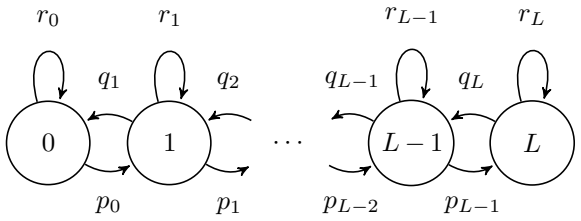


Figure 1: A birth-death chain with $\mathcal{X} = \{0, \ldots, L\}$

3 Imprecise Birth-Death Chains

Imprecise birth-death chains are similar to precise birth death chains. The main difference is that the probability mass functions that make up the matrix P do not need to be specified exactly. They are only known to belong to convex closed sets of probability mass functions, called credal sets. Formally, for every finite set $\mathcal{Y}$, a credal set on $\mathcal{Y}$ is a closed and convex subset of the set

$$\Sigma_{\mathcal{Y}} := \left\{ \pi \in \mathbb{R}^{\mathcal{Y}} \colon \sum_{y \in \mathcal{Y}} \pi(y) = 1, (\forall y \in \mathcal{Y})\, \pi(y) \geq 0 \right\}$$

of all probability mass functions on $\mathcal{Y}$.

For every $i \in \mathcal{X} \setminus \{0, L\}$, we consider a credal set $\mathcal{Q}_i$ on $\mathcal{X}_m := \{\ell, e, u\}$, where—for reasons that should become clear soon—m stands for middle and ℓ, e and u stand for lower, equal and upper, respectively. For the individual probability mass functions $\pi_i \in Q_i$, we will make frequent use of the notational convention that

$$(p_i, r_i, q_i) = \big(\pi_i(\ell), \pi_i(e), \pi_i(u)\big),$$

thereby establishing an intuitive connection with the matrix P that characterises a precise birth-death chain. Similarly, $\mathcal{Q}_0$ and $\mathcal{Q}_L$ are taken to be credal sets on $\mathcal{X}_0 := \{e, u\}$ and $\mathcal{X}_L := \{\ell, e\}$, respectively. For their elements $\pi_0 \in \mathcal{Q}_0$ and $\pi_L \in \mathcal{Q}_L$, we adopt the following notational conventions:

$$(r_0, p_0) = \big(\pi_0(e), \pi_0(u)\big)$$

and

$$(q_L, r_L) = \big(\pi_L(\ell), \pi_L(e)\big).$$

Since $\mathcal{X}_0$ is binary, $\mathcal{Q}_0$ is fully determined by the minimum and maximum value of p_0, as π_0 ranges over the elements of $\mathcal{Q}_0$. We denote this minimum and maximum by $\underline{p}_0$ and $\overline{p}_0$, respectively. Similarly, $\mathcal{Q}_L$ is fully determined by $\underline{q}_L$ and $\overline{q}_L$.

For reasons of mathematical convenience, we adopt the following positivity assumption.

Assumption 1 (Positivity assumption)**.** The local credal sets $\mathcal{Q}_i$, $i \in \mathcal{X}$, consist of strictly positive probability mass functions.

This assumption implies—amongst many other useful properties, such as Theorem 1—that the lower probabilities $\underline{p}_0$ and $\underline{q}_L$ are strictly positive.

We now use the credal sets $\mathcal{Q}_i$ to define corresponding credal sets $\mathcal{M}_i$ on $\mathcal{X}$. For all $i \in \mathcal{X} \setminus \{0, L\}$, a probability mass function $\phi_i \in \Sigma_{\mathcal{X}}$ belongs to $\mathcal{M}_i$ if and only if there is some $\pi_i \in \mathcal{Q}_i$ such that

$$\phi_i(j) = \begin{cases} q_i & \text{if } j = i-1 \\ r_i & \text{if } j = i \\ p_i & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } j \in \mathcal{X}.$$

Similarly, ϕ_0 belongs to $\mathcal{M}_0$ if and only if there is some $\pi_0 \in \mathcal{Q}_0$ such that

$$\phi_0(j) = \begin{cases} r_0 & \text{if } j = 0 \\ p_0 & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } j \in \mathcal{X}$$

and ϕ_L belongs to $\mathcal{M}_L$ if and only if there is some $\pi_L \in \mathcal{Q}_L$ such that

$$\phi_L(j) = \begin{cases} q_L & \text{if } j = L-1 \\ r_L & \text{if } j = L \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } j \in \mathcal{X}.$$

For any real-valued function f on $\mathcal{X}$ and any state i in $\mathcal{X}$, we now consider the corresponding lower and upper expectation of f, defined by

$$\underline{E}(f|i) \coloneqq \min_{\phi_i \in \mathcal{M}_i} E_{\phi_i}(f) = \min_{\phi_i \in \mathcal{M}_i} \left\{ \sum_{j \in \mathcal{X}} \phi_i(j) f(j) \right\}$$

and

$$\overline{E}(f|i) \coloneqq \max_{\phi_i \in \mathcal{M}_i} E_{\phi_i}(f) = \max_{\phi_i \in \mathcal{M}_i} \left\{ \sum_{j \in \mathcal{X}} \phi_i(j) f(j) \right\},$$

where $E_{\phi_i}(f) \coloneqq \sum_{j \in \mathcal{X}} \phi_i(j) f(j)$. The resulting lower and upper expectation operators are connected by conjugacy: $\overline{E}(f|i) = -\underline{E}(-f|i)$. For that reason, without loss of generality, we can focus on the lower expectation operators $\underline{E}(\cdot|i)$, $i \in \mathcal{X}$.

An imprecise birth-death chain is now simply a time-homogeneous imprecise Markov chain [5] that has these lower previsions $\underline{E}(\cdot|i)$—or equivalently, the credal sets $\mathcal{M}_i$—as its local transition models. The corresponding global uncertainty models are derived from the conditional lower expectation operators $\underline{E}_{n+1}(\cdot|x_{1:n})$, defined for all $n \in \mathbb{N}$ and $x_{1:n} \in \mathcal{X}^n$ by[3]

$$\underline{E}_{n+1}(\cdot|x_{1:n}) = \underline{E}_{n+1}(\cdot|x_n) \coloneqq \underline{E}(\cdot|x_n), \qquad (2)$$

where the first equality follows from the so-called imprecise Markov condition and the second equality follows from time-homogeneity.

We want to stress here that the imprecise Markov condition that is imposed by Equation (2) is *not* equivalent to an element-wise application of the (precise) Markov condition in Equation (1). We do *not* require $E_{n+1}(\cdot|x_{1:n})$ and $E_{n+1}(\cdot|x_n)$ to be equal; we only require the bounds on these expectations to be equal. Imposing Equation (1) element-wise would be equivalent to considering a set of precise birth-death chains, each of which is required to satisfy the usual precise Markov assumption. Our approach imposes less stringent constraints. Using imprecise-probabilistic terminology: we impose epistemic irrelevance rather than strong independence; more information can be found in Reference [1].

From the local assessments that are provided by Equation (2), we now derive global uncertainty models for our imprecise Markov chain. For any $i \in \mathcal{X}$ and $n' \in \mathbb{N}$ such that $n' > n$, the global uncertainty model for the variables $X_{n+1:n'}$, conditional on $X_n = i$, is a lower expectation operator $\underline{E}_{n+1:n'}(\cdot|i)$ that takes real-valued functions on $\mathcal{X}^{n'-n}$ as its argument. It is given by the natural extension [10] of the models that were defined in Equation (2); see Reference [4] for more details and alternative interpretations. For the purposes of this paper, we need global uncertainty models that are even more general. In particular, for every $i \in \mathcal{X}$ and $n \in \mathbb{N}$, we need an uncertainty model for the infinite sequence of variables $X_{n+1:\infty}$, conditional on $X_n = i$, in the form of a lower expectation operator $\underline{E}_{n+1:\infty}(\cdot|i)$, defined for all extended real-valued functions on $\mathcal{X}^{\mathbb{N}}$.

These more general global models can be defined in multiple ways, and typically require some additional technical continuity argument; see Reference [3] for a definition in terms of submartingales, which is the one that we will adopt here. However, for our present purposes, the exact definition is only relevant for Theorem 1, which—due to the page limit constraint—is stated without proof. Therefore, and in order to avoid having to introduce the technical concept of a submartingale, we choose not to provide a definition for the global models $\underline{E}_{n+1:\infty}(\cdot|i)$. All that is important for the developments in this paper is that these global models are time-homogeneous and satisfy—a specific version of—the law of iterated expectation. For every $n \in \mathbb{N}$ and every extended real-valued function g on $\mathcal{X}^{\mathbb{N}}$, it holds that

$$\underline{E}_{n+1:\infty}(g(X_{n+1:\infty})|i) = \underline{E}_{n+2:\infty}(g(X_{n+2:\infty})|i). \quad (3)$$

Furthermore, if we define the—possibly extended—real-valued function f' on $\mathcal{X}$ by

$$f'(i') \coloneqq \underline{E}_{n+2:\infty}(g(i', X_{n+2:\infty})|i') \text{ for all } i' \in \mathcal{X},$$

then, if f' is real-valued, we have that

$$\underline{E}_{n+1:\infty}(g(X_{n+1:\infty})|i) = \underline{E}_{n+1}(f'|i) = \underline{E}(f'|i), \quad (4)$$

where the second equality follows from Equation (2).

4 Return and First Passage Times

Consider a timepoint $n \in \mathbb{N}$ and two—possibly identical—states i and j in $\mathcal{X}$. If the variable X_n has i as its value, then the corresponding first passage time to j—the number of time-steps required to reach j—is a function $\tau_{i \to j}(i, X_{n+1:\infty})$ of the infinite sequence of variables $X_{n+1:\infty}$, defined by the following

[3] In general, an initial model $\underline{E}_1(\cdot)$ is required as well. However, for our present purposes, it is not necessary to specify one.

recursion equation:

$$\begin{aligned}\tau_{i\to j}(i, X_{n+1:\infty}) &\coloneqq \begin{cases} 1 & \text{if } X_{n+1} = j \\ 1 + \tau_{X_{n+1}\to j}(X_{n+1}, X_{n+2:\infty}) & \text{if } X_{n+1} \neq j \end{cases} \\ &= 1 + \mathbb{I}_{j^c}(X_{n+1})\tau_{X_{n+1}\to j}(X_{n+1}, X_{n+2:\infty}) \qquad (5)\end{aligned}$$

where $\mathbb{I}_{j^c}$ is the indicator of $j^c \coloneqq \mathcal{X} \setminus \{j\}$, defined by

$$\mathbb{I}_{j^c}(x) \coloneqq \begin{cases} 0 & \text{if } x = j \\ 1 & \text{if } x \neq j \end{cases} \quad \text{for all } x \in \mathcal{X}.$$

If $i = j$, the corresponding first passage time is referred to as the return time to i. The so-called upward and downward first passage times correspond to the cases $i < j$ and $i > j$, respectively.

Due to Equation (3), we know that the lower expected value

$$\underline{\tau}_{i\to j,n} \coloneqq \underline{E}_{n+1:\infty}(\tau_{i\to j}(i, X_{n+1:\infty})|i)$$

and upper expected value

$$\begin{aligned}\overline{\tau}_{i\to j,n} &\coloneqq \overline{E}_{n+1:\infty}(\tau_{i\to j}(i, X_{n+1:\infty})|i) \\ &\coloneqq -\underline{E}_{n+1:\infty}(-\tau_{i\to j}(i, X_{n+1:\infty})|i)\end{aligned}$$

of the first passage time from i to j do not depend on the specific timepoint $n \in \mathbb{N}$ that is chosen. For that reason, we can simply denote them by $\underline{\tau}_{i\to j}$ and $\overline{\tau}_{i\to j}$, respectively.

Theorem 1. *If Assumption 1 is satisfied, then for all $i, j \in \mathcal{X}$, the lower and upper first passage times $\underline{\tau}_{i\to j}$ and $\overline{\tau}_{i\to j}$ are real-valued and strictly positive.*

By combining Equations (4) and (5) with Theorem 1, we find that

$$\underline{\tau}_{i\to j} = 1 + \underline{E}(\mathbb{I}_{j^c}\underline{\tau}_{\bullet\to j}|i) \qquad (6)$$

and

$$\overline{\tau}_{i\to j} = 1 + \overline{E}(\mathbb{I}_{j^c}\overline{\tau}_{\bullet\to j}|i), \qquad (7)$$

where $\underline{\tau}_{\bullet\to j}$ and $\overline{\tau}_{\bullet\to j}$ are functions on $\mathcal{X}$, defined for all $x \in \mathcal{X}$ by

$$\underline{\tau}_{\bullet\to j}(x) \coloneqq \underline{\tau}_{x\to j} \text{ and } \overline{\tau}_{\bullet\to j}(x) \coloneqq \overline{\tau}_{x\to j}.$$

Taking into account our definition for $\underline{E}(\cdot|i)$, Equation (6) results in the following system of non-linear equalities: for all $j \in \mathcal{X}$, we have that

$$\underline{\tau}_{0\to j} = 1 + \min_{\pi_0 \in \mathcal{Q}_0} \left\{ r_0 \mathbb{I}_{j^c}(0)\underline{\tau}_{0\to j} + p_0 \mathbb{I}_{j^c}(1)\underline{\tau}_{1\to j} \right\}, \quad (8)$$

$$\underline{\tau}_{L\to j} = 1 + \min_{\pi_L \in \mathcal{Q}_L} \left\{ q_L \mathbb{I}_{j^c}(L-1)\underline{\tau}_{L-1\to j} + r_L \mathbb{I}_{j^c}(L)\underline{\tau}_{L\to j} \right\}$$

and, for all $i \in \mathcal{X}/\{0, L\}$, that

$$\underline{\tau}_{i\to j} = 1 + \min_{\pi_i \in \mathcal{Q}_i} \left\{ q_i \mathbb{I}_{j^c}(i-1)\underline{\tau}_{i-1\to j} + r_i \mathbb{I}_{j^c}(i)\underline{\tau}_{i\to j} + p_i \mathbb{I}_{j^c}(i+1)\underline{\tau}_{i+1\to j} \right\}. \quad (9)$$

A similar system of non-linear equalities can be derived from Equation (7) as well. In the remainder of this paper, we will solve these non-linear systems, leading to simple expressions that can be used to compute $\underline{\tau}_{i\to j}$ and $\overline{\tau}_{i\to j}$, for any $i, j \in \mathcal{X}$.

5 Lower and Upper Expected Upward First Passage Times

We start by computing lower expected values of upward first passage times, that is, for any $i, j \in \mathcal{X}$ such that $i < j$, we will compute $\underline{\tau}_{i\to j}$. We initially focus on calculating $\underline{\tau}_{i\to i+1}$, for $i \in \mathcal{X} \setminus \{L\}$, and then show that any lower expected upward first passage time can be obtained as a sum of such terms. Similar results are obtained for upper expected upward first passage times.

Finding $\underline{\tau}_{0\to 1}$ is easy. It follows from Equation (8), with $j = 1$, that

$$\begin{aligned}\underline{\tau}_{0\to 1} &= 1 + \min_{\pi_0 \in \mathcal{Q}_0} \{r_0 \underline{\tau}_{0\to 1}\} \\ &= 1 + \min_{\pi_0 \in \mathcal{Q}_0} \{(1 - p_0)\underline{\tau}_{0\to 1}\} \\ &= 1 + \underline{\tau}_{0\to 1} - \max_{\pi_0 \in \mathcal{Q}_0} \{p_0 \underline{\tau}_{0\to 1}\} \\ &= 1 + \underline{\tau}_{0\to 1} - \overline{p}_0 \underline{\tau}_{0\to 1}, \qquad (10)\end{aligned}$$

where the second equality holds because π_0 is a probability mass function and the last equality holds because we know from Theorem 1 that $\underline{\tau}_{0\to 1}$ is real-valued and therefore finite. Since Theorem 1 also tells us that $\underline{\tau}_{0\to 1}$ is strictly positive, we infer from Equation (10) that

$$\underline{\tau}_{0\to 1} = \frac{1}{\overline{p}_0}. \qquad (11)$$

Finding $\underline{\tau}_{0\to j}$, for $j \in \{2, \ldots, L\}$, is more involved. We start by establishing a connection with $\underline{\tau}_{1\to j}$. By applying Equation (8), we find that

$$\begin{aligned}\underline{\tau}_{0\to j} &= 1 + \min_{\pi_0 \in \mathcal{Q}_0} \{r_0 \underline{\tau}_{0\to j} + p_0 \underline{\tau}_{1\to j}\} \\ &= 1 + \min_{\pi_0 \in \mathcal{Q}_0} \{(1 - p_0)\underline{\tau}_{0\to j} + p_0 \underline{\tau}_{1\to j}\} \\ &= 1 + \underline{\tau}_{0\to j} + \min_{\pi_0 \in \mathcal{Q}_0} \{p_0(\underline{\tau}_{1\to j} - \underline{\tau}_{0\to j})\},\end{aligned}$$

which implies, due to Theorem 1, that

$$\min_{\pi_0 \in \mathcal{Q}_0} \{p_0(\underline{\tau}_{1\to j} - \underline{\tau}_{0\to j})\} = -1. \tag{12}$$

Since the minimum in Equation (12) is negative and p_0 is a probability and therefore non-negative, it must be that $\underline{\tau}_{1\to j} - \underline{\tau}_{0\to j} < 0$. Therefore, Equation (12) is minimised for $p_0 = \overline{p}_0$ and we find that

$$\underline{\tau}_{0\to j} = \frac{1}{\overline{p}_0} + \underline{\tau}_{1\to j}. \tag{13}$$

By combining Equations (11) and (13), we see that

$$\underline{\tau}_{0\to j} = \underline{\tau}_{0\to 1} + \underline{\tau}_{1\to j} \text{ for all } j \in \{2, \ldots, L\}. \tag{14}$$

Since we already know $\underline{\tau}_{0\to 1}$—see Equation (11)—we are now left to find $\underline{\tau}_{1\to j}$.

We first consider the case $j = 2$. It follows from Equation (9), with $i = 1$ and $j = 2$, that

$$\begin{aligned}\underline{\tau}_{1\to 2} &= 1 + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to 2} + r_1 \underline{\tau}_{1\to 2}\} \\ &= 1 + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to 2} + (1 - q_1 - p_1)\underline{\tau}_{1\to 2}\} \\ &= 1 + \underline{\tau}_{1\to 2} + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1(\underline{\tau}_{0\to 2} - \underline{\tau}_{1\to 2}) - p_1 \underline{\tau}_{1\to 2}\},\end{aligned}$$

which implies, due to Theorem 1, that

$$\min_{\pi_1 \in \mathcal{Q}_1} \{q_1(\underline{\tau}_{0\to 2} - \underline{\tau}_{1\to 2}) - p_1 \underline{\tau}_{1\to 2}\} = -1.$$

By applying Equation (14), for $j = 2$, we find that

$$\min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to 1} - p_1 \underline{\tau}_{1\to 2}\} = -1. \tag{15}$$

Since we already know $\underline{\tau}_{0\to 1}$, it follows from Assumption 1 and the following lemma that $\underline{\tau}_{1\to 2}$ is the unique solution to Equation (15).

Lemma 2. *Consider a credal set $\mathcal{Q}$ on $\mathcal{X}_m$ that consists of strictly positive probability mass functions and let c be a real constant. Then*

$$\min_{\pi \in \mathcal{Q}} \{qc - p\mu\}$$

is a strictly decreasing function of μ.

This unique solution $\underline{\tau}_{1\to 2}$ is furthermore easy to compute. It follows from Lemma 2 that a simple bisection method suffices.

Next, we consider the case $j \in \{3, \ldots, L\}$. By applying Equation (9), for such a j and with $i = 1$, we find that

$$\begin{aligned}\underline{\tau}_{1\to j} &= 1 + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to j} + r_1 \underline{\tau}_{1\to j} + p_1 \underline{\tau}_{2\to j}\} \\ &= 1 + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to j} + (1 - q_1 - p_1)\underline{\tau}_{1\to j} + p_1 \underline{\tau}_{2\to j}\} \\ &= 1 + \underline{\tau}_{1\to j} + \min_{\pi_1 \in \mathcal{Q}_1} \{q_1(\underline{\tau}_{0\to j} - \underline{\tau}_{1\to j}) \\ &\qquad + p_1(\underline{\tau}_{2\to j} - \underline{\tau}_{1\to j})\},\end{aligned}$$

which implies, due to Theorem 1, that

$$\min_{\pi_1 \in \mathcal{Q}_1} \{q_1(\underline{\tau}_{0\to j} - \underline{\tau}_{1\to j}) + p_1(\underline{\tau}_{2\to j} - \underline{\tau}_{1\to j})\} = -1.$$

In combination with Equation (14), this results in

$$\min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to 1} + p_1(\underline{\tau}_{2\to j} - \underline{\tau}_{1\to j})\} = -1. \tag{16}$$

Since we know from Assumption 1 and Lemma 2 that the equation

$$\min_{\pi_1 \in \mathcal{Q}_1} \{q_1 \underline{\tau}_{0\to 1} + p_1 \mu\} = -1$$

has a unique solution μ, it follows directly from Equations (15) and (16) that

$$\underline{\tau}_{1\to j} = \underline{\tau}_{1\to 2} + \underline{\tau}_{2\to j} \text{ for all } j \in \{3, \ldots, L\}. \tag{17}$$

At this point, we already know how to compute $\underline{\tau}_{0\to 1}$ and $\underline{\tau}_{1\to 2}$ and we have also established the following additivity property:

$$\underline{\tau}_{i\to j} = \underline{\tau}_{i\to i+1} + \underline{\tau}_{i+1\to j}$$

for all $i \in \{0, 1\}$ and $j \in \{i+2, \ldots, L\}$. Continuing in a similar way, we now derive an expression for computing $\underline{\tau}_{2\to 3}$ and prove that the above additivity property holds for $i = 2$ as well. By applying Equation (9), for $i = 2$ and $j = 3$, we find that

$$\begin{aligned}\underline{\tau}_{2\to 3} &= 1 + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2 \underline{\tau}_{1\to 3} + r_2 \underline{\tau}_{2\to 3}\} \\ &= 1 + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2 \underline{\tau}_{1\to 3} + (1 - q_2 - p_2)\underline{\tau}_{2\to 3}\} \\ &= 1 + \underline{\tau}_{2\to 3} + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2(\underline{\tau}_{1\to 3} - \underline{\tau}_{2\to 3}) - p_2 \underline{\tau}_{2\to 3}\},\end{aligned}$$

which implies, due to Theorem 1, that

$$\min_{\pi_2 \in \mathcal{Q}_2} \{q_2(\underline{\tau}_{1\to 3} - \underline{\tau}_{2\to 3}) - p_2 \underline{\tau}_{2\to 3}\} = -1.$$

By applying Equation (17), for $j = 3$, we find that

$$\min_{\pi_2 \in \mathcal{Q}_2} \{q_2 \underline{\tau}_{1\to 2} - p_2 \underline{\tau}_{2\to 3}\} = -1. \tag{18}$$

Since we have already computed $\underline{\tau}_{1\to 2}$, it follows from Assumption 1 and Lemma 2 that $\underline{\tau}_{2\to 3}$ is the unique solution to Equation (18) and that this unique solution can furthermore easily be computed by means of a bisection method.

Next, by applying Equation (9), for $i = 2$ and j in $\{4, \ldots, L\}$, we find that

$$\begin{aligned}\underline{\tau}_{2\to j} &= 1 + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2 \underline{\tau}_{1\to j} + r_2 \underline{\tau}_{2\to j} + p_2 \underline{\tau}_{3\to j}\} \\ &= 1 + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2 \underline{\tau}_{1\to j} + (1 - q_2 - p_2)\underline{\tau}_{2\to j} + p_2 \underline{\tau}_{3\to j}\} \\ &= 1 + \underline{\tau}_{2\to j} + \min_{\pi_2 \in \mathcal{Q}_2} \{q_2(\underline{\tau}_{1\to j} - \underline{\tau}_{2\to j}) \\ &\qquad + p_2(\underline{\tau}_{3\to j} - \underline{\tau}_{2\to j})\},\end{aligned}$$

which implies, due to Theorem 1, that

$$\min_{\pi_2 \in \mathcal{Q}_2} \{q_2(\underline{\tau}_{1\to j} - \underline{\tau}_{2\to j}) + p_2(\underline{\tau}_{3\to j} - \underline{\tau}_{2\to j})\} = -1.$$

In combination with Equation (17), this results in

$$\min_{\pi_2 \in \mathcal{Q}_2} \{q_2\underline{\tau}_{1\to 2} + p_2(\underline{\tau}_{3\to j} - \underline{\tau}_{2\to j})\} = -1. \tag{19}$$

It now follows from Equations (18) and (19), Assumption 1 and Lemma 2, that

$$\underline{\tau}_{2\to j} = \underline{\tau}_{2\to 3} + \underline{\tau}_{3\to j} \text{ for all } j \in \{4, \ldots, L\}.$$

At this point, it should be clear that, by continuing in this way, we obtain the following two results.

Proposition 3. *For all $i \in \mathcal{X} \setminus \{0, L\}$, we have that*

$$\min_{\pi_i \in \mathcal{Q}_i} \{q_i\underline{\tau}_{i-1\to i} - p_i\underline{\tau}_{i\to i+1}\} = -1. \tag{20}$$

Proposition 4. *For all $i, j \in \mathcal{X}$ such that $i + 1 < j$, we have that*

$$\underline{\tau}_{i\to j} = \underline{\tau}_{i\to i+1} + \underline{\tau}_{i+1\to j}.$$

For any $i \in \mathcal{X} \setminus \{L\}$, the value of $\underline{\tau}_{i\to i+1}$ can therefore be computed recursively. For $i = 0$, we simply apply Equation (11). For any other $i \in \mathcal{X} \setminus \{0, L\}$, it follows from Assumption 1, Lemma 2 and Proposition 3 that $\underline{\tau}_{i\to i+1}$ is the unique solution to Equation (20), which can be obtained by means of a bisection method. In this equation, the value of $\underline{\tau}_{i-1\to i}$ has already been computed earlier on in this recursive procedure.

The following additivity result is a direct consequence of Proposition 4.

Corollary 5. *For any $i, j \in \mathcal{X}$ such that $i < j$, we have that*

$$\underline{\tau}_{i\to j} = \sum_{k=i}^{j-1} \underline{\tau}_{k\to k+1}.$$

It implies that the recursive techniques that we developed in this section can be used to compute any lower expected upward first passage time.

Similar results can be proved for upper expected values of upward first passage times. We only provide the final expressions; the derivation is completely analogous. In this case, the starting point is that

$$\overline{\tau}_{0\to 1} = \frac{1}{\underline{p}_0} \tag{21}$$

For any $i \in \mathcal{X} \setminus \{0, L\}$, the value of $\overline{\tau}_{i\to i+1}$ can then be computed recursively, due to Assumption 1 and the following two results.

Proposition 6. *For all $i \in \mathcal{X} \setminus \{0, L\}$, we have that*

$$\max_{\pi_i \in \mathcal{Q}_i} \{q_i\overline{\tau}_{i-1\to i} - p_i\overline{\tau}_{i\to i+1}\} = -1.$$

Corollary 7. *Consider a credal set $\mathcal{Q}$ on $\mathcal{X}_m$ that consists of strictly positive probability mass functions and let c be a real constant. Then*

$$\max_{\pi \in \mathcal{Q}} \{qc - p\mu\}$$

is a strictly decreasing function of μ.

Due to the next result, this recursive technique allows us to compute arbitrary upper expected upward first passage times.

Proposition 8. *For any $i, j \in \mathcal{X}$ such that $i < j$, we have that*

$$\overline{\tau}_{i\to j} = \sum_{k=i}^{j-1} \overline{\tau}_{k\to k+1}$$

6 Lower and Upper Expected Downward First Passage Times

Lower and upper expected values of downward first passage times can be computed in more or less the same way. The main difference is that the recursive expressions now start from the other side, that is, from $i = L$. We find that

$$\underline{\tau}_{L\to L-1} = \frac{1}{\overline{q}_L} \tag{22}$$

and

$$\overline{\tau}_{L\to L-1} = \frac{1}{\underline{q}_L} \tag{23}$$

For any $i \in \mathcal{X} \setminus \{0, L\}$, due to Assumption 1, the values of $\underline{\tau}_{i\to i-1}$ and $\overline{\tau}_{i\to i-1}$ can now be computed recursively, using the following two results.

Proposition 9. *For all $i \in \mathcal{X} \setminus \{0, L\}$, we have that*

$$\min_{\pi_i \in \mathcal{Q}_i} \{-q_i\underline{\tau}_{i\to i-1} + p_i\underline{\tau}_{i+1\to i}\} = -1$$

and

$$\max_{\pi_i \in \mathcal{Q}_i} \{-q_i\overline{\tau}_{i\to i-1} + p_i\overline{\tau}_{i+1\to i}\} = -1$$

Corollary 10. *Consider a credal set $\mathcal{Q}$ on $\mathcal{X}_m$ that consists of strictly positive probability mass functions and let c be a real constant. Then*

$$\min_{\pi \in \mathcal{Q}} \{-q\mu + pc\} \text{ and } \max_{\pi \in \mathcal{Q}} \{-q\mu + pc\}$$

are strictly decreasing functions of μ.

Once we have computed $\underline{\tau}_{i\to i-1}$ and $\overline{\tau}_{i\to i-1}$ for all $i \in \mathcal{X} \setminus \{L\}$, the following result enables us to easily obtain all other lower and upper expected downward first passage times.

Proposition 11. *For any $i, j \in \mathcal{X}$ such that $i > j$, we have that*

$$\underline{\tau}_{i\to j} = \sum_{k=j}^{i-1} \underline{\tau}_{k+1\to k} \text{ and } \overline{\tau}_{i\to j} = \sum_{k=j}^{i-1} \overline{\tau}_{k+1\to k}$$

7 Lower and Upper Expected Return Times

Lower and upper expected return times can now be computed very easily. By applying Equations (8)–(9), with j equal to 0, L and i, respectively, we find that

$$\underline{\tau}_{0\to 0} = 1 + \min_{\pi_0 \in \mathcal{Q}_0} \{p_0 \underline{\tau}_{1\to 0}\} = 1 + \underline{p}_0 \underline{\tau}_{1\to 0}, \tag{24}$$

$$\underline{\tau}_{L\to L} = 1 + \min_{\pi_L \in \mathcal{Q}_L} \{q_L \underline{\tau}_{L-1\to L}\} = 1 + \underline{q}_L \underline{\tau}_{L-1\to L} \tag{25}$$

and, for all $i \in \mathcal{X} \setminus \{0, L\}$, that

$$\underline{\tau}_{i\to i} = 1 + \min_{\pi_i \in \mathcal{Q}_i} \{q_i \underline{\tau}_{i-1\to i} + p_i \underline{\tau}_{i+1\to i}\} \tag{26}$$

In these expressions, the lower expected first passage times $\underline{\tau}_{1\to 0}$, $\underline{\tau}_{L-1\to L}$, $\underline{\tau}_{i-1\to i}$ and $\underline{\tau}_{i+1\to i}$ can be computed using the recursive techniques that we developed in the previous two sections. Similarly, for the upper case, we find that

$$\overline{\tau}_{0\to 0} = 1 + \max_{\pi_0 \in \mathcal{Q}_0} \{p_0 \overline{\tau}_{1\to 0}\} = 1 + \overline{p}_0 \overline{\tau}_{1\to 0}, \tag{27}$$

$$\overline{\tau}_{L\to L} = 1 + \max_{\pi_L \in \mathcal{Q}_L} \{q_L \overline{\tau}_{L-1\to L}\} = 1 + \overline{q}_L \overline{\tau}_{L-1\to L} \tag{28}$$

and, for all $i \in \mathcal{X} \setminus \{0, L\}$, that

$$\overline{\tau}_{i\to i} = 1 + \max_{\pi_i \in \mathcal{Q}_i} \{q_i \overline{\tau}_{i-1\to i} + p_i \overline{\tau}_{i+1\to i}\}. \tag{29}$$

Again, the upper expected first passage times $\overline{\tau}_{1\to 0}$, $\overline{\tau}_{L-1\to L}$, $\overline{\tau}_{i-1\to i}$ and $\overline{\tau}_{i+1\to i}$ that appear in these expressions can be computed with the recursive techniques that were introduced above.

8 Linear-Vacuous Mixtures

We now apply our results to the special case where all the local models are linear-vacuous mixtures. In that case, the computation of lower and upper expected first passage and return times becomes even simpler.

We start from given strictly positive probability mass functions $\pi_0^* = (r_0^*, p_0^*) \in \Sigma_{\mathcal{X}_0}$, $\pi_L^* = (q_L^*, r_L^*) \in \Sigma_{\mathcal{X}_L}$ and, for all $i \in \mathcal{X} \setminus \{0, L\}$, $\pi_i^* = (q_i^*, r_i^*, p_i^*) \in \Sigma_{\mathcal{X}_m}$. Furthermore, for all $i \in \mathcal{X}$, we consider some real-valued $\varepsilon_i \in [0, 1)$. We use these parameters to define the following so-called linear-vacuous [10, Section 2.9.2] local credal sets:

$$\mathcal{Q}_0 = \mathcal{Q}_{\pi_0^*}^{\varepsilon_0} := \{(1-\varepsilon_0)\pi_0^* + \varepsilon_0 \pi_0' : \pi_0' \in \Sigma_{\mathcal{X}_0}\},$$

$$\mathcal{Q}_L = \mathcal{Q}_{\pi_L^*}^{\varepsilon_L} := \{(1-\varepsilon_L)\pi_L^* + \varepsilon_L \pi_L' : \pi_L' \in \Sigma_{\mathcal{X}_L}\}$$

and, for all $i \in \mathcal{X} \setminus \{0, L\}$,

$$\mathcal{Q}_i = \mathcal{Q}_{\pi_i^*}^{\varepsilon_i} := \{(1-\varepsilon_i)\pi_i^* + \varepsilon_i \pi_i' : \pi_i' \in \Sigma_{\mathcal{X}_m}\},$$

which can be regarded as neighbourhood models for the probability mass functions π_i^*, $i \in \mathcal{X}$. Furthermore, for all $i \in \mathcal{X} \setminus \{0\}$, we define

$$\underline{q}_i := (1-\varepsilon_i) q_i^* \text{ and } \overline{q}_i := (1-\varepsilon_i) q_i^* + \varepsilon_i$$

and, for all $i \in \mathcal{X} \setminus \{L\}$,

$$\underline{p}_i := (1-\varepsilon_i) p_i^* \text{ and } \overline{p}_i := (1-\varepsilon_i) p_i^* + \varepsilon_i,$$

which are the minimum and maximum values of q_i and p_i, for $\pi_i \in \mathcal{Q}_i$, respectively.

In this special case, Equation (20) can be solved analytically. For all $i \in \mathcal{X} \setminus \{0, L\}$, we find that

$$\begin{aligned}
&\min_{\pi_i \in \mathcal{Q}_i} \{q_i \underline{\tau}_{i-1\to i} - p_i \underline{\tau}_{i\to i+1}\} \\
&= \min_{\pi' \in \Sigma_{\mathcal{X}_m}} \big\{[(1-\varepsilon_i)q_i^* + \varepsilon_i q_i'] \underline{\tau}_{i-1\to i} \\
&\qquad\qquad - [(1-\varepsilon_i)p_i^* + \varepsilon_i p_i'] \underline{\tau}_{i\to i+1}\big\} \\
&= (1-\varepsilon_i)(q_i^* \underline{\tau}_{i-1\to i} - p_i^* \underline{\tau}_{i\to i+1}) \\
&\qquad + \varepsilon_i \min_{\pi_i' \in \Sigma_{\mathcal{X}_m}} \{q_i' \underline{\tau}_{i-1\to i} - p_i' \underline{\tau}_{i\to i+1}\} \\
&= (1-\varepsilon_i)(q_i^* \underline{\tau}_{i-1\to i} - p_i^* \underline{\tau}_{i\to i+1}) - \varepsilon_i \underline{\tau}_{i\to i+1} \\
&= \underline{q}_i \underline{\tau}_{i-1\to i} - \overline{p}_i \underline{\tau}_{i\to i+1},
\end{aligned}$$

where the third equation holds because we know from Theorem 1 that $\underline{\tau}_{i-1\to i}$ and $\underline{\tau}_{i\to i+1}$ are real-valued and positive. Therefore, for all $i \in \mathcal{X} \setminus \{0, L\}$, it follows directly from Equation (20) that

$$\underline{\tau}_{i\to i+1} = \frac{1}{\overline{p}_i} + \frac{\underline{q}_i}{\overline{p}_i} \underline{\tau}_{i-1\to i}.$$

By combining this recursive expression with Equation (11), we can derive explicit expressions. For all $i \in \mathcal{X} \setminus \{L\}$, we find that:

$$\underline{\tau}_{i\to i+1} = \sum_{k=0}^{i} \frac{\prod_{\ell=k+1}^{i} \underline{q}_\ell}{\prod_{m=k}^{i} \overline{p}_m}. \tag{30}$$

In combination with Corollary 5, this equation allows us to easily compute all lower expected upward first passage times for the linear-vacuous case.

Similar results can be obtained for upper expected upward first passage times and for lower and upper expected downward first passage times. For all $i \in \mathcal{X} \setminus \{0, L\}$, we find that

$$\overline{\tau}_{i\to i+1} = \frac{1}{\underline{p}_i} + \frac{\overline{q}_i}{\underline{p}_i}\overline{\tau}_{i-1\to i},$$

$$\underline{\tau}_{i\to i-1} = \frac{1}{\overline{q}_i} + \frac{\underline{p}_i}{\overline{q}_i}\underline{\tau}_{i+1\to i}$$

and

$$\overline{\tau}_{i\to i-1} = \frac{1}{\underline{q}_i} + \frac{\overline{p}_i}{\underline{q}_i}\overline{\tau}_{i+1\to i}.$$

By combining these recursive equations with Equations (21), (22) and (23), respectively, we can obtain explicit expressions. For all $i \in \mathcal{X} \setminus \{L\}$, we find that

$$\overline{\tau}_{i\to i+1} = \sum_{k=0}^{i} \frac{\prod_{\ell=k+1}^{i} \overline{q}_\ell}{\prod_{m=k}^{i} \underline{p}_m}$$

and, for all $i \in \mathcal{X} \setminus \{0\}$, we find that

$$\underline{\tau}_{i\to i-1} = \sum_{k=i}^{L} \frac{\prod_{\ell=i}^{k-1} \underline{p}_\ell}{\prod_{m=i}^{k} \overline{q}_m} \tag{31}$$

and

$$\overline{\tau}_{i\to i-1} = \sum_{k=i}^{L} \frac{\prod_{\ell=i}^{k-1} \overline{p}_\ell}{\prod_{m=i}^{k} \underline{q}_m}.$$

In combination with Proposition 8 and 11, these equations allow us to easily compute all upper expected upward first passage times and all lower and upper expected downward first passage times for the linear-vacuous case.

For the lower and upper return times, we still use Equations (24) and (25) if $i = 0$ and Equations (27) and (28) if $i = L$. If $i \in \mathcal{X} \setminus \{0, L\}$, then, for this linear-vacuous case, Equations (26) and (29) can be simplified. We find that

$$\begin{aligned}
\underline{\tau}_{i\to i} &= 1 + \min_{\pi_i \in \mathcal{Q}_i} \{q_i \underline{\tau}_{i-1\to i} + p_i \underline{\tau}_{i+1\to i}\} \\
&= 1 + \min_{\pi_i' \in \Sigma_{\mathcal{X}_m}} \big\{[(1-\varepsilon_i)q_i^* + \varepsilon_i q_i']\underline{\tau}_{i-1\to i} \\
&\qquad\qquad + [(1-\varepsilon_i)p_i^* + \varepsilon_i p_i']\underline{\tau}_{i+1\to i}\big\} \\
&= 1 + (1-\varepsilon_i)(q_i^* \underline{\tau}_{i-1\to i} + p_i^* \underline{\tau}_{i+1\to i}) \\
&= 1 + \underline{q}_i \underline{\tau}_{i-1\to i} + \underline{p}_i \underline{\tau}_{i+1\to i}.
\end{aligned} \tag{32}$$

and that

$$\begin{aligned}
\overline{\tau}_{i\to i} &= 1 + (1-\varepsilon_i)(q_i^* \overline{\tau}_{i-1\to i} + p_i^* \overline{\tau}_{i+1\to i}) \\
&\qquad + \varepsilon_i \max\{\overline{\tau}_{i-1\to i}, \overline{\tau}_{i+1\to i}\} \\
&= 1 + \max\{\overline{q}_i \overline{\tau}_{i-1\to i} + \underline{p}_i \overline{\tau}_{i+1\to i}, \\
&\qquad\qquad \underline{q}_i \overline{\tau}_{i-1\to i} + \overline{p}_i \overline{\tau}_{i+1\to i}\}.
\end{aligned}$$

9 Numerical Results

We end by computing lower and upper expected first passage and return times for two examples of imprecise birth-death chains. The first is a general example of an imprecise birth-death chain and the second one is an imprecise birth-death chain with linear-vacuous local models. In both examples, we take $\mathcal{Q}_i$ to be identical for all $i \in \mathcal{X} \setminus \{0, L\}$, and simply denote it by $\mathcal{Q}$, which is a credal set on $\mathcal{X}_m$. Some of the lower and upper expected values that we compute have many decimal points; we present them up to the third decimal point.

General Example

Consider an imprecise birth-death chain with state space $\mathcal{X} = \{0, 1, 2, 3, 4\}$, that is, $L = 4$. Let $\mathcal{Q}_0$ be determined by $\underline{p}_0 = 0.15$ and $\overline{p}_0 = 0.4$ and let $\mathcal{Q}_L$ be determined by $\underline{q}_L = 0.2$ and $\overline{q}_L = 0.6$. The credal set $\mathcal{Q}$ is taken to be the convex hull of the following 10 extreme points, which are of the form $\pi = (q, r, p)$.

$$\begin{aligned}
&(0.65, 0.15, 0.2), (0.6, 0.25, 0.15), (0.5, 0.4, 0.1), \\
&(0.43, 0.45, 0.12), (0.33, 0.5, 0.17), (0.27, 0.43, 0.3), \\
&(0.25, 0.35, 0.4), (0.3, 0.25, 0.45), (0.4, 0.17, 0.43), \\
&(0.55, 0.1, 0.35)
\end{aligned}$$

Figure 2 provides a graphical representation of this credal set $\mathcal{Q}$.[4]

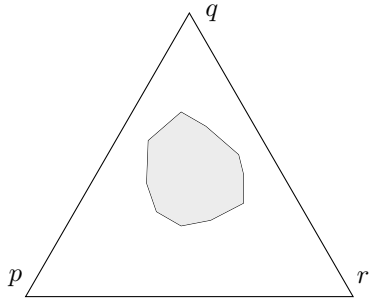


Figure 2: The grey zone depicts the credal set $\mathcal{Q}$ from the birth-death chain in the general example.

For this particular example, we now compute $\underline{\tau}_{0\to 4}$, $\overline{\tau}_{0\to 4}$, $\underline{\tau}_{4\to 0}$ and $\overline{\tau}_{4\to 0}$.

Due to Corollary 5, we know that

$$\underline{\tau}_{0\to 4} = \underline{\tau}_{0\to 1} + \underline{\tau}_{1\to 2} + \underline{\tau}_{2\to 3} + \underline{\tau}_{3\to 4}, \tag{33}$$

[4] We represent $\Sigma_{\mathcal{X}_m}$ by means of a equilateral triangle of height one. The elements $\pi = (q, r, p)$ of $\Sigma_{\mathcal{X}_m}$ correspond to points in this triangle. For every such π, the value of q, r, p is equal to the perpendicular distance from that point to the edge that opposes the corresponding corner.

$\underline{\tau}_{0\to4}$	16.635
$\overline{\tau}_{0\to4}$	1420
$\underline{\tau}_{4\to0}$	8.093
$\overline{\tau}_{4\to0}$	81.32

Table 1: Final results for the general example.

$\underline{\tau}_{0\to1}$	2.5	$\underline{\tau}_{4\to3}$	1.666
$\underline{\tau}_{1\to2}$	3.889	$\underline{\tau}_{3\to2}$	2.051
$\underline{\tau}_{2\to3}$	4.814	$\underline{\tau}_{2\to1}$	2.169
$\underline{\tau}_{3\to4}$	5.432	$\underline{\tau}_{1\to0}$	2.206
$\overline{\tau}_{0\to1}$	6.666	$\overline{\tau}_{4\to3}$	5
$\overline{\tau}_{1\to2}$	43.333	$\overline{\tau}_{3\to2}$	12
$\overline{\tau}_{2\to3}$	226.666	$\overline{\tau}_{2\to1}$	23.2
$\overline{\tau}_{3\to4}$	1143.333	$\overline{\tau}_{1\to0}$	41.12

Table 2: Intermediate results for the general example.

where, using Equation (11),

$$\underline{\tau}_{0\to1} = 1/\overline{p}_0 = 2.5.$$

By plugging this value for $\underline{\tau}_{0\to1}$ in Equation (20), for $i = 1$, we find that

$$\min_{\pi_1\in\mathcal{Q}}\{2.5q_1 - p_1\underline{\tau}_{1\to2}\} = -1$$

As we know from Lemma 2, this equality has a unique solution that can for example be obtained by means of a bisection method. We find that $\underline{\tau}_{1\to2} = 3.889$. Similarly, in a recursive fashion, we find that $\underline{\tau}_{2\to3} = 4.814$ and $\underline{\tau}_{3\to4} = 5.432$. A final application of Equation (33) tells us that $\underline{\tau}_{0\to4} = 16.635$. $\overline{\tau}_{0\to4}$, $\underline{\tau}_{4\to0}$ and $\overline{\tau}_{4\to0}$ can be computed analogously; the results are given in Table 1. Intermediate results can be found in Table 2.

Linear-Vacuous Example

Consider a precise birth-death chain with state space $\mathcal{X} = \{0, 1, 2, 3, 4\}$—$L = 4$—and the following probability matrix:

$$P^* = \begin{pmatrix} 0.55 & 0.45 & 0 & 0 & 0 \\ 0.3 & 0.5 & 0.2 & 0 & 0 \\ 0 & 0.3 & 0.5 & 0.2 & 0 \\ 0 & 0 & 0.3 & 0.5 & 0.2 \\ 0 & 0 & 0 & 0.6 & 0.4, \end{pmatrix}$$

which is completely characterised by the probability mass functions $\pi_0^* = (0.55, 0.45)$, $\pi_L^* = (0.6, 0.4)$ and, for all $i \in \mathcal{X} \setminus \{0, L\}$, $\pi_i^* = \pi^* = (0.3, 0.5, 0.2)$.

We now let $\varepsilon_i = \varepsilon = 0.4$ for all $i \in \mathcal{X}$ and consider the imprecise birth-death chain that has the corresponding linear-vacuous credal sets as its local models.

In this way, we obtain the following lower and upper probabilities:

$$\underline{p}_0 = 0.27,\ \overline{p}_0 = 0.67,\ \underline{q}_L = 0.36,\ \overline{q}_L = 0.76$$

and, for all $i \in \mathcal{X} \setminus \{0, L\}$:

$$\underline{q}_i = 0.18,\ \overline{q}_i = 0.58,\ \underline{p}_i = 0.12,\ \overline{p}_i = 0.52.$$

For all $i \in \mathcal{X} \setminus \{0, L\}$, the credal set $\mathcal{Q}_i$ is equal to $\mathcal{Q}^{\varepsilon}_{\pi^*}$, which is the convex hull of the following three extreme points:

$$(0.58, 0.3, 0.12), (0.18, 0.7, 0.12), (0.18, 0.3, 0.52).$$

Figure 3 provides a graphical representation of this credal set $\mathcal{Q}^{\varepsilon}_{\pi^*}$.

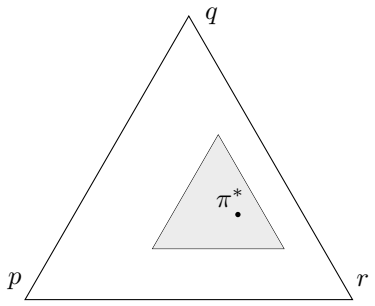


Figure 3: The grey zone depicts the credal set $\mathcal{Q}^{\varepsilon}_{\pi^*}$ from the birth-death chain in the the linear-vacuous example.

The lower and upper expected return times that correspond to this particular example can be found in Table 3. For the sake of this example, we compute $\underline{\tau}_{1\to1}$ explicitly.

We start by applying Equation (32) for $i = 1$, which tells us that

$$\begin{aligned}\underline{\tau}_{1\to1} &= 1 + \underline{q}_1\underline{\tau}_{0\to1} + \underline{p}_1\underline{\tau}_{2\to1} \\ &= 1 + 0.18\underline{\tau}_{0\to1} + 0.12\underline{\tau}_{2\to1}.\end{aligned}$$

Therefore, since we know from Equations (30) and (31) that

$$\underline{\tau}_{0\to1} = \frac{1}{\overline{p}_0} = 1.492$$

and

$$\underline{\tau}_{2\to1} = \frac{1}{\overline{q}_2} + \frac{\underline{p}_2}{\overline{q}_2\overline{q}_3} + \frac{\underline{p}_2\underline{p}_3}{\overline{q}_2\overline{q}_3\overline{q}_4} = 2.154,$$

we find that $\underline{\tau}_{1\to1} = 1.526$.

i	$\underline{\tau}_{i\to i}$	$\overline{\tau}_{i\to i}$
0	1.584	91.41
1	1.526	24.956
2	1.678	17.845
3	1.656	79.71
4	2.037	503.724

Table 3: Lower and upper expected return times for the birth-death chain in the linear-vacuous mixture example.

10 Summary and Future Work

We have presented a simple method for computing lower and upper expected—upward and downward—first passage times and return times in imprecise birth-death chains, have presented numerical results, and have discussed a special case for which our method simplifies even more.

In future research, we plan to try and apply similar methods to (a) other simple types of imprecise Markov chains—different from birth-death chains—such as, for example, the Bonus-Malus systems that are described in Reference [6] and (b) continuous—rather than discrete—time models.

Acknowledgements

The research of Stavros Lopatatzidis and Gert de Cooman was funded through project number 3G012512 of the Research Foundation Flanders (FWO). Jasper De Bock is a PhD Fellow of the FWO and wishes to acknowledge its financial support. The authors would also like to thank three anonymous referees for their many helpful suggestions.

References

[1] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5(02):165–181, 2000.

[2] Richard J. Crossman, Pauline Coolen-Schrijner, and Frank P.A. Coolen. Time-homogeneous birth-death processes with probability intervals and absorbing state. *Journal of Statistical Theory and Practice*, 3(1):103–118, 2009.

[3] Gert de Cooman, Jasper De Bock, and Stavros Lopatatzidis. A pointwise ergodic theorem for imprecise Markov chains. *Accepted for publication in ISIPTA '15 – Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*. SIPTA, 2015.

[4] Gert de Cooman and Filip Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008.

[5] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, 2009.

[6] Michel Denuit, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, 2007.

[7] Igor O. Kozine and Lev V. Utkin. Interval-valued finite Markov chains. *Reliable Computing*, 8(2):97–113, 2002.

[8] Ushio Sumita and Yasushi Masuda. On first passage time structure of random walks. *Stochastic processes and their applications*, 20(1):133–147, 1985.

[9] D. Škulj. Regular finite Markov chains with interval probabilities. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *ISIPTA '07 – Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 405–413. SIPTA, 2007.

[10] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[11] P. Whittle. *Probability via Expectation*. Springer, New York, fourth edition, 2000.

A Prior Near-Ignorance Gaussian Process Model for Nonparametric Regression

Francesca Mangili
IDSIA, USI-SUPSI, Lugano, Switzerland
francesca@idsia.ch

Abstract

A Gaussian Process (GP) defines a distribution over functions and thus it is a natural prior distribution for learning real-valued functions from a set of noisy data. GPs offer a great modeling flexibility and have found widespread application in many regression problems. A GP is fully defined by a mean function that represents our prior belief about the shape of the regression function and a covariance function, relating the function values at different covariates. In the absence of prior information, one typically assumes a GP with zero mean function. Therefore, a priori, it is assumed that the regression function is constantly equal to zero. The aim of this paper is to model a situation of prior near-ignorance about the GP mean function. For this we consider the set of all GPs with fixed covariance function and constant mean function free to vary from $-\infty$ to $+\infty$. We apply the model with constant mean function to hypothesis testing; in particular we test the equality of two regression functions and show that the use of a prior near-ignorance model allows the test to automatically detect when a reliable decision cannot be made based on the available data. Finally, we propose a generalization of this model that allows considering other sets of prior mean functions.

Keywords. Gaussian Process, prior near-ignorance, nonparametric regression, hypothesis testing, Bayesian nonparametrics.

1 Introduction

Gaussian processes (GPs) extend multivariate Gaussian distributions to infinite dimensionality, thus defining a distribution over functions that can be used as prior distribution for inferences about an unknown function $f(x)$. GPs have found widespread use in different application domains such as classification, regression etc. [9, 8, 6, 12, 11, 4]. The reason of such success can be attributed to the great modelling flexibility of GPs, which are often used in situations where little is known about $f(x)$. However, GPs are not completely free-form, since a GP is completely specified by its mean function and covariance function. The covariance function describes the relation between observations from the same process. A multitude of possible families exists for the covariance function, including squared exponential, polynomial, periodic, etc. (see [12]), among which the squared exponential family is by far the most popular. On the other side, the mean function represents our prior belief about the form of the regression function. In the absence of prior knowledge, which is typically the case, the mean function is assumed to be zero everywhere and, to comply with this assumption, data are transformed to have zero mean. However, this seems quite a poor representation of the condition of prior ignorance about $f(x)$. In this work we improve this representation by considering a set of GP priors with mean functions free to vary in the set of all constant functions. As the expectation of $f(x^*)$ at the covariate x^* w.r.t. the prior GPs can vary in $[-\infty, +\infty]$, this set of priors is a model of prior ignorance about $f(x)$. Prior ignorance and learning from data are usually conflicting properties [13, Sec. 7.4],[10, 14]. However, in [3, 2] it is shown that, for Gaussian distributions, if we let the variance to depend on the mean, prior near-ignorance and learning from data can be guaranteed at the same time. In this work, we apply this idea to GPs. In order for the GP model to be able to learn from data, we add to the covariance function a constant term increasing with the prior mean function.

We will use this set of priors to test the difference between two regression function given two samples of noisy observations. A nonparametric Bayesian test for the equality of regression functions based on GPs is described in [1]. In that work it is assumed that the covariates of the two samples cover the same range of values, and the comparison between the regression functions is limited to that range of values, assuming that, having no data outside of it, nothing can be stated about the difference or equality of the two

functions. Using the Imprecise GP (IGP) it is possible to perform the equality test without worrying about the distribution of the covariates, as the imprecise approach is able to identify those instances where the decision is prior dependent and thus it automatically detects when a reliable decision cannot be made.

Finally, we introduce a IGP model that generalizes the previous one by considering all GPs with mean function $Mh(x)$ where $M > 0$ and $h(x)$ belongs to a set of functions $\mathcal{H}$. We derive the conditions that $\mathcal{H}$ has to satisfy to make prior near-ignorance and leaning hold for the IGP model. From this model we can derive the IGP with constant mean as well as well as other models considering different/larger sets of prior mean functions.

2 Gaussian Process

Consider the regression model

$$y = f(x) + v, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $v \sim \mathcal{N}(0, \sigma_n^2)$ is a white noise, and assume that we observe the data (x_i, y_i) for $i = 1, \dots, n$. Our goal is to employ these observations to make inferences about the unknown function $f(x)$. Following the Bayesian estimation approach, we place a prior distribution on $f(x)$, and employ the observations to compute its posterior distribution; finally we use this posterior to make inferences about $f(x)$. Since $f(x)$ is a function, the Gaussian process is a natural prior distribution for it [6, 12]. Formally,

Definition 1. *Let $\mu : \mathbb{R} \to \mathbb{R}$ and $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a positive definite symmetric function.*[1] *A function $f(x)$ with $x \in \mathbb{R}$ is said to be distributed according to a Gaussian process with mean function μ and covariance kernel k if for any finite set of covariates $x_1^*, \dots, x_m^*$, the vector $[f(x_1^*), \dots, f(x_m^*)]^T$ has a multivariate m-dimensional Gaussian distribution with mean $[\mu(x_1^*), \dots, \mu(x_m^*)]^T$ and covariance matrix with (i, j)-th entry $k(x_i^*, x_j^*)$, $i, j = 1, \dots, m$.*

In the following, $GP(\mu(x), k_{\boldsymbol{\theta}}(x, x'))$ will denote a GP with mean function $\mu(x)$ and covariance function $k_{\boldsymbol{\theta}}(x, x') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$. The subscript $\boldsymbol{\theta}$ has been introduced to highlight that the covariance function usually depends on a vector of hyperparameters $\boldsymbol{\theta}$ [12]. If $f(x) \sim GP(\mu(x), k_{\boldsymbol{\theta}}(x, x'))$, then, for any fixed m points $\mathbf{x}^* = [x_1^*, \dots, x_m^*]^T$, the vector $\mathbf{f}^* = [f(x_1^*), \dots, f(x_m^*)]^T$ is Gaussian distributed:

$$p(\mathbf{f}^*|\mathbf{x}^*, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, K^{**}), \tag{2}$$

[1] A symmetric function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is said to be positive definite if for any $\mathbf{x} = [x_1^*, \dots, x_m^*]^T$ with $x_i^* \in \mathbb{R}$, the $m \times m$ matrix $[k(x_i^*, x_j^*)]_{ij}$ is positive definite.

with mean $\boldsymbol{\mu}^* = \mu(\mathbf{x}^*)$ and covariance matrix $K^{**} = [k_{\boldsymbol{\theta}}(x_i^*, x_j^*)]_{ij}$ for each $i, j = 1, \dots, m$. Consider a set of n inputs $\mathbf{x} = [x_1, \dots, x_n]^T$ and a vector of noisy output data $\mathbf{y} = [y_1, \dots, y_n]^T$. Based on the training data (x_i, y_i) for $i = 1, \dots, n$, and given a test input $\mathbf{x}^*$, we wish to find the posterior distribution of $\mathbf{f}^* = [f(x_1^*), \dots, f(x_m^*)]^T$. From (1) and the properties of the Gaussian distribution, it follows that [12, Sec. 2.2]:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}^* \end{bmatrix}, \begin{bmatrix} K + \sigma_n^2 \mathbf{I} & K^{*T} \\ K^* & K^{**} \end{bmatrix} \right), \tag{3}$$

where $\boldsymbol{\mu} = \mu(\mathbf{x})$, $K = [k_{\boldsymbol{\theta}}(x_i, x_j)]_{ij}$, $i, j = 1 \dots, n$ and $K^* = [k_{\boldsymbol{\theta}}(x_i^*, x_j)]_{ij}$, $i = 1 \dots, m$, $j = 1 \dots, n$. When σ_n^2 is not known, it can also be considered a hyperparameter. Hence, we introduce the extended vector $\boldsymbol{\theta}_n = [\boldsymbol{\theta}, \sigma_n^2]$ of all model hyperparameters, including the noise variance. The posterior distribution of $\mathbf{f}^*$ is then

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n) = \mathcal{N}(\mathbf{f}^*; \hat{\boldsymbol{\mu}}^*, \hat{K}^{**}), \tag{4}$$

with posterior mean and covariance given by:

$$\hat{\boldsymbol{\mu}}^* = \boldsymbol{\mu}^* + K^*(K + \sigma_n^2 \mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{5}$$

$$\hat{K}^{**} = K^{**} - K^*(K + \sigma_n^2 \mathbf{I})^{-1} K^{*T}. \tag{6}$$

Once we have computed $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n)$ we can make any inference about $\mathbf{f}^*$.

GP models use a kernel to define the covariance between any two function values: $Cov(f(x), f(x')) = k_{\boldsymbol{\theta}}(x, x')$. A popular choice is the squared exponential kernel:

$$k_{\boldsymbol{\theta}}(x, x') = \sigma_k^2 \exp\left[-\frac{1}{2} \frac{(x - x')^2}{\ell^2} \right], \tag{7}$$

with hyperparameters $\boldsymbol{\theta} = (\sigma_k, \ell) > 0$. This kernel assumes that the correlation between two function values decreases with the distance of their covariates. Observations whose covariates have a distance much larger than the lengthscale ℓ are almost uncorrelated. A multitude of other possible families of covariance functions exists (polynomial, periodic, etc.), and more can be obtained by kernel composition, as positive definite kernels (i.e. those which define valid covariance functions) are closed under addition and multiplication. Once we have selected a kernel or a particular kernel composition, we must determine the values of the hyperparameters $\boldsymbol{\theta}_n$. The proper Bayesian procedure is to choose a prior for $\boldsymbol{\theta}_n$ and then determine the posterior distribution of the quantities of interest. For instance, inferences on $\mathbf{f}^*$ can be carried out by marginalizing out $\boldsymbol{\theta}_n$:

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = \int p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n|\mathbf{x}^*, \mathbf{x}, \mathbf{y})) d\boldsymbol{\theta}_n.$$

No closed form solution exists for $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y})$ or for the posterior of the hyperparameters and, therefore,

inferences must be computed numerically by Markov Chain Monte Carlo methods (MCMC). The convergence of MCMC methods can be quite slow when the dimension of $\boldsymbol{\theta}_n$ is high and, therefore, when we are not interested in the posterior distribution of $\boldsymbol{\theta}_n$, we can approximate the marginal of $\mathbf{f}^*$ by plugging the maximum a posteriori (MAP) estimate for $\boldsymbol{\theta}_n$ into (4). In other words, we maximize w.r.t. $\boldsymbol{\theta}_n$ the joint marginal probability of $\mathbf{y}$ and $\boldsymbol{\theta}_n$, whose logarithm can be computed analytically [12, Ch.2]:

$$\begin{aligned}\log p(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) = &-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}^*)^T(K_n)^{-1}(\mathbf{y}-\boldsymbol{\mu}^*)+\\ &-\frac{1}{2}\log|K_n| - \frac{n}{2}\log 2\pi + \log p(\boldsymbol{\theta}_n)\end{aligned} \tag{8}$$

where $K_n = K + \sigma_n^2 \boldsymbol{I}$.

3 Imprecise Gaussian Process with Constant Mean Function

In this section we relax the assumption of zero-mean function and consider a set of GPs with constant mean function varying from $-\infty$ to $+\infty$.

Definition 2. *Given a covariance kernel $k_{\boldsymbol{\theta}}(x, x')$ and a constant $c > 0$, we define the constant mean Imprecise Gaussian Process (c-IGP) as the set of GPs:*

$$\mathcal{G}_c = \left\{ GP\left(Mh, k_{\boldsymbol{\theta}}(x,x') + \tfrac{1+M}{c}\right) :\ h = \pm 1,\ M \geq 0 \right\}.$$

As discussed below, the constant parameter c determines the degree of posterior imprecision.
The c-IGP includes all GPs with constant mean function and covariance function made of two components: a first one, $k_{\boldsymbol{\theta}}(x, x')$, hereafter referred to as *base kernel*, which is chosen according to the specific application and is identical for all GPs in $\mathcal{G}_c$, and a constant component $(M+1)/c$ proportional to $M+1$. The constant component allows the model to learn from data, as it forces the covariance to increase with the prior mean. In [2, pag. 22], it is shown for the one-parameter exponential family that if the product $n_0|y_0|$ of the number of pseudo-observations n_0 (which represent the strength of the prior and for a Gaussian prior is given by the inverse of its variance) and the absolute value of the pseudo-observation y_0 (which represent our prior opinion about the parameter value and for a Gaussian prior is given by its mean) is bounded, then learning from data is guaranteed. For the c-IGP model, this holds for each individual x^* because the prior about $f(x^*)$ is a Gaussian distribution with mean Mh (corresponding to y_0) and variance $k_{\boldsymbol{\theta}}(x^*, x^*) + \frac{M+1}{c}$ (corresponding to $1/n_0$), and thus:

$$n_0|y_0| = \frac{M|h|}{k_{\boldsymbol{\theta}}(x^*, x^*) + \frac{M+1}{c}} \leq c.$$

Notice however that this guarantees only learning from data with covariate equal to x^*.

Proposition 1. *The c-IGP is a model of prior ignorance about the expectation of $f(x^*)$ in the sense that for any covariate x^* it holds*

$$\inf_{M,h} E[f(x^*)] = -\infty, \quad \sup_{M,h} E[f(x^*)] = +\infty.$$

The proof of this and the following propositions and theorems can be found in the Appendix.

A posteriori we have the following result.

Theorem 1. *Let $\mathbf{x}$ be a vector of inputs and $\mathbf{y}$ a set of noisy observations of $f(\mathbf{x})$ with $f(x) \sim GP(Mh, k_{\boldsymbol{\theta}} + \frac{M+1}{c})$, and let $\mathbf{k}_x = [k_{\boldsymbol{\theta}}(x, x_1), \ldots, k_{\boldsymbol{\theta}}(x, x_n)]^T$. The posterior distribution of f is a GP with mean function*

$$\hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1}(\mathbf{y} - \hat{y}\mathbb{1}_n) + \hat{y} \tag{9}$$

with $\hat{y} = \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c + (M+1)S_k}$, and covariance function

$$\begin{aligned}\hat{k}(x, x') = &k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \\ &\frac{(M+1)(1 - \mathbf{k}_x^T\mathbf{s}_k)(1 - \mathbf{k}_{x'}^T\mathbf{s}_k)}{c + (M+1)S_k},\end{aligned} \tag{10}$$

where $\mathbf{s}_k = K_n^{-1}\mathbb{1}_n$, $S_k = \mathbb{1}_n^T K_n^{-1}\mathbb{1}_n$, and $\mathbb{1}_n$ is a n-dimensional vector of ones.

The posterior mean function is the same that would have been obtained from the prior $GP(\hat{y}, k_\theta(x, x'))$. We can interpret $\hat{y}$ as an adjusted mean obtained by combining the prior mean Mh and a weighted average of the observations $\mathbf{y}$. This is due to the constant term in the covariance function which introduces a correlation between all function values (does not matter how distant their covariates are). As this constant term goes to infinity, that is, as $c \to 0$, the adjusted mean becomes $\hat{y} \to \frac{\mathbf{s}_k^T}{S_k}\mathbf{y}$ which is independent of h and M and the posterior distribution is a GP with mean and covariance functions

$$\begin{aligned}\lim_{c\to 0}\hat{\mu}(x) &= \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T s_k)\frac{s_k^T}{S_k}\mathbf{y},\\ \lim_{c\to 0}\hat{k}(x, x') &= k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \\ &\qquad \frac{(1 - \mathbf{k}_x^T\mathbf{s}_k)(1 - \mathbf{k}_{x'}^T\mathbf{s}_k)}{S_k}.\end{aligned}$$

Notice that, for $c \to 0$ we have a precise model, as IGP posterior inferences are not influenced by the mean function of the prior and converge to a single GP. This prior can, thus, be interpreted as a partially uninformative prior (inferences still depend on the base kernel).

As discussed in Section 2, MAP estimates of the hyperparameters $\boldsymbol{\theta}_n$ are used in the model. However

the different priors in the IGP set produce different estimates, whereas the IGP model here proposed requires the same set of hyperparameters for all priors. The issue is, then, which of the IGP priors should be used to estimate $\boldsymbol{\theta}_n$. Notice that posterior inferences obtained for any c always encompass those obtained for $c \to 0$ (see Theorem 3 below). Hence, we use the MAP estimate of $\boldsymbol{\theta}_n$ given by this prior.

Theorem 2. *MAP estimates of the hyperparameters of the GP$(Mh, k_{\boldsymbol{\theta}} + \frac{M+1}{c})$ with $c \to 0$ are obtained by maximizing $L(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) + \log p(\boldsymbol{\theta}_n)$ where*

$$L(\mathbf{y}, \boldsymbol{\theta}_n|\mathbf{x}) = \frac{1}{2}\left(\mathbf{y}^T K_n^{-1}\mathbf{y} - \frac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k} - \log S_k|K_n|\right) \tag{11}$$

is, up to an additive constant, the logarithm of the joint marginal likelihood of $\mathbf{y}, \boldsymbol{\theta}_n$

From (9) we can derive the upper and lower expectations of $f(x)$.

Theorem 3. *Under the c-IGP model, if $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$, the upper and lower bounds, $\overline{\mu}(x)$ and $\underline{\mu}(x)$, of $\hat{\mu}(x)$ are*

$$\overline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} + c\,\frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k} \tag{12}$$

$$\underline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} - c\,\frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k}, \tag{13}$$

which, if $1 - \mathbf{k}_x^T\mathbf{s}_k \geq 0$, are obtained for $M \to \infty$, $h = 1$ (upper) and $M \to \infty, h = -1$ (lower), while, if $1 - \mathbf{k}_x^T\mathbf{s}_k < 0$, are obtained for $M \to \infty$, $h = -1$ (upper) and $M \to \infty, h = 1$ (lower).

If, instead, $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| > 1 + \frac{c}{S_k}$ and $(1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k\mathbf{y}}{S_k} > 0$, the upper bound is found for $M \to \infty$ and $h = 1$ and the lower for $M = 0$; they are given by

$$\overline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T}{S_k}\mathbf{y} + c\,\frac{1 - \mathbf{k}_x^T\mathbf{s}_k}{S_k}$$

$$\underline{\mu}(x) = \mathbf{k}_x^T K_n^{-1}\mathbf{y} + \ (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k^T\mathbf{y}}{c + S_k}.$$

Finally, if $(1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\mathbf{s}_k\mathbf{y}}{S_k} < 0$, the upper bound is found for $M = 0$ and the lower for $M \to \infty$ and $h = 1$.

From Theorem 3 we can see that the imprecision of the model verifies

$$\overline{\mu}(x) - \underline{\mu}(x) \geq 2c\,\frac{|1 - \mathbf{k}_x^T\mathbf{s}_k|}{S_k},$$

where the equality holds if the condition $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$ is verified. In this case, we can see that the imprecision is symmetric with respect to the posterior mean of the prior with $c \to 0$. Parameter c determines the degree of imprecision of the model. A large value of c implies a large imprecision. For $c \to \infty$ we have a vacuous model that cannot learn from data.

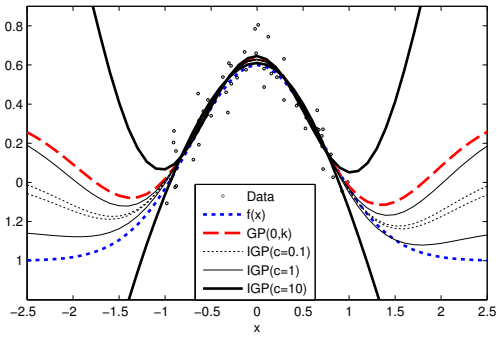


Figure 1: GP and c-IGP estimates of the function $f(x)$ given $n = 50$ observations.

When the condition $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$ is verified, by a simple rewriting of equations (12)-(13) it can be seen that $\overline{\mu}(x)$ and $\underline{\mu}(x)$ are equivalent to the posterior mean given the prior GP$(\hat{y}, k_\theta(x, x'))$ with adjusted mean

$$\hat{y}_b = \frac{\mathbf{s}_k^T}{S_k}\mathbf{y} \pm \frac{c}{S_k}.$$

Example 1. *A sample of $n = 50$ observations affected by Gaussian noise with $\sigma_n = 0.1$ is drawn from the function $f(x) = \exp(-x^2)$. The covariates $x_1, \ldots, x_n$ are uniformly distributed in $[-1, 1]$, i.e., $x \sim U[-1, 1]$. The function is modelled by the precise GP process GP$(0, k_\theta)$ and the c-IGP with the squared-exponential kernel in (7) as base kernel k_θ. Figure 1 shows the posterior expectation of the GP and the upper and lower expectations of the c-IGP for different values of c. Notice that in the region where there are observations ($x \in [-1, 1]$) the imprecision remains very small even when c is large, whereas it increases significantly outside this region.*

It is often useful to compute pointwise credible intervals $CI_f(x, \alpha) = [\underline{f}_{x,\alpha}, \overline{f}_{x,\alpha}]$ for the value of $f(x)$. Using a GP prior $f(x) \sim GP(\mu(x), k_\theta(x, x'))$, a posterior $(1 - \alpha)\%$ credible interval for the value of $f(x)$ is $CI_f(x, \alpha) = [\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x, x)}, \hat{\mu}(x) + z_{\alpha/2}\sqrt{\hat{k}(x, x)}]$ with $z_{\alpha/2}$ the $1 - \alpha/2$ percentile of the standard normal distribution. Hence we have that the posterior probability $P\big(f(x) \in CI_f(x, \alpha)\big)$ is $1 - \alpha$. In the imprecise case, we define the credible interval by imposing that the upper posterior probabilities $\overline{P}\big(f(x) < \underline{f}_{x,\alpha}\big)$ and $\overline{P}\big(f(x) > \overline{f}_{x,\alpha}\big)$ are equal to $\alpha/2$. This implies that $P\big(f(x) \in CI_f(x, \alpha)\big) \geq 1 - \alpha$ for all GPs in $\mathcal{G}$.

Theorem 4. *Under the c-IGP model, the interval*

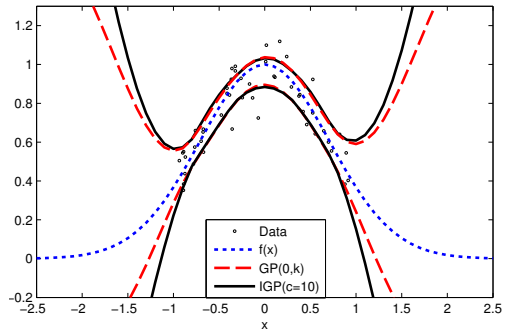


Figure 2: GP and c-IGP estimates of pointwise credible intervals for the value of $f(x)$ in Example 1.

$$CI_\alpha = [\underline{f}_{x,\alpha} = \underline{\mu}(x) - z_{\alpha/2}\sigma_{f_x}, \overline{f}_{x,\alpha} = \overline{\mu}(x) + z_{\alpha/2}\sigma_{f_x}]$$

with

$$\sigma^2_{f_x} = k_{\boldsymbol{\theta}}(x,x) - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_x + \frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k},$$

verifies

$$\overline{P}\big(f(x) < \underline{f}_x\big) \leq \alpha/2, \quad \overline{P}\big(f(x) > \overline{f}_x\big) \leq \alpha/2$$

where the equality holds if $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$.

Notice that with respect to the precise model $GP(0, k_{\boldsymbol{\theta}}(x,x))$ the value of $\sigma^2_{f_x}$ increases only for the term $\frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k}$. Moreover, $\sigma^2_{f_x}$ does not depend on c and is the same given by the precise model with $c \to 0$. Then, the width of the pointwise CIs for $c > 0$ increases only for a term equal to the difference between the upper and lower expectation of $f(x)$. Figure 2 compares the credible intervals obtained using the prior $GP(0, k_{\boldsymbol{\theta}}(x,x))$ and the c-IGP model. As for the expectation, the width of the CIs remains small for $x \in [-1,1]$ and increases significantly outside this region.

Data analyst are often interested also in simultaneous credible regions (SCR) for the value of f at multiple covariate values. Given a vector of m covariates $\mathbf{x}^*$, a $(1-\alpha)\%$ SGR for $f(\mathbf{x}^*)$ includes all vectors $\mathbf{f}^*$ that verify

$$(\mathbf{f} - \hat{\boldsymbol{\mu}}^*)^T(\hat{K}^{**})^{-1}(\mathbf{f} - \hat{\boldsymbol{\mu}}^*) < \chi^{-1}(1-\alpha|m), \quad (14)$$

where $\chi^{-1}(1-\alpha|m)$ is the $(1-\alpha)$-quantile of a Chi-squared distribution with m degrees of freedoms. For the the condition 14 to be verified by all priors in the c-IGP, it has to be verified by the upper bound of $(\mathbf{f} - \hat{\boldsymbol{\mu}}^*)^T(\hat{K}^{**})^{-1}(\mathbf{f} - \hat{\boldsymbol{\mu}}^*)$, which can be found by solving numerically an optimization problem. An example of such optimization is given in the next section in the context of hypothesis testing.

4 Application: Hypothesis Test for the Equality of two Functions

An equality test is used to detect differences between two regression functions $f_1(x)$ and $f_2(x)$ given the two independent samples $D_1 = (\mathbf{x}^{(1)}, \mathbf{y}^{(1)})$ and $D_2 = (\mathbf{x}^{(2)}, \mathbf{y}^{(2)})$ of, respectively, n_1 and n_2 observations. Our aim is to extend the Bayesian test based on the GP presented in [1] using the c-IGP model. The approach in [1] assumes the same GP prior $GP(0, k_{\boldsymbol{\theta}})$ for the two functions f_1 and f_2; the two posterior distributions share the same hyperparameters. Here, we assume the same c-IGP set of priors $\mathcal{G}$ for the two functions, that is,

$$f_i \sim GP\left(M_i h_i, k_{\boldsymbol{\theta}}(x,x') + \frac{M_i+1}{c}\right),$$

with $i = 1, 2$, $h_i = \pm 1$ and $M_i \geq 0$. As a consequence, we are assuming that f_1 and f_2 are two GPs with the same base kernel $k_{\boldsymbol{\theta}}(x,x')$. Instead, their prior mean functions $M_1 h_1$ and $M_2 h_2$ can be different, as they are free to vary in the set of all constant functions. We assume that the two samples $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are affected by Gaussian noise with variance, respectively, σ_1^2 and σ_2^2. The hyperparameters $\boldsymbol{\theta}, \sigma_1, \sigma_2$ are obtained considering for both f_1 and f_2 the prior with $c \to 0$. Then, after combining the two datasets $\{D_1, D_2\}$, we maximize the joint marginal probability of $(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2)$ with respect to $\boldsymbol{\theta}, \sigma_1, \sigma_2$. Assuming that f_1 and f_2 are independent Gaussian processes, we have that

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_1, \sigma_2) = p(\mathbf{y}^{(1)}|\mathbf{x}^{(1)}, \boldsymbol{\theta}, \sigma_1)p(\mathbf{y}^{(2)}|\mathbf{x}^{(2)}, \boldsymbol{\theta}, \sigma_2).$$

Then, the logarithm of the joint marginal of $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \boldsymbol{\theta}, \sigma_1^2, \sigma_2^2$ is

$$\sum_{i=1}^{2} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i) + \log p(\boldsymbol{\theta}, \sigma_1, \sigma_2)$$

where $\log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}, \sigma_i)$, is given in (11), up to an additive constant.

In the precise approach, given the prior $GP(0, k_\theta)$, we compute from (4) the posterior marginal GPs $p(\mathbf{f}_1^*|\mathbf{x}^*, D_1)$ and $p(\mathbf{f}_2^*|\mathbf{x}^*, D_2)$ at the $m = n_1 + n_2$ test inputs $\mathbf{x}^* = \{x_i^* : i = 1, \ldots, m,\ x_i^* \in [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}]\}$. In this way, the equality of the two functions is tested at the covariates of the observations, that is, where we have the experimental evidence. Moreover, it is assumed that the observation covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ cover the same region of the covariate space. This is done to avoid testing the equality in regions where there are no observations for one or both functions, as in these region we do not expect to be able to state

any conclusion about equality or difference of the two functions. If applied in such regions, the precise test would always assign very large posterior probability to the hypothesis that there is no difference between the functions. Using a IGP model, we can test the equality assumption in any subset $\mathcal{X}_T$ of the covariate space $\mathcal{X}$ by taking the m test inputs $\mathbf{x}^*$ so to cover uniformly the region of interest $\mathcal{X}_T$. If all priors in the IGP set entail the same decision, we retain it, if instead they lead to different decisions we conclude that a robust decision cannot be made in $\mathcal{X}_T$. This way, we can automatically identify a situation where data do not allow to state any conclusion.

Let us denote the mean and covariance functions of the posterior distributions of $\mathbf{f}_1^*$ and $\mathbf{f}_2^*$ as $\hat{\mu}^{(i)}(x)$ and $\hat{k}^{(i)}(x,x')$, $i=1,2$. Since the difference of two Gaussian variables is Gaussian, it follows that the posterior of the GP $\Delta f(x) = f^1(x) - f^2(x)$ is also a GP with mean and covariance functions $\Delta\hat{\mu}(x) = \hat{\mu}^{(1)}(x) - \hat{\mu}^{(2)}(x)$ and $\hat{k}_\Delta(x,x') = \hat{k}^{(1)}(x,x') + \hat{k}^{(2)}(x,x')$. Let $\Delta\mathbf{f}^*$, $\Delta\hat{\boldsymbol{\mu}}^*$ and $\hat{K}_\Delta^*$ be the difference, the mean and the covariance functions evaluated at the test covariates $\mathbf{x}^*$, then, we say that the two functions are equal with posterior probability $1-\alpha$ if the credible region for $\Delta\mathbf{f}^*$ includes the zero vector or, in other words, if:

$$(\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}_\Delta^*)^{-1}\Delta\hat{\boldsymbol{\mu}}^* \leq \chi^{-1}(1-\alpha|\nu), \qquad (15)$$

where ν is the number of positive eigenvalues of $\hat{K}_\Delta^*$. In practice, as the number m of test inputs is likely to be considerably larger than the dimensionality of the covariance function, the matrix $\hat{K}_\Delta^*$ is not full rank. Thus, we decompose it as PDP^T, where D is the diagonal matrix of the eigenvalues $\lambda_1,\dots,\lambda_m$ (sorted in descending order), and retain only the sub-matrices $P_\nu D_\nu P_\nu^T$ corresponding to the eigenvalues $\lambda_1,\dots,\lambda_\nu$ which verify the condition $\lambda_{\nu+1}/\sum_{i=1}^m \lambda_i < \epsilon$, where ϵ is a small, positive constant. In the example below, we use $\epsilon = 0.0001$.

In the c-IGP model, the inference about $\chi_s^2(M_1,M_2,h_1,h_2) = (\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}_\Delta^*)^{-1}\Delta\hat{\boldsymbol{\mu}}^*$ depends on the choice of the prior, that is on the value of M_1, M_2, and of h_1, h_2.

Proposition 2. *The c-IGP model is a prior ignorance model for inferences about χ_s^2, i.e.,*

$$\underline{\chi}_s^2 = 0 \quad \overline{\chi}_s^2 \to = +\infty.$$

A posteriori let $\hat{\mu}_0^{(i)}(x) = k_x^{(i)T}K_n^{(i)-1}\mathbf{y}^{(i)}$ be the posterior mean functions obtained from a GP with zero mean and covariance function $k_\theta(x,x')$ when $\mathbf{x}=\mathbf{x}^{(i)}$, $\mathbf{y}=\mathbf{y}^{(i)}$, and let $\hat{k}_0^{(i)}(x,x')$ be the covariance function obtained from (10) when $c\to\infty$. For a given value of M_1 and M_2 the posterior expectation of the GP $\Delta f(x)$, that is $\Delta\hat{\mu}(x)$, can be derived from (9) and is

$$\Delta\hat{\mu}_{M_1,M_2}(x) = \hat{\mu}_0^{(1)}(x) - \hat{\mu}_0^{(2)}(x) + \mu_c^{(1)}(x) + \mu_c^{(2)}(x)$$

where $\mu_c^{(i)}(x) = (1-\mathbf{k}_x^{(i)T}K_n^{(i)-1})\frac{\mathbf{s}_k^{(i)T}\mathbf{y}^{(i)}+t_i c}{c(1-|t_i|)+S_k^{(i)}}$, $t_i = \frac{Mh}{M+1}$ and $\mathbf{k}_x^{(i)}$ and $K_n^{(i)}$ are obtained by evaluating the covariance functions at the training covariates $\mathbf{x}^{(i)}$, $i=1,2$. The lower/upper bounds for $\chi_s^2(M_1,M_2,h_1,h_2)$ are obtained by minimizing/maximizing w.r.t. $t_i \in [-1,1]$ the statistic:

$$\chi_s^2 = \Delta\hat{\boldsymbol{\mu}}^{*T}(\hat{K}_{M_1,M_2}^\Delta)^{-1}\Delta\hat{\boldsymbol{\mu}}^*, \qquad (16)$$

where $\hat{K}_{M_1,M_2}^\Delta = [\hat{k}_0^{(1)}(x_i,x_j) + \hat{k}_0^{(2)}(x_i,x_j) + \hat{k}_c^{(1)}(x_i,x_j) + \hat{k}_c^{(2)}(x_i,x_j)]_{i,j}$, and $\hat{k}_c^{(i)}(x,x') = \frac{(M_i+1)[1-\mathbf{k}_x^{(i)T}\mathbf{s}_k^{(i)}][1-\mathbf{k}_{x'}^{(i)T}\mathbf{s}_k^{(i)}]^T}{c+(M_i+1)S_k^{(i)}}$, $i=1,2$.

4.1 Numerical Example

Let us consider two samples D_1 and D_2 that we wish to compare on the subset $\mathcal{X}_T=[a,b]$ of the covariate space. Assuming an observation noise $v\sim\mathcal{N}(0,\sigma_n=0.2)$, we sample D_1 and D_2 from:

Case A: $x_i^{(1,2)} \sim U[-2,2]$, $y_i^{(1,2)} = f(x_i)+v_i$,

Case B: $x_i^{(1)} \sim U[-2,2]$, $y_i^{(1)} = f(x_i)+v_i$,
$x_i^{(2)} \sim U[-2,2]$ $y_i^{(2)} = g(x_i)+v_i$,

Case C: $x_i^{(1)} \sim U[-2,0]$, $y_i^{(1)} = f(x_i)+v_i$,
$x_i^{(2)} \sim U[-2,4]$, $y_i^{(2)} = g(x_i)+v_i$,

Case D: $x_i^{(1)} \sim U[-2,2]$, $y_i^{(1)} = f(x_i)+v_i$,
$x_i^{(2)} \sim U[-2,4]$, $y_i^{(2)} = g(x_i)+v_i$,

where $f(x)=\exp(-x^2)$ and $g(x)=f(x)+0.5f(x-2)$ (see Figure 3). For each scenario the two datasets D_1 and D_2 have been simulated only once. More extensive simulations are left to future work. We have tested the difference between the two samples for different test subsets $\mathcal{X}_T\in[-2,b]$. The difference $f(x)-g(x)$ is about zero for $x<0$, is large ($>\sigma_n$) in the interval $[1,3]$ and is small ($<\sigma_n$) in $[0,1]$. Therefore, we expect to easily detect a difference between the two samples when $b>1$, whereas for $b<1$ the decision is more difficult and for $b<0$ we can assume that the two functions are equal. Table 1 shows the decisions for the precise and the imprecise tests at different values of c and b. One can notice that for $c=10$ we are most often undecided (save when the decision is simple, e.g., in cases B and D when $b>1$ and thus all tests recognize the difference) as the imprecision is very large in this case.

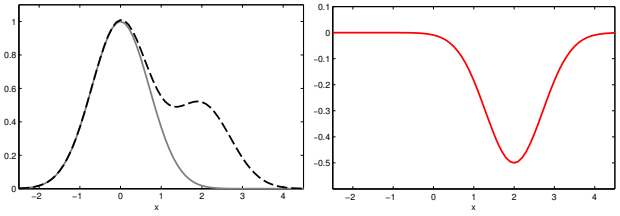


Figure 3: Left: Functions f (continuous line) and g (dashed line). Right: difference $f - g$.

		GP		IGP	
Case	b	n=50	n=200	n=50	n=200
A	2	0	0	0/0/2	0/0/2
A	4	0	0	0/2/2	0/2/2
B	0	0	0	0/0/2	0/0/2
B	1	0	1	0/2/2	1/1/1
B	2	1	1	1/1/1	1/1/1
B	4	1	1	1/1/1	1/1/1
C	0	0	0	0/0/2	0/0/2
C	1	0	0	0/0/2	0/0/2
C	2	0	0	0/2/2	0/2/2
C	4	0	0	0/2/2	0/2/2
D	0	0	0	0/0/0	0/0/2
D	1	0	1	2/2/2	1/1/1
D	2	1	1	1/1/1	1/1/1
D	4	1	1	1/1/1	1/1/1

Table 1: Decisions of the precise test for $c = 1/5/10$, where 0 indicates that the two functions are equal with posterior probability $1 - \alpha$, 1 indicates that the two functions are different (i.e., the posterior probability that they are equal is less than α), 2 indicates indecision (i.e., the decision depends on the prior).

On the other side, for $c = 1$ the test makes almost always the same decision as the precise test, as the imprecision is very small in this case. When $c = 5$ we have a better balance between robustness and power: the IGP test makes the same decision as the precise one when there is enough information to make a robust decision, whereas it is undecided when the decision is difficult due to the lack of information. For instance, in case A with $b = 2$ the precise test always issues a no difference decision. The same happens in case C, although the two situations are very different, because in the first case $f_1 = f_2$ and we can observe both functions on the entire set $\mathcal{X}_T$, whereas in the second case $f_1 \neq f_2$ but we cannot see it as we observe f_1 only in the range $[-2, 0]$ where the two function are almost identical. On the other side, the imprecise test detects the difference of the two situations, and in case A it correctly issues a no difference decision, whereas in case C it is undecided, thus acknowledging that there is not enough information to make a decision.

Something similar can be observed also in case D: when $b = 0$ both the precise and imprecise tests issue a no difference decision as in this range the two function can be actually considered identical; when, instead, $b = 1$, the functions are different, but, since the difference is small, it cannot be clearly detected with only $n = 50$ data. However, the imprecise test recognizes that the decision is somehow difficult and is undecided, whereas the precise test can only decide that there is no difference. For $n = 200$, the information is enough to make both tests detect a difference.

5 A Generalization of the IGP Model

In this Section we generalize the IGP with constant mean by considering an IGP with mean function proportional to an arbitrary function $h(x)$.

Definition 3. *Given a function $h(x)$ and a constant $c > 0$, we define an Imprecise Gaussian Process with base mean function $h(x)$ (hIGP) the set of GPs:*

$$\mathcal{G}_{h(x)} = \left\{ GP\left(Mh(x), k_{\boldsymbol{\theta}}(x, x') + \tfrac{M+1}{c} h(x)h(x')\right),\ M \geq 0 \right\}.$$

A posteriori we have the following result.

Theorem 5. *Let $\mathbf{h} = h(\mathbf{x})$ and*

$$f(x) \sim GP(Mh(x), k_{\boldsymbol{\theta}} + \tfrac{M+1}{c} h(x), h(x')).$$

The posterior distribution of f is a GP with mean function

$$\hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1}(\mathbf{y} - \hat{y}(\mathbf{x})) + \hat{y}(x) \qquad (17)$$

with $\hat{y}(x) = \frac{(M+1)\mathbf{h}^T K_n^{-1}\mathbf{y} + cM}{c + (M+1)\mathbf{h}^T K_n^{-1}\mathbf{h}} h(x)$, and covariance function

$$\hat{k}(x, x') = k_{\boldsymbol{\theta}}(x, x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \frac{(M+1)(h(x) - \mathbf{k}_x^T K_n^{-1}\mathbf{h})^T (h(x') - \mathbf{k}_{x'}^T K_n^{-1}\mathbf{h})}{c + (M+1)\mathbf{h}^T K_n^{-1}\mathbf{h}}.$$

We can further generalize the IGP model in Definition 3 by letting $h(x)$ free to vary in a set of functions $\mathcal{H}$.

Definition 4. *Given a set of functions $\mathcal{H}$ and a constant $c > 0$, we define an Imprecise Gaussian Process with set of base mean functions $\mathcal{H}$ ($\mathcal{H}$-IGP) the set of GPs:*

$$\mathcal{G}_{\mathcal{H}} = \left\{ \mathcal{G}_{h(x)} : h(x) \in \mathcal{H} \right\}. \qquad (18)$$

From Theorem 5 we can see that not all $\mathcal{H}$-IGP verify learning. In fact, if $\mathcal{H}$ include functions that are zero at all training covariates $\mathbf{x}$ so that $\mathbf{h}^T K_n^{-1}\mathbf{h} = 0$, posterior inferences are vacuous.

Proposition 3. *Any set $\mathcal{H}$-IGP such that* $\mathbf{h}$ *is a nonzero vector for all $h(x) \in \mathcal{H}$ can learn from the observations* $\mathbf{x}, \mathbf{y}$.

Moreover, not all $\mathcal{H}$-IGP verify prior-ignorance about $E[f(x^*)]$ for all x^*. If, for instance, $h(x^*) = 0$ all $h(x) \in \mathcal{H}$, then a priori $E[f(x_i^*)] = Mh(x^*) = 0$ for all M.

Proposition 4. *If it exist $h^+(x^*) \in \mathcal{H} : h^+(x^*) > 0$ and $h^-(x^*) \in \mathcal{H} : h^-(x^*) < 0$, then the $\mathcal{H}$-IGP is a model of prior ignorance about the expectation of $f(x^*)$ in the sense that it verifies*

$$\inf_{M,h(x)} E[f(x_i^*)] = -\infty, \quad \sup_{M,h(x)} E[f(x_i^*)] = +\infty.$$

By properly selecting the set $\mathcal{H}$ one can obtain IGP models that verify both prior near-ignorance and learning.

Example 2. *The c-IGP model presented in Section 3 is an $\mathcal{H}$-IGP model with set of base mean functions $\mathcal{H}_c = \{h(x) = -1,\ h(x) = 1\}$. This set verifies the conditions of both Proposition 4 and 3 an thus verifies both prior near-ignorance and learning.*

Example 3. *Let us consider the set $\mathcal{H} = \{h(x) = -x,\ h(x) = x\}$. It verifies the conditions of Proposition 4 for all covariates except $x = 0$ and verifies the condition of Proposition 3 provided that* $\mathbf{x}$ *is a nonzero vector. The corresponding $\mathcal{H}$-IGP includes GPs with mean function varying in the set of all linear functions with intercept in 0. It is a model of prior ignorance about $f(x)$ for all $x \neq 0$ and can learn from data with covariate $x \neq 0$.*

6 Conclusions

In this paper we have presented a model of prior near ignorance about the value of a regression function based on the Gaussian process. We have shown that this IGP model can be used to make inferences about the regression functions which are more robust with respect to the choice of the prior. In fact, for those subsets of $\mathcal{X}$ where there are many observations inferences almost coincide with the precise model, whereas in those subset with no observations the imprecision of the prediction is very high, thus reflecting the actual lack of knowledge. As a consequence of this, decisions based on this model are more reliable. For instance, we have applied the IGP to test the difference between regression functions, and shown that the IGP model allows us to acknowledge when the available data are not informative enough to make a robust decision. Although in this paper we have only consider univariate functions, the c-IGP model can be straightforward extended to the multivariate case where x is a vector of covariates.

A generalization of the IGP with constant mean has also been proposed, based on which it will be possible to develop other prior near ignorance models that consider different sets of prior mean functions. The study of these models and their properties will be the object of future work. Moreover, as a strong prior information is introduced in the model also by the base kernel, further research should focus on the development of models allowing for a weaker specification of the kernel function.

There are many techniques other than GPs available for nonparametric regression, e.g., splines, relevance vector machines, kernel smoothers, etc., that have not been considered in this work. Their relative strengths and weaknesses w.r.t. GPs are discussed in [12, Sec. 7]. As they are all precise methods, we can expect them to suffer from the same weaknesses of the precise GPs. The probabilistic formulation of GPs and the simple closed form expression of their posterior inferences, have made them a good starting point to develop an imprecise approach to nonparametric regression that, in the future, could be extended to other regression techniques, taking advantage also from the connections they have with GPs [12, Sec. 6].

Appendix

6.1 Proof of Proposition 1

This can be seen by considering that a priori $E[f(x_i^*)] = Mh(x_i^*)$ so that for $h(x_i^*) = \pm 1$ and $M \to \infty$ we have $E[f(x_i^*)] \to \pm\infty$.

6.2 Proof of Theorem 1

Miller in [7] proves the following

Lemma 1. *If A and $A + B$ are invertible, and B has rank 1, then*

$$(A + B)^{-1} = A^{-1} - \frac{1}{1 + g} A^{-1} B A^{-1},$$

where $g = trace(BA^{-1})$ with $g \neq -1$.

From this, it follows that

$$(K_n + \frac{M+1}{c} \mathbb{1}_{nn})^{-1} = K_n^{-1} - \frac{M+1}{c + (M+1)S_k} \mathbf{s}_k \mathbf{s}_k^T, \tag{19}$$

where $\mathbb{1}_{nn}$ is a $n \times n$ dimensional matrix of ones. Then,

$$\begin{aligned}
&\hat{\mu}(x) \\
&= Mh + \left(\mathbf{k}_x + \frac{M+1}{c}\mathbb{1}_n\right)^T \\
&\qquad \left(K_n^{-1} - \frac{M+1}{c+(M+1)S_k}\mathbf{s}_k\mathbf{s}_k^T\right)(\mathbf{y} - Mh\mathbb{1}_n) \\
&= Mh + \left[\mathbf{k}_x^T K_n^{-1}\left(1 - \frac{(M+1)\mathbb{1}_n\mathbf{s}_k^T}{c+(M+1)S_k}\right) + \right. \\
&\qquad \left. \frac{(M+1)\mathbf{s}_k^T}{c+(M+1)S_k}\right](\mathbf{y} - Mh\mathbb{1}_n) \\
&= \mathbf{k}_x^T K_n^{-1}\left(\mathbf{y} - Mh(\mathbf{x}) - \frac{(M+1)\mathbb{1}_n\mathbf{s}_k^T}{c+(M+1)S_k}(\mathbf{y} - Mh(\mathbf{x}))\right) \\
&\qquad + Mh + \frac{(M+1)\mathbf{s}_k^T}{c+(M+1)S_k}(\mathbf{y} - Mh(\mathbf{x})) \\
&= \mathbf{k}_x^T K_n^{-1}\left(\mathbf{y} - \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c+(M+1)S_k}\mathbb{1}_n\right) + \\
&\qquad \frac{(M+1)\mathbf{s}_k^T\mathbf{y} + cMh}{c+(M+1)S_k}.
\end{aligned}$$

Similarly for the covariance function we obtain:

$$\begin{aligned}
&\hat{k}(x,x') \\
&= k_\theta(x,x') + \frac{M+1}{c} - \left(\mathbf{k}_x + \frac{M+1}{c}\mathbb{1}_n\right)^T \\
&\qquad \left(K_n^{-1} - \frac{M+1}{c+(M+1)S_k}\mathbf{s}_k\mathbf{s}_k^T\right)\left(\mathbf{k}_{x'} + \frac{M+1}{c}\mathbb{1}_n\right) \\
&= k_\theta(x,x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_x + \\
&\qquad \frac{M+1}{c+(M+1)S_k}\left(\mathbf{k}_x^T\mathbf{s}_k\mathbf{s}_k^T\mathbf{k}_{x'} - \mathbf{k}_x^T\mathbf{s}_k - \mathbf{s}_k^T\mathbf{k}_{x'} + 1\right).
\end{aligned}$$

6.3 Proof of Theorem 2

From (8), the logarithm of the marginal probability of $\mathbf{y}, \theta_\mathbf{n}$ given the prior $GP\left(Mh, k_{\boldsymbol{\theta}}(x,x') + \frac{M+1}{c}\right)$ is

$$\begin{aligned}
&\log p(\mathbf{y}, \theta_\mathbf{n}) = \\
&-\tfrac{1}{2}(\mathbf{y} - Mh\mathbb{1}_n)^T\left(K_n + \frac{M+1}{c}\mathbb{1}_{nn}\right)^{-1}(\mathbf{y} - Mh\mathbb{1}_n) + \\
&-\tfrac{1}{2}\log|K_n + \tfrac{M+1}{c}\mathbb{1}_{nn}| - \tfrac{n}{2}\log 2\pi + \log p(\theta_\mathbf{n}).
\end{aligned} \tag{20}$$

From (19) we obtained that the first term on the r.h.s. of (20) is equal to

$$\frac{cM^2h^2S_k - 2cMh\mathbf{y}^T\mathbf{s}_k - (M+1)(\mathbf{y}^T\mathbf{s}_k)^2}{2c + 2(M+1)S_k} - \frac{1}{2}\mathbf{y}^T K_n^{-1}\mathbf{y}. \tag{21}$$

Based on the matrix determinant Lemma [5] which states:

Lemma 2. *Given a $n \times n$ invertible matrix A and two n dimensional vectors $\mathbf{u}, \mathbf{v}$*

$$|A + \mathbf{u}\mathbf{v}^T| = |A|(1 + \mathbf{v}^T A^{-1}\mathbf{u}),$$

we obtain

$$|K_n + \frac{M+1}{c}\mathbb{1}_{nn}| = |K_n|\frac{c+(M+1)S_k}{c}. \tag{22}$$

Then, from (21), (20) and (22), it follows

$$\begin{aligned}
\log p(\mathbf{y}, \theta_\mathbf{n}) \quad &\xrightarrow{c\to 0} -\tfrac{1}{2}\left[\mathbf{y}^T K_n^{-1}\mathbf{y} - \tfrac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k}\right] + \\
&-\tfrac{1}{2}\log|K_n|\tfrac{M}{c}S_k - \tfrac{n}{2}\log 2\pi + \log p(\theta_\mathbf{n}).
\end{aligned} \tag{23}$$

Finally, by considering that the terms $-\frac{1}{2}\log\frac{M}{c}$ and $-\frac{n}{2}\log 2\pi$ are constant with θ, we have that

$$\begin{aligned}
&\text{argmax}_{\boldsymbol{\theta}}[\log p(\mathbf{y}, \boldsymbol{\theta}_n)] \\
&= \text{argmax}_{\boldsymbol{\theta}}\left[-\frac{1}{2}\left(\mathbf{y}^T K_n^{-1}\mathbf{y} - \frac{(\mathbf{y}^T\mathbf{s}_k)^2}{S_k} - \log S_k|K_n|\right)\right. \\
&\qquad \left. + \log p(\theta_\mathbf{n})\right].
\end{aligned}$$

6.4 Proof of Theorem 3

The derivative of (9) with respect to M is

$$\frac{\partial\hat{\mu}(x)}{\partial M} = (1 - \mathbf{k}_x^T\mathbf{s}_k)\frac{\pm c \pm S_k + \mathbf{s}_k^T\mathbf{y}}{(c+(M+1)S_k)^2}$$

If $\left|\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}\right| \leq \frac{c}{S_k} + 1$, the second term of the derivative above is positive for $h = 1$ and negative for $h = -1$. Notice also that, for $M = 0$, the values of $\hat{\mu}(x)$ in the two cases $h = \pm 1$ coincide. Then, if the first term $1 - \mathbf{k}_x^T\mathbf{s}_k$ is positive, we have that $\hat{\mu}(x)$ increases with M for $h = 1$ and decreases for $h = -1$ so that the upper is found for $h = 1$ and $M \to \infty$ and the lower for $h = -1$ and $M \to \infty$. Vice versa, if the first term is negative, the upper is found for $h = -1$ and $M \to \infty$ and the lower for $h = 1$ and $M \to \infty$.

If, instead, $\left|\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}\right| > \frac{c}{S_k} + 1$, the second term of the derivative above is always positive (if $\frac{\mathbf{s}_k^T\mathbf{y}}{S_k} > 0$) or negative (otherwise); then, $\hat{\mu}(x)$ increases if $1 - \mathbf{k}_x^T\mathbf{s}_k$ and $\frac{\mathbf{s}_k^T\mathbf{y}}{S_k}$ have the same sign, and decreases otherwise. In the first case, the upper is found for $h = 1$ and $M \to \infty$ and the lower for $M = 0$; vice versa, in the second case, the upper is found for $M = 0$ and the lower for $h = -1$ and $M \to \infty$.

The value of the upper and lower can be derived from (9).

6.5 Proof of Theorem 4

For each GP in $\mathcal{G}$ the lower bound of a $(1-\alpha)\%$ credible interval is

$$\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)} \leq \underline{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)}.$$

Moreover, from (10), it can be seen that $\hat{k}(x,x)$ increases with M, so that it maximum is found at $M \to \infty$ and is $\sigma_{f_x}^2 = k_{\boldsymbol{\theta}}(x,x') - \mathbf{k}_x^T K_n^{-1}\mathbf{k}_{x'} + \frac{(1-\mathbf{k}_x^T\mathbf{s}_k)^2}{S_k}$ so that

$$\hat{\mu}(x) - z_{\alpha/2}\sqrt{\hat{k}(x,x)} \leq \underline{\mu}(x) - z_{\alpha/2}\sigma_{f_x}.$$

where the equality holds when the lower expectation $\underline{\mu}(x)$ is found for $M \to \infty$, that is when $\left|\frac{\mathbf{s}_k\mathbf{y}}{S_k}\right| \leq 1 + \frac{c}{S_k}$ (Theorem 3).

6.6 Proof of Proposition 2

It can be verified that the lower bound is found by choosing $M_1 = M_2 = 0$. The upper is found, for instance, for $M_1 = 0$, $M_2 \to \infty$, as we have

$$\begin{aligned}\chi^2_s(0, M_2, h_1, h_2) &= M_2^2 \mathbb{1}_m^T (2 * K^{**} + \frac{M_2+2}{c})^{-1} \mathbb{1}_m \\ &= \frac{M_2^2}{2} \mathbb{1}_m^T \left[(K^{**})^{-1} - \frac{M_2+2}{2c + (M_2+2)S_{k^*}} \mathbf{s}_{k^*} \mathbf{s}_{k^*}^T \right] \mathbb{1}_m \\ &= M_2^2 \frac{1}{2} \left[S_{k^*} - \frac{M_2+2}{2c + (M_2+2)S_{k^*}} S_{k^*}^2 \right] \\ &= M_2^2 \frac{cS_{k^*}}{2c + (M_2+2)S_{k^*}} \xrightarrow{M_2 \to \infty} cM_2 \to \infty,\end{aligned}$$

where $S_{k^*} = \mathbb{1}_m^T (K^{**})^{-1} \mathbb{1}_m$, $\mathbf{s}_{k^*} = (K^{**})^{-1} \mathbb{1}_m$ and where we have used Lemma 1.

6.7 Proof of Theorem 5

Let us define $M' = \frac{M+1}{c}$, $S_h = \mathbf{h}^T K_n^{-1} \mathbf{h}$, $\mathbf{s}_h = K_n^{-1} \mathbf{h}$ and $D = 1 + M' S_h$. From Lemma 1, it follows that

$$(K_n + M' \mathbf{h}\mathbf{h}^T)^{-1} = K_n^{-1} - \frac{M'}{D} \mathbf{s}_h \mathbf{s}_h^T. \tag{24}$$

Then, $\hat{\mu}(x) =$

$$\begin{aligned}&= Mh(x) + \left(\mathbf{k}_x + M' h(x) \mathbf{h}^T \right)^T \left(K_n^{-1} - \frac{M'}{D} \mathbf{s}_h \mathbf{s}_h^T \right) (\mathbf{y} - M\mathbf{h}) \\ &= Mh(x) + \mathbf{k}_x^T K_n^{-1} \mathbf{y} - \mathbf{k}_x^T \frac{M'}{D} \mathbf{s}_h \mathbf{s}_h^T \mathbf{y} + \frac{M'}{D} h(x) \mathbf{s}_h^T \mathbf{y} + \\ &\quad - \frac{M}{D} \mathbf{k}_x^T \mathbf{s}_h - \frac{MM'}{D} h(x) S_h \\ &= \mathbf{k}_x^T K_n^{-1} \left(\mathbf{y} - \hat{y}(\mathbf{x}) \right) + \hat{y}(x)\end{aligned}$$

Similarly for the covariance function we obtain:

$$\begin{aligned}\hat{k}(x, x') =& k_\theta(x, x') + M' h(x) h(x') - \left(\mathbf{k}_x + M' h(x) \mathbf{h} \right)^T \\ & \left(K_n^{-1} - \frac{M'}{D} \mathbf{s}_h \mathbf{s}_h^T \right) \left(\mathbf{k}_{x'} + M' h(x') \mathbf{h} \right) \\ =& k_\theta(x, x') - \mathbf{k}_x^T K_n^{-1} \mathbf{k}_x + \frac{M'}{D} \left(\mathbf{k}_x^T \mathbf{s}_h \mathbf{s}_h^T \mathbf{k}_{x'} + \right. \\ & \left. - h(x) \mathbf{k}_x^T \mathbf{s}_h - h(x') \mathbf{s}_h^T \mathbf{k}_{x'} + h(x) h(x') \right).\end{aligned}$$

6.8 Proof of Proposition 3

From (17) it follows that

$$\lim_{M \to +\infty} \hat{\mu}(x) = \mathbf{k}_x^T K_n^{-1} \mathbf{y} + (h(x) - \mathbf{k}_x^T K_n^{-1} \mathbf{h}) \frac{\mathbf{h}^T K_n^{-1} \mathbf{y} + c}{\mathbf{h}^T K_n^{-1} \mathbf{h}},$$

which, if $\mathbf{h}$ is a nonzero vector, is bounded.

6.9 Proof of Proposition 4

This can be seen by considering that a priori $E[f(x_i^*)] = Mh(x_i^*)$ so that for $h(x) = h^+(x)$ and $M \to \infty$ we have $E[f(x_i^*)] \to +\infty$ and for $h(x) = h^-(x)$ and $M \to \infty$ we have $E[f(x_i^*)] \to -\infty$.

Acknowledgements

The author would like to thank Alessio Benavoli for his valuable comments and suggestions.

References

[1] A. Benavoli and F. Mangili. Gaussian Processes for Bayesian hypothesis tests. In *Proc 18th AIstat Conference*. Society for Artificial Intelligence and Statistics, 2015.

[2] A. Benavoli and M. Zaffalon. A model of prior ignorance for inferences in the one-parameter exponential family. *J of Stat Planning and Inference*, 142(7):1960 – 1979, 2012.

[3] A. Benavoli and M. Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 2014. in-press.

[4] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[5] D. A. Harville. *Matrix algebra from a statistician's perspective*. Springer, 1997.

[6] D. J. MacKay. Introduction to Gaussian processes. *In Bishop, C. M., editor, Neural Networks and Machine Learning*, pages 133–166, 1998.

[7] K. S. Miller. On the inverse of the sum of matrices. *Mathematics Magazine*, 54(2):67–72, 1981.

[8] R. M. Neal. Regression and classification using gaussian process priors. In *Bernardo, et al. eds., Bayesian Statistics 6: Proc of the 6th Valencia international meeting*, volume 6, page 475, 1998.

[9] A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42, 1978.

[10] L. R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, 58:1–23, 1991.

[11] C. E. Rasmussen. The Gaussian Processes Web Site. http://www.gaussianprocess.org/, February 2011.

[12] C. E. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA, 2006.

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

[14] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57, 1996.

Conformity and Independence with Coherent Lower Previsions

Enrique Miranda
University of Oviedo
Oviedo (Spain)
mirandaenrique@uniovi.es

Marco Zaffalon
IDSIA
Lugano (Switzerland)
zaffalon@idsia.ch

Abstract

We study the conformity of marginal unconditional and conditional models with a joint model under assumptions of epistemic irrelevance and independence, within Walley's theory of coherent lower previsions. By doing so, we make a link with a number of prominent models within this theory: the marginal extension, the irrelevant natural extension, the independent natural extension and the strong product.

Keywords. Coherent lower previsions, sets of desirable gambles, epistemic irrelevance, epistemic independence, marginal extension, strong product.

1 Introduction

The theory of coherent lower previsions was developed by Peter Walley [22], with some influence from earlier work by Peter Williams [23], as a generalisation to the imprecise case of the behavioural approach to probability championed by de Finetti [10]. One of its advantages is that it includes as particular cases most of the models of non-additive measures existing in the literature, such as Choquet capacities [3], belief functions [21] or possibility measures [12].

Coherent lower previsions can be used to express both unconditional and conditional information, and several coherent lower previsions can be used to build a joint model that puts together the assessments present in each of the underlying sources. This is usually done by means of the notion of *natural extension*, which in some cases can be combined with other structural assessments, such as independence or exchangeability [1, Chapter 3].

The conformity of some marginal lower previsions with a joint model is easy to understand (it means simply that the joint model produces this marginals when restricted to gambles that depend on one of the variables); however, the relationship with the conditional models is more problematic. This is due to two reasons: on the one hand, there are several ways in which we can consider that a number of conditional models are consistent with a joint model, as the different notions of coherence by Williams and Walley testify. In this paper, we are going to use Walley's theory of coherent lower previsions, which makes use of the notion of *conglomerability*. This is an assumption that is not considered in de Finetti and Williams' approaches, and that has been subject to some controversy.

On the other hand, even if we stick to Walley's approach (but also in the finite case, where conglomerability is not an issue), there are several ways in which we can derive conditional models from an unconditional ones, so it is not immediate how to tell which conditional assessments are the ones derived from the unconditional model.

Our choice in this paper is to consider the notion of conditional natural extension, which, according to Walley, provides the most conservative behavioural implications of the assessments present in the unconditional model.

Under this setting, we are going to define a notion of conformity of marginal and conditional assessments with the meaning of the existence of a joint that induces them with the procedures of marginalization and natural extension mentioned above. We shall consider three different scenarios: that where we start from two marginal models and make an assessment of epistemic irrelevance, that where we make an assessment of epistemic independence, and that where our starting point is a marginal and a conditional lower prevision. In each of these cases we shall show that the notion of conformity does not always hold, we shall give necessary and sufficient conditions for its existence, and determine the least conservative model satisfying this notion.

Interestingly, we shall prove that this so-called conforming natural extension coincides under some conditions with some well-known models within the theory of coherent lower previsions: the marginal extension, the irrelevant natural extension, and the independent natural extension. This has led us to deepen our study of independent models, by completing some recent work in [19, 25]. In particular, we study in detail two properties that we have recently linked with independent products, and more specifically

with the strong product. We investigate to which extent they are satisfied by other independent products and also by the marginal extension. The properties are in some cases formulated in terms of sets of desirable gambles, which provide the behavioural interpretation underlying coherent lower previsions.

The paper is organized as follows: in Section 2, we introduce the basics of the theory of coherent lower previsions that we shall use in the rest of the paper. Our study begins in Section 3 with the definition of conformity for a marginal and a conditional model, when the latter is defined by means of an assessment of epistemic irrelevance. This is completed in Section 4 with a study of the relationship between conformity and independent products. Then in Section 5 we consider the general case of conformity of a marginal and a conditional lower prevision, where the latter need not satisfy the property of epistemic irrelevance. The paper ends in Section 6 with some additional comments and remarks.

2 Preliminaries

2.1 Coherent Lower Previsions

Let us give the basics of the theory of coherent lower previsions necessary to follow the remainder of this paper. An in-depth study with details on the behavioural interpretation of the following notions may be found in [22].

Consider a possibility space Ω. A *gamble* on Ω is a bounded real-valued function $f : \Omega \to \mathbb{R}$. The set of all gambles on Ω is denoted $\mathcal{L}(\Omega)$. In particular, we shall let $\mathcal{L}^+(\Omega) \coloneqq \{f \in \mathcal{L}(\Omega) : f \geq 0, f \neq 0\}$. For any subset B of Ω, we use I_B to denote its indicator gamble, that takes the value 1 on the elements of B and 0 otherwise.

Definition 1. A coherent lower prevision on $\mathcal{L}(\Omega)$ is a function $\underline{P} : \mathcal{L}(\Omega) \to \mathbb{R}$ satisfying the following properties:

(C1) $\underline{P}(f) \geq \inf f$;

(C2) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$;

(C3) $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$

for every $f, g \in \mathcal{L}(\Omega)$ and every $\lambda > 0$.

One example of a coherent lower prevision is the vacuous prevision with respect to a subset B of Ω, given by $\underline{P}(f) \coloneqq \inf_{\omega \in B} f(\omega)$.

A coherent lower prevision satisfying (C3) with equality for every $f, g \in \mathcal{L}(\Omega)$ is called a *linear prevision*. Linear previsions can be used to characterise coherence: a lower prevision $\underline{P}$ is coherent if and only if it is the lower envelope of its associated *credal set* $\mathcal{M}(\underline{P}) \coloneqq \{P \text{ linear prevision} : P(f) \geq \underline{P}(f)\ \forall f\}$, meaning that $\underline{P}(f) = \min\{P(f) : P \in \mathcal{M}(\underline{P})\}$ for every gamble f.

Given a partition $\mathcal{B}$ of Ω, a gamble $f \in \mathcal{L}(\Omega)$ is called *$\mathcal{B}$-measurable* when it is constant on the elements of $\mathcal{B}$. A *separately coherent conditional lower prevision* is a map $\underline{P}(\cdot|\mathcal{B})$ such that for every $B \in \mathcal{B}$, $\underline{P}(\cdot|B)$ is a coherent lower prevision satisfying $\underline{P}(B|B) = 1$, and where $\underline{P}(f|\mathcal{B})$ is the $\mathcal{B}$-measurable gamble given by $\underline{P}(f|\mathcal{B}) = \sum_{B \in \mathcal{B}} I_B \underline{P}(f|B)$.

Definition 2. Given a coherent lower prevision $\underline{P}$ and a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$, they are called (jointly) coherent when

(GBR) $\underline{P}(I_B(f - \underline{P}(f|B))) = 0$;

(CNG) $\underline{P}(f - \underline{P}(f|\mathcal{B})) \geq 0$

for every gamble f and every $B \in \mathcal{B}$.

The first of these conditions is called the *Generalised Bayes rule*, and determines $\underline{P}(f|B)$ uniquely when $\underline{P}(B) > 0$. The second is usually referred to as a *conglomerability* condition, and follows from the first when the partition $\mathcal{B}$ is finite.

In this paper, we will focus on the case where $\Omega \coloneqq \mathcal{X}_1 \times \mathcal{X}_2$. By an abuse of notation, we shall use $\underline{P}(\cdot|\mathcal{X}_1)$ to refer to a lower prevision conditional on the partition $\{\{x_1\} \times \mathcal{X}_2\}$ of $\mathcal{X}_1 \times \mathcal{X}_2$, and we shall say that a gamble is $\mathcal{X}_1$-measurable when it is measurable with respect to this partition. There is a one-to-one correspondence between $\mathcal{L}(\mathcal{X}_1)$ and the class of $\mathcal{X}_1$-measurable gambles (and also between $\mathcal{L}(\mathcal{X}_2)$ and the class of $\mathcal{X}_2$-measurable gambles); we shall use it throughout the paper to alleviate the notation.

In particular, we shall mention the notion of coherence of a joint lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ with conditional lower previsions $\underline{P}(\cdot|\mathcal{X}_1), \underline{P}(\cdot|\mathcal{X}_2)$; for the purposes of this paper, we only need that it implies the coherence of $\underline{P}$ with each of $\underline{P}(\cdot|\mathcal{X}_1), \underline{P}(\cdot|\mathcal{X}_2)$. A more detailed account can be found in [22, Section 7.1].

2.2 Sets of Desirable Gambles

A more general model than coherent lower previsions are *coherent sets of desirable gambles*:

Definition 3. A subset $\mathcal{R} \subseteq \mathcal{L}(\Omega)$ is *coherent* when it is a convex cone that includes $\mathcal{L}^+(\Omega)$ and does not include 0.

If $\mathcal{R}$ is a coherent set of desirable gambles, then the lower prevision given by

$$\underline{P}(f) \coloneqq \sup\{\mu : f - \mu \in \mathcal{R}\} \tag{1}$$

is coherent. On the other hand, there are several coherent sets of desirable gambles that induce the same coherent lower prevision. The smallest such set is called the set of *strictly desirable gambles*, and it is given by $\underline{\mathcal{R}} \coloneqq \mathcal{L}^+(\Omega) \cup \{f : \underline{P}(f) > 0\}$. On the other hand, the closure of any of these sets in the topology of uniform convergence is given

by $\overline{\mathcal{R}} := \{f : \underline{P}(f) \geq 0\}$, and this is called the set of *almost-desirable gambles* associated with $\underline{P}$.

Similarly, a coherent set of desirable gambles can also be used to define a separately coherent conditional lower prevision, by means of the formula

$$\underline{P}(f|B) := \sup\{\mu : I_B(f-\mu) \in \mathcal{R}\} \tag{2}$$

for every gamble f and every conditioning event B.

3 Irrelevant Products

Consider two possibility spaces $\mathcal{X}_1, \mathcal{X}_2$ and let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. Its marginal lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ are defined on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$ as the restriction of $\underline{P}$ to $\mathcal{X}_1$- and $\mathcal{X}_2$-measurable gambles, respectively.

The conditional information encompassed by $\underline{P}$ can be determined by many different updating rules (see for instance [22, Chapter 6] or [16]). In this paper we are using the updating rule determined by the natural extension:

Definition 4. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. For any $x_1 \in \mathcal{X}_1$, the *conditional natural extension* $\underline{E}(\cdot|x_1)$ is defined as

$$\underline{E}(f|x_1) := \begin{cases} \sup\{\mu : \underline{P}(I_{x_1}(f-\mu)) \geq 0\} & \text{if } \underline{P}(x_1) > 0 \\ \inf_{x \in \mathcal{X}_2} f(x_1, x) & \text{otherwise.} \end{cases} \tag{3}$$

Recall that when $\underline{P}(x_1) > 0$ the conditional lower prevision $\underline{E}(\cdot|x_1)$ is uniquely determined by (GBR). In general, $\underline{P}, \underline{E}(\cdot|\mathcal{X}_1)$ need not be coherent; when they are, $\underline{E}(\cdot|\mathcal{X}_1)$ is the smallest, or least committal, conditional lower prevision that is jointly coherent with $\underline{P}$. This is equivalent to the *conglomerability* of $\underline{P}$, a notion discussed in much detail in [22, Section 6.8].

Definition 5. $\underline{P}$ is said to *model* $\mathcal{X}_1$-$\mathcal{X}_2$ *irrelevance* when its conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ satisfies epistemic irrelevance, meaning that $\underline{E}(f|x_1) = \underline{E}(f|x_1')$ for every $\mathcal{X}_2$-measurable f and every $x_1, x_1' \in \mathcal{X}_1$.

Note that given marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$, we can always make an assessment of irrelevance and obtain a conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ by means of the formula

$$\underline{P}(f|x_1) := \underline{P}_{\mathcal{X}_2}(f(x_1, \cdot)) \ \forall f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2), \forall x_1 \in \mathcal{X}_1. \tag{4}$$

However, there may be no joint $\underline{P}$ modelling $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance and inducing some given $\underline{P}_{\mathcal{X}_1}, \underline{P}(\cdot|\mathcal{X}_1)$: the reason is that as soon as $\underline{P}_{\mathcal{X}_1}(x_1) = 0$ for some x_1 it follows from Eq. (3) that $\underline{P}(\cdot|x_1)$ should be vacuous, and then by irrelevance $\underline{P}_{\mathcal{X}_2}$ should be vacuous too.

When instead we can find such a $\underline{P}$, we say that $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ are conforming with an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model, or, more briefly, that they are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance:

Definition 6. We say that $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ are *conforming with* $\mathcal{X}_1$-$\mathcal{X}_2$ *irrelevance* when there is a coherent lower prevision $\underline{P}$ with marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ and whose conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ satisfies Eq. (4).

We shall denote $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ the set of coherent lower previsions satisfying the conditions of the definition above for given $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$.

It is interesting to remark that $\underline{P}$ may be an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model with marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ while its conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ does not satisfy Eq. (4):

Example 1. Consider $\mathcal{X}_1 := \mathcal{X}_2 := \{0,1\}$, and let $\underline{P}$ be the lower envelope of the linear previsions $\{P_1, P_2\}$ associated with the mass functions $\mathcal{X}_1 \times \mathcal{X}_2 = \{(0,0,0.5,0.5), (0.5,0.5,0,0)\}$ on $\{(0,0),(0,1),(1,0),(1,1)\}$. Then $\underline{P}_{\mathcal{X}_1}(0) = \underline{P}_{\mathcal{X}_1}(1) = 0$, so $\underline{E}(\cdot|\mathcal{X}_1)$ is vacuous and as a consequence it satisfies $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance. However, the $\mathcal{X}_2$-marginal of $\underline{P}$ is the linear prevision given by $P_{\mathcal{X}_2}(f) = (f(0) + f(1))/2$ for every $f \in \mathcal{L}(\mathcal{X}_2)$. Thus, Eq. (4) is not satisfied. ◊

In other words, the conditional natural extension may satisfy an irrelevance condition with respect to an unconditional coherent lower prevision different from the marginal of $\underline{P}$. This is why we are explicitly requiring this to hold in Definition 6.

Note also that, given the marginal coherent lower prevision $\underline{P}_{\mathcal{X}_1}$ on $\mathcal{L}(\mathcal{X}_1)$ and the conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ derived from $\underline{P}_{\mathcal{X}_2}$ by Eq. (4), Walley models their behavioural implications by means of the smallest lower prevision $\underline{E}$ that is coherent with them, and which in this case is given by the concatenation $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ [22, Theorem 8.1.7]. However, the conditional natural extension of $\underline{E}$ may not agree with $\underline{P}(\cdot|\mathcal{X}_1)$, whence it is arguable that with $\underline{E}$ we encompass assessments different from the ones we started with. Indeed, the notion of conformity differs from that of coherence of conditional and unconditional lower previsions considered by Walley (although the latter follows from conformity in the finite case). Conformity can be characterised as follows:

Proposition 1. *Let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ be two coherent lower previsions with respective domains $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$. Then $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})} \neq \varnothing$ if and only if either $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ for every x_1 or $\underline{P}_{\mathcal{X}_2}$ is vacuous.*

Proof. Let us start with the direct implication. Assume that $\underline{P}$ is a coherent lower prevision with marginal $\underline{P}_{\mathcal{X}_1}$ and whose conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ coincides with the conditional lower prevision that the marginal $\underline{P}_{\mathcal{X}_2}$ induces by means of (4). If there is some $x_1 \in \mathcal{X}_1$ such that $\underline{P}_{\mathcal{X}_1}(x_1) = 0$, it follows from Eq. (3) that $\underline{E}(\cdot|x_1)$ must be vacuous, and from Eq. (4) that $\underline{P}_{\mathcal{X}_2}$ must then be the vacuous lower prevision.

Conversely, consider the coherent lower prevision

$\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$, where $\underline{P}(\cdot|\mathcal{X}_1)$ is induced from $\underline{P}_{\mathcal{X}_2}$ by Eq. (4). It follows by definition that its marginals are $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. Thus, to see that it belongs to $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ under any of the conditions of the proposition statement, it suffices to show that in those cases $\underline{P}(\cdot|\mathcal{X}_1)$ coincides with the conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ of $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$.

By [22, Theorem 6.7.2], $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ is coherent with $\underline{P}(\cdot|\mathcal{X}_1)$. In particular, this means that $\underline{E}(\cdot|\mathcal{X}_1)$ is dominated by $\underline{P}(\cdot|\mathcal{X}_1)$. If $\underline{P}_{\mathcal{X}_2}$ is vacuous, then so is $\underline{P}(\cdot|\mathcal{X}_1)$, and as a consequence it is equal to $\underline{E}(\cdot|\mathcal{X}_1)$. On the other hand, if $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ for every $x_1 \in \mathcal{X}_1$, it follows from the coherence of $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ and $\underline{P}(\cdot|\mathcal{X}_1)$ that the latter is equal to $\underline{E}(\cdot|\mathcal{X}_1)$, because this is the only conditional lower prevision that satisfies (GBR) with $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$. □

This shows that conformity is quite a stringent notion, because the assumption $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ for every $x_1 \in \mathcal{X}_1$ can only hold when the space $\mathcal{X}_1$ is countable.

In other words, for most pairs of marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ there is no joint $\underline{P}$ whose conditional natural extension is the one that $\underline{P}_2$ induces by epistemic irrelevance. There are two reasons for this: one is the use of the natural extension as an updating rule; the other is that, as we have showed in Example 1, a $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model may induce a different conditional prevision than the one determined by its marginal and the notion of irrelevance.

Next, we are going to compare our definition above with another notion that has been considered in the literature [9]: it may be considered that $\underline{P}$ models irrelevance with respect to $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ when it is coherent (in the sense considered in Definition 2) with the conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ given by Eq. (4) and has marginal $\underline{P}_{\mathcal{X}_1}$.

In order to make this comparison, we prove first of all that the set $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ is closed under lower envelopes:

Proposition 2. *The lower envelope $\underline{P}$ of a family $\{\underline{P}^\lambda : \lambda \in \Lambda\}$ of elements of $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ also belongs to $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$.*

Proof. It follows trivially that the marginals of $\underline{P}$ are $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. Moreover, the conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$ of $\underline{P}$ must be dominated by that of each $\underline{P}^\lambda$, that is, $\underline{P}(\cdot|\mathcal{X}_1)$. Now we apply Proposition 1.

If $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ then $\underline{E}(f|x_1) = \sup\{\mu : \underline{P}(I_{x_1}(f-\mu)) \geq 0\} = \sup\{\mu : \underline{P}^\lambda(I_{x_1}(f-\mu)) \geq 0 \ \forall \lambda \in \Lambda\} = \underline{P}(f|x_1)$: it suffices to note that for every $\mu < \underline{P}(f|x_1)$ we obtain that $\underline{P}(I_{x_1}(f-\mu)) = \inf_{\lambda\in\Lambda} \underline{P}^\lambda(I_{x_1}(f-\mu)) \geq 0$, whence $\underline{E}(f|x_1) \geq \mu$ and as a consequence $\underline{E}(f|x_1) \geq \underline{P}(f|x_1)$.

On the other hand, if $\underline{P}_{\mathcal{X}_2}$ is vacuous then so is $\underline{P}(\cdot|\mathcal{X}_1)$, and as a consequence it coincides with $\underline{E}(\cdot|\mathcal{X}_1)$. Applying Proposition 1, we conclude that $\underline{P}$ is an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model. □

In this manner, we may define a *conforming natural extension*, and regard conformity as a structural assessment that can be made together with irrelevance. It models the implications of the assessments present in the marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ and our notion of conformity. In the finite case, it is easy to establish the following:

Proposition 3. *Consider marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$, and assume that $\mathcal{X}_1$ is finite. If $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})} \neq \emptyset$, then the smallest model in this set is $\underline{P}_{\mathcal{X}_1}(\underline{P}_{\mathcal{X}_2}) := \underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$, where $\underline{P}(\cdot|\mathcal{X}_1)$ is derived from $\underline{P}_{\mathcal{X}_2}$ by* (4)*.*

Proof. When $\mathcal{X}_1$ is finite, any coherent lower prevision $\underline{P}$ is coherent with its conditional natural extension $\underline{E}(\cdot|\mathcal{X}_1)$. Applying [22, Section 6.7.2], we deduce that any element of $\mathbb{P}^{irr}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ must dominate the marginal extension $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$.

Moreover, from the proof of Proposition 1 we see that when there is an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model that is conforming with $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$, then $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ is one such irrelevant model. These two facts imply that $\underline{P}_{\mathcal{X}_1}(\underline{P}_{\mathcal{X}_2})$ is the smallest irrelevant model. □

We shall refer to $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ as the *irrelevant natural extension* of $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. What Proposition 3 shows is that this coincides with the conforming natural extension in the finite case, *whenever the latter exists*.

Note that when $\mathcal{X}_1$ is infinite the above result may not hold, as we see from the next example. The key is that in our definition of conformity with $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance, we are not requiring the joint model to be coherent with the conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ that $\underline{P}_{\mathcal{X}_2}$ induces by means of Eq. (4):

Example 2. Let $\mathcal{X}_1 := \mathbb{N}, \mathcal{X}_2 := \{0,1\}$ and let P_1 be the linear prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ whose restriction to events is the σ-additive probability given by $P_1(n,0) = P_1(n,1) = \frac{1}{2^{n+1}} \ \forall n$. Let P_2 be a linear prevision whose restriction to events satisfies $P_2(\{2n : n \in \mathbb{N}\} \times \{0\}) = P_2(\{2n+1 : n \in \mathbb{N}\} \times \{1\}) = 0.5$ and $P_2(n,0) = P_2(n,1) = 0$ for every n. Let $P = 0.5(P_1 + P_2)$. Then $P(n) > 0$ for every $n \in \mathbb{N}$, and the conditional natural extension of P is given by $P(f|n) = 0.5f(n,0) + 0.5f(n,1) \ \forall f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. This shows that P is an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant model that is conforming with its marginals $P_{\mathcal{X}_1}, P_{\mathcal{X}_2}$.

To see that it does not coincide with the marginal extension $P_{\mathcal{X}_1}(P_{\mathcal{X}_2})$, take $A = \{2n : n \in \mathbb{N}\} \times \{1\}$. Then $P(A) = \frac{1}{8} \neq P_{\mathcal{X}_1}(P(A|\mathcal{X}_1)) = 0.5P_{\mathcal{X}_1}(\{2n : n \in \mathbb{N}\}) = 0.25$. This means that P is not coherent with the conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$, because it does not satisfy the notion of conglomerability. ◊

We can immediately characterise conformity in the precise case when $\mathcal{X}_1$ is finite:

Proposition 4. *When $\mathcal{X}_1$ is finite, two linear previsions $P_{\mathcal{X}_1}, P_{\mathcal{X}_2}$ are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance if and only if $P_{\mathcal{X}_1}(x_1) > 0$ for every x_1. In that case, the only conforming model is given by $P_{\mathcal{X}_1}(P_{\mathcal{X}_2})$.*

4 Independent Products

Similarly to the previous section, we say that $\underline{P}$ models $\mathcal{X}_1$-$\mathcal{X}_2$ independence when each of its conditional natural extensions satisfies epistemic irrelevance:

Definition 7. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ and let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ denote its marginals on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$. We say that $\underline{P}$ models $\mathcal{X}_1$-$\mathcal{X}_2$ *independence* when $\underline{E}(f|x_1) = \underline{E}(f|x_1')$ for every $\mathcal{X}_2$-measurable f and $x_1, x_1' \in \mathcal{X}_1$, and $\underline{E}(g|x_2) = \underline{E}(g|x_2')$ for every $\mathcal{X}_1$-measurable g and $x_2, x_2' \in \mathcal{X}_2$.

Similarly to what we did in the previous section, we may also study which pairs of marginals are conforming with a $\mathcal{X}_1$-$\mathcal{X}_2$ independent model, in the sense that its conditional natural extensions coincide with the conditionals $\underline{P}(\cdot|\mathcal{X}_1), \underline{P}(\cdot|\mathcal{X}_2)$ determined by $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ and the assumption of irrelevance. This produces the following definition:

Definition 8. Given two marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$ we say that they are *conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ independence* when there is a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ with marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ satisfying

$$\underline{E}(f|x_1) = \underline{P}_{\mathcal{X}_2}(f(x_1, \cdot)) \ \forall \mathcal{X}_2\text{-measurable } f$$
$$\underline{E}(f|x_2) = \underline{P}_{\mathcal{X}_1}(f(\cdot, x_2)) \ \forall \mathcal{X}_1\text{-measurable } f.$$

We shall denote $\mathbb{P}^{ind}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$ the set of coherent lower previsions satisfying the conditions of the definition above for given $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$.

From Proposition 1, we immediately derive the following:

Proposition 5. *Let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ be two coherent lower previsions with respective domains $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$. Then $\mathbb{P}^{ind}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})} \neq \varnothing$ if and only if (a) either $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ for every x_1 or $\underline{P}_{\mathcal{X}_2}$ is vacuous; and (b) either $\underline{P}_{\mathcal{X}_2}(x_2) > 0$ for every x_2 or $\underline{P}_{\mathcal{X}_1}$ is vacuous.*

In particular, it follows that if $\mathcal{X}_1, \mathcal{X}_2$ are uncountable, then the only marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ that are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ independence are the vacuous ones.

When two marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ independence, there are in general many different coherent lower previsions in $\mathbb{P}^{ind}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$. One example is given by the following family:

Definition 9. A coherent lower prevision $\underline{P}$ on $\mathcal{X}_1 \times \mathcal{X}_2$ is called an *independent product* of the marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ if and only if $\underline{P}, \underline{P}(\cdot|\mathcal{X}_1), \underline{P}(\cdot|\mathcal{X}_2)$ are coherent, where $\underline{P}(f|x_2) \coloneqq \underline{P}_{\mathcal{X}_1}(f(\cdot, x_2))$ and $\underline{P}(f|x_1) \coloneqq \underline{P}_{\mathcal{X}_2}(f(x_1, \cdot))$ for every $f \in \mathcal{X}_1 \times \mathcal{X}_2$.

Note, however, that independent products of given marginals always exist when $\mathcal{X}_1, \mathcal{X}_2$ are finite, whereas this is not the case for $\mathcal{X}_1$-$\mathcal{X}_2$ independent models, as shown by Proposition 5.

Proposition 6. *Assume that $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance (resp., independence). Then any independent product of $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ belongs to $\mathbb{P}^{ind}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})}$.*

Proof. Let us establish the result for $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevance; the proof for $\mathcal{X}_1$-$\mathcal{X}_2$ independence is analogous.

One the one hand, any independent product $\underline{P}$ is coherent with $\underline{P}(\cdot|\mathcal{X}_1)$ and has marginal $\underline{P}_{\mathcal{X}_1}$, so it suffices to show that it must induce $\underline{P}(\cdot|\mathcal{X}_1)$ by means of Eq. (3). If $\underline{P}_{\mathcal{X}_1}(x_1) > 0$ for every x_1, it follows from coherence that $\underline{P}(\cdot|x_1)$ coincides with $\underline{E}(\cdot|x_1)$; if $\underline{P}_{\mathcal{X}_1}(x_1) = 0$ for some x_1, then if $\underline{P}'$ is an $\mathcal{X}_1$-$\mathcal{X}_2$ irrelevant conforming joint it must be $\underline{P}'(\cdot|x_1)$ vacuous, whence $\underline{P}_{\mathcal{X}_2}$ is the vacuous lower prevision. But then since $\underline{P}$ is coherent with the vacuous conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$, it follows that it must be $\underline{E}(\cdot|x_1)$ vacuous even if $\underline{P}(x_1) > 0$. Thus, $\underline{P}(\cdot|\mathcal{X}_1)$ agrees with the conditional natural extension of $\underline{P}$. □

Independent products were studied in [9] in the case when $\mathcal{X}_1, \mathcal{X}_2$ are finite and in [19] when they are infinite. In particular, in [19, Theorem 3] it is proved that, if two marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ have an independent product, the smallest one corresponds to the smallest coherent lower prevision that dominates the two concatenations $\underline{P}_{\mathcal{X}_1}(\underline{P}_{\mathcal{X}_2}), \underline{P}_{\mathcal{X}_2}(\underline{P}_{\mathcal{X}_1})$. This coherent lower prevision is called the *independent natural extension*, and it is denoted by $\underline{P}_{\mathcal{X}_1} \otimes \underline{P}_{\mathcal{X}_2}$. Two interesting surveys of the notion of independence within imprecise probabilities are [5, 7].

The result above, together with Proposition 3, allows us to establish the following:

Proposition 7. *Assume that $\mathcal{X}_1, \mathcal{X}_2$ are finite, and let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ be two coherent lower previsions with respective domains $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$. If $\mathbb{P}^{ind}_{(\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2})} \neq \varnothing$ then the smallest element of this set is the independent natural extension $\underline{P}_{\mathcal{X}_1} \otimes \underline{P}_{\mathcal{X}_2}$.*

When both $\mathcal{X}_1, \mathcal{X}_2$ are infinite, independent products may not exist [19, Example 1]. Taking into account that this case is also problematic with respect to conformity, that reduces in most cases just to the combination of the vacuous marginal coherent lower previsions, in the remainder of this section we shall assume that at least one of $\mathcal{X}_1, \mathcal{X}_2$ is finite.

One particular family of independent products are the lower envelopes of factorising linear previsions:

Definition 10. A coherent lower prevision $\underline{P}$ with marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ is called an *independent envelope* when there is some $\mathcal{M} \subseteq \text{ext}(\mathcal{M}(\underline{P}_{\mathcal{X}_1})) \times \text{ext}(\mathcal{M}(\underline{P}_{\mathcal{X}_2}))$ such that $\underline{P} = \min\{P : P \in \mathcal{M}\}$.

The smallest independent envelope corresponds to the case where $\mathcal{M} = \text{ext}(\mathcal{M}(\underline{P}_{\mathcal{X}_1})) \times \text{ext}(\mathcal{M}(\underline{P}_{\mathcal{X}_2}))$. It is called the *strong product* of $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$, and we shall denote it $\underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}$. The strong product may not coincide with the independent natural extension [22, Section 9.3.4]; moreover, it is only guaranteed to exist when at least one of the possibility spaces is finite [19]: otherwise, the two products $P_1 \times P_2$ and $P_2 \times P_1$ of any marginal linear previsions may not coincide.

To see that the strong product is not the only independent envelope with given marginals, consider the following example:

Example 3. Consider $\mathcal{X}_1 := \{\omega_1, \omega_2\}, \mathcal{X}_2 := \{x_1, x_2\}$ and let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ be determined by $\underline{P}_{\mathcal{X}_1}(\omega_1) := 0.4$, $\overline{P}_{\mathcal{X}_1}(\omega_1) := 0.5$, $\underline{P}_{\mathcal{X}_2}(x_1) := 0.4$, $\overline{P}_{\mathcal{X}_2}(x_1) := 0.5$. Let $\underline{P}$ be the coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ given by $\underline{P} := \min\{P_1, P_2\}$, where P_1, P_2 are associated with the mass functions $(0.25, 0.25, 0.25, 0.25)$ and $(0.16, 0.24, 0.24, 0.36)$ on $\{(\omega_1, x_1), (\omega_1, x_2), (\omega_2, x_1), (\omega_2, x_2)\}$, respectively.

It follows from [22, Section 9.3.4] that $\underline{P}$ dominates the strong product of $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. Moreover, it is the lower envelope of two factorising previsions, and as a consequence [9, Section 4.3] it is an independent product of its marginals, that coincide with $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. To see that it does not coincide with the strong product $\underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}$, note that $\underline{P}((\omega_1, x_2)) = 0.24 > 0.2 = \underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}((\omega_1, x_2))$. ◊

A similar example (involving zero lower probabilities) can be found in [9, Example 3].

The independent natural extension and the strong product are the two most important independent products in the literature of imprecise probabilities: the first one, because it corresponds to the most conservative product under the notion of *epistemic independence*; and the second because it is the one that models adequately the notion of *strong independence* [5]. Indeed, if we want to give a sensitivity analysis interpretation, we see that if $\underline{P}$ is a lower envelope of precise models P that are conforming with $\mathcal{X}_1$-$\mathcal{X}_2$ independence then Proposition 4 implies that $\underline{P}$ must be an independent envelope; and the smallest such envelope is given by the strong product.

In [25, Theorem 28], we established that when $\mathcal{X}_2$ is finite a coherent lower prevision $\underline{P}$ with marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ is dominated by the strong product if and only if it satisfies the following condition:

$$\underline{P}(f) \leq \underline{P}(P_{\mathcal{X}_2}(f|\mathcal{X}_1)) \ \forall f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2), P_{\mathcal{X}_2} \geq \underline{P}_{\mathcal{X}_2}, \quad (5)$$

where $P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1)$ is derived from $P_{\mathcal{X}_2}$ using Eq. (4).

In particular, this property is satisfied by the independent natural extension. However, it is not a sufficient condition for independence, as we show next:

Example 4. Consider $\mathcal{X}_1 := \{0,1\} =: \mathcal{X}_2$, and let $\underline{P}$ be the coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ given by $\underline{P}(f) := \min\left\{\frac{f(0,0)+f(0,1)}{2}, \frac{f(1,0)+f(1,1)}{2}, \frac{f(0,0)+f(1,1)}{2}\right\}$. Then the marginals of $\underline{P}$ are $\underline{P}_{\mathcal{X}_1}(f) = \min\{f(0), f(1)\}, P_{\mathcal{X}_2}(g) = \frac{g(0)+g(1)}{2}$ for every $f \in \mathcal{L}(\mathcal{X}_1), g \in \mathcal{L}(\mathcal{X}_2)$. The strong product of these marginals is given by $\underline{P}_{\mathcal{X}_1} \boxtimes P_{\mathcal{X}_2}(f) = \min\left\{\frac{f(0,0)+f(0,1)}{2}, \frac{f(1,0)+f(1,1)}{2}\right\}$ for every $f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. Moreover, it follows from [9, Proposition 25] that $\underline{P}_{\mathcal{X}_1} \boxtimes P_{\mathcal{X}_2}$ is the only independent product of these marginals. Since it dominates $\underline{P}$, we deduce from [25, Theorem 28] that $\underline{P}$ satisfies (5). However, they do not coincide, since $\underline{P}(\{(0,1),(1,0)\}) = 0 < 0.5 = (\underline{P}_{\mathcal{X}_1} \boxtimes P_{\mathcal{X}_2})((0,1),(1,0))$, and as a consequence $\underline{P}$ is not an independent product. ◊

Note also that in general being an independent product is not sufficient for condition (5), since there exist independent products that dominate the strong product; one instance is given in Example 3.

Next we discuss another condition that has been linked with independent products, as an attempt to give a behavioural interpretation of the strong product. For every $\omega \in \mathcal{X}_1$ and $f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$, let us define

$$\begin{aligned} f^\omega : \mathcal{X}_1 \times \mathcal{X}_2 &\to \mathbb{R} \\ (\omega', x) &\mapsto f(\omega, x). \end{aligned}$$

Proposition 8. *[25, Proposition 30] Let $\mathcal{X}_1, \mathcal{X}_2$ be finite spaces, and let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. Then $\underline{P}$ is an independent envelope if and only if*

$$\underline{P}(g - f) \geq \min_{\omega \in \mathcal{X}_1} \underline{P}(g - f^\omega) \ \forall g, f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2). \quad (6)$$

The condition above can be regarded as a kind of dissociation between the marginal and conditional beliefs, and in the case of linear previsions it can be written more simply as $P(f) \leq \max_{\omega \in \mathcal{X}_1} P(f^\omega) = \max_{\omega \in \mathcal{X}_1} P_{\mathcal{X}_2}(f(\omega, \cdot)) \ \forall f \in \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. From Proposition 8 it follows that we can regard the strong product $\underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}$ as the smallest coherent lower prevision that satisfies (6).

Note however, not every coherent lower prevision that dominates the strong product satisfies (6), as we can see from the following example:

Example 5. Consider the spaces $\mathcal{X}_1, \mathcal{X}_2$ and the marginal coherent lower previsions $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ from Example 3. By [22, Example 9.3.4] the strong product $\underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}$ of $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ is given by $\underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2} = \min\{P_1, P_2, P_3, P_4\}$, where these are associated with the mass functions $(0.16, 0.24, 0.24, 0.36), (0.25, 0.25, 0.25, 0.25)$, $(0.2, 0.3, 0.2, 0.3)$ and $(0.2, 0.2, 0.3, 0.3)$ on $\{(\omega_1, x_1), (\omega_1, x_2), (\omega_2, x_1), (\omega_2, x_2)\}$, respectively.

Consider the coherent lower prevision $\underline{P} = \min\{P_1, P_2, 0.5P_3 + 0.5P_4\}$, that obviously dominates the strong product. To see that it does not satisfy (6), consider the gambles $g = (1,0,-9,-0.81), f = (0,0,1,-0.81)$, where the vector is made up of the images of $(\omega_1,x_1),(\omega_1,x_2),(\omega_2,x_1),(\omega_2,x_2)$, respectively. Then it holds that:

	(ω_1,x_1)	(ω_1,x_2)	(ω_2,x_1)	(ω_2,x_2)
$g-f$	1	0	-10	0
$g-f^{\omega_1}$	1	0	-9	-0.81
$g-f^{\omega_2}$	0	0.81	-10	0

and from this we derive that $\underline{P}(g-f) = -2.3, \underline{P}(g-f^{\omega_1}) = -2.293$ and $\underline{P}(g-f^{\omega_2}) = -2.2975$. Thus, we see that $\underline{P}(g-f) < \min_{\omega\in\mathcal{X}_1}\underline{P}(g-f^{\omega})$, so (6) does not hold. ◊

This is another way of showing that not every model dominating the strong product of its marginals is an independent envelope.

Condition (6) can be equivalently expressed in terms of the set of *almost-desirable* gambles associated with the coherent lower prevision $\underline{P}$. More generally, when $\underline{P}$ is induced by a set of desirable gambles $\mathcal{R}$, then the analogous condition would be given by

$$g - f^{\omega} \in \mathcal{R}\ \forall \omega \in \mathcal{X}_1 \Rightarrow g - f \in \mathcal{R}.$$

Note that this is not equivalent to Eq. (6), as we can tell from the fact that it is not always satisfied by the strong product:

Example 6. Let P be the uniform linear prevision on $\{\omega_1,\omega_2\}\times\{x_1,x_2\}$, which is trivially the strong (and only) product of its marginals. Consider the set of gambles $\mathcal{R} := \{f : P(f) > 0\} \cup \{f : P(f) = 0, f(\omega_1,x_2) + f(\omega_2,x_1) > 0\}$. It is easy to check that $\mathcal{R}$ is a coherent set of desirable gambles and that its associated coherent lower prevision coincides with P.

Consider now the gambles f, g given by:

	(ω_1,x_1)	(ω_1,x_2)	(ω_2,x_1)	(ω_2,x_2)
f	-1	3	3	-1
g	1	2	1	0.

Then the gambles $g-f^{\omega_1}, g-f^{\omega_2}$ and $g-f$ are given by:

	(ω_1,x_1)	(ω_1,x_2)	(ω_2,x_1)	(ω_2,x_2)
$g-f^{\omega_1}$	2	-1	2	-3
$g-f^{\omega_2}$	-2	3	-2	1
g-f	2	-1	-2	1

It follows that $g-f^{\omega_1}, g-f^{\omega_2} \in \mathcal{R}$ but $g-f$ does not. ◊

On the other hand, under some conditions the strong product satisfies an analogous condition for sets of strictly desirable gambles:

Proposition 9. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1\times\mathcal{X}_2)$, where $\mathcal{X}_1,\mathcal{X}_2$ are finite spaces, and let $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$ denote the marginals of $\underline{P}$, respectively. If $\underline{P} = \underline{P}_{\mathcal{X}_1} \boxtimes \underline{P}_{\mathcal{X}_2}$ and $\underline{P}_{\mathcal{X}_1}(\omega) > 0$ for every $\omega \in \mathcal{X}_1$, then it satisfies*

$$g - f^{\omega} \in \underline{\mathcal{R}}\ \forall \omega \in \mathcal{X}_1 \Rightarrow g - f \in \underline{\mathcal{R}}.$$

Proof. Consider gambles $f, g \in \mathcal{L}(\mathcal{X}_1\times\mathcal{X}_2)$ such that $g - f^{\omega} \in \underline{\mathcal{R}}$ for every $\omega \in \mathcal{X}_1$, and let us define $\mathcal{X}_1' := \{\omega \in \mathcal{X}_1 : g - f^{\omega} \not\geq 0\}$.

If $\mathcal{X}_1' = \varnothing$, then $g - f^{\omega} \geq 0$ for all ω, whence $I_{\omega}(g-f^{\omega}) \geq 0\ \forall\omega$ and therefore $g - f = \sum_{\omega} I_{\omega}(g-f^{\omega}) \geq 0$. But it cannot be $g = f$ because in that case the inequality $g - g^{\omega} \geq 0$ for every ω would imply that g is $\mathcal{X}$-measurable and then $g = g^{\omega}$ for all ω, a contradiction with the coherence of $\underline{\mathcal{R}}$.

Next, if $\mathcal{X}_1' \neq \varnothing$, then given $\omega \in \mathcal{X}_1'$ it must be $\underline{P}(g-f^{\omega}) > 0$, whence there is some $\delta_{\omega} > 0$ such that $\underline{P}(g-f^{\omega}) > \delta_{\omega} > 0$. This means that for every $P_{\mathcal{X}_1} \geq \underline{P}_{\mathcal{X}_1}$ and every $P_{\mathcal{X}_2} \geq \underline{P}_{\mathcal{X}_2}$ it holds that $(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(g) \geq (P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(f^{\omega}) + \delta_{\omega} = P_{\mathcal{X}_2}(f(\omega,\cdot)) + \delta_{\omega}$. On the other hand, given $\omega \notin \mathcal{X}_1'$, $g - f^{\omega} \geq 0$, whence $(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(g) \geq (P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(f^{\omega}) = P_{\mathcal{X}_2}(f(\omega,\cdot))$. We deduce that

$$\begin{aligned}
(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(f) &= \sum_{\omega\in\mathcal{X}_1, x\in\mathcal{X}_2} P_{\mathcal{X}_1}(\omega)P_{\mathcal{X}_2}(x)f(\omega,x)\\
&\leq \sum_{\omega\in\mathcal{X}_1'} P_{\mathcal{X}_1}(\omega)(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2}(g) - \delta_{\omega})\\
&\quad + \sum_{\omega\notin\mathcal{X}_1'} P_{\mathcal{X}_1}(\omega)(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2}(g))\\
&= (P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(g) - \sum_{\omega\in\mathcal{X}_1'} \delta_{\omega}P_{\mathcal{X}_1}(\omega),
\end{aligned}$$

whence $(P_{\mathcal{X}_1}\times P_{\mathcal{X}_2})(g-f) \geq \sum_{\omega\in\mathcal{X}_1'}\delta_{\omega}P_{\mathcal{X}_1}(\omega) \geq \sum_{\omega\in\mathcal{X}_1'}\delta_{\omega}\underline{P}_{\mathcal{X}_1}(\omega)$; from this $\underline{P}(g-f) \geq \sum_{\omega\in\mathcal{X}_1'}\delta_{\omega}\underline{P}_{\mathcal{X}_1}(\omega) > 0$, and therefore $g - f \in \underline{\mathcal{R}}$. □

To see that this does not always hold without the assumption of positive lower probabilities in $\mathcal{X}_1$, consider the following example:

Example 7. Let $\mathcal{X}_1 := \{\omega_1,\omega_2\}, \mathcal{X} := \{x_1,x_2\}$ and let $P_{\mathcal{X}_1}, P_{\mathcal{X}_2}$ be the linear previsions on $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_2)$ associated with the mass functions $P_{\mathcal{X}_1}(\omega_1) = 1 = P_{\mathcal{X}_2}(x_1)$. Consider the gambles f, g given by

	(ω_1,x_1)	(ω_1,x_2)	(ω_2,x_1)	(ω_2,x_2)
g	1	2	2	2
f	1	1	0	4

We obtain

	(ω_1,x_1)	(ω_1,x_2)	(ω_2,x_1)	(ω_2,x_2)
$g-f^{\omega_1}$	0	1	1	1
$g-f^{\omega_2}$	1	-2	2	-2
$g-f$	0	1	2	-2

Then $g - f^{\omega_1}$ is strictly desirable because it is non-negative; $g - f^{\omega_2}$ is strictly desirable because $(P_{\mathcal{X}_1} \times P_{\mathcal{X}_2})(g - f^{\omega_2}) = (g - f^{\omega_2})(\omega_1, x_1) = 1 > 0$; and $g - f$ is not strictly desirable because $(P_{\mathcal{X}_1} \times P_{\mathcal{X}_2})(g - f) = 0$ and it is not a non-negative gamble. Hence, the product $P_{\mathcal{X}_1} \times P_{\mathcal{X}_2}$ does not satisfy

$$g - f^{\omega} \in \mathcal{R}\ \forall \omega \in \mathcal{X}_1 \Rightarrow g - f \in \mathcal{R}. \lozenge$$

5 Conformity of Marginal and Conditional Models

In the previous sections we have studied to which extent the notion of conformity can be imposed together with a structural assessment of epistemic irrelevance or independence. Next we study in more detail the properties of conformity, without making any structural assumptions.

Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$, and let $\underline{P}(\cdot|\mathcal{X}_1)$ be its conditional natural extension. It follows from [22, Theorem 6.8.2] that $\underline{P}$ is coherent with $\underline{P}(\cdot|\mathcal{X}_1)$ if and only if it is $\mathcal{X}_1$-conglomerable, and that this condition holds trivially when $\mathcal{X}_1$ is finite.

Similarly to what we discussed in Section 3, given a coherent lower prevision $\underline{P}_{\mathcal{X}_1}$ on $\mathcal{L}(\mathcal{X}_1)$ and a conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$, we say that they are *conforming* when there is a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$ with marginal $\underline{P}_{\mathcal{X}_1}$ and conditional natural extension $\underline{P}(\cdot|\mathcal{X}_1)$.

It is easy to see that not every marginal and conditional models are conforming with a joint model $\underline{P}$ in the manner depicted above: this follows from the fact that if $\underline{P}(x_1) = 0$ then the conditional lower prevision $\underline{P}(\cdot|x_1)$ determined by natural extension must be vacuous. In fact, with a similar proof to that of Proposition 1, it is possible to show the following:

Proposition 10. *Let $\underline{P}_{\mathcal{X}_1}, \underline{P}(\cdot|\mathcal{X}_1)$ be a marginal and a conditional lower prevision with respective domains $\mathcal{L}(\mathcal{X}_1), \mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. Then $\underline{P}_{\mathcal{X}_1}, \underline{P}(\cdot|\mathcal{X}_1)$ are conforming if and only if $\underline{P}(\cdot|x_1)$ is vacuous whenever $\underline{P}_{\mathcal{X}_1}(x_1) = 0$.*

Moreover, when $\underline{P}_{\mathcal{X}_1}$ and $\underline{P}(\cdot|\mathcal{X}_1)$ are conforming with some joint model, they they may be conforming with more than one. In the finite case, the smallest of these is determined by the notion of marginal extension.

Proposition 11. *Assume $\mathcal{X}_1$ is finite, and let $\underline{P}_{\mathcal{X}_1}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1)$ and $\underline{P}(\cdot|\mathcal{X}_1)$ a conditional lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$. If there is some $\underline{P}$ conforming with $\underline{P}_{\mathcal{X}_1}, \underline{P}(\cdot|\mathcal{X}_1)$, then the smallest such model is given by $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$.*

Proof. It follows from [22, Section 6.7.2] that $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ is coherent with $\underline{P}(\cdot|\mathcal{X}_1)$ and has marginal $\underline{P}_{\mathcal{X}_1}$, so it only remains to show that it induces $\underline{P}(\cdot|\mathcal{X}_1)$ by means of Eq. (3).

If $\underline{P}_{\mathcal{X}_1}(x_1) = 0$, then by Proposition 10 $\underline{P}(\cdot|x_1)$ is vacuous, whence by Eq. (3) it coincides with the conditional natural extension $\underline{E}(\cdot|x_1)$ of $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$. On the other hand, if $\underline{P}(x_1) > 0$ then it follows from the coherence of $\underline{P}(\underline{P}(\cdot|\mathcal{X}_1))$ and $\underline{P}(\cdot|\mathcal{X}_1)$ that $\underline{P}(\cdot|x_1)$ is uniquely determined by $\underline{P}(f|x_1) := \sup\{\mu : \underline{P}(I_{\{x_1\}\times\mathcal{X}_2}(f - \mu)) \geq 0\}$, and so it coincides with $\underline{E}(\cdot|x_1)$.

Finally, note that if $\mathcal{X}_1$ is finite then any joint model $\underline{P}$ that is conforming with $\underline{P}_{\mathcal{X}_1}, \underline{P}(\cdot|\mathcal{X}_1)$ is in particular coherent with $\underline{P}(\cdot|\mathcal{X}_1)$, and as a consequence it must dominate the marginal extension $\underline{P}_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$. □

To see that the result may not hold when $\mathcal{X}_1$ is infinite, we refer to Example 2.

In the precise case, the conditional natural extension $\underline{P}(\cdot|\mathcal{X}_1)$ of a linear prevision is precise if and only if $P(x_1) > 0$ for every $x_1 \in \mathcal{X}_1$, and then $P = P_{\mathcal{X}_1}(\underline{P}(\cdot|\mathcal{X}_1))$ if and only if P is $\mathcal{X}_1$-disintegrable [11] (which holds trivially when $\mathcal{X}_1$ is finite).

With respect to the two conditions we discussed in the previous section, in general a marginal extension may not satisfy Eq. (5). To see this, it suffices to consider a linear prevision that is not the product of its marginals, as in the following example:

Example 8. Consider $\mathcal{X}_1 := \{\omega_1, \omega_2\}, \mathcal{X}_2 := \{x_1, x_2\}$ and let P be the linear prevision associated with the mass function $P(\{\omega_1, x_2\}) := P(\{\omega_2, x_1\}) := P(\{\omega_2, x_2\}) := \frac{1}{3}$. Since it is a linear prevision then it is a marginal extension model.

Consider $f := -I_{\{(\omega_1,x_1),(\omega_2,x_2)\}}$. Then $P(f) = -\frac{1}{3}$. Since the marginal of P is given by $P_{\mathcal{X}_1}(\omega_1) = \frac{1}{3}, P_{\mathcal{X}_1}(\omega_2) = \frac{2}{3}$, we obtain $P(P_{\mathcal{X}_2}(f)) = P(\omega_1) \cdot (-\frac{1}{3}) + P(\omega_2) \cdot (-\frac{2}{3}) = -\frac{5}{9} < P(f). \lozenge$

In fact, for a linear prevision P on a finite space $\mathcal{X}_1 \times \mathcal{X}_2$ it can be checked that conditions (5) and (6) are each of them equivalent to P being the product of its marginals [25].

When just one of the marginals is linear, condition (6) can be used to characterise the equality between the marginal extension and the strong product, as we show next:

Proposition 12. *Consider finite spaces $\mathcal{X}_1, \mathcal{X}_2$, and let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\mathcal{X}_1 \times \mathcal{X}_2)$, with respective marginals $\underline{P}_{\mathcal{X}_1}, \underline{P}_{\mathcal{X}_2}$. If $\underline{P}_{\mathcal{X}_2}$ is linear, then $\underline{P}$ satisfies* (6) *if and only if $\underline{P} = \underline{P}_{\mathcal{X}_1}(P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1))$.*

Proof. By [9, Proposition 25], when $P_{\mathcal{X}_2}$ is linear there is only one independent product of $\underline{P}_{\mathcal{X}_1}, P_{\mathcal{X}_2}$: the marginal extension $\underline{P}_{\mathcal{X}_1}(P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1))$, which coincides thus with the strong product.

Now, by Proposition 8, if $\underline{P}$ satisfies (6) then it is an independent envelope, and as a consequence it is an independent product. This means that $\underline{P}$ must agree with $\underline{P}_{\mathcal{X}_1}(P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1))$. Conversely, if $\underline{P}$ is equal to $\underline{P}_{\mathcal{X}_1}(P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1))$ then it also coincides with the strong product, and by Proposition 8 it satisfies Eq. (6). □

On the other hand, if $P_{\mathcal{X}_2}$ is linear then Eq. (5) holds if and only if $\underline{P} \leq \underline{P}_{\mathcal{X}_1}(P_{\mathcal{X}_2}(\cdot|\mathcal{X}_1))$. This would mean that the marginal extension satisfies Eq. (5), although it may not be the only one to do so: for a counterexample, consider again the coherent lower prevision $\underline{P}$ from Example 4.

We conclude this section by giving a property of the marginal extension in terms of sets of desirable gambles:

Proposition 13. *Assume $\mathcal{X}_1$ is finite, and let $\mathcal{R}$ be a coherent set of gambles on $\mathcal{X}_1 \times \mathcal{X}_2$, and let $\underline{P}, \underline{P}(\cdot|\mathcal{X}_1)$ be the lower previsions it induces by means of Eqs.* (1)*,* (2)*. Then $\underline{P} \geq \underline{P}(\underline{P}(\cdot|\mathcal{X}_1))$, and they coincide only if $\mathcal{R}$ is negatively additive, meaning that*

$$(\forall \omega \in \mathcal{X}_1) I_{\{\omega\}} g \notin \mathcal{R} \Rightarrow g \notin \mathcal{R}. \tag{7}$$

Proof. The inequality $\underline{P} \geq \underline{P}(\underline{P}(\cdot|\mathcal{X}_1))$ holds because $\underline{P}$ is coherent with the conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ (use for instance [18, Thm. 8]), and applying then [22, Thm. 6.7.2].

Assume that $\mathcal{R}$ is not negatively additive, so that Eq. (7) is violated for some gamble f. Then it follows that $\underline{P}(f) > \max_{\omega \in \mathcal{X}_1} \underline{P}(f|\omega)$, whence for every $P \geq \underline{P}$ it holds that

$$P(\underline{P}(f|\mathcal{X}_1)) \leq \max_{\omega \in \mathcal{X}_1} \underline{P}(f|\omega) < \underline{P}(f),$$

whence $\underline{P}(\underline{P}(f|\mathcal{X}_1)) \leq \overline{P}(\underline{P}(f|\mathcal{X}_1)) < \underline{P}(f)$. □

To see that negative additivity is not sufficient for $\underline{P}$ to be a marginal extension, consider the following example:

Example 9. Let $\mathcal{X}_1 := \{\omega_1, \omega_2\}, \mathcal{X}_2 := \{x_1, x_2\}$ and let $\mathcal{H} = \{A \subseteq \mathcal{X}_1 \times \mathcal{X}_2 : |A| = 2\}$. For any $A \in \mathcal{H}$, define P_A as $P_A(f) := (\sum_{z \in A} f(z))/2$, and let $\underline{P}$ be the lower envelope of the family $\{P_A : A \in \mathcal{H}\}$. Consider $\mathcal{R}$ its associated set of strictly desirable gambles. To see that it is negatively additive, note that its associated conditional lower prevision $\underline{P}(\cdot|\mathcal{X}_1)$ is vacuous and that $\underline{P}, \underline{P}(\cdot|\mathcal{X}_1)$ satisfy the condition

$$\underline{P}(f) \leq \max\{\underline{P}(f|\omega_1), \underline{P}(f|\omega_2)\} \ \forall f; \tag{8}$$

It is not difficult to show that Eq. (8) is equivalent to the negative additivity of $\mathcal{R}$.

However, if we consider the gamble f given by $f(\omega_1, x_1) = 1, f(\omega_1, x_2) = 2, f(\omega_2, x_1) = 3, f(\omega_2, x_2) = 4$, we obtain that

$$\underline{P}(f) = 1.5 > 1 = \underline{P}(\underline{P}(f|\mathcal{X}_1))$$

and as a consequence $\underline{P}$ is not a marginal extension. ◊

A slightly related result can be found in [24, Proposition 13]: it is shown there that negative additivity implies the notion of *temporal consistency* between a coherent set of gambles $\mathcal{R}$ and its associated set of conditional beliefs, when the set $\mathcal{R}$ is defined by means of marginal extension.

6 Conclusions

The results in this paper show that the notion of conformity clashes (except in the vacuous case) with the existence of zero lower probabilities, which appear quite often within the theory of imprecise probabilities, and which have been the object or quite some discussion (see for instance [22, Section 6.10], [4]); note moreover that within our framework we obtain zero lower probabilities as soon as the conditioning space is uncountable.

One possible alternative that may help to deal better with this issue would be to use a different updating rule, such as regular extension, that produces more informative inferences and that in particular does not imply that the marginal lower previsions are vacuous as soon as one element has lower probability zero. The study of conformity under this scenario is left as an open problem.

If we restrict our attention to finite spaces, then it is easy to see that a structural assessment of conformity gives rise to the marginal extension, and in case it is combined with assessments of epistemic irrelevance and independence it produces the notions of irrelevant and independent natural extension. This has led us to study in more detail the properties of independent products, and more particularly those that are lower envelopes of factorising linear previsions: the independent envelopes. This allows us to give our notion a sensitivity analysis interpretation that usually gets lost when we move from the unconditional to the conditional case. We have considered two properties that imply that a lower prevision is dominated, and dominates, the strong product, respectively, and have shown that under some conditions they are satisfied by the marginal extension, too.

There are several lines of research that we can derive from the results in this paper: on the one hand, we should study the conformity of more than two (marginal or conditional) models. This has been studied from the point of view of coherence in [17, Section 8.2], where it was shown that Walley's *weak coherence* is quite related to the works in [2, 13] and also to the notion of *satisfiability* [14, 15].

Note also that in our treatment of epistemic irrelevance and independence we have only considered conditional information on the singletons; it would be interesting to consider a more general setting where we condition on arbitrary subsets of the possibility spaces $\mathcal{X}_1, \mathcal{X}_2$.

On the other hand, it would also be interesting to make a similar study using the more general language of *sets of desirable gambles*, that may help overcome some of the issues related to conditioning on sets of (lower) probability zero. We expect that links with the irrelevant and the independent natural extension for sets of gambles considered in [6, 8, 20] should arise in this context.

Acknowledgments

The authors acknowledge financial support from project TIN2014-59543-P and from the Swiss NSF n. 200021_146606 / 1. We would like also to thank especially one of the reviewers for pointing out a mistake in a proof in the first version of this paper.

References

[1] T. Augustin, F. Coolen, G. de Cooman and M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, 2014.

[2] G. Boole. *The Laws of Thought*. Dover Publications, New York, 1847, reprint 1961.

[3] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953–1954.

[4] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting*. Kluwer, 2002.

[5] I. Couso, S. Moral and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[6] J. de Bock and G. de Cooman. Credal networks under epistemic irrelevance: The sets of desirable gambles approach. *International Journal of Approximate Reasoning*, 56(B):178–207, 2015.

[7] L. M. de Campos and S. Moral. Independence concepts for convex sets of probabilities. In *Proceedings of 11th. Conference on Uncertainty in Artificial Intelligence*, pages 108–115, San Mateo (EEUU), 1995.

[8] G. de Cooman and E. Miranda. Irrelevance and independence for sets of desirable gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012.

[9] G. de Cooman, E. Miranda and M. Zaffalon. Independent natural extension. *Artificial Intelligence*, 175(12–13):1911–1950, 2011.

[10] B. de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester, 1974–1975.

[11] L. E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3:88–99, 1975.

[12] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

[13] M. Fréchet. Sur les tableaux de correlation dont les marges sont données. *Ann. Univ. Lyon, Section A, Series 3*, 14:53–77, 1951.

[14] P. Hansen, B. Jaumard, M. Poggi de Aragão, F. Chauny and S. Perron. Probabilistic satisfability with imprecise probabilities. *International Journal of Approximate Reasoning*, 24(2–3):171–189, 2000.

[15] B. Jaumard, H. Hansen, and M. Poggi de Aragão. Column generation methods for probabilistic logic. *ORSA Journal on Computing*, 3:135–148, 1991.

[16] E. Miranda. Updating coherent previsions on finite spaces. *Fuzzy Sets and Systems*, 160(9):1286–1407, 2009.

[17] E. Miranda and M. Zaffalon. Coherence graphs. *Artificial Intelligence*, 173(1):104–144, 2009.

[18] E. Miranda and M. Zaffalon. Notes on desirability and conditional lower previsions. *Annals of Mathematics and Artificial Intelligence*, 60(3-4):251–309, 2010.

[19] E. Miranda and M. Zaffalon. Independent products in infinite spaces. *Journal of Mathematical Analysis and Applications*, 425(1):460–488, 2015.

[20] S. Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45:197–214, 2005.

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

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

[23] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975.

[24] M. Zaffalon and E. Miranda. Probability and time. *Artificial Intelligence*, 198(1):1–51, 2013.

[25] M. Zaffalon and E. Miranda. Desirability and the birth of incomplete preferences. 2015. Submitted for publication.

Comonotone Lower Probabilities for Bivariate and Discrete Structures

Ignacio Montes and **Sebastien Destercke**
Technologic University of Compiegne, France
ignacio.montes@hds.utc.fr sebastien.destercke@hds.utc.fr

Abstract

Two random variables are called comonotone when there is an increasing relation between them, in the sense that when one of them increases (decreases), the other one also increases (decreases). This notion has been widely investigated in probability theory, and is related to the theory of copulas. This contribution studies the notion of comonotonicity in an imprecise setting. We define comonotone lower probabilities and investigate its characterizations. Also, we provide some sufficient conditions allowing to define a comonotone belief function with fixed marginals and characterize comonotone bivariate p-boxes.

Keywords. Comonotonicity, copulas, lower probabilities, belief functions, p-boxes.

1 Introduction

Random variables are usual tools in probability theory when modeling uncertainty. When dealing with two random variables, Sklar's Theorem [10] tells us that the joint distribution function can be expressed in terms of the marginals by means of a function called copula [7]. Thus, the copula gathers the information concerning the possible dependence between the random variables. When there is no dependence between them, we talk about independent random variables, and the copula associated with those variables is the product. The extreme cases of dependence between variables are related to situations in which either there is an increasing or decreasing relation between them. In the former case, this means that when the value of one variable increases the other variable also increases, while in the second scenario when the value of one variable increases, the value of the other variable decreases. They are referred as comonotone and countermonotone random variables, respectively, and the associated copulas are the minimum and the Łukasiewicz operator.

In this work we shall assume the existence of an imprecisely known probability and we shall use coherent lower probabilities to model it. Lower probabilities are one of the models within the theory of Imprecise Probabilities introduced by Walley [12], as well as belief functions [9], possibilities [3, 13] or uni- and bivariate p-boxes [11, 8], all of them particular families of coherent lower probabilities.

The aim of this paper is to extend the notion of comonotonicity to coherent lower probabilities and to investigate the particular cases in which the lower probability is a belief function or is associated with a bivariate p-box.

After introducing some preliminary notions related to lower probabilities and copulas in Section 2, Section 3 investigates the definition and characterizations of comonotonicity for coherent lower probabilities. We shall see that, in contrast with the precise framework, not any two marginal coherent lower probabilities allow us to define a joint comonotone coherent lower probability. Thus, in Section 4 we investigate some conditions under which this property is satisfied for the particular case of belief functions. In Section 5 we consider bivariate p-boxes and we characterize the conditions they must satisfy to ensure that its associated lower probability is comonotone.

2 Preliminaries

In this section we introduce some preliminary notions that will be useful throughout the paper. First of all we introduce lower probabilities [12], which are very useful to model situations in which a probability is imprecisely defined. Other model related to lower probabilities is that of p-boxes [4], which are used to model the imprecise knowledge to cumulative distribution functions. Univariate p-boxes are connected to belief functions, which play a key role in Shafer's Theory of Evidence [9]. Finally, we also introduce possibility measures [3], which can be embedded both

into the Theory of Evidence and the Fuzzy Set Theory.

Secondly we introduce the main notions of the Theory of Copulas [7] and we explain the problem we are dealing in this paper.

2.1 Lower Probabilities

A *lower probability* [12] is a function $\underline{P} : \mathcal{K} \to [0,1]$, where $\mathcal{K} \subseteq \mathcal{P}(\Omega)$. $\underline{P}(A)$ can be interpreted as the subject's supremum acceptable buying price for the bet A, in the sense that we obtain 1 if A happens and 0 otherwise. Any lower probability defines, using a conjugacy relation, an *upper probability* $\overline{P} : \mathcal{K}^c \to [0,1]$, where $\mathcal{K}^c = \{A^c : A \in \mathcal{K}\}$, by:

$$\overline{P}(A) = 1 - \underline{P}(A^c) \quad \forall A \in \mathcal{K}^c.$$

Any lower probability defines a set of probabilities, usually called *credal set*, given by:

$$\mathcal{M}(\underline{P}) = \{P \text{ prob.} \mid \underline{P}(A) \leq P(A) \leq \overline{P}(A)\}.$$

Some consistency requirements are usually imposed on lower probabilities. The most usual one is *coherence*: a lower probability $\underline{P}$ is coherent when

$$\underline{P}(A) = \min_{P \in \mathcal{M}(\underline{P})} P(A) \quad \forall A \subseteq \Omega,$$

It is well-known that any coherent lower probability satisfies $\underline{P}(A) \leq \overline{P}(A)$ whenever $A \in \mathcal{K} \cap \mathcal{K}^c$. Furthermore, any coherent lower probability defined on $\mathcal{K}$ can be extended to a greater domain $\mathcal{K} \subseteq \mathcal{K}'$ by using the natural extension [12]:

$$\underline{E}(A) = \min\{P(A) \mid P \in \mathcal{M}(\underline{P})\}, \ \forall A \in \mathcal{K}'.$$

In this work we consider lower probabilities defined on finite and ordered possibility spaces, denoted by $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$, called *marginal* lower probabilities, or defined on the cartesian product of two finite and ordered sets, denoted by $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^2$, called *joint* lower probabilities, where both $\mathcal{X}$ and $\mathcal{Y}$ are finite. In particular, if $\underline{P}_{\mathrm{X,Y}}$ is a joint lower probability defined on $\mathcal{X} \times \mathcal{Y}$, it defines two marginals on $\mathcal{X}$ and $\mathcal{Y}$ by:

$$\underline{P}_{\mathrm{X}}(A) = \underline{P}_{\mathrm{X,Y}}(A \times \mathcal{Y}), \ \forall A \subseteq \mathcal{X}.$$
$$\underline{P}_{\mathrm{Y}}(B) = \underline{P}_{\mathrm{X,Y}}(\mathcal{X} \times B), \ \forall B \subseteq \mathcal{Y}.$$

Uni- and bivariate p-boxes are specific instances of lower probabilities, defined as follows.

Definition 1. *A discrete univariate p-box defined on the ordered*[1] *finite set $\mathcal{X} = \{x_1, \ldots, x_n\}$ is a pair of increasing functions $\underline{F}, \overline{F} : \mathcal{X} \to [0,1]$ such that $\underline{F} \leq \overline{F}$ and $\underline{F}(x_n) = \overline{F}(x_n) = 1$.*

[1] We assume the elements in $\mathcal{X}$ are indexed according to this order, that is, $x_1 < \ldots < x_n$.

A discrete bivariate p-box defined on the Cartesian product of finite ordered[2] *sets $\mathcal{X} \times \mathcal{Y} = \{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_m\}$ is a pair of component-wise increasing functions*[3] *$\underline{F}, \overline{F} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ such that $\underline{F} \leq \overline{F}$ and $\underline{F}(x_n, y_m) = \overline{F}(x_n, y_m) = 1$.*

In what remains, and for the sake of simplicity, we avoid the term "discrete" so we will speak about uni- and bivariate p-boxes.

Remark 1. *Note that in the definition of univariate p-box we do not require $\underline{F}, \overline{F}$ to satisfy $\underline{F}(x_1) = \overline{F}(x_1) = 0$. The reason is that we interpret $(\underline{F}, \overline{F})$ as the imprecise observation of a cumulative distribution function F. However, cumulative distribution functions F defined on a finite space $\mathcal{X} = \{x_1, \ldots, x_n\}$ satisfy the properties: F is increasing and $F(x_n) = 1$. Nevertheless, as soon as x_1 has strictly positive probability, $F(x_1)$ will be strictly positive. For this reason the property $\underline{F}(x_1) = \overline{F}(x_1) = 0$ is not required for univariate p-boxes.*

With a similar reasoning we can justify why $\underline{F}(x_1, y_1) = \overline{F}(x_1, y_1) = 0$ is not required for bivariate p-boxes $(\underline{F}, \overline{F})$.

Univariate [11] and bivariate [8] p-boxes can be used to model the imprecise information about (univariate or bivariate) cumulative distribution functions.

Definition 2. *For any $x \in \mathcal{X} = \{x_1, \ldots, x_n\}$ and $y \in \mathcal{Y} = \{y_1, \ldots, y_m\}$, consider the following notation:*

$$A_x = [x_1, x], \text{ and } A_{x,y} = A_x \times A_y.$$

A univariate p-box defines a coherent lower probability on the domain $\mathcal{K}_1 = \{A_x, A_x^c : x \in \mathcal{X}\}$ by:

$$\underline{P}(A_x) = \underline{F}(x) \text{ and } \underline{P}(A_x^c) = 1 - \overline{F}(x).$$

A bivariate p-box defines a lower probability on the domain $\mathcal{K}_2 = \{A_{x,y}, A_{x,y}^c : (x,y) \in \mathcal{X} \times \mathcal{Y}\}$ by:

$$\underline{P}(A_{x,y}) = \underline{F}(x,y) \text{ and } \underline{P}(A_{x,y}^c) = 1 - \overline{F}(x,y). \quad (1)$$

Belief functions are another particular case of lower probabilities.

Definition 3. *A lower probability $\underline{P}$ on $\mathcal{P}(\Omega)$ is called* n-monotone *if and only if:*

$$\underline{P}\left(\cup_{i=1}^{p} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,p\}} (-1)^{|I|+1} \underline{P}\left(\cap_{i \in I} A_i\right)$$

for any $2 \leq p \leq n$ and any $A_1, \ldots, A_p \subseteq \Omega$. A lower probability that is n-monotone for any n is called

[2] Again, we assume the elements in $\mathcal{X}$ and $\mathcal{Y}$ are indexed according to this order: $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_m$.

[3] A function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is component-wise increasing when $F(x, y_i) \leq F(x, y_j)$ for any $x \in \mathcal{X}$ and $i, j \in \{1, \ldots, m\}$ such that $i < j$ and $F(x_i, y) \leq F(x_j, y)$ for any $y \in \mathcal{Y}$ and $i, j \in \{1, \ldots, n\}$ such that $i < j$.

completely monotone *or* belief function*, and its upper probability is called* plausibility*. Belief and plausibility functions are usually denoted by* Bel *and* Pl*, and they are coherent lower and upper probabilities.*

Using the so-called Möbius inverse, they define a *mass distribution* [3] in the following way:

$$m(A) = \sum_{E \subseteq A} (-1)^{|A \setminus E|} \mathrm{Bel(E)} \quad \forall \mathrm{A} \subseteq \Omega. \qquad (2)$$

A mass distribution $m : \mathcal{P}(\Omega) \to [0,1]$ satisfies $m(\emptyset) = 0$ and $\sum_{E \subseteq \Omega} m(E) = 1$. Conversely, any mass function defines a belief and plausibility functions

$$\mathrm{Bel(A)} = \textstyle\sum_{\mathrm{E} \subseteq \mathrm{A}} \mathrm{m(E)} \quad \forall \mathrm{A} \subseteq \Omega,$$
$$\mathrm{Pl(A)} = \textstyle\sum_{\mathrm{E}:\mathrm{E} \cap \mathrm{A} \neq \emptyset} \mathrm{m(E)} \quad \forall \mathrm{A} \subseteq \Omega.$$

The positivity of the mass m is characteristic of belief functions, in the sense that Eq. (2) is positive if and only if it is applied to a completely monotone lower probability.

Definition 4. *[9] Given a belief function* Bel *with mass distribution m, the elements $E \subseteq \Omega$ with positive mass, $m(E) > 0$, are called* focal elements*, and we will denote by $\mathcal{F}$ the set of focal elements. The union of all the focal sets is called the* core *of* Bel*, and it is denoted by* Core(Bel).

As for lower probabilities, we shall also use the terminology of *marginal* and *joint* to refer to belief functions defined on $\mathcal{X}, \mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. Any joint belief function Bel defined on $\mathcal{X} \times \mathcal{Y}$ with mass distribution m defines two *marginal belief functions* $\mathrm{Bel_X}$ and $\mathrm{Bel_Y}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with associated mass distributions m_X and m_Y:

$$m_\mathrm{X}(A) = \sum_{E:E^{\downarrow\mathcal{X}}=A} m(E) \text{ and } m_\mathrm{Y}(B) = \sum_{E:E^{\downarrow\mathcal{Y}}=B} m(E)$$

for any $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, and where $E^{\downarrow\mathcal{X}}$ and $E^{\downarrow\mathcal{X}}$ denote the projection of E on spaces $\mathcal{X}$ and $\mathcal{Y}$. Two important models to which we will devote particular attention and that induce belief functions are univariate p-boxes and possibility measures.

From now on, given E with a finite number of elements we will use the following notation:

$$\underline{e} = \min E, \quad \overline{e} = \max E. \qquad (3)$$

Kriegler and Held [5] showed that the lower probability induced by a p-box in the following way:

$$\underline{P}(A) = \inf\{P(A) : \underline{F} \leq F_P \leq \overline{F}\}, \quad (4)$$

where F_P is the cumulative distribution function associated with P, is indeed a belief function. Such belief function can be computed as Figure 1 shows. Thus, from now on we shall use the term *focal elements of a p-box* to refer to the focal elements of the belief function associated with a p-box using Eq. (4). According to [5], the focal elements of a p-box, named $E_1, \ldots, E_n$, can be ordered such that $\underline{e}_i \leq \underline{e}_{i+1}$ and $\overline{e}_i \leq \overline{e}_{i+1}$. When dealing with focal sets of p-boxes, we will consider that they are indexed according to this ordering. Furthermore, any joint belief function Bel defines a bivariate p-box in the following way:

$$\underline{F}(x,y) = \inf\{F_P(x,y) : P \in \mathcal{M}(\mathrm{Bel})\} = \mathrm{Bel}(\mathrm{A_{x,y}});$$
$$\overline{F}(x,y) = \sup\{F_P(x,y) : P \in \mathcal{M}(\mathrm{Bel_Y})\} = \mathrm{Pl}(\mathrm{A_{x,y}});$$

whereas any marginal belief function Bel defines an univariate p-box:

$$\underline{F}_\mathrm{X}(x) = \inf\{F_P(x) : P \in \mathcal{M}(\mathrm{Bel_X})\} = \mathrm{Bel_X}(\mathrm{A_x});$$
$$\overline{F}_\mathrm{X}(x) = \sup\{F_P(x) : P \in \mathcal{M}(\mathrm{Bel_X})\} = \mathrm{Pl_X}(\mathrm{A_x}).$$

A possibility measure constitutes another important specific case of plausibility function.

Definition 5. *A possibility measure* $\Pi : \mathcal{P}(\Omega) \to [0,1]$ *is a supremum-preserving map:* $\Pi(\cup_{i \in I} A_i) = \sup_{i \in I} \Pi(A_i)$ *for any I, $A_i \subseteq \Omega$.*

The conjugate of a possibility, $N(A) = 1 - \Pi(A^c)$ $\forall A \subseteq \Omega$, is a belief function. Its focal elements are nested: if E_1 and E_2 are focal elements, then either $E_1 \subseteq E_2$ or $E_2 \subseteq E_1$. Since we are dealing with finite referentials, there are only a finite number of focal sets $E_1, \ldots, E_n$, and for possibility measures we can assume they are indexed such that $E_1 \subseteq \ldots \subseteq E_n$.

2.2 Sklar's Theorem

Sklar's Theorem is an important tool in probability theory that allows a joint cumulative distribution function (cdf for short) to be expressed in terms of the marginals by means of a function called copula.

Definition 6. *[7] A* copula *is a commutative binary operator $C : [0,1]^2 \to [0,1]$ satisfying:*

1. $C(x,0) = 0$, $C(x,1) = x$ $\forall x \in [0,1]$.
2. $C(x_1,y_1) + C(x_2,y_2) \geq C(x_2,y_1) + C(x_1,y_2)$ $\forall x_1, x_2, y_1, y_2 \in [0,1]$ *such that* $x_1 \leq x_2$, $y_1 \leq y_2$.

Some classical copulas are the product copula, $\Pi(x,y) = x \cdot y$, the minimum, $M(x,y) = \min(x,y)$, and the Łukasiewicz operator $W(x,y) = \max(x+y-1, 0)$. The minimum and Łukasiewicz operators are also called the Fréchet-Hoeffding bounds because any copula satisfies the so-called Fréchet-Hoeffding inequality $M(x,y) \leq C(x,y) \leq W(x,y)$. Copulas play an important roll in the famous Sklar's Theorem.

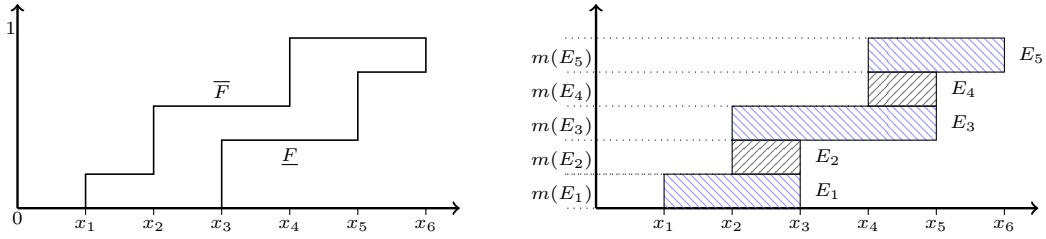


Figure 1: P-box (left) and its associated belief function (right), with focal elements $E_1 = \{x_1, x_2, x_3\}$, $E_2 = \{x_2, x_3\}$, $E_3 = \{x_2, x_3, x_4, x_5\}$, $E_4 = \{x_4, x_5\}$ and $E_5 = \{x_4, x_5, x_6\}$.

Theorem 1. *[10][Sklar's Theorem] Let $F_{\mathrm{X,Y}}$ be a joint cdf with marginals F_{X} and F_{Y}. Then, there exists a copula C such that*

$$F_{\mathrm{X,Y}}(x, y) = C(F_{\mathrm{X}}(x), F_{\mathrm{Y}}(y)) \ \forall (x, y) \in [0, 1]^2. \quad (5)$$

Conversely, given two marginal cdfs F_{X} and F_{Y} and a copula C, they define a joint cdf $F_{\mathrm{X,Y}}$ using Eq. (5).

Possibly the most usual application of Sklar's Theorem concerns independent random variables. Two variables X and Y are independent if $F_{\mathrm{X,Y}}(x, y) = F_{\mathrm{X}}(x) \cdot F_{\mathrm{Y}}(y)$, that is, when the copula linking the marginals is the product. Also very important are the cases in which the random variables are coupled by the Fréchet-Hoeffding bounds. Random variables coupled by the minimum (resp., Łukasiewicz operator) are called *comonotone* (resp., *countermonotone*). Comonotone random variables can be characterized in many different ways. For this aim, we first introduce the following notion.

Definition 7. *A subset S of $\mathbb{R}^2$ is* increasing *when for any $(x, y), (u, v) \in S$, $x < u$ implies $y \leq v$, and $y < v$ implies $x \leq u$.*

Then, a pair of random variables (X, Y) is comonotone if it satisfies one, and therefore all, of the following equivalent conditions:

- The copula that links the marginals is the minimum: $F_{\mathrm{X,Y}}(x, y) = \min(F_{\mathrm{X}}(x), F_{\mathrm{Y}}(y)) \ \forall (x, y)$.
- The support of (X, Y) is an increasing set on $\mathbb{R}^2$.
- $\forall (x, y) \in \mathbb{R}^2$, either $P(X \leq x, Y > y) = 0$ or $P(X > x, Y \leq y) = 0$.

Remark 2. *The notion of comonotonicity can also be defined for discrete probabilities $P_{\mathrm{X,Y}}$, just by substituting the support of (X, Y) by the support of $P_{\mathrm{X,Y}}$.*

For example, consider a finite Ω where all its elements have positive probability. Consider the random variables X, Y defined by:

	$\omega \in \Omega_1 \subset \Omega$	$\omega \in \Omega_1^c \subset \Omega$
X	1	2
Y	0	3

Then, the support of (X, Y) is given by $\{(1, 0), (2, 3)\}$, which is an increasing subset of $\mathbb{R}^2$ and therefore (X, Y) is comonotone. In this case, we can also consider the support of $P_{(X,Y)}$, which are the elements (x, y) with positive possibility. In this case, the support of $P_{(X,Y)}$ coincides with the support of (X, Y) and therefore $P_{(X,Y)}$ is comonotone.

When we have imprecise information about the joint or the marginal cdfs or about the copula, Sklar's Theorem cannot be applied. The next Theorem adapts Sklar's Theorem to the imprecise setting, using p-boxes, both uni- and bivariate, and sets of copulas.

Theorem 2. *[6][Imprecise Sklar's Theorem]*

1. *Given two univariate p-boxes $(\underline{F}_{\mathrm{X}}, \overline{F}_{\mathrm{X}})$ and $(\underline{F}_{\mathrm{Y}}, \overline{F}_{\mathrm{Y}})$ and a set of copulas $\mathcal{C}$, consider:*

$$\underline{F}(x, y) = \inf_{C \in \mathcal{C}} C(\underline{F}_{\mathrm{X}}(x), \underline{F}_{\mathrm{Y}}(y)) \text{ and}$$
$$\overline{F}(x, y) = \sup_{C \in \mathcal{C}} C(\overline{F}_{\mathrm{X}}(x), \overline{F}_{\mathrm{Y}}(y)).$$

   Then, they define a bivariate p-box $(\underline{F}, \overline{F})$ whose associated lower probability is coherent.

2. *Given a bivariate p-box $(\underline{F}, \overline{F})$, it could not be possible to express it in terms of the univariate p-boxes and a set of copulas, even when the lower probability associated with $(\underline{F}, \overline{F})$ is coherent.*

In the framework of imprecise probabilities, the notion of independence has been widely investigated [1, 2]. However, those satisfying the factorizing property have the same associated bivariate p-box, and it is obtained by applying the product copula to the marginals p-boxes [6, Prop. 6]:

$$\underline{F}(x, y) = \underline{F}_{\mathrm{X}}(x) \cdot \underline{F}_{\mathrm{Y}}(y) \text{ and } \overline{F}(x, y) = \overline{F}_{\mathrm{X}}(x) \cdot \overline{F}_{\mathrm{Y}}(y)$$

for any x, y. The question now is: what is the meaning of comonotonicity in the imprecise probability setting? As far as we know, this remains unexplored. Thus, the aim of this paper is to define the notion of comonotonicity when dealing with coherent lower probabilities.

3 Comonotone Lower Probabilities

We have seen that comonotonicity in the precise framework can be expressed in three equivalent ways. Now, we shall try to investigate to what extent these conditions, or similar ones, also hold in the case of coherent lower probabilities.

In our framework, we consider a coherent lower probability $\underline{P}$ defined on the power set of $\mathcal{X}\times\mathcal{Y}$. We assume that $\underline{P}$ models the imprecise information about a joint probability $P_{\mathrm{X,Y}}$. The question is: how can we model the additional information that $P_{\mathrm{X,Y}}$ is comonotone?

Definition 8. *A lower probability $\underline{P}$ defined on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is called* comonotone *when any $P\in\mathcal{M}(\underline{P})$ is comonotone.*

This is a straightforward definition, since if $\underline{P}$ models the imprecise information about a comonotone probability $P_{\mathrm{X,Y}}$, all the probabilities compatible with the lower probability should be comonotone.

Example 1. *Consider the lower probability $\underline{P}$ defined on $\{0,1\}\times\{1,2\}$ such that:*

$\underline{P}(\{(1,2)\})=\alpha\in(0,0'5),\ \underline{P}(\{(0,1)\})=\beta\in(0,0'5)$
$\underline{P}(\{(0,1),(0,2),(1,2)\})=1,$
$\underline{P}(\{(0,1),(0,2),(1,2),(1,1)\})=1,$
$\underline{P}(A)=0$ *otherwise.*

This lower probability is coherent and its credal set is formed by all the convex combinations of the following precise probabilities:

	$\{(0,1)\}$	$\{(0,2)\}$	$\{(1,2)\}$
P_1	β	$1-\alpha-\beta$	α
P_2	β	0	$1-\beta$
P_3	$1-\alpha$	0	α

Then, the support of any $P\in\mathcal{M}(\underline{P})$ is included in $\{(0,1),(0,2),(1,2)\}$, that is an increasing set, and therefore all the probabilities in $\mathcal{M}(\underline{P})$ are comonotone, and then also is $\underline{P}$.

We now investigate how comonotone coherent lower probabilities can be equivalently expressed. We first express it by means of sets $\{X>x,Y\leq y\}$ and $\{X\leq x,Y>y\}$.

Theorem 3. *A coherent lower probability $\underline{P}$ defined on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is comonotone if and only if any $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$ either*

$$\overline{P}(\{(u,v):u\leq x,v>y\})=0 \text{ or}$$
$$\overline{P}(\{(u,v):u>x,v\leq y\})=0.$$

This theorem shows that the characterization of comonotone random variables in terms of event probabilities also holds in the imprecise case. Now, we are going to see that, if we define the support $\mathrm{supp}(\underline{P})$ of a lower probability $\underline{P}$ by:

$$\mathrm{supp}(\underline{P})=\bigcup_{P\in\mathcal{M}(\underline{P})}\mathrm{supp}(P),$$

its comonotonocity can also be equivalently expressed in terms of the increasingness of $\mathrm{supp}(\underline{P})$.

Theorem 4. *A coherent lower probability $\underline{P}$ defined on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is comonotone if and only if its support* $\mathrm{supp}(\underline{P})$ *is an increasing set.*

Therefore, this second equivalent expression also holds for lower probabilities. Now, it only remains to check whether or not the comonotonicity of lower probabilities is related to the copula that links the marginals. The next result shows one implication.

Theorem 5. *Let $\underline{P}$ be a coherent comonotone lower probability defined on $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. If $(\underline{F},\overline{F})$, $(\underline{F}_{\mathrm{X}},\overline{F}_{\mathrm{X}})$ and $(\underline{F}_{\mathrm{Y}},\overline{F}_{\mathrm{Y}})$ denote the bivariate and the marginal univariate p-boxes, respectively, then for any (x,y):*

$$\underline{F}(x,y)=\min(\underline{F}_{\mathrm{X}}(x),\underline{F}_{\mathrm{Y}}(y)) \text{ and}$$
$$\overline{F}(x,y)=\min(\overline{F}_{\mathrm{X}}(x),\overline{F}_{\mathrm{Y}}(y)).$$

The next example shows that, unfortunately, the converse implication does not hold in general.

Example 2. *Consider the joint coherent lower probability $\underline{P}$ defined on $\{1,2\}^2$ by:*

$$\underline{P}(\{(1,1),(1,2),(2,2)\})=\alpha>0,$$
$$\underline{P}(\{(1,1),(2,1),(2,2)\})=1-\alpha>0,$$
$$\underline{P}(\{(1,1),(1,2),(2,1),(2,2)\})=1,$$
$$\underline{P}(A)=0 \text{ otherwise.}$$

Then, regardless of α, $\underline{F}=I_{\{(x,y):x,y\geq 2\}}$ and $\overline{F}=I_{\{(x,y):x,y\geq 1\}}$. Furthermore:

$$\underline{F}_{\mathrm{X}}(x)=\underline{F}_{\mathrm{Y}}(x)=I_{\{x\geq 2\}}(x) \text{ and}$$
$$\overline{F}_{\mathrm{X}}(x)=\overline{F}_{\mathrm{Y}}(x)=I_{\{x\geq 1\}}(x).$$

Then:

$$\underline{F}(x,y)=\min(\underline{F}_{\mathrm{X}}(x),\underline{F}_{\mathrm{Y}}(y)) \text{ and}$$
$$\overline{F}(x,y)=\min(\overline{F}_{\mathrm{X}}(x),\overline{F}_{\mathrm{Y}}(y)).$$

However, $\underline{P}$ is not comonotone because the support of $\underline{P}$ contains the elements $(1,2)$ and $(2,1)$, and this contradicts Theorem 4.

Thus, going from a precise to an imprecise setting, comonotonicity can only be characterized by two equivalent ways: by means of the increasingness of the support or by means of the upper probability of the adequate sets. Indeed, the bivariate p-box of a comonotone lower probability is the minimum of the marginals, but the minimum of two marginal p-boxes will not necessarily generate a comonotone lower probabilities. Figure 2 summarizes the conditions we have seen along this section.

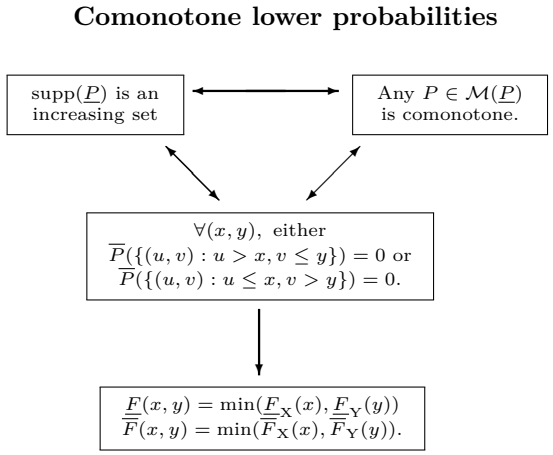


Figure 2: Summary of the conditions for joint comonotone lower probabilities.

4 Comonotone Belief Functions

We now focus on the comonotonicity of belief functions. In this case we have to note that supp(Bel) coincides with Core(Bel), and therefore Theorem 4 can be directly adapted.

Corollary 1. *A belief function* Bel *defined on* $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ *is comonotone if and only if* Core(Bel) *is an increasing set.*

We may think that the converse of Theorem 5 could hold when dealing with belief functions. However, this is not the case, since the lower probability given in Example 2 is in fact a belief function with focal elements $E_1 = \{(1,1),(1,2),(2,2)\}$ and $E_2 = \{(1,1),(2,1),(2,2)\}$ with $m(E_1) = \alpha$ and $m(E_2) = 1-\alpha$, respectively.

Although Section 3 characterized comonotone lower probabilities, did not explore important questions: when and how can we build a comonotone lower probability $\underline{P}$ from marginals $\underline{P}_{\mathrm{X}}$, $\underline{P}_{\mathrm{X}}$? These are the questions we address in this section, for the specific case of belief functions.

Note that those questions can always be answered positively in the precise framework, as it is always possible to define a joint comonotone probability from two marginal probabilities P_{X} and P_{Y}, by simply defining $F_{\mathrm{X,Y}}$ as the minimum of the marginals and then considering the associated probability. Unfortunately, not every marginal lower probabilities allow us to define a comonotone lower probability with the given marginals, even when the lower probabilities are belief functions, as the next example shows.

Example 3. *Let* $\mathrm{Bel}_{\mathrm{X}}$ *and* $\mathrm{Bel}_{\mathrm{Y}}$ *be the marginal belief functions, defined over* $\mathcal{X} = \{1,2,3\}$ *and* $\mathcal{Y} = \{1,2\}$ *with mass distributions*

$$m_{\mathrm{X}}(\{1,2\}) = 0.7,\ m_{\mathrm{X}}(\{1,2,3\}) = 0.3;$$
$$m_{\mathrm{Y}}(\{1\}) = 0.3,\ m_{\mathrm{Y}}(\{2\}) = 0.7.$$

Let us assume that there is a comonotone joint belief function Bel *whose marginals are the belief functions* $\mathrm{Bel}_{\mathrm{X}}, \mathrm{Bel}_{\mathrm{Y}}$ *induced by* m_{X}, m_{Y}. *If this is the case, using Theorem 5, the bivariate p-box induced by* Bel *is the minimum of the marginals* $\underline{F}_X, \overline{F}_X$ *and* $\underline{F}_Y, \overline{F}_Y$. *Then:*

$$\overline{F}(1,2) = \min(\overline{F}_{\mathrm{X}}(1), \overline{F}_{\mathrm{Y}}(2)) = 1.$$

This implies that any focal set E *of* Bel *satisfies* $E \cap \{(1,1),(1,2)\} \neq \emptyset$ *because* $\overline{F}(1,2) = Pl(\{(1,1),(1,2)\})$. *Furthermore,* $(1,2) \in \mathrm{Core(Bel)}$, *because:*

$$Pl(\{(1,1)\}) = \overline{F}(1,1) = \min(\overline{F}_{\mathrm{X}}(1), \overline{F}_{\mathrm{Y}}(1)) = 0.3 < 1 = \overline{F}(1,2),$$

which means that there is a focal element E *such that* $(1,2) \in E$ *and* $(1,1) \notin E$. *Now, since* Bel *is comonotone,* Core(Bel) *is increasing by Corollary 1, and then* $(2,1) \notin \mathrm{Core(Bel)}$, *hence there is no focal element* E *such that* $(2,1) \in E$. *Yet, we have*

$$\underline{F}(2,1) = \mathrm{Bel}(\{(1,1),(2,1)\}) = 0.3,$$

which implies that there is a focal set E *such that* $E \subseteq \{(1,1),(2,1)\}$. *Since* $(2,1) \notin E$, $E = \{(1,1)\}$, *what implies that* $\mathrm{Bel}(\{(1,1)\}) > 0$. *However, if* Bel *is comonotone, it follows that*

$$\mathrm{Bel}(\{(1,1)\}) = \underline{\mathrm{F}}(1,1) = \min(\underline{\mathrm{F}}_{\mathrm{X}}(1), \underline{\mathrm{F}}_{\mathrm{Y}}(1)) = 0,$$

a contradiction showing that there are no comonotone belief functions with marginals $\mathrm{Bel}_{\mathrm{X}}, \mathrm{Bel}_{\mathrm{Y}}$.

This shows that our problem is trickier to answer in the imprecise setting. Below we provide some situations under which a joint comonotone belief function exists with given marginals.

The first case we investigate is when the marginals are possibility measures. Before introducing the main result, note that for any two possibilities having $A_1, \ldots, A_m$ and $B_1, \ldots, B_\ell$ as focal elements, we can always duplicate those elements to build an equivalent mass function (in terms of induced belief function) with focal elements $C_1, \ldots, C_n$ and $D_1, \ldots, D_n$ such that

- $C_i \subseteq C_{i+1}$ and $D_i \subseteq D_{i+1}$ for any $i = 1, \ldots, n-1$.
- $C_i \in \{A_1, \ldots, A_m\}$ and $D_i \in \{B_1, \ldots, B_l\}$ for any $i \in \{1, \ldots, n\}$.

- $m_{\mathrm{X}}(C_i) = m_{\mathrm{Y}}(D_i)$.

The next example illustrates this procedure.

Example 4. *Consider the possibility measures with the following focal sets:*

$$A_1 = \{2\}, \quad A_2 = \{1,2\}, \quad A_3 = \{1,2,3\},$$
$$B_1 = \{1,2\}, \quad B_2 = \{1,2,3,4\},$$

with the following masses:

$$m_{\mathrm{X}}(A_1) = 0.3, \quad m_{\mathrm{X}}(A_2) = 0.5, \quad m_{\mathrm{X}}(A_3) = 0.2.$$
$$m_{\mathrm{Y}}(B_1) = 0.5, \quad m_{\mathrm{Y}}(B_2) = 0.5.$$

Now, we rewrite the focal sets in the following way

	A_1	A_2	A_2	A_3
	B_1	B_1	B_2	B_2
	C_1	C_2	C_3	C_4
	D_1	D_2	D_3	D_4
m	0.3	0.2	0.3	0.2.

We can therefore assume, without loss of generality, that any two possibilities have the same number of focal sets and that their masses coincide.

Proposition 1. *Given two marginal possibility measures, there exists a joint comonotone possibility whose marginals are the original possibility measures.*

This result gives a constructive method for building the joint comonotone possibility. If $A_1 \subseteq \ldots \subseteq A_n$ and $B_1 \subseteq \ldots \subseteq B_n$ denotes the focal elements of m_{X} and m_{Y} such that $m_{\mathrm{X}}(A_i) = m_{\mathrm{Y}}(B_i)$ for $i = 1, \ldots, n$. Using the notation of Eq. (3), Algorithm 1 shows how to define the focal elements and mass function associated with the joint comonotone possibility.

The next example shows how to apply this procedure.

Example 5. *Consider the possibility measures of Example 4. We define the following focal sets for the joint comonotone possibility:*

$$E_1 = \{(2,1),(2,2)\}, \quad E_2 = \{(1,1),(2,1),(2,2)\},$$
$$E_3 = \{(1,1),(2,1),(2,2),(2,3),(2,4)\},$$
$$E_4 = \{(1,1),(2,1),(2,2),(2,3),(2,4),(3,4)\},$$

and the masses are:

	E_1	E_2	E_3	E_4
m	0.3	0.2	0.3	0.2

Let us now look at the case where the focal elements $A_1, \ldots, A_m$ and $B_1, \ldots, B_\ell$ of the marginal belief functions $\mathrm{Bel}_{\mathrm{X}}$ and $\mathrm{Bel}_{\mathrm{Y}}$ can be ordered such that, following notation of Eq. (3), $\underline{a}_i \leq \underline{a}_{i+1}$, $\overline{a}_i \leq \overline{a}_{i+1}$ and $\underline{b}_j \leq \underline{b}_{j+1}$, $\overline{b}_j \leq \overline{b}_{j+1}$ for any $i = 1, \ldots, m-1$ and $j = 1, \ldots, \ell$ and are intervals, in the sense that any

Algorithm 1 Procedure defining focal elements of the joint comonotone possibility

1: **for** $i = 2, \ldots, n$ **do**

$$\mathcal{I}_i = \{(x, \underline{b}_i) : x \in [\underline{a}_i, \underline{a}_{i-1}]\} \cup \{(\underline{a}_{i-1}, y) : y \in [\underline{b}_i, \underline{b}_{i-1}\}$$
$$\cup\{(c, \overline{b}_{i-1}) : x \in [\overline{a}_{i-1}, \overline{a}_i]\} \cup \{(\overline{a}_i, y) : y \in [\overline{b}_{i-1}, \overline{b}_i]\}$$

2: **end for**
3: Define

$$\mathcal{G}_1 = \{(x, \overline{b}_1) : x \in [\underline{a}_1, \overline{a}_1]\} \cup \{(\overline{a}_1, y) : y \in [\underline{b}_1, \overline{b}_{i-1}]\}$$

4: **for** i=2,. . . ,n **do**

$$\mathcal{G}_i = \mathcal{I}_i \cup \mathcal{G}_{i-1}$$

5: **end for**
6: **for** i=1,. . . ,n **do**

$$F_i = \mathcal{G}_i \cap (A_i \times \mathbb{R}) \cap (\mathbb{R} \times B_i)$$
$$m(F_i) = m_{\mathrm{X}}(A_i) = m_{\mathrm{Y}}(B_i)$$

7: **end for**

A_i, B_j contains all elements in $\mathcal{X}, \mathcal{Y}$ between $\underline{a}_i, \overline{a}_i$ and $\underline{b}_j, \overline{b}_j$, respectively. Similarly to focal elements of possibility distributions, those focal elements can be expressed as $\{C_1, \ldots, C_n\}$ and $\{D_1, \ldots, D_n\}$ simply by duplicating elements. Then, they satisfy:

- $\underline{c}_i \leq \underline{c}_{i+1}$, $\overline{c}_i \leq \overline{c}_{i+1}$, $\underline{d}_i \leq \underline{d}_{i+1}$ and $\overline{d}_i \leq \overline{d}_{i+1}$ for any $i \in \{1, \ldots, n\}$.
- $C_i \in \{A_1, \ldots, A_m\}$ and $D_i \in \{B_1, \ldots, B_\ell\}$ for any $i \in \{1, \ldots, n\}$.
- $m_{\mathrm{X}}(C_i) = m_{\mathrm{Y}}(D_i)$ for any $i = 1, \ldots, n$.

Example 6. *Consider the belief functions* $\mathrm{Bel}_{\mathrm{X}}$ *and* $\mathrm{Bel}_{\mathrm{Y}}$ *whose focal elements are:*

$$A_1 = \{0,1\}, \quad A_2 = \{1,2\}, \quad A_3 = \{2,3\} \text{ and}$$
$$B_1 = \{0,1\}, \quad B_2 = \{1,2\},$$

whose masses are:

$$m_{\mathrm{X}}(A_1) = 0.4, \quad m_{\mathrm{X}}(A_2) = 0.3, \quad m_{\mathrm{X}}(A_3) = 0.3;$$
$$m_{\mathrm{Y}}(B_1) = 0.6, \quad m_{\mathrm{Y}}(B_2) = 0.4.$$

We rewrite the focal elements in the following way:

	A_1	A_2	A_2	A_3
	B_1	B_1	B_2	B_2
	C_1	C_2	C_3	C_4
	D_1	D_2	D_3	D_4
m	0.4	0.2	0.1	0.3

Then, from now on we will assume that given two marginal belief functions whose focal sets are intervals ordered through the lattice ordering, both belief

functions have the same number of focal sets and their masses coincide.

Proposition 2. *Consider two marginal belief functions* Bel_X *and* Bel_Y *with mass distributions* m_X, m_Y *whose focal elements* $\mathcal{A} = \{A_1, \ldots, A_n\}$, $\mathcal{B} = \{B_1, \ldots, B_n\}$ *are such that* A_i *and* B_i *are intervals and* $m_\text{X}(A_i) = m_\text{Y}(B_i)$ *for any* $i = 1, \ldots, n$. *If* $\mathcal{A}$ *and* $\mathcal{B}$ *satisfy the following constraints:*

I) $\underline{a}_i \leq \underline{a}_{i+1}$ *and* $\overline{a}_i \leq \overline{a}_{i+1}$ *for any* $i = 1, \ldots, n$.

II) $\underline{b}_i \leq \underline{b}_{i+1}$ *and* $\overline{b}_i \leq \overline{b}_{i+1}$ *for any* $i = 1, \ldots, n$.

III) If $\overline{a}_i < \underline{a}_j$, *then* $\overline{b}_i \leq \underline{b}_j$.

IV) If $\overline{b}_i < \underline{b}_j$, *then* $\overline{a}_i \leq \underline{a}_j$

then, there exists a joint comonotone belief function Bel *such that its marginal masses coincide with* m_X *and* m_Y.

Using the notation of Eq. (3), Algorithm 2 shows how to build the focal elements and the mass of the joint comonotone belief function.

Algorithm 2 Procedure defining focal elements of the joint comonotone belief function

1: Define

$$\mathcal{G} = \{(\underline{a}_i, \underline{b}_i), (\overline{a}_i, \overline{b}_i) : i = 1, \ldots, n\}$$

2: Name the elements on $\mathcal{G}$ by:

$$\mathcal{G} = \{(c_1, d_1), \ldots, (c_{2n}, d_{2n})\}$$
$$c_i \leq c_{i+1} \text{ and } d_i \leq d_{i+1} \text{ for } i = 1, \ldots, 2n-1$$

3: **for** $i = 1, \ldots, 2n$-1 **do**

$$\mathcal{I}_i = \{(x, d_k) : x \in [c_k, c_{k+1}]\} \cup \{c_{k+1}, y) : y \in [d_k, d_{k+1}\}$$

4: **end for**

5: **for** i=1,. . . ,n **do**

$$E_i = \cup_{(\underline{a}_i, \underline{b}_i) \leq (c_k, d_k) < (\overline{a}_i, \overline{a}_i)} \mathcal{I}_k$$
$$m(E_i) = m_\text{X}(A_i) = m_\text{Y}(B_i)$$

6: **end for**

The next example shows how this algorithm is applied.

Example 7. *Let us continue Example 6. We build the following focal sets for the joint belief function:*

$$E_1 = \{(0,0), (1,0), (1,1)\},\ E_2 = \{(1,0), (1,1), (2,1)\},$$
$$E_3 = \{(1,1), (2,1), (2,2)\},\ E_4 = \{(2,1), (2,2), (2,3)\}.$$

Now, we assigns the following masses:

	E_1	E_2	E_3	E_4
m	0.4	0.2	0.1	0.3

This joint belief function is comonotone and its marginals coincide with Bel_X *and* Bel_Y.

The condition in Proposition 2 that focal sets should be intervals is essential, as the next example shows.

Example 8. *Consider two mass functions* m_X *and* m_Y *with* $\mathcal{A} = \{A_1, A_2\}$ *and* $\mathcal{B} = \{B\}$, *where:*

$$A_1 = \{1, 3\}, \quad A_2 = \{2, 4\}, \quad B = \{1, 1, \ldots, n-1, n\}$$

for $n > 3$. $\mathcal{A}$ *and* $\mathcal{B}$ *satisfy all the conditions of Proposition 2, except for being intervals. However, there is no joint comonotone belief functions having those marginals. Indeed, following Algorithm 2, such a joint would have two focal elements* E_1, E_2 *with projections* A_1, B *and* A_2, B, *respectively, and such that* $E_1 \cup E_2$ *is increasing. Now, for any* $x \in \{1, 1, \ldots, n-1, n\}$, $E_1 \cup E_2$ *must contain, at least for one* x, *any the following pair:* $(x, 1)$ *and* $(x, 2)$, $(x, 1)$ *and* $(x, 4)$, $(x, 3)$ *and* $(x, 2)$, *or* $(x, 3)$ *and* $(x, 4)$, *for* E_1, E_2 *to have the required projections. If we take any two of those pairs for two different* $x \leq y$ *in* $\{1, 1, \ldots, n-1, n\}$, *then they form a non-increasing set. For example, take* $(x, 1), (x, 4)$ *and* $(y, 3), (y, 4)$, *we have* $(x, 4) \not\leq (y, 3)$. *Hence it is not possible to build a comonotone joint belief from* m_X *and* m_Y.

We have seen conditions under which, given marginal belief functions, it is possible to define a joint comonotone belief function. However, the next example shows that this joint comonotone belief function is not unique.

Example 9. *Consider the marginal belief functions* Bel_X *and* Bel_Y *with mass distributions* m_X *and* m_Y, *given by:*

$$m_\text{X}(\{1, 2\}) = m_\text{Y}(\{1, 2\}) = 1.$$

In this case, we can define three joint belief functions that are comonotone: if we denote their masses by m, m' *and* m'', *they are given by:*

$$m(\{(1,1), (2,2)\}) = m'(\{(1,1), (2,2), (1,2)\}) = m''(\{(1,1), (2,2), (2,1)\}) = 1.$$

5 Comonotone p-boxes

Consider now a bivariate p-box $(\underline{F}, \overline{F})$ defined on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$. We have already said that bivariate p-boxes define a lower probability $\underline{P}$ on the set $\mathcal{K}_2$ following Eq. (1).

Consider the natural extension of $\underline{P}$ to $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and we are going to investigate whether it is comonotone or not. During this section, and for the sake of simplicity, we shall assume that $n, m > 1$, $\overline{P}(\{x_i\}) > 0$ and $\overline{P}(\{y_j\}) > 0$ for any $i = 1, \ldots, n$ and $j = 1, \ldots, m$. This implies that $\overline{F}_{\mathrm{X}}(x_i) > \underline{F}_{\mathrm{X}}(x_{i-1})$ and $\overline{F}_{\mathrm{Y}}(y_j) > \underline{F}_{\mathrm{Y}}(y_{j-1})$ for $i = 2, \ldots, n$ and $j = 2, \ldots, m$.

In [8] it is argued that some notions like "avoiding sure loss", "coherence" or "2-mononicity" of a bivariate p-box is given in terms of its associated lower probability (given in Eq. (1)).

Definition 9. *A coherent bivariate p-box is comonotone when its associated lower probability is comonotone.*

Next results give two characterizations of comonotone bivariate p-boxes. The first one establishes the form of the bivariate p-box.

Proposition 3. *Let $(\underline{F}, \overline{F})$ be a coherent bivariate p-box defined on $\mathcal{X} \times \mathcal{Y}$. Then, it is comonotone if and only if there is an increasing set $S \subseteq \mathcal{X} \times \mathcal{Y}$, named $S = \{(u_1, v_1), \ldots, (u_k, v_k)\}$, such that:*

S.1 *The X and Y projections of S are $\mathcal{X}$ and $\mathcal{Y}$.*

S.2 *If $(x_i, y_j) \in S$ and $(x_{i+1}, y_j) \notin S$, then*

$$\underline{F}(x_i, y_j) = \underline{F}(x_{i+1}, y_j) = \cdots = \underline{F}(x_n, y_j) = \overline{F}(x_i, y_j) = \overline{F}(x_{i+1}, y_j) = \cdots = \overline{F}(x_n, y_j).$$

S.3 *If $(x_i, y_j) \in S$ and $(x_i, y_{j+1}) \notin S$, then*

$$\underline{F}(x_i, y_j) = \underline{F}(x_i, y_{j+1}) = \cdots = \underline{F}(x_i, y_m) = \overline{F}(x_i, y_j) = \overline{F}(x_i, y_{j+1}) = \cdots = \overline{F}(x_i, y_m).$$

The second result characterizes comonotone coherent bivariate p-boxes in terms of the belief functions associated with its marginal p-boxes.

Theorem 6. *Let $(\underline{F}, \overline{F})$ be a coherent bivariate p-box defined on $\mathcal{X} \times \mathcal{Y}$. Denote by $(\underline{F}_{\mathrm{X}}, \overline{F}_{\mathrm{X}})$ and $(\underline{F}_{\mathrm{Y}}, \overline{F}_{\mathrm{Y}})$ its marginal p-boxes, and by $\mathrm{Bel}_{\mathrm{X}}$ and $\mathrm{Bel}_{\mathrm{Y}}$ the belief functions associated with the marginal p-boxes. Then, $(\underline{F}, \overline{F})$ is comonotone if and only if one of the following conditions are satisfied:*

1. $\mathrm{Bel}_{\mathrm{X}}$ *is precise with positive probability in* $\{x_1\}, \ldots, \{x_n\}$. $\mathrm{Bel}_{\mathrm{X}}$ *and* $\mathrm{Bel}_{\mathrm{Y}}$ *satisfy the following conditions:*
   - *The focal elements of* $\mathrm{Bel}_{\mathrm{Y}}$ *are* $\{y_1\}, \ldots, \{y_{l-1}\}$, *where* $l \in \{1, \ldots, m\}$, *and* $B_1, \ldots, B_s$, *where* $y_l = \min_{i=1,\ldots,s} \min B_i$ *and,* $\cup_{i=1}^{s} B_i = \{y_{l+1}, \ldots, y_m\}$.
   - $m_{\mathrm{X}}(\{x_n\}) \geq \sum_{i=1}^{s} m_{\mathrm{Y}}(B_i) - m_{\mathrm{Y}}(\{y_l\})$.

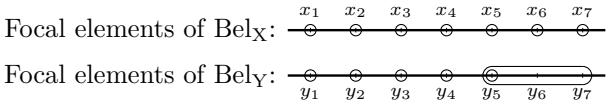


Figure 3: Example of belief functions that allow to build a comonotone bivariate p-box. According to Theorem 6, $m_{\mathrm{Y}}(\{y_5, y_6, y_7\}) \leq m_{\mathrm{X}}(\{x_7\})$ must hold.

2. *Condition 1 holds when we exchange the role of* $\mathrm{Bel}_{\mathrm{X}}$ *and* $\mathrm{Bel}_{\mathrm{Y}}$.

Using the previous theorem, we can state the following corollary.

Corollary 2. *If a bivariate p-box is comonotone, its associated lower probability is a belief function.*

From this result we know that any comonotone coherent bivariate p-box can be built with the adequate belief functions. We can also deduce that most bivariate p-boxes will not be comonotone. Thus, under the interpretation of Definitions 8 and 9, bivariate p-boxes do not seem to be adequate to model comonotonicity.

Example 10. *Figure 3 shows an example of marginal belief functions satisfying the conditions of Theorem 6. Assume that the masses are the following:*

	$\{x_1\}$	$\{x_2\}$	$\{x_3\}$	$\{x_4\}$	$\{x_5\}$	$\{x_6\}$	$\{x_7\}$
m_{X}	0.12	0.15	0.22	0.13	0.1	0.08	0.2
	$\{y_1\}$	$\{y_2\}$	$\{y_3\}$	$\{y_4\}$	$\{y_5\}$	$\{y_5, y_6, y_7\}$	
m_{Y}	0.17	0.15	0.15	0.18	0.2	0.15	

Then, the comonotone bivariate p-box has the following focal elements:

$$E_1 = \{(x_1, y_1)\}, \quad E_2 = \{(x_2, y_1)\}, \quad E_3 = \{(x_2, y_2)\},$$
$$E_4 = \{(x_3, y_2)\}, \quad E_5 = \{(x_3, y_3)\}, \quad E_6 = \{(x_3, y_4)\},$$
$$E_7 = \{(x_4, y_4)\}, \quad E_8 = \{(x_5, y_4)\}, \quad E_9 = \{(x_5, y_5)\},$$
$$E_{10} = \{(x_6, y_5)\}, \quad E_{11} = \{(x_7, y_5)\},$$
$$E_{12} = \{x_7\} \times \{y_5, y_6, y_7\}.$$

Their masses are:

	E_1	E_2	E_3	E_4	E_5	E_6
m	0.12	0.05	0.1	0.05	0.15	0.02
	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}
m	0.13	0.03	0.07	0.08	0.05	0.2

Note again that the set S of Proposition 3 is the core of Bel. *It can be seen in Figure 4.*

6 Conclusions

This paper investigates the notion of comonotonicity for coherent lower probabilities. We have seen that

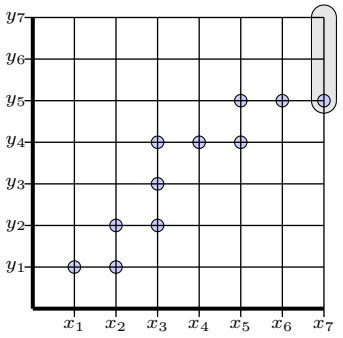


Figure 4: Core of the belief function that defines a comonotone bivariate p-box.

the comonotonicity of a coherent lower probability can be expressed in two equivalent ways: by means of the increasingness of its support or by means of the upper probability of the sets $\{(u,v) : u > x, v \leq y\}$ and $\{(u,v) : u \leq x, v > y\}$. Furthermore, the bivariate p-box associated with a comonotone coherent lower probability can be expressed as the minimum of the marginal p-boxes. However, in contrast to the precise setting, the converse does not hold in general.

Another important difference between precise and imprecise frameworks is that in the former any pair of marginal probabilities admits the definition of a joint comonotone probability with the fixed marginals. This is not the case of lower probabilities, not even when they are belief functions. Nevertheless, such a property does hold for possibility measures and for univariate p-boxes satisfying some additional restrictions.

Unfortunately, we have also seen that bivariate p-boxes, except in very special cases, do not seem to be adequate to model comonotonicity because they impose very strong conditions, like for instance one of the marginals must be precise. Then, in contrast to the precise framework where bivariate distribution functions express the information about comonotonicity, this is not the case of bivariate p-boxes.

One interesting open problem is to investigate the meaning of comonotonicity for a more general framework, that of lower previsions. Although independent products satisfying the factorizing property have the same associated bivariate p-box, in the general framework of lower prevision they are no longer equivalent. It would not be surprisingly that comonotonicity could be extended in many different ways.

Acknowledgements

This work was carried out and funded in the framework of the Labex MS2T. It was supported by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). We would like to thank the anonymous reviewers for their helpful comments.

References

[1] I. Couso, S. Moral and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[2] G. de Cooman, E. Miranda and M. Zaffalon. Independent natural extension. *Artificial Intelligence*, 175(12-13):1911–1950, 2011.

[3] D. Dubois and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York. 1988.

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

[5] E. Kriegler and H. Held. Utilizing random sets for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[6] I. Montes, E. Miranda, R. Pelessoni and P. Vicig. Sklar's Theorem in an imprecise setting. *Fuzzy Sets and Systems*, DOI: 10.1016/j.fss.2014,10.007, 2014.

[7] R. Nelsen. *An introduction to copulas*. Springer, New York. 1999.

[8] R. Pelessoni, P. Vicig, I. Montes and E. Miranda. Bivariate p-boxes. *Submitted to publication*, 2014.

[9] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princerton, NJ. 1976.

[10] A. Sklar. Fonctions de répartition à n-dimensions et leurs marges. *Publications de l'Institute de Statistique de l'Université de Paris*, 8:229–231, 1959.

[11] M. C. M. Troffaes and S. Destercke. Probability boxes on totally preordered space for multivariate modelling. *International Journal of Approximate Reasoning*, 52(6):767–791, 2011.

[12] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, New York. 1991.

[13] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

A Robust Bayesian Analysis of the Impact of Policy Decisions on Crop Rotations

Lewis Paton
Durham University, UK
l.w.paton@durham.ac.uk

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Nigel Boatman, Mohamud Hussein
The Food and Environment Research Agency, UK
{nigel.boatman, mohamud.hussein}@fera.gsi.gov.uk

Abstract

We analyse the impact of a policy decision on crop rotations, using the imprecise land use model that was developed by the authors in earlier work. A specific challenge in crop rotation models is that farmer's crop choices are driven by both policy changes and external non-stationary factors, such as rainfall, temperature and agricultural input and output prices. Such dynamics can be modelled by a non-stationary stochastic process, where crop transition probabilities are multinomial logistic functions of such external factors. We use a robust Bayesian approach to estimate the parameters of our model, and validate it by comparing the model response with a non-parametric estimate, as well as by cross validation. Finally, we use the resulting predictions to solve a hypothetical yet realistic policy problem.

Keywords. multinomial logistic regression, stochastic process, robust Bayesian, conjugate, maximum likelihood, crop, decision

1 Introduction

This paper investigates a specific actual real-world problem, namely how imprecise probability can be used to inform policy, in a way that reflects limited data and lack of information to policy makers. In general, policy decisions aim to balance the greater good to society with the welfare of the individual, in terms of economic costs and benefits from a policy implementation. For example, farmers typically grow crops to maximise their profits, however governments can influence this decision through policy interventions to meet the needs of society, such as biodiversity, economic resilience, and security of supply.

An issue which has received a lot of attention recently concerns changes in crop rotations, which are linked to negative environmental impact, reduced diversification of crops and reduced self-sufficiency in feed and food. Concerning animal feed, protein demand has increased a lot, due to increasing meat demand from developing countries. Also, the use of European legumes such as peas and beans [11] has declined. At the moment, the UK imports most of its protein; however, these prices are going up due to growing global demand for soya [7]. Simultaneously, growing more protein can improve diversity, and thereby increase resistance against disease and climate change, and improve supply security [8]. For these reasons, reforms of the Common Agricultural Policy that are now being implemented includes two measures specifically aimed at increasing the amount of protein crops grown [2].

We will look at a hypothetical scenario to see how nitrogen price affects the amount of legumes being grown. Legumes produce their own nitrogen, and so require little nitrogen based fertiliser. As such, one expects that farmers tend to grow less fertiliser dependant crops as nitrogen prices increase. We will formulate and answer a hypothetical decision problem which illustrates the types of problems that can be solved using land use models.

Farmers generally grow crops in rotation to prevent build-up of pests and diseases, and thereby to maximise yields and profit margins. The optimal crop choices vary with soil type and climate conditions. The rotation is generally driven by the length of the period required between successive plantings of the most valuable crop that can be grown, in order to allow pests and diseases to decline to non-damaging or readily controllable levels. Rotating crops also spreads risk in the face of weather variability and annual fluctuations in commodity prices.

Modelling crop distributions across time and space is highly non-trivial. Building a statistical model for farmers' crop choices is difficult, because there are so many factors that influence a farmer's choice. We need to take care in picking the relevant major influencing factors. Moreover, although we have a reasonably sized database, some crop types and factor levels are quite

rare. Furthermore, prior expert information is difficult to obtain. Thus, building a model capable of making reasonable inferences about future crop distributions is a difficult problem.

Building on the work of Luo [9], and Chen and Ibrahim [4], we previously developed a land use model that accurately captures uncertainty in the modelling process [16, 13]. In that work, a non-stationary stochastic process models crop choice, where crop transition probabilities are multinomial logistic functions, and predictions are based on sets of conjugate priors and MAP estimates for efficient sensitivity analysis. Here, we will use this model to answer the hypothetical policy question discussed earlier.

Compared to our earlier work in this domain [16, 13], the novel contributions of this paper are: (i) We train our model on a much larger data set, and handle a larger number of crop types. (ii) We deal with numerical stability issues resulting from near-zero counts. (iii) We propose a non-parametric estimation method, which is, as far as we know, new in the literature. (iv) We validate our model, using two different approaches: formally through classification based accuracy measures, and heuristically through comparison with non-parametric estimates. (v) We propose a new method for the decision analysis based on MAP estimation. (vi) We apply our model to a hypothetical policy decision problem.

The paper is structured as follows. Section 2 describes the land use model from [13]. Section 3 explains the set of priors and posterior inferences. Section 4 shows some of the results from the model. Section 5 describes the model validation. Section 6 analyses a decision problem. Section 7 concludes the paper.

2 The Model

We model crop rotations on a particular field as a non-stationary stochastic process, with J states, corresponding to J crop choices. The crop grown at time k is denoted by Y_k. The choice of Y_{k+1} is influenced by regressors $X_k = (X_{k0}, X_{k1}, \ldots, X_{kM})$, as well as by Y_k, but is otherwise independent of the history of the system. As usual in a regression analysis, we set $X_{k0} = 1$. We denote the transition probabilities by

$$\pi_{ij}(x) = P(Y_{k+1} = j \mid Y_k = i, X_k = x) \qquad (1)$$

We assume a multinomial logistic regression model for $\pi_{ij}(x)$, with $J^2(M+1)$ model parameters β_{ijm}, where $i \in \{1, \ldots, J\}$, $j \in \{1, \ldots, J\}$, and $m \in \{0, \ldots, M\}$:

$$\pi_{ij}(x) = \frac{\exp(\beta_{ij}x)}{\sum_{h=1}^{J} \exp(\beta_{ih}x)} \qquad (2)$$

i	x_1	x_2	$n_i(x)$	$k_{i1}(x)$	$k_{i2}(x)$	$k_{i3}(x)$	$k_{i4}(x)$
1	93	112	2	0	1	0	1
2	56	154	1	0	0	1	0
1	85	110	1	0	0	0	1
3	30	90	1	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Crop rotation data for a particular soil type, where i is the previous crop grown, x_1 is the observed rainfall, x_2 is the nitrogen price, $n_i(x)$ is the current crop total for i and x, and $k_{ij}(x)$ is the number of crop j being grown.

with $\beta_{ij}x \coloneqq \sum_{m=0}^{M} \beta_{ijm}x_m$. Without loss of generality we can set $\beta_{iJm} = 0$ for all i and m, and call this the *baseline category logit model* [3].

Soil type is a significant driver of crop choice. Following [9], we split our data by soil type, and perform a separate analysis for each soil type. For ease, we do not index our model parameters by soil type.

For estimation, we have $n_i(x)$ observations where the previous crop was i, and the regressors were x. Obviously $n_i(x)$ will be zero at all but a finite number of $x \in \mathcal{X}$, where $\mathcal{X} = \{1\} \times \mathbb{R}^M$. Of these $n_i(x)$ observations, the crop choice was j in $k_{ij}(x)$ cases. Obviously, $n_i(x) = \sum_{j=1}^{J} k_{ij}(x)$ for each i. Table 1 shows an extract from the data set.

The following conjugate prior for the model parameters β was proposed in [13]:

$$f_0(\beta|s_0, t_0) \propto \exp\left(\sum_{i=1}^{J} \sum_{x\in\mathcal{X}} s_{0i}(x) \Big[\sum_{j=1}^{J} t_{0ij}(x)\beta_{ij}x - \log \sum_{j=1}^{J} \exp(\beta_{ij}x)\Big]\right) \qquad (3)$$

where s_{0i} and t_{0ij} are non-negative functions such that $s_{0i}(x) = t_{0ij}(x) = 0$ for all but a finite number of $x \in \mathcal{X}$, with $0 \le t_{0ij}(x) \le 1$ and $\sum_{j=1}^{J} t_{0ij}(x) = 1$ on those points x where $s_{0i}(x) > 0$. This conjugate prior matches the form of the likelihood, and the posterior distribution and parameters are [13]:

$$f(\beta|k, n, s_0, t_0) = f_0(\beta|s_n, t_n) \qquad (4)$$

$$s_{ni}(x) = s_{0i}(x) + n_i(x) \qquad (5)$$

$$t_{nij}(x) = \frac{s_{0i}(x)t_{0ij}(x) + k_{ij}(x)}{s_{0i}(x) + n_i(x)} \qquad (6)$$

3 Inference

Because prior expert opinion is very difficult to obtain in our problem, we use sets of prior densities, similarly to Walley's IDM [18]. Here, we study inferences resulting from a fixed prior function for $s_{0i}(x)$:

$$s_{0i}(x) = \begin{cases} s & \text{if } x \in \mathfrak{X}, \\ 0 & \text{otherwise,} \end{cases} \qquad (7)$$

for some $\mathfrak{X} \subset \mathcal{X}$ and a near vacuous set $\mathfrak{T}$ of prior functions for t_0. Note that in earlier work [13] we used a full vacuous set, however we found that we need to bound the $t_{0ij}(x)$ parameters away from zero at those points where $s_{0i}(x) > 0$ in order to maintain numerical stability in cases where we have very few observations; we chose this bound $\epsilon = 0.01 > 0$ small enough to have no observable impact on the analysis.

$\mathfrak{X}$ is the set of regressor values where we specify prior beliefs. It can be any finite subset of $\mathcal{X}$, but we note that the inferences appear more intuitive if $\mathfrak{X}$ is chosen to sensibly cover the range of observed x values [16]. As in the imprecise Dirichlet model [18, Section. 2.5], smaller values of s typically produce tighter posterior predictive bounds. For further discussion of why this choice of priors makes sense, we refer to [13].

A standard way to do the inference now would go via MCMC. However, as we wish to perform a sensitivity analysis against the prior, and the dimension of the parameter space is very large, MCMC is too slow for our purpose. Therefore, we simply use MAP estimation. If we can find a MAP estimate for all $t_0 \in \mathfrak{T}$, we obtain a set B^* of solutions β^*, one for each $t_0 \in \mathfrak{T}$. Each member of B^* corresponds to an estimate of the posterior transition probability. Therefore,

$$\underline{\hat{\pi}}_{ij}(x) \approx \inf_{\beta^* \in B^*} \frac{\exp(\beta^*_{ij} x)}{\sum_{h=1}^{J} \exp(\beta^*_{ih} x)} \tag{8}$$

$$\overline{\hat{\pi}}_{ij}(x) \approx \sup_{\beta^* \in B^*} \frac{\exp(\beta^*_{ij} x)}{\sum_{h=1}^{J} \exp(\beta^*_{ih} x)} \tag{9}$$

are the desired lower and upper posterior probability estimates of the transition probability.

4 Case Study

We have crop rotation data from two separate regions in the UK, detailing which crop was grown in every field in each region from 1993 until 2004 [15].

We have data available for a variety of regressors: here we look at rainfall [10] before sowing and the nitrogen price [1]. Rainfall is important as some crops grow better when it is wetter, and some soil types deal with heavy rainfall better. We can assume farmers are interested in maximising their profit margin. Most fertilisers are nitrogen based, and as such a high nitrogen price will impact profit margins for crops which require large amounts of fertiliser.

We will assume a farmer is faced with a choice of $J = 4$ types of crops: wheat, legumes, rapeseed and all other crops. A common practice is to grow wheat (generally the most profitable crop) followed by a *break crop*, such as legumes or rapeseed. Transitions between legumes and rapeseed are very rare (this only occurred 3 times in roughly 30000 observations). We could leave these transitions in, but they make negligible difference to the inferences, and experts have no interest in these transitions anyway. Therefore, we remove them from the model. Figure 1 depicts all crops and transitions in our model.

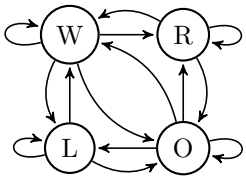


Figure 1: Possible crop transitions, where W is wheat, R is rapeseed, L is legumes, and O is other. Transitions between R and L do not occur in practice so have been excluded from the model.

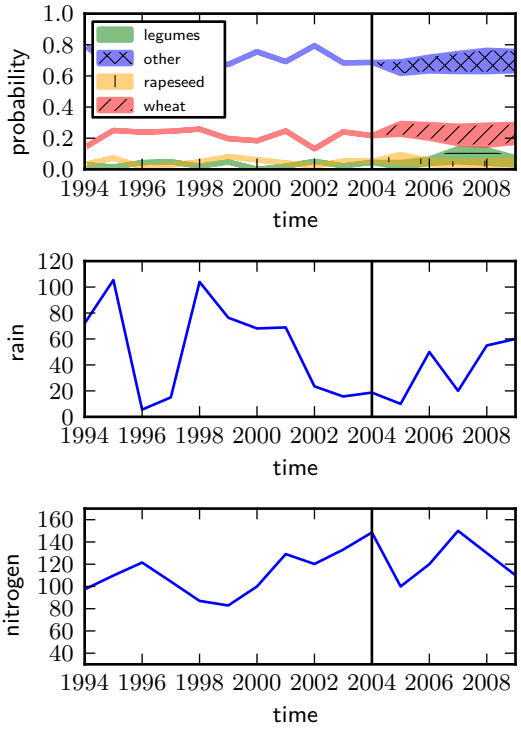


Figure 2: Prediction of future crop distributions on heavy soil given a future scenario.

An important use of land use models is to predict what may happen in the future, given a future scenario for the regressors. For future crop distributions, we use the methodology for imprecise Markov chains developed in [6]. Our initial distribution is calculated empirically from the data and is 23% wheat, 5% rapeseed, 4% legumes and 68% others. Figure 2 shows the results for heavy soil.

The figure shows the historical crop distribution up to 2004 and the values of the regressors until that point. It then shows the predicted crop distributions for the next five years, given a future scenario for the regressors. Here, we have analysed what would happen if future rainfall will be quite low (compared to the observed historical values), and future nitrogen prices will be high. We can see that, although nothing drastic is predicted, legumes seem to increase somewhat. We could compare different scenarios to analyse the impact of changes in, say, nitrogen price. However, a government can influence regressors such as nitrogen price through policy. Thus, it is of interest to study a decision problem which aims to advise this policy. We do this analysis in section 6.

Note that our data runs from 1993 to 2004, so in fact the prediction is until 2009. It would be interesting to compare predictions with actually observed crop distributions, however field level data was no longer being collected from 2005 onwards. It may be possible in the future to validate against satellite data (we currently do not have such data in this study), and thereby to gauge the predictive power of the model.

5 Validation

We discuss two methods for validating the model. A first naive but simple way is to graphically compare the predicted transition probabilities with a non-parametric estimate from the data. A second way is to cross validate the model's predicted best response with parameters estimated from training data against the response as in the test data; this is similar to what is done in classification.

5.1 Non-Parametric Estimates

A simple non-parametric estimate of $\pi_{ij}(x)$ takes a weighted average of the observations around x:

$$\tilde{\pi}_{ij}(x) \coloneqq \frac{\sum_{x' \in \mathcal{X}} w(x - x') k_{ij}(x')}{\sum_{x' \in \mathcal{X}} w(x - x') n_i(x')} \tag{10}$$

where w is some suitably chosen kernel, that is, a non-negative symmetrical function centred around the origin. A key choice in this function is the so-called bandwidth, which quantifies the smoothness of the estimate. We took a multivariate Gaussian kernel:

$$w(x) \coloneqq |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right) \tag{11}$$

with

$$\Sigma^2 \coloneqq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 20^2 & 0 \\ 0 & 0 & 20^2 \end{bmatrix} \tag{12}$$

Note that the first component of x is always taken to be the constant 1, hence only the lower right 2×2 submatrix of Σ^2 is relevant. The choice of 20 for both components was done by trial and error to get sufficiently smooth estimates.

Figure 3 depicts $\tilde{\pi}_{ij}(x)$ as calculated from eq. (10) and $[\hat{\underline{\pi}}_{ij}(x), \hat{\overline{\pi}}_{ij}(x)]$ as calculated from eqs. (8) and (9), for all cases of previous crop i and soil type, as a function of nitrogen price and for a fixed value of rainfall (we chose the historic mean, 55mm). We can see that our model predictions and the non-parametric estimates coincide quite well. The most notable differences are located at the extremes of our observed nitrogen data.

Figure 4 shows a smoothed version of $n_i(x)$, that is:

$$\sum_{x' \in \mathcal{X}} w(x - x') n_i(x') / w(0) \tag{13}$$

These plots give an idea of the size of the denominator in eq. (10), and thereby how much data is near each point x. The lowest data densities are observed from legumes on heavy soil type, where the average number of observations lies around 20. The highest data density is observed from other on light soil type, where we see numbers between 1000 and 2300. This difference in data density is well reflected in the robust Bayesian estimates. The data density decreases substantially as nitrogen price increases, and interestingly our robust Bayesian intervals also become wider in this direction, as desired: we built a robust Bayesian model to capture exactly this sort of feature.

The worst fits are observed in the two bottom right plots, where the robust Bayesian model seems to slightly overestimate the slopes of the curves. We currently have no good explanation as to why this behaviour occurs.

5.2 Cross Validation

A typical method for validating classifiers is to split the data into training and test data, and then to compare the predicted class (or set of classes) from the model based on the training data, with the actual classes in the test data. We can consider our model as a classifier, in the following sense: we compare the farmer's actual choice with the most likely predicted crop. For example, for the predictions in fig. 3, for that particular value of rainfall, the most likely crop from other is other, wheat from legumes and rapeseed, and either other or wheat from wheat, depending on nitrogen price. Of course, in the test data, rainfall will vary as well; fig. 3 just shows a particular slice of the model. Note that our model sometimes produces a set of most likely crops, as we do a sensitivity analysis over all $\beta^* \in B^*$.

For credal classification, there are a number of performance measures [5]. The *determinacy* is the percent-

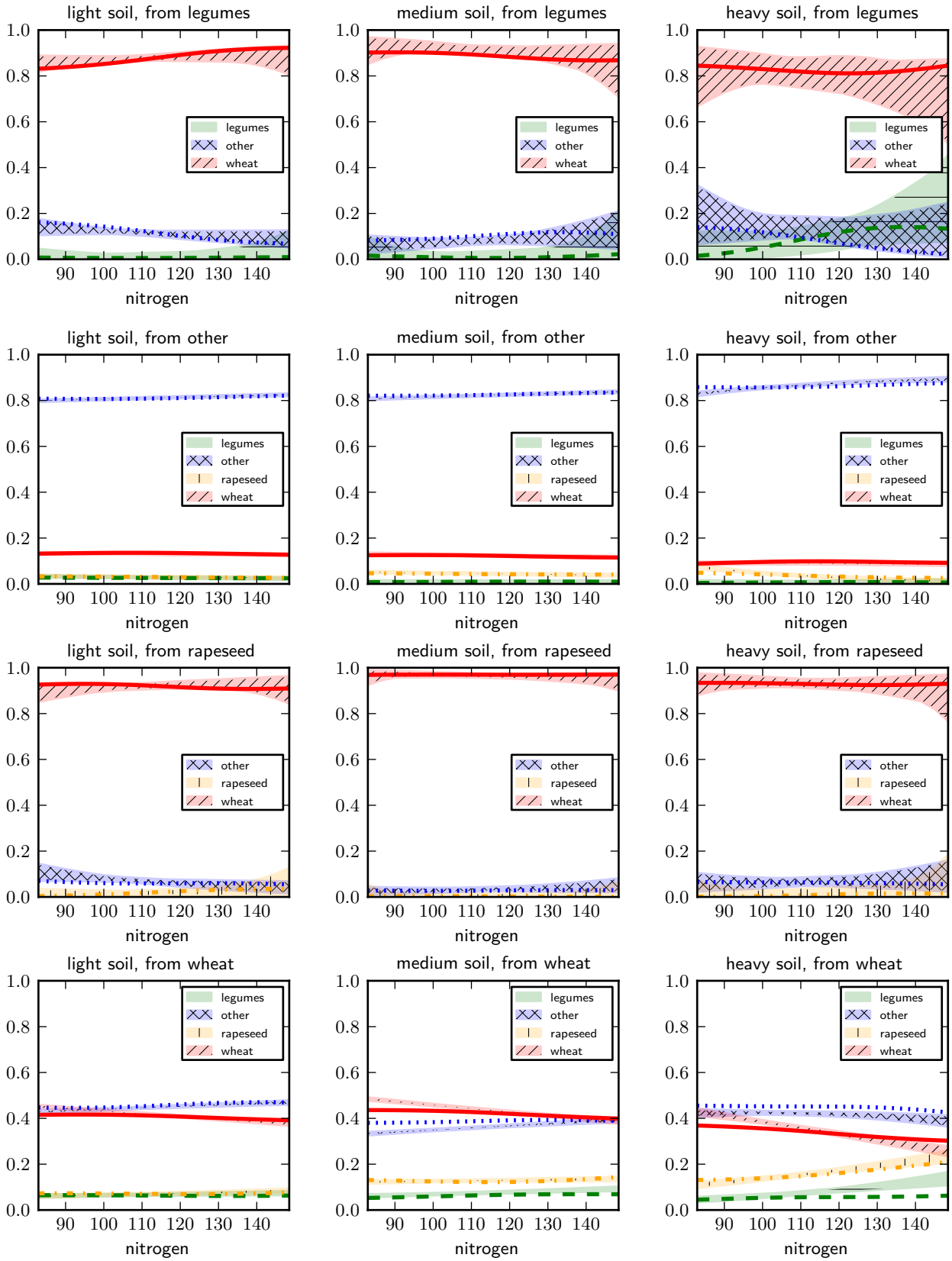


Figure 3: Non-parametric estimates $\tilde{\pi}_{ij}(x)$ and robust Bayesian interval estimates $[\hat{\underline{\pi}}_{ij}(x), \hat{\overline{\pi}}_{ij}(x)]$ for all previous crops i, soil types, as a function of nitrogen price, for fixed rainfall. Probability lies on the y axis.

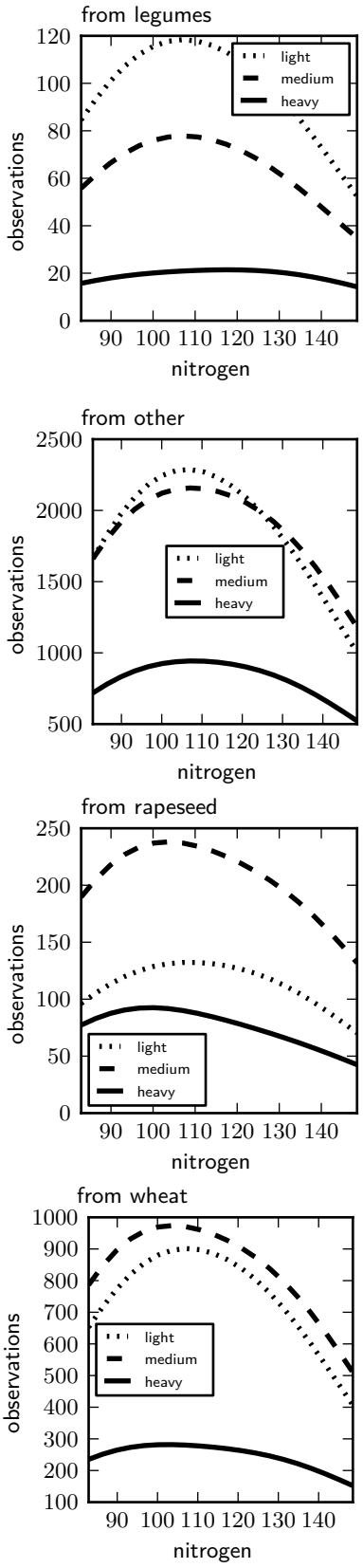


Figure 4: Smoothed $n_i(x)$ as a function of x. This gives an idea of the accuracy of the non-parametric estimates plotted in fig. 3.

region	determinacy	single accuracy	indeterminate output size	set accuracy
Anglia	0.968	0.722	2.008	0.855
Mease	0.988	0.758	2.140	0.929
All	0.976	0.734	2	0.795

Table 2: Cross validation results

age of classifications where the output class is unique. The *single accuracy* is then the accuracy of those predictions. The *indeterminate output size* is the average number of classes when the output class is not unique. Finally, the *set-accuracy* is the percentage of times an indeterminate set contains the correct classification.

To ensure that all the data is used for testing, the analysis is typically repeated, say 10 times, by splitting the original data set into 10 parts, and then repeatedly testing on each of these parts, based on training on the complement of the testing part.

We can use a similar approach to validate our model. We have two distinct geographical regions in our dataset. We perform cross validation within each region, and also combine the two regions together and perform cross validation on the entire data set.

Table 2 presents the results. The single accuracy is quite excellent: our model predicts the correct crop in 70–75% of the cases. The set accuracy is even better, around 80–90%. We note that the determinacy is quite high as a result of the large data set used. This is mostly due to the fact that there is a clear dominant crop type for most combinations of soil and previous crop growing. If we split our analysis by soil and previous crop, we find certain combinations where the determinacy is much lower. Due to space constraints we omit this analysis here, as it only affects determinacy in a substantial way. Indeed, we can already now tell that the set accuracy will on average remain about 80–90%, which indicates that the model performs very well. Finally, we note that the set accuracy is at its lowest for the full data. A logical explanation for this is that the regions are geographically quite distinct. Even despite these differences, the model copes well.

To assess the predictive power of the model, we compared our multicategorical logistic model with a much simpler multinomial model, without covariates, using the imprecise Dirichlet model [18] with $s = 2$ for our priors. Due to the amount of observations in our data set, this model always predicts a single crop type. In regions where data is abundant, the logistic model also outputs a single predicted crop, and the models perform similarly (around 73% accuracy in both cases). However, in regions where the data is sparse and where therefore the logistic model produces a set of predictions, the logistic model has 84% set accuracy,

whereas the multinomial model has only 43% accuracy. This shows the benefits of our logistic model in regions of sparse data.

Note that this method for validation assesses only whether the farmer grows the most likely predicted crop. If this is what we are interested in, then, in regions where there is abundant data, the multinomial model is preferable: it produces similar performance as generally one crop dominates the others, and it is a much simpler model. However, we are interested in understanding the drivers behind farmer's crop choices, and obviously the multinomial model cannot capture this, unlike the logistic model. Consequently, in our view, the traditional classification performance measures are not entirely suitable to assess model performance. This also raises an interesting question in how classification performance measures could be adapted to capture model performance not only related to the most likely predicted class.

6 Policy Example

An important use of land use modelling is to aid policy makers. Changes in policy affect farmer's decisions, and so land use models can predict the impact of these changes. As mentioned in Section 1, there is an interest in the UK in increasing the amount of legumes being grown. Changes in government policy can help to achieve this.

To inform policy makers, we consider a series of scenarios with varying nitrogen price, and thereby investigate the hypothetical impact on crop transitions. Because legumes require far less fertiliser than rapeseed, we expect that an increase in nitrogen price leads to an increased growing of legumes. We emphasize that we have not built a causal model [14], thus one must be wary not to give too strong an interpretation to the inferences presented here.

Both legumes and rapeseed are break crops, so we are particularly interested in transitions from wheat, depicted in the bottom three plots of fig. 3. We see that, for all soil types, as nitrogen prices increase, the amount of legumes grown after wheat increases too. We use these three plots in our policy example.

There is perhaps a more obvious way to approach this problem. The usual way a government would aim to increase levels of legumes is by offering a subsidy to grow them. We have the data available to us to attempt this. By including profit margin as a regressor, we performed an analysis where we altered the profit margin of legumes relative to rapeseed. One would expect that as legumes became relatively more profitable, for example through increased subsidy, more farmers would plant legumes as a break crop instead of rapeseed. However, the results in fact showed the opposite happening.

One potential explanation for this is the format of the data. The profit data we use [12] is actually the predicted profit for the next year. We use this as that is the information farmers have available when making their decision. As such, if there is expected to be an increase in legumes for the next year, then because of supply and demand, there may be a predicted decrease in the profitability of legumes. As such, we suspect there is a confounding variable. In fact, using nitrogen price directly produces more sensible results. Although this makes the analysis less intuitive, for this reason, we proceed with nitrogen price directly.

We are interested in analysing how a farmer's decision responds to changes in nitrogen price. Thus, we assume that the policy maker has some control over the nitrogen price, and we analyse the decision problem from the policy maker's point of view (rather than the farmer's). If the policy maker can specify utilities for different outcomes, then we can use these utilities to make a specific recommendation as to which nitrogen price achieves the best expected utility. In our robust Bayesian setting, we investigate the effect of a wide range of priors on the optimal decision. As legumes are fairly rare in some cases, this allows us to identify situations where we do not have sufficient information in order to arrive at a conclusion.

For the purpose of this paper, we choose a very simple form for the utility function:

$$U(a, b) = 100a - \kappa b \tag{14}$$

where a is the fraction of legumes across all farms, b is the nitrogen price, and κ is chosen to control how this price is weighed against the level of legumes. Note that a is multiplied by 100. This ensures a reasonable scale for the utility, but otherwise makes no technical difference as utility functions are unique up to positive affine transformations. Also, we do not fix any particular value for κ; instead, we investigate our decision problem across a range of κ values.

As before, we do not actually calculate the expected utility, as this is computationally too expensive. Instead, we directly use the MAP estimate for β, and calculate the corresponding value for a

$$a(\beta^*, b) \coloneqq \frac{\exp(\beta^*_{ij} \cdot (1, r, b))}{\sum_{h=1}^{J} \exp(\beta^*_{ih} \cdot (1, r, b))} \tag{15}$$

where $(1, r, b)$ is x; r is rainfall, which for the purpose of this analysis is kept fixed. Varying r makes no substantial difference to the conclusions of our study. As here we are only interested in transitions from

wheat to legumes, i represents wheat and j represents legumes. The (approximate) optimal decision is then

$$\arg \max_{b\in[b_1,b_2]} U(a(\beta^*, b), b) \tag{16}$$

where $a(\beta^*, b)$ is the fraction of legumes in the model with MAP parameter β^* and nitrogen price b.

In our robust setting, we actually have a set B^* of β^* values. We use interval dominance, due to the simplicity by which it can be computed and graphically represented. Specifically, with

$$\underline{U}(b) \coloneqq \inf_{\beta^*\in B^*} U(a(\beta^*, b), b) \tag{17}$$

$$\overline{U}(b) \coloneqq \sup_{\beta^*\in B^*} U(a(\beta^*, b), b) \tag{18}$$

all $b \in [b_1, b_2]$ that satisfy

$$\overline{U}(b) \geq \max_{b\in[b_1,b_2]} \underline{U}(b) \tag{19}$$

are deemed optimal. We have taken the values b_1 and b_2 to be the lowest and highest observed historical nitrogen price. These are the values our model is built on. Therefore, in our decision problem we vary nitrogen price over the range of values we have previously observed. Figure 5 shows $[\underline{U}(b), \overline{U}(b)]$ when moving from wheat on each soil type and for various values of κ. The horizontal black line represents $\max_{b\in[b_1,b_2]} \underline{U}(b)$. Values of b for which $\overline{U}(b)$ lies above this line are optimal by interval dominance. Of course, in reality, a government would not base policy on previous crop or soil. However, we present this analysis as it shows a variety of interesting features, and also compares well with the validation plots in fig. 3.

The same trends are observable across all soil types. When $\kappa = 0$, we are saying that the policy maker is indifferent to changes in nitrogen price. As such a high nitrogen price is desirable, as the model predicts this leads to an increase in legume growth. Thus, the values of b which are optimal are high.

As we increase κ, eq. (14) says that a higher nitrogen price is becoming more detrimental to society. As such, we expect lower values of b to become optimal. Eventually, we reach a point for which all b are optimal. For example, on light soil this occurs at $\kappa = 0.02$.

Eventually we reach a stage where a high nitrogen price is highly undesirable for society, regardless of the benefits that it brings with respect to increased legume growth. For example, on medium soil and $\kappa = 0.07$, only b values less than 100 are optimal.

For a policy maker, once decided on a value of κ (which would be determined by the policy maker determining what scenarios they are indifferent between), then the job would be to determine how to alter the nitrogen price to suit society's needs. For example, on heavy soil with $\kappa = 0.06$, a high nitrogen price is beneficial to society. As such, a government could increase tax on nitrogen to increase the price of it. On the other hand, for heavy soil and $\kappa = 0.2$ government could decrease tax on nitrogen.

We stress again that the above analysis is purely hypothetical. We made unrealistic assumptions, and made no attempt at modelling causal relationships, so the conclusions drawn above in no way represent realistic policy proposals. Instead we demonstrated mathematical techniques for aiding policy making. Only if we had suitable data, a suitable utility function, and a suitable choice of causal covariates, could we draw hard policy conclusions from the results.

7 Summary and Conclusions

In this paper we further developed the previously proposed land use model from [13]. The model uses multinomial imprecise logistic regression with sets of conjugate prior distributions, on a non-stationary stochastic process. We obtained robust Bayesian bounds on the posterior transition probabilities of growing wheat, legumes, rapeseed or anything else, as functions of rainfall and nitrogen price. Compared to previous work we trained our model on a much larger data set. We addressed numerical stability issues by use of a near vacuous set of priors to bound probabilities away from zero.

We validated our model in two ways: comparing a non-parametric estimate with the robust Bayesian interval estimate, and by performing cross-validation. The results show that our model performs well, particularly in areas where there are few observations.

We formulated and answered a hypothetical decision problem with real-world relevance. We investigated what level of nitrogen price is most beneficial to society to promote legume growth. We used interval dominance to identify optimal policies due to its graphical representability and computational simplicity. We demonstrated how land use modelling can aid policy makers, and how imprecise probability can help to solve real world problems.

On a critical note, we may wonder about what is the advantage of using an imprecise probability model as opposed to a precise non-parametric model, or a precise Bayesian model? Indeed, confidence intervals on the parameters could be easily obtained through the non-parametric model that we introduced in eq. (10)—albeit with all the issues that come with such estimates particularly in regions where the data is sparse and where we do not believe that eq. (10) is accurate. Similarly, credible intervals could be obtained

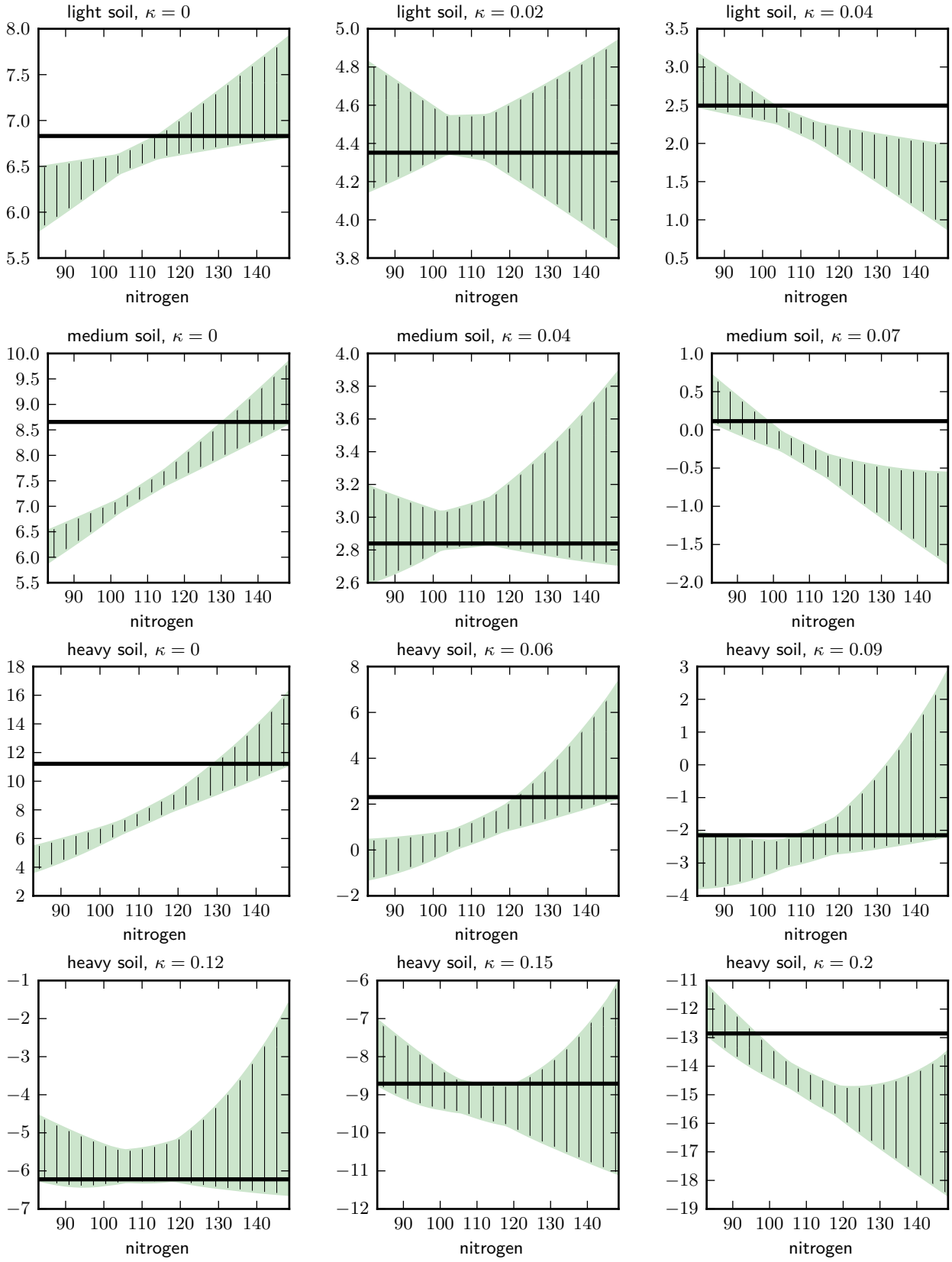


Figure 5: $[\underline{U}(b), \overline{U}(b)]$ when moving from wheat to legumes on all soil types, for various values of κ. Utility lies on the y axis.

through MCMC on just a single prior. However, for decision making, we need expected utility (or, loss), not confidence intervals or credible intervals. A precise Bayesian model always gives an exact expectation, and one would still worry about sensitivity against the prior, thereby ending up doing exactly what we do in the paper. Moreover, it is well known that the simplest way to find admissible frequentist decisions goes through a robust Bayesian analysis [17]. So, frequentists should find our analysis also quite appealing, provided they accept the parametric model.

Future work will concentrate on analysing decision problems in a more realistic way. Our data set is quite old—after 2004 field level data was not collected in the UK. However, it is planned to start again in the near future, meaning the model can be built on more relevant data. We plan on obtaining legume subsidy price, and including that as a regressor to see if that stops the confounding error discussed in section 6. The profit margin of a crop is simply a function of various factors, including subsidy level. Thus, including subsidy directly in the model as a regressor will be straightforward. We also plan to investigate other decision criteria, such as maximality and E-admissibility, particularly when interval dominance leads to vacuous decisions. The utility function could also be enhanced to account for risk aversion, and other factors that influence the benefits to society.

Acknowledgements

The first two authors are supported by the Food and Environment Research Agency (FERA) and EPSRC. We also thank Andy Hart (FERA) for his input.

References

[1] Data collected by DEFRA for the index of the purchase prices of the means of agricultural production.

[2] Legume futures report. `http://www.legumefutures.de/`. Accessed: 18/02/2015.

[3] Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, third edition, 2013.

[4] Ming-Hui Chen and Joseph G. Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, 13:461–476, 2003.

[5] Giorgio Corani and Marco Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(4):581–621, 2008.

[6] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, October 2009. doi: 10.1017/S0269964809990039.

[7] C. L. Gilbert and C. W. Morgan. Food price volatility. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1554):3023–3034, 2010. doi: 10.1098/rstb.2010.0139.

[8] Erik Steen Jensen, Mark B. Peoplese, Robert M.Boddey, Peter M. Gresshoff, Henrik Hauggaard-Nielsen, Bruno J.R. Alves, and Malcolm J. Morrison. Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries. a review. *Agronomy for Sustainable Development*, 32 (2):329–364, 2012. doi: 10.1007/s13593-011-0056-7.

[9] Weiqi Luo. Land use modelling. Internal Report, Food and Envirnoment Research Agency, 2010.

[10] MET. Data collected by the Met Office. `http://www.metoffice.gov.uk/climate/uk/stationdata/`. Accessed: 11/02/2013.

[11] Thomas Nemecek, Julia-Sophie von Richthofen, Gaëtan Dubois, Pierre Casta, Raphaël Charles, and Hubert Pahl. Environmental impacts of introducing grain legumes into european crop rotations. *European Journal of Agronomy*, 28:380–393.

[12] John Nix. *Farm Management Pocketbook*. Agro Business Consultants Ltd., 1993–2004.

[13] Lewis Paton, Matthias C. M. Troffaes, Nigel Boatman, Mohamud Hussein, and Andy Hart. Multinomial logistic regression on Markov chains for crop rotation modelling. In Anne Laurent, Oliver Strauss, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Proceedings of the 15th International Conference IPMU 2014 (Information Processing and Management of Uncertainty in Knowledge-Based Systems, 15–19 July 2014, Montpellier, France)*, volume 444 of *Communications in Computer and Information Science*, pages 476–485. Springer, 2014. doi: 10.1007/978-3-319-08852-5_49.

[14] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. doi: 10.1214/09-SS057.

[15] RPA. Data collected by the Rural Payments Agency under the integrated administration and control system for the administration of subsidies under the common agricultural policy.

[16] Matthias C. M. Troffaes and Lewis Paton. Logistic regression on Markov chains for crop rotation modelling. In F. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA'13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pages 329–336, Compiègne, France, July 2013. SIPTA. URL `http://www.sipta.org/isipta13/index.php?id=paper&paper=033.html`.

[17] Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, December 1939. doi: 10.1214/aoms/1177732144.

[18] Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58(1):3–34, 1996. URL `http://www.jstor.org/stable/2346164`.

Dilation, Disintegrations, and Delayed Decisions

Arthur Paul Pedersen
Center for Adaptive Rationality
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin
pedersen@mpib-berlin.mpg.de

Gregory Wheeler
Munich Center for Mathematical Philosophy
Ludwig Maximilians University
Geschwister-Scholl-Platz 1, 80539 Munich
gregory.wheeler@lrz.uni-muenchen.de

Abstract

Both dilation and non-conglomerability have been alleged to conflict with a fundamental principle of Bayesian methodology that we call *Good's Principle*: one should always delay making a terminal decision between alternative courses of action if given the opportunity to first learn, at zero cost, the outcome of an experiment relevant to the decision. In particular, both dilation and non-conglomerability have been alleged to permit or even mandate choosing to make a terminal decision in deliberate ignorance of relevant, cost-free information. Although dilation and non-conglomerability share some similarities, some authors maintain that there are important differences between the two that warrant endorsing different normative positions regarding dilation and non-conglomerability. This article reassesses the grounds for treating dilation and non-conglomerability differently. Our analysis exploits a new and general characterization result for dilation to draw a closer connection between dilation and non-conglomerability.

1 Introduction

Good's Principle is considered by I. J. Good [8], among others before him [24, 19, 25], to be a fundamental principle of rational decision making. Good's Principle recommends to delay making a terminal decision between alternative courses of action if the opportunity arises to learn, at no cost, the outcome of an experiment relevant to the decision.

Dilation [34, 31, 21] occurs when an interval estimate of an event E is properly included in the interval estimate of E conditional on the occurrence of every event of a measurable partition $\mathcal{B}$. In such circumstances merely running the experiment to determine the value of $\mathcal{B}$, whatever the outcome, suffices to render your initial estimate of E less precise. Should you update your estimate of E to the less precise estimate? Should you refuse a free offer to learn the outcome of such an experiment? Is it rational for you to pay someone to *not* tell you?

A probability function p is *non-conglomerable* [4, 5] for an event E in a measurable partition $\mathcal{B}$ if the marginal probability of E fails to be included in the closed interval determined by the infimum and supremum of the set of conditional probabilities of E given each cell of $\mathcal{B}$. When $\mathcal{B}$ is denumerable, any probability function is non-conglomerable for E in $\mathcal{B}$ only if it fails to be countably additive [4, 5, 26]. In such circumstances merely running the experiment to determine the value of $\mathcal{B}$, whatever the outcome, suffices to uniformly increase (or decrease) your initial estimate of E. Is your initial estimate of E coherent? Is it rational to forgo the opportunity to learn the experimental outcome of $\mathcal{B}$?

Even though both dilation [9, 6] and non-conglomerability [26] have been alleged to conflict with Good's Principle, there is a tradition within the imprecise probability community to treat each differently. Walley, for example, argues that conglomerability is a requirement of rationality in the course of extending coherent lower previsions to conditional lower previsions. More recently, Zaffalon and Miranda argue that conglomerability is a requirement of rationality when an agent's future commitments and current conditional beliefs are established together [36]. Either way, instances of non-conglomerability generate violations of salient dominance principles and allow for the devaluation of cost-free information and thus violations of Good's Principle [14]. Even so, instances of dilation do not preclude violations of salient dominance principles and of Good's Principle – see §5, below – but dilation is viewed as a reasonable, even if surprising, feature of conditional lower previsions [34, §6.4.3]. For Seidenfeld et al. [26], non-conglomerability raises a challenge for those who concede that sometimes rationality permits credal states to be representable by numerically precise probabilities failing to be countable additive. More specifically, Seidenfeld et al. observe that every

instance of non-conglomerability can be transformed into a violation of admissibility, a dominance principle at the heart of the Bayesian enterprise (e.g., Wald, de Finetti, Savage), and that expected utility maximization in such cases admits the devaluation of cost-free information. However, for some decision rules proposed for imprecise probabilities, such as Γ-maximin, dilation also invites a devaluation of cost-free information. But in this case Seidenfeld recommends to reject the decision rule rather than dilation [29].

It is true that while failures of conglomerability can only occur only in infinite partitions, dilation can occur with respect to finite partitions. This observation alone, of course, fails to provide an adequate explanation for adopting a view that treats dilation and non-conglomerability differently with respect to similar problems. In this paper we challenge the practice of treating dilation and non-conglomerability differently. Our analysis appeals to a new and general characterization result for dilation to draw a closer connection between dilation and non-conglomerability

The structure of the paper is as follows. In §2 we review dilation and present our general characterization result purely in terms of distance from independence. Then, in §3 we review the conglomerability principle and rehearse a standard example of non-conglomerability. In §4, we discuss Good's Principle in more detail and introduce a general framework within which to express Good's Principle, as it is commonly understood, in terms of subjective expected utility. Then, in §5 we discuss various violations of Good's Principle, with special attention to two examples in particular, one involving dilation and the other involving non-conglomerability. In particular, we argue that the normative standing of Good's Principle in the dilation case depends on particular features of the uncertainty model and the decision rules used, both of which depend ultimately on the decision maker's beliefs, values and goals. We then turn to an example involving non-conglomerability to argue that such examples should be treated in the same fashion, that is, that the normative standing of conglomerability likewise depends on the features of the uncertainty model and decision rules the decision maker uses.

2 Dilation

A *lower probability space* is a quadruple $(\Omega, \mathcal{A}, \mathbb{P}, \underline{\mathrm{P}})$ such that Ω is a set of states, $\mathcal{A}$ is an algebra over Ω, $\mathbb{P}$ is a nonempty set of probability functions on $\mathcal{A}$, and $\underline{P}$ is a *lower probability function* on $\mathcal{A}$ with respect to $\mathbb{P}$—that is, $\underline{\mathrm{P}}(E) = \inf\{p(E) : p \in \mathbb{P}\}$ for each $E \in \mathcal{A}$. The value $\underline{\mathrm{P}}(E)$ is called the *lower probability* of E. The *upper probability function* $\overline{\mathrm{P}}$ is then defined in the usual manner by stipulating that $\overline{\mathrm{P}}(E) = 1 - \underline{\mathrm{P}}(E^c)$ for each $E \in \mathcal{A}$; the value $\overline{\mathrm{P}}(E)$ is called the *upper probability* of E. If $\underline{\mathrm{P}}(H) > 0$, then conditional lower and upper probabilities are defined as $\underline{\mathrm{P}}(E \mid H) = \inf\{p(E \mid H) : p \in \mathbb{P}\}$ and $\overline{\mathrm{P}}(E \mid H) = \sup\{p(E \mid H) : p \in \mathbb{P}\}$, respectively. In the following, we call a collection of events $\mathcal{B}$ from $\mathcal{A}$ a *positive measurable partition* (of Ω) if $\mathcal{B}$ is a partition of Ω such that $\underline{\mathrm{P}}(H) > 0$ for each $H \in \mathcal{B}$.

Let $\mathcal{B}$ be a positive measurable partition of Ω. We say that $\mathcal{B}$ *dilates* E just in case for each $H \in \mathcal{B}$:

$$\underline{\mathrm{P}}(E \mid H) \;<\; \underline{\mathrm{P}}(E) \;\leq\; \overline{\mathrm{P}}(E) \;<\; \overline{\mathrm{P}}(E \mid H).^{1}$$

In other words, $\mathcal{B}$ dilates E just in case the closed interval $[\underline{\mathrm{P}}(E), \overline{\mathrm{P}}(E)]$ is contained in the open interval $(\underline{\mathrm{P}}(E \mid H), \overline{\mathrm{P}}(E \mid H))$ for each $H \in \mathcal{B}$.

What is remarkable about dilation is the specter of turning a more precise estimate of E into a less precise estimate, no matter what event from the partition occurs.

Next, in §2.1, we rehearse an example from [28] involving a maximally uncertain event, G, a flip of a fair coin (whose outcomes form a partition, $\mathcal{B}$) and a pivotal quantity, E, defined in terms of G and the outcome of the coin toss. Then, in §2.3, we provide a simple characterization of dilation in terms of distance from stochastic independence, followed by a short discussion of the result.

2.1 Example of Dilation

Suppose G is a highly uncertain event, one with upper probability close to 1, $\overline{P}(G) = .9$, and lower probability close to 0, $\underline{P}(G) = .1$. So,

$$\overline{P}(G) - \underline{P}(G) = 0.8. \tag{1}$$

Suppose now that $\mathcal{B} = \{H, H^c\}$ is a partition representing the outcomes of a fairly tossed coin,

$$\underline{P}(H) = \overline{P}(H) = \frac{1}{2} = \underline{P}(H^c) = \overline{P}(H^c). \tag{2}$$

In addition to being positively measurable, suppose the outcomes of the toss are *stochastically independent* of our maximally uncertain event. In particular, the event of the coin landing heads, H, is stochastically independent of G occurring; hence, for each $p \in \mathbb{P}$,

$$p(G \cap H) = p(G)p(H) = \frac{p(G)}{2}. \tag{3}$$

Next let E be the event of either G and H both occurring or both failing to occur, namely $E := (G \cap$

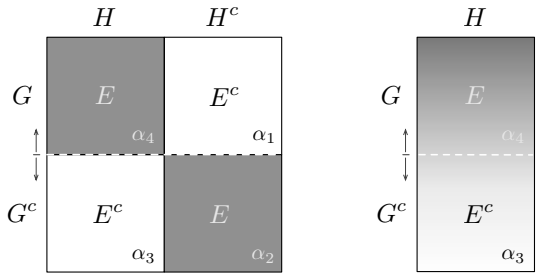


Figure 1: (a) 2x2 Table for an uncertain event (row) and a fair coin randomizer (column); (b) The event of learning that the outcome of the coin toss is 'heads'.

$H) \cup (G^c \cap H^c)$. The probability of E is determinate: $p(E) = \frac{1}{2}$. Similarly, for each $p \in \mathbb{P}$, $p(E^c) = 1/2$.

These conditions are represented by the two-by-two table in Figure 1(a). Here the *columns* H and H^c represent the two possible outcomes of the fair coin toss; the *rows* G and G^c represent the two outcomes of our maximally uncertain event; the *diagonals* E and E^c describe the two events defined in terms of the possible outcomes of row and column: E is the "matching" event of either H and E both obtaining or neither obtaining, and E^c is the "unmatched" event of one but not the other obtaining.

Observe that E is dilated by $\mathcal{B} = \{H, H^c\}$: although the initial estimate of E is precisely one-half, learning the outcome of the coin toss, whether heads or tails, dilates the probability estimate of E to $[.1, .9]$.

Proof - We show that $0.1 = \underline{P}(E \mid H) < \underline{P}(E) = 1/2$.

$$\begin{aligned}
\underline{P}(E \mid H) &= \inf \{\, p(E \mid H) \,:\, p \in \mathbb{P} \,\} \\
&= \inf \left\{ \frac{p\left([(G \cap H) \cup (G^c \cap H^c)] \cap H\right)}{p(H)} : p \in \mathbb{P} \right\} \\
&= \inf \left\{ \frac{p(G \cap H)}{p(H)} : p \in \mathbb{P} \right\} \\
&= \inf \left\{ \frac{p(G)p(H)}{p(H)} : p \in \mathbb{P} \right\} \\
&= 0.1
\end{aligned}$$

A similar argument establishes $0.9 = \overline{P}(E \mid H) > 1/2$, and the argument holds if instead the coin lands tails, i.e., $\underline{P}(E \mid H^c) = 0.1$ and $\overline{P}(E \mid H^c) = 0.9$. Thus, E is dilated by the coin toss, $\mathcal{B} = \{H, H^c\}$. ◇

The specific case where the coin lands H is illustrated in Figure 1(b). Here conditioning on H reduces the probability that E obtains to the probability that G obtains, which is highly uncertain.

[1]While our terminology agrees with that of [11, p. 252], it differs from that of [31, p. 1141] and [12, p. 412], who call dilation in our sense *strict dilation*.

2.2 Measuring Distance from Independence

Given a single probability function p on $\mathcal{A}$ and events E and H with positive probability, the degree to which two events E and H diverge from stochastic independence, if they diverge at all, may be characterized by a simple measure of distance from stochastic independence:

$$\mathrm{S}_p(E, H) \;:=\; \frac{p(E \cap H)}{p(E)p(H)}.$$

The measure S_p is simply the covariance of E and H, $Cov(E, H) = p(E \cap H) - p(E)p(H)$, put in ratio form. Therefore, $\mathrm{S}_p(E, H) = 1$ just in case E and H are stochastically independent; $\mathrm{S}_p(E, H) > 1$ when E and H are positively correlated; and $\mathrm{S}_p(E, H) < 1$ when E and F are negatively correlated. The measure S_p naturally extends to a set of probability functions $\mathbb{P}$ as follows:

$$\begin{aligned}
\mathrm{S}^+_{\mathbb{P}}(E, H) &:= \{p \in \mathbb{P} : \mathrm{S}_p(E, H) > 1\}; \\
\mathrm{S}^-_{\mathbb{P}}(E, H) &:= \{p \in \mathbb{P} : \mathrm{S}_p(E, H) < 1\}; \\
\mathrm{I}_{\mathbb{P}}(E, H) &:= \{p \in \mathbb{P} : \mathrm{S}_p(E, H) = 1\}.
\end{aligned}$$

The set of probability functions $\mathrm{I}_{\mathbb{P}}(E, H)$ from $\mathbb{P}$ with E and H stochastically independent is called *the surface of independence* for E and H with respect to $\mathbb{P}$. In the remainder subscripts will be dropped when there is no danger of confusion.

2.3 A Simple Characterization of Dilation

In this section, we present simple necessary and sufficient conditions for dilation formulated in terms of deviation from stochastic independence, which improves upon previous results in [21]. We illustrate an immediate application of such a characterization with measures of dilation. To begin, we introduce the notion of a neighborhood.

Given a lower probability space $(\Omega, \mathcal{A}, \mathbb{P}, \underline{\mathrm{P}})$, events $E, H \in \mathcal{A}$ with $\underline{\mathrm{P}}(H) > 0$, and $\epsilon > 0$ define:

$$\begin{aligned}
\underline{\mathbb{P}}(E|H, \epsilon) &:= \{p \in \mathbb{P} : \; |p(E|H) - \underline{\mathrm{P}}(E|H)| < \epsilon\}; \\
\overline{\mathbb{P}}(E|H, \epsilon) &:= \{p \in \mathbb{P} : \; |p(E|H) - \overline{\mathrm{P}}(E|H)| < \epsilon\}.
\end{aligned}$$

We call the sets $\underline{\mathbb{P}}(E|H, \epsilon)$ and $\overline{\mathbb{P}}(E|H, \epsilon)$ lower and upper *neighborhoods* of E conditional on H with radius ϵ, respectively. Thus, a probability function p is an element of $\underline{\mathbb{P}}(E|H, \epsilon)$ if $p(E|H)$ is within ϵ of $\underline{\mathrm{P}}(E|H)$, and similarly for an upper neighborhood.

For the sake of readability in what follows, given a nonempty set of probabilities $\mathbb{P}$, let $\mathbb{P}_*$ denote $\overline{\mathrm{co}}(\mathbb{P})$, the weak*-closed convex hull of $\mathbb{P}$. Thus, $\underline{\mathbb{P}}_*(E|F, \epsilon) =$

$\underline{\text{co}}(\mathbb{P})(E|F,\epsilon)$ and $\overline{\mathbb{P}}_*(E|F,\epsilon) = \overline{\text{co}(\mathbb{P})}(E|F,\epsilon)$. Similarly, let $\text{S}_*^+(E,F)$ and $\text{S}_*^-(E,F)$ be defined by:

$$\text{S}_*^+(E,F) := \{p \in \overline{\text{co}}(\mathbb{P}) : S_p(E,F) > 1\}$$
$$\text{S}_*^-(E,F) := \{p \in \overline{\text{co}}(\mathbb{P}) : S_p(E,F) < 1\}.$$

Given a nonempty set I, we let $\mathbb{R}_+^I$ denote the set of elements $(r_i)_{i\in I}$ of $\mathbb{R}^I$ such that $r_i > 0$ for each $i \in I$. We now state a result characterizing dilation and then report an immediate corollary.

Theorem 1 Let $(\Omega, \mathcal{A}, \mathbb{P}, \underline{\mathbb{P}})$ be a lower probability space, let $\mathcal{B} = \{H_i : i \in I\}$ be a positive measurable partition, and let $E \in \mathcal{A}$. Then the following are equivalent:

(i) $\mathcal{B}$ dilates E;

(ii) There is $(\epsilon_i)_{i\in I} \in \mathbb{R}_+^I$ such that for every $i \in I$:

$$\underline{\mathbb{P}}_*(E|H_i,\epsilon_i) \subseteq \text{S}_*^-(E,H_i) \text{ and}$$
$$\overline{\mathbb{P}}_*(E|H_i,\epsilon_i) \subseteq \text{S}_*^+(E,H_i);$$

(iii) There is $(\epsilon_i)_{i\in I} \in \mathbb{R}_+^I$ such that for every $i \in I$:

$$\underline{\mathbb{P}}(E|H_i,\epsilon_i) \subseteq \text{S}^-(E,H_i) \text{ and}$$
$$\overline{\mathbb{P}}(E|H_i,\epsilon_i) \subseteq \text{S}^+(E,H_i),$$

where for each $i \in I$, $\epsilon_i \leq \min(\underline{\epsilon}_i, \overline{\epsilon}_i)$ and $\underline{\epsilon}_i$ is the unique minimum of $|p(E|H_i) - \underline{\text{P}}(E|H_i)|$ attained on $C_i^+ =_{df} \{p \in \mathbb{P}^* : \text{S}_p(E,H_i) \geq 1\}$, and $\overline{\epsilon}_i$ is the unique minimum of $|p(E|H_i) - \overline{\text{P}}(E|H_i)|$ attained on $C_i^- =_{df} \{p \in \mathbb{P}^* : \text{S}_p(E,H_i) \leq 1\}$. ◇

Theorem 1 implies that a positive measurable partition $\mathcal{B}$ dilates an event E just in case for each partition cell H, there are upper and lower neighborhoods of E conditional on H such that the lower neighborhood of E conditional on H lies entirely within the subset of the set of probabilities in question for which E and H are negatively correlated, while the upper neighborhood of E given H lies entirely within the subset of the set of probabilities in question for which E and H are positively correlated. We remark that Theorem 1 holds for *arbitrary* nonempty sets of probabilities.

When $\mathcal{B}$ is a finite positive measurable partition, the preceding theorem may be simplified.

Corollary 1 Let $(\Omega, \mathcal{A}, \mathbb{P}, \underline{\mathbb{P}})$ be a lower probability space, let $\mathcal{B} = (H_i)_{i=1}^n$ be a finite positive measurable partition, and let $E \in \mathcal{A}$. Then the following are equivalent:

(i) $\mathcal{B}$ dilates E;

(ii) There is $\epsilon > 0$ such that for each $i = 1, \ldots, n$:

$$\underline{\mathbb{P}}_*(E|H_i,\epsilon) \subseteq \text{S}_*^-(E,H_i) \text{ and}$$
$$\overline{\mathbb{P}}_*(E|H_i,\epsilon) \subseteq \text{S}_*^+(E,H_i);$$

(iii) There is $\epsilon > 0$ such that for each $i = 1, \ldots, n$:

$$\underline{\mathbb{P}}(E|H_i,\epsilon) \subseteq \text{S}^-(E,H_i) \text{ and}$$
$$\overline{\mathbb{P}}(E|H_i,\epsilon) \subseteq \text{S}^+(E,H_i),$$

where $\epsilon \leq \min(\underline{\epsilon}_i, \overline{\epsilon}_i : i = 1, \ldots, n)$ and $\underline{\epsilon}_i$ is the unique minimum of $|p(E|H_i) - \underline{\text{P}}(E|H_i)|$ attained on $C_i^+ =_{df} \{p \in \mathbb{P}^* : \text{S}_p(E,H_i) \geq 1\}$, and $\overline{\epsilon}_i$ is the unique minimum of $|p(E|H_i) - \overline{\text{P}}(E|H_i)|$ attained on $C_i^- =_{df} \{p \in \mathbb{P}^* : \text{S}_p(E,H_i) \leq 1\}$. ◇

Thus, when the positive measurable partition $\mathcal{B}$ is finite, the radii ϵ_i of Theorem 1 may be replaced by a *single* positive radius ϵ. The preceding corollary also improves upon a similar result in [21].

Discussion. Theorem 1 and Corollary 1 should hardly be surprising. The correlation properties that entail dilation are rather straightforward consequences of the definition. Moreover, these correlation properties entail that each dilating partition cell and dilated event live on the surface of independence under *some* probability function from the closed convex hull of the set of probabilities in question. Although straightforward, Theorem 1 shows that by looking *beyond* the upper and lower supporting hyperplanes $\overline{\mathbb{P}}_*(E|H)$ and $\underline{\mathbb{P}}_*(E|H)$ to the upper and lower supporting *neighborhoods* $\overline{\mathbb{P}}_*(E|H,\epsilon)$ and $\underline{\mathbb{P}}_*(E|H,\epsilon)$, it becomes possible to characterize dilation completely in terms of positive and negative correlation, achieving a longstanding goal. The results also show that dilation, properly understood, is a property of the *convex closure* of a set of probabilities.

One may see the generality of Theorem 1 by comparing it to an earlier result in [35, Result 1]. Observe that (1) Theorem 1 applies to *arbitrary* positive measurable partitions, whereas [35, Result 1] applies only to binary partitions; (2) Theorem 1 applies to *arbitrary* sets of probabilities, whereas Result 1 just applies to weak*-closed convex sets of probabilities; and (3) Theorem 1 presents characterizing conditions—property (ii) and property (iii)—formulated in terms of upper and lower neighborhoods, whereas Result 1 gives a characterizing condition formulated in terms of a patchwork of infimums and supremums—a point we discuss further in [21, §4]. Of course, Theorem 1, given its generality, entails that the characterizing condition of Result 1 in [35] is logically equivalent to property (ii)—or property (iii)—of Theorem 1 in the very special case for binary partitions and weak*-closed convex sets of

probabilities. Yet, in our judgment, the characterizing condition of Result 1, even with its narrow scope, is periphrastic. The upshot is that Theorem 1, in spite of its wide scope, delivers characterizing conditions which succinctly express the *wherefore* of dilation.

Last, returning to the simple heuristic example of dilation we presented in §2, we remark that a straightforward calculation of the relevant radii $\underline{\epsilon}_1, \overline{\epsilon}_1, \underline{\epsilon}_2, \overline{\epsilon}_2$ corresponding to H and H^c, respectively, yields $\frac{2}{5}$.

2.4 Proper and Improper Dilation

It is well known that the familiar univocal notion of probabilistic independence splinters into a plurality of logically distinct independence concepts [34, 2]. Thus, if a decision modeler knows that one event is epistemically independent of another – that is, that each event is epistemically irrelevant to the other – then he knows that observing the outcome of one event does not change the estimate in the other, even though the two events may fail to be stochastically independent, and thus may admit dilation. In other words, our characterization results hold for a variety of extensions—including unknown interaction, irrelevant natural extensions, and independent natural extensions [2]—without discriminating between models which correctly or incorrectly encode knowledge of either epistemic irrelevance or epistemic independence. However, our proposal is that a correctly parameterized extension *can* provide principled grounds for avoiding the loss of precision by dilation that may otherwise come from updating. So, even if the conditions for Theorem 1 hold, there may be enough knowledge about the relationship between the two events in question to support a parameterization that defuses the diluting effect that dilation has from updating. We therefore distinguish between two kinds of dilation phenomenon: *proper dilation*, which occurs within a model that correctly parameterizes the set of distributions to reflect what is known about how the events are interrelated, if anything is known at all, and *improper dilation*, which occurs within a model whose parameterization does not correctly represent what is known about how the events interact.

3 Non-Conglomerability

Given a real-valued finitely additive probability function p on an (σ-) algebra $\mathcal{A}$ over a set of states Ω, a positive measurable partition $\mathcal{B}$ of Ω, and an event E of $\mathcal{A}$, we say that p is *conglomerable* for E in $\mathcal{B}$ if

$$\inf\{\,p(E|H)\colon H\in\mathcal{B}\}\leq p(E)\leq\sup\{\,p(E|H)\colon H\in\mathcal{B}\}$$

Otherwise we say that p is *non-conglomerable* for E in $\mathcal{B}$. So p is non-conglomerable for E in $\mathcal{B}$ just in case $p(E)$ fails to lie in the closed interval $[\inf\{\,p(E|H)\colon H\in\mathcal{B}\}, \sup\{\,p(E|H)\colon H\in\mathcal{B}\}]$.

Of course, every probability function is conglomerable for all events and finite $\mathcal{B}$. Cases of non-conglomerability only arise for infinite $\mathcal{B}$. It is well-known that any probability function with an infinite range is conglomerable for each event E and denumerable $\mathcal{B}$ just in case it is countably additive. In addition, any such probability function is non-conglomerable for some event E and denumerable $\mathcal{B}$ just in case it fails to be *disintegrable* for E in $\mathcal{B}$—that is, if fails to satisfy the law of total probability for E in $\mathcal{B}$. These concepts and results can be extended to bounded random quantities [5] and unbounded random quantities [27]. Further, it should be noted that some probability functions that fail to be countably additive may nonetheless be conglomerable in arbitrary positive measurable partitions. Moreover, in some cases, a nontrivial convex combination of probability functions, each of which fails to be conglomerable in a positive measurable partition, may very well be conglomerable in the partition. Indeed, a nontrivial convex combination of probability functions, each of which *is* conglomerable in a positive measurable partition, may very well fail to be conglomerable in the partition. To gain control over these cases, authors investigated conglomerability within the setting of primitive conditional probability, which accommodates conditioning events with zero probability [1, 26, 30]. Next we give an example of non-conglomerability for a denumerable partition.

3.1 Example of Non-Conglomerability

Following [5],[2] let $\mathcal{A}$ be the collection of all subsets of $\Omega = \{0,1\}\times\mathbb{N}_{>0}$, let $E = \{(1,n) : n\in\mathbb{N}_{>0}\}$, and let $\mathcal{B} = \{H_n : n\in\mathbb{N}_{>0}\}$, where $H_n = \{(0,n),(1,n)\}$ for each $n\in\mathbb{N}_{>0}$. Let p be a finitely-additive probability function on $\mathcal{A}$ such that:

(i) $p(E) = \frac{1}{2}$;

(ii) $p(E\cap H_n) = \frac{1}{2^{n+1}}$ for each $n\in\mathbb{N}_{>0}$; and

(iii) $p(E^c\cap H_n) = 0$.

Then $p(E) < \inf\{p(E|H_n) : n\in\mathbb{N}_{>0}\} = 1$, so p is non-conglomerable for E in the denumerable partition $\mathcal{B}$.

[2] This example seems to have entered the literature in [3, p. 205], although de Finetti there reports that Lester Dubins presented the example in a letter to L.J. Savage.

4 Good's Principle and Expected Utility

In *Foundations of Statistics*, Savage considers the difference between a **basic** decision problem, in which an agent is to choose to perform one action from among several he judges to be available for choice, and a **derived** decision problem, in which the agent is to choose from the same basic actions, but only after considering the associated conditional expected utilities for the basic action given each possible outcome of some experiment. "It is almost obvious," Savage remarks, "that the value of a derived problem cannot be less than, and typically is greater, than the value of the basic problem from which it is derived" [25, §6.2]. Savage thereupon formulates and proves what has become a fundamental principle of Bayesian methodology [25, Chapter 7]. Although Ramsey [24] aired the idea of this result in unpublished work and many others have reaffirmed it following Savage's seminal work (e.g., [22], [19]), Good famously defended the principle in a short article published in the 1960s [8] – and there has been a rich discussion ever since [7, 33, 20, 28, 9, 32, 13]. Following Stigler's law of eponymy, let us briefly explain the basic idea of *Good's Principle*.

4.1 Formalizing Good's Principle

Here is the set up. Suppose that at some time t_1 you are to face a choice among several courses of action $a_1, \ldots, a_n$. Prior to this choice, however, you face a decision at some time t_0 before t_1 between (i) choosing from among several courses of action $a_1, \ldots, a_n$ at time t_1 or (ii) choosing from among the same courses of action $a_1, \ldots, a_n$ at some later time t_2 after you have observed, at no cost, the outcome of an experiment $\mathfrak{E}$.

According to Good's Principle, Bayesian standards prohibit you from rejecting the opportunity to choose from among $a_1, \ldots, a_n$ at t_2 after observing the outcome of the experiment $\mathfrak{E}$. In addition, if the experimental outcome might affect your choice from among the courses of action, then Bayesian standards prohibit you from deciding to choose from among $a_1, \ldots, a_n$ at t_1. In short, to be a Good Bayesian, take Good's advice: accessible cost-free information relevant to a decision should never be ignored.

As a piece of Bayesian legislation, Good's Principle is expressed in the legalese for codifying norms of classical subjective expected utility theory. In order to express Good's Principle in the language of subjective expected utility, we first introduce the formal framework we shall use in our discussion. This framework is sufficiently expressive for our purposes and will enable us to carry out our discussion while remaining neutral over further controversial matters unrelated to our concerns.

Let Ω be a set of states corresponding to a collection of hypotheses which are individually consistent, mutually exclusive, and collectively exhaustive relative to your state of certainty at time t_0. A set of actions A is said to be a *decision problem* for you at time t if it consists of all actions you judge to be available for you to choose. Suppose that for each action a from A and each state ω in Ω, you have identified a unique consequence $\sigma(a, \omega)$ to be relevant for evaluating the action's success in promoting the goals and values you endorse. So, you recognize that if you augment your state of certainty with the hypothetical supposition that you have implemented action a and state ω obtains, your transformed state of certainty commits you to being certain that consequence $\sigma(a, \omega)$ prevails.

We presume that you endorse a standard for decision making that commits you to identifying a nonempty subset $\mathbf{c}(\mathrm{A})$ of your feasible actions A you judge to be *admissible*, or acceptable for choice, given your beliefs, values and goals.

Turn now to Good's Principle illustrated in Figure 2. Suppose that you endorse subjective expected utility maximization as your standard. To sidestep some technical issues, suppose in particular that your judgments of admissibility can be represented in terms of subjective expected utility maximization with respect to a real-valued expectation $\mathbb{E}_p[\,\cdot\,]$ agreeing with a real-valued probability function p defined on a Boolean algebra over the set of states and a real-valued utility function u defined over the set of consequences.[3]

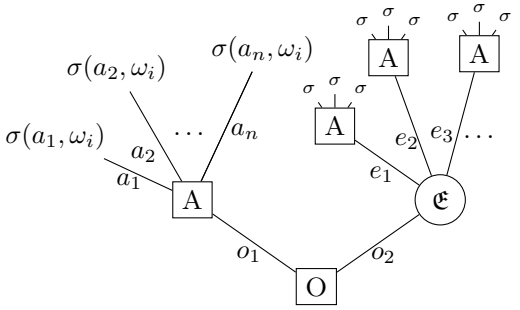


Figure 2: Illustration of Good's Principle

At time t_0 you confront a decision problem $\mathrm{O} = \{o_1, o_2\}$. If you implement option o_1 at time t_0, then at time t_1 you will face a decision problem $\mathrm{A} = \{a_1, \ldots, a_n\}$ without observing the outcome of experiment $\mathfrak{E}$. If you implement option o_2 at time t_0, then at time t_2 you will face the same decision problem

[3] Often a uniqueness result for probabilities and utilities accompanies the representation result (asserting, for example, that the probability function is unique and that the utility function is unique up to a positive affine transformation).

A after observing the outcome of experiment $\mathfrak{E}$. Now under the hypothesis N that at t_1 you face decision problem A after implementing option o_1 at t_0 (not observing the outcome of experiment $\mathfrak{E}$), let $\mathbf{c}(\mathrm{A}|N)$ denote the set of admissible options given N (where $\circ$ denotes functional composition):

$$\mathbf{c}(\mathrm{A}|N) = \arg\max_{a\in\mathrm{A}} \mathbb{E}_{p(\,\cdot\,|N)}\Big[(u\circ\sigma)(a,\, p(d\omega|N)\,)\Big].$$

Similarly, under the hypothesis K_i that at t_2 you face decision problem A after implementing option o_2 at t_0 and observing outcome e_i of experiment $\mathfrak{E}$ at t_1, let $\mathbf{c}(\mathrm{A}|K_i)$ denote the set of admissible options given K_i:

$$\mathbf{c}(\mathrm{A}|K_i) = \arg\max_{a\in\mathrm{A}} \mathbb{E}_{p(\,\cdot\,|K_i)}\Big[(u\circ\sigma)(a,\, p(d\omega|K_i)\,)\Big].$$

Good's principle assumes that at t_0 you are certain, regardless of whether or not you choose to observe the outcome of experiment $\mathfrak{E}$, that you will choose an option A which maximizes your expected utility, that your preferences over consequences remain unchanged, and that your beliefs given hypotheses accord with Bayesian conditionalization. Your expectation of (1) *your maximum conditional expected utility of choosing from* A *given experiment* $\mathfrak{E}$ is not less than your expectation of (2) *your maximum conditional expected utility of choosing from* A *under option* o_1. That is, $o_2 \in \mathbf{c}(\mathrm{O})$, the set of admissible options from O. Moreover, your expectation of (1) is strictly greater than (2) unless there is an action from A that maximizes conditional expected utility from A regardless of the experimental outcome of $\mathfrak{E}$. In other words, unless the experiment is irrelevant, $\mathbf{c}(\mathrm{O}) = \{o_2\}$.

4.2 Remarks on Conditional Probabilities

We wish to remark that your conditional probability judgments, whether precise or imprecise, concern only your commitments at the initial time t_0. In our analysis we adopt a distinction made by Isaac Levi, and suggested, at least roughly, by many others.[4] Specifically, we interpret your conditional probability judgments in one of two ways. First, according to the *called-off* interpretation, your conditional probability judgment given H expresses your commitment at time t_0 to specific *unconditional* attitudes contingent on the occurrence of H. According to de Finetti's theory of previsions, for example, your conditional probability assessment of an event E given an event H at a particular time t_0 expresses your unconditional commitment at time t_0 to judge contracts concerning E that are "called-off" if H does not occur, where they are posited to be nil. Alternatively, according to the *hypothetical* interpretation, your conditional probability judgment given H expresses your commitment at time t_0 to specific attitudes on the hypothetical supposition that H obtains. We contrast these two interpretations of conditional probability judgment with a third *temporal* interpretation which expresses your future commitment to attitudes upon observing that H obtains. In the sequential decision problems discussed in this paper, your current (at t_0) conditional probability judgments given a (possibly) future event H express your assessments on the hypothetical supposition that H is true. Similar remarks apply to other conditional judgments you endorse, such as your conditional value judgments and your conditional assessments of admissibility.

In our view, the question whether conglomerability is an appropriate normative standard for evaluating probability judgments in the senses of interest in this paper remains unsettled.

5 What's so good about Good's Principle?

Although Good's Principle continues to be thought of as a cornerstone of orthodox Bayesianism by critics and champions alike, we maintain that the principle is not ironclad. In this section we consider two examples of violations of Good's Principle in some detail, one involving dilation, another, non-conglomerability.

5.1 Good's Principle and Dilation

Return to the dilation example from [28] that we began in §2.1. Recall that E is defined as the event of *either* the highly uncertain event G and the fair coin toss yielding the event H both occurring *or* both G and H failing to occur, that is $E := (G \cap H) \cup (G^c \cap H^c)$. Recall too that the probability of E and the probability of H are each determinate, namely $p(E) = \frac{1}{2} = p(H)$, whereas the probability of G is highly uncertain, namely $\overline{P}(G) = 0.9$ and $\underline{P}(G) = 0.1$.

Now suppose that at t_0 you face a decision problem $\mathrm{O} = \{o_1, o_2\}$, where option o_1 is a **basic** decision problem A whereby you are to choose at t_1 between two acts: a_1, which pays you \$1 if E occurs and 'pays' you $-$\$1 if E^c, i.e., $\sigma(a_1, E) = \$1$ and $\sigma(a_1, E^c) = -\$1$;[5] or the act a_2 which 'pays' you a constant $-$\$0.50. Assume that your utility is linear in dollar amounts with $u(\$x) = x$. See Figure 3.

In this basic decision problem A, which is the result

[4] See, for example, [17, 18] and [23, 22, 10, 16, 34]; for a summary of Levi's ideas, see [15, Appendix A].

[5] Here we abuse our notation by writing $\sigma(a, E) = \$1$, for instance, to express that $\sigma(a, \cdot)$ is a constant \$1 on E.

of implementing option o_1, the subjective expected utility of a_1 is \$0 and the subjective expected utility of a_2 is $-$\$0.50. So, a_1 is uniquely admissible from A: receiving nothing is better than paying 50 cents.

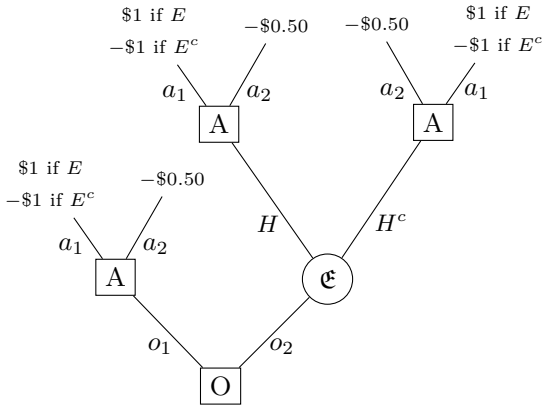


Figure 3: A Sequential Decision Example.

Turn now to option o_2, whereby at t_2 you face a **derived** decision problem conditional on the outcome of experiment $\mathfrak{E}$. That is, you are confronted with the same decision problem A at t_2 after learning (only) that H obtains or H^c obtains at t_1. But the derived decision problem A, which is the result of implementing option o_2, is different from the basic decision problem A: in the derived decision problem the act a_1 is inadmissible against a_2. Why? Because in the basic decision problem $p(E) = 1/2$, but in the derived decision problem E is dilated by $\mathfrak{E}$ to 0.1 and 0.9: whether the outcome of the fair coin toss is heads or tails, E conditional on that outcome is highly uncertain. Thus, in the derived decision problem, there are probability mass functions $p \in \mathbb{P}$ whereby $p(E^c)$ is .9, in which case the minimum expected utility of a_1 is $-$\$0.80. So, in the derived decision problem, by Γ-Maximin, a_2 has a higher minimum expected value than a_1 regardless of the outcome of the experiment, $\mathfrak{E}$.

Assume that a decision maker is certain that she will not change her preferences, will update her belief state by Generalized Bayesian conditionalization, and that she will choose to maximize her minimal expected utility (Γ-Maximin). Then, in a pairwise choice between a_1 of the basic decision problem determined by option o_1, which has an expected value of zero, and a_2 of the derived decision problem determined by option o_2, which has an expected value of $-$\$0.50, observing cost-free information at t_1, i.e., learning the outcome of the fair coin toss $\mathfrak{E}$, is devalued. Here we have a case where the decision maker would strictly prefer *not* to receive cost-free information!

Discussion. Although in finite spaces some decision rules, including Γ-Maximin, *require* decision makers to reject the opportunity to observe cost-free information before making a decision, others merely *permit* decision makers to reject the opportunity to observe cost-free information before making a decision. For example, E-Admissibility permits, but never requires, you to reject an opportunity to observe cost-free information before making a decision. Even so, E-admissibility supplemented with a secondary criterion for selecting among E-Admissible options—namely, to maximize expected utility with respect to a least informative distribution from among E-Admissibility options—respects the value of (cost-free) information, and therefore mandates that decision makers abide by Good's Principle. So, the first point to note is that dilating probabilities can be paired with a variety of decision rules, some abide by Good's Principle, others do not.

The second point to emphasize is that E and H are *not* stochastically independent, so the basic (o_1) and derived (o_2) forms of the decision problem A are importantly different. (If the uncertainty in G were represented by a single probability rather than a set of probabilities, then the two forms would be equivalent.) From Theorem 1 we see that the association between E and H is the key to dilation; the effects one sees from evaluating conditional judgments merely are a consequence. Performing the experiment $\mathfrak{E}$ reveals to you the extent of your uncertainty about the dependence of E on the experimental outcomes of $\mathfrak{E}$. How knowledge of this particular form of uncertainty affects decision making will depend on the decision maker's beliefs, values and goals.

5.2 Good's Principle and Non-Conglomerability

Suppose that at t_0 you face a decision problem $\mathrm{O} = \{o_1, o_2\}$ as in the previous section, here with decision problem $\mathrm{A} = \{a_1, a_2\}$ and experiment $\mathfrak{E} = \{H_n : n \in \mathbb{N}_{>0}\}$. Action a_1 pays you \$1 if E occurs and 'pays' you $-$\$1 if E^c, while action a_2 'pays' you a constant $-$\$0.50 i.e., $\sigma(a_1, E) = \$1$ and $\sigma(a_1, E^c) = -\$1$ and $\sigma(a_2, E) = -\$0.50$ and $\sigma(a_2, E^c) = -\$0.50$.

Now, under the hypothesis that at t_1 you face the decision problem A without observing the outcome of experiment $\mathfrak{E}$, your subjective expected utility of a_1 is \$0, while your subjective expected utility of a_2 is $-$\$0.50. So you judge a_1 to be uniquely admissible from the **basic** decision problem A. That is, $\mathbf{c}(\mathrm{A}|N) = \{a_1\}$, where N is the hypothesis that at t_1 you have implemented option o_1.

Under the hypothesis that at t_2 you face the decision problem A after observing outcome H_i of $\mathfrak{E}$, your

subjective expected utility of a_1 is $-\$1$, while your subjective expected utility of a_2 remains $-\$0.50$. So you judge a_2 to be uniquely admissible from the **derived** decision problem A. That is, $\mathbf{c}(\mathrm{A}|K_i) = \{a_2\}$, where K_i is the hypothesis that at t_2 you face the decision problem A after implementing option o_2 at t_0 and observing outcome e_i of experiment $\mathfrak{E}$ at t_1. Thus, assuming that you are certain you will not change your preferences, that you will update your belief state by Bayesian conditionalization, and that you will maximize subjective expected utility, option o_1 has constant utility \$0 and option o_2 has constant utility $-\$.50$. In other words, $\mathbf{c}(\mathrm{A}) = \{o_1\}$: at t_0 you judge that choosing from A without observing the outcome of $\mathfrak{E}$ to be exclusively admissible for choice.

Discussion. One might argue that there are significant differences between failures of Good's Principle due to dilation and failures due to non-conglomerability. For instance, in the case of dilation, some decision rules respect Good's Principle and some do not, which has been cited as grounds for modifying particular decision rules rather than the modifying the uncertainty model. In the case of non-conglomerability, it may appear that there is a disanalogy. The standard reply to cases of non-conglomerability is to modify the uncertainty model, namely by imposing countable additivity, rather than to modify the expected utility maximization, which many take for granted. What are the grounds for adjudicating between these two cases?

6 Conclusion

In closing, there are three general points to make. First, notice that there are several familiar approaches that do not countenance imprecise probabilities but which nevertheless require decision makers to forgo the opportunity to observe cost-free information before making a decision, and some of those approaches do so even in finite spaces. Second, while Good's Principle is often implicated in learning or sequential decision making, Good's Principle itself is a synchronic, confirmational rule about an agent's state of belief at a particular time, rather than a temporal rule regulating updating of an agent's state of belief in light of an observation. Similarly, dilation is likewise characterized synchronically, rather than dynamically.

Finally, what is the normative standing of Good's Principle? We believe it is not an obvious general principle of rationality, and that the classical argument strategies for establishing the principle rest on strong structural assumptions, not only about a decision maker's adherence to expected utility maximization, but also about the decision maker's beliefs about her future preferences, future belief states, and future decision strategies. Although Good's Principle is familiar, the foundations for its (still) wide acceptance are not; indeed, there appear to be a host of reasonable exceptions to Good's Principle, even within the standard setting of utility maximization. For these reasons, we are puzzled why some authors still elevate Good's Principle to a general normative principle while remaining indecisive about the normative status of (merely) finitely additive probabilities.

Acknowledgements

This research was supported in part by the Alexander von Humboldt Foundation.

References

[1] Thomas E. Armstrong. Conglomerability of probability measures on boolean algebras. *Journal of Mathematical Analysis and Applications*, 150:335 – 358, 1990.

[2] Inés Couso, Serafín Moral, and Peter Walley. Examples of independence for imprecise probabilities. In Gert de Cooman, editor, *Proceedings of the First Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, Ghent, Belgium, 1999.

[3] Bruno de Finetti. Probability, statistics, and induction: Their relationship according to the various points of view. In Isotta Cesari and Leonard J. Savage, editors, *Probability, Induction, and Statistics, 1979*. John Wiley and Sons, 1959.

[4] Bruno de Finetti. *Theory of Probability*, volume I. Wiley, 1990 edition, 1974.

[5] Lester E. Dubins. Finitely additive conditional probability, conglomerability, and disintegrations. *Annals of Probability*, 3:89–99, 1975.

[6] Adam Elga. Subjective probabilities should be sharp. *Philosophers Imprint*, 10(5), May 2010.

[7] Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1:107–43, 2009.

[8] I. J. Good. On the principle of total evidence. *The British Journal for the Philosophy of Science*, 17(4):319–321, 1967.

[9] Peter Grünwald and Joseph Y. Halpern. When ignorance is bliss. In Joseph Y. Halpern, editor, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04)*, pages 226–234, Arlington, Virginia, 2004. AUAI Press.

[10] Ian Hacking. Slightly more realistic personal probability. *Philosophy of Science*, 34(4):311–325, 1967.

[11] Timothy Herron, Teddy Seidenfeld, and Larry Wasserman. The extent of dilation of sets of probabilities and the asymptotics of robust bayesian inference. In *PSA 1994 Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1, pages 250–259, 1994.

[12] Timothy Herron, Teddy Seidenfeld, and Larry Wasserman. Divisive conditioning: further results on dilation. *Philosophy of Science*, 64:411–444, 1997.

[13] Brian Hill. Dynamic consistency and ambiguity: A reappraisal. Technical Report ECO/SCD-2013-983, HEC Paris, Paris, 2013.

[14] Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, 91(435):1228–1235, 1996.

[15] Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. Is ignorance bliss? *Journal of Philosophy*, 105(1):5–36, 2008.

[16] Henry E. Kyburg, Jr. Bets and beliefs. *American Philosophical Quarterly*, 1:54–63, 1968.

[17] Isaac Levi. Confirmational conditionalization. *Journal of Philosophy*, 75(12):730–737, 1978.

[18] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, MA, 1980.

[19] David V. Lindley. *Introduction to Probability and Statistics*. Cambridge University Press, Cambridge, 1965.

[20] Mark J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4):1622–68, 1989.

[21] Arthur Paul Pedersen and Gregory Wheeler. Demystifying dilation. *Erkenntnis*, 79(6):1305–1342, 2014.

[22] Howard Raiffa and Robert Schlaiffer. *Applied Statistical Decision Theory*, volume 1 of *Studies in Managerial Economics*. Harvard Business School Publications, 1961.

[23] Frank P. Ramsey. Truth and probability. In H. E. Kyburg and H. E. Smokler, editors, *Studies in subjective probability*, pages 23–52. Robert E. Krieger Publishing Company, Huntington, New York, second (1980) edition, 1926.

[24] Frank P. Ramsey. *The Foundations of Mathematics and Other Essays*, volume 1. Humanities Press, New York, 1931.

[25] Leonard J. Savage. *Foundations of Statistics*. Dover, New York, 1972.

[26] Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane. The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(2):205–226, 1984.

[27] Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane. On the equivalence of conglomerability and disintegrability for unbounded random variables. *Statistical Methods and Applications*, 23:501–518, 2014.

[28] Teddy Seidenfeld. When normal and extensive form decisions differ. In D. Prawitz, B. Skyrms, and D. Westerstahl, editors, *Logic, Methodology and Philosophy of Science*. Elsevier Science B. V., 1994.

[29] Teddy Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: Γ-Maxmin versus E-admissibility. *Synthese*, 140:69–88, 2004.

[30] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Non-conglomerability for finite-valued, finitely additive probability. *he Indian Journal of Statistics, Series A*, 60(3):476 – 491, 1998.

[31] Teddy Seidenfeld and Larry Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21:1139–154, 1993.

[32] Marciano Siniscalchi. Dynamic choice under ambiguity. *Theoretical Economics*, 6:379–421, 2011.

[33] Peter Wakker. Nonexpected utility as aversion of information. *Journal of Behavioral Decision Making*, 1:169–75, 1988.

[34] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[35] Larry Wasserman and Teddy Seidenfeld. The dilation phenomenon in robust Bayesian inference. *Journal of Statistical Planning and Inference*, 40:345–356, 1994.

[36] Marco Zaffalon and Enrique Miranda. Probability and time. *Artificial Intelligence*, 198:1–51, 2013.

Weak Consistency for Imprecise Conditional Previsions

Renato Pelessoni
University of Trieste, Italy
renato.pelessoni@econ.units.it

Paolo Vicig
University of Trieste, Italy
paolo.vicig@econ.units.it

Abstract

In this paper we explore relaxations of (Williams) coherent and convex conditional previsions that form the families of n-coherent and n-convex conditional previsions, at the varying of n. We investigate which such previsions are the most general one may reasonably consider, suggesting (centered) 2-convex or, if positive homogeneity and conjugacy is needed, 2-coherent lower previsions. Basic properties of these previsions are studied. In particular, centered 2-convex previsions satisfy the Generalized Bayes Rule and always have a 2-convex natural extension. We discuss then the rationality requirements of 2-convexity and 2-coherence from a desirability perspective. Among the uncertainty concepts that can be modelled by 2-convexity, we mention generalizations of capacities and niveloids to a conditional framework.

Keywords. Williams coherence, 2-coherent previsions, 2-convex previsions, Generalized Bayes Rule.

1 Introduction

In his influential book [16], P. Walley developed a behavioural approach to imprecise probabilities (and previsions) extending de Finetti's [4] interpretation of precise previsions in terms of coherence. Operationally, this was achieved through a relaxation of de Finetti's betting scheme.

In fact, following de Finetti, P is a coherent precise prevision on a set $\mathcal{S}$ of gambles if and only if for all m, $n \in \mathbb{N}_0$, $s_1, \ldots, s_m, r_1, \ldots, r_n \geq 0$, $X_1, \ldots, X_m, Y_1, \ldots, Y_n \in \mathcal{S}$, defining $G = \sum_{i=1}^{m} s_i(X_i - P(X_i)) - \sum_{j=1}^{n} r_j(Y_j - P(Y_j))$, it holds that $\sup G \geq 0$. The terms $s_i(X_i - P(X_i))$, $r_j(Y_j - P(Y_j))$ are proportional (with coefficients or *stakes* s_i, r_j) to the *gains* arising from, respectively, buying X_i at $P(X_i)$ or selling Y_j at $P(Y_j)$. A coherent lower prevision $\underline{P}$ on $\mathcal{S}$ may be defined in a similar way, just restricting n to belong to $\{0, 1\}$. This means that the betting scheme is modified to allow selling at most one gamble. Several other betting scheme variants have been investigated in the literature, either extending coherence for lower previsions (conditional lower previsions) or weakening it (previsions that are convex, or avoid sure loss). In particular, a convex lower prevision is defined introducing a convexity constraint $n = 1, \sum_{i=1}^{m} s_i = r_1 = 1$ in the betting scheme. In [16, Appendix B] n-coherent previsions are studied, as a different relaxation of coherence.

In this paper, we explore further variations of the behavioural approach/betting scheme: n-coherent and n-convex conditional lower previsions, formally defined later on as generalisations of the n-coherent (unconditional) previsions in [16]. Our major aims are:

a) to explore the flexibility of the behavioural approach and its capability to encompass different uncertainty models;

b) to point out which are the basic axioms/properties of coherence which hold even for much looser consistency concepts.

Referring to b) and with a view towards the utmost generality, we shall mainly concentrate on the extreme quantitative models that can be incorporated into a (modified) behavioural approach. This does not imply that these models should be regarded as preferable to coherent lower previsions. On the contrary they will not, as far as certain questions are concerned. For instance, inferences will typically be rather vague. However, it is interesting and somehow surprising to detect that certain properties like the Generalised Bayes Rule must hold even for such models, or that they can be approached in terms of desirability.

N-coherence and n-convexity may be naturally seen as relaxations of, respectively, (Williams) coherence and convexity. These and other preliminary concepts are recalled in Section 2. Starting from the weakest

reasonably sound consistency concepts, we explore basic properties of 2-convex lower previsions in Section 3. We supply a characterisation by means of axioms, on a special set of conditional gambles generalising a linear space and termed $\mathcal{D}_{LIN}$ (Definition 2). Interestingly, it turns out that n-convexity with $n \geq 3$ and convexity are equivalent on $\mathcal{D}_{LIN}$. 2-convex previsions exhibit some drawbacks: a 2-convex natural extension may be defined, but its finiteness is not guaranteed; the property of internality may fail, as well as agreement with conditional implication (the Goodman-Nguyen relation). In Section 4, we show that the special subset of centered 2-convex previsions is not affected by these problems. In Section 5, 2-coherent lower previsions are discussed and characterised on $\mathcal{D}_{LIN}$ (Proposition 8). Again, n-coherence ($n \geq 3$) and coherence are equivalent on $\mathcal{D}_{LIN}$. On generic sets of gambles, n-coherent previsions ($n \geq 3$) have no n-coherent extension on sufficiently large supersets whenever the equivalence does not hold. We show also that 2-coherence should be preferred to 2-convexity when positive homogeneity and conjugacy are required. In Section 6 we analyse 2-convexity and 2-coherence in a desirability approach. Generalising prior work by Williams [17, 18] for coherence, we focus on the correspondence between these previsions and sets of desirable gambles, and on establishing the ensuing desirability rules. Models that can be accommodated into the framework of 2-convexity, but not of coherence, are presented in Section 7. These are conditional versions of capacities and niveloids. Section 8 concludes the paper. Due to spacing constraints, proofs of the results are omitted (some can be partly derived from results in [10, 12]).

2 Preliminaries

The starting points for our investigation are the known consistency concepts of coherent and convex lower conditional prevision [10, 11, 17, 18]. They both refer to an arbitrary set $\mathcal{D}$ of conditional gambles, that is of conditional bounded random variables. We denote with $X|B$ a generic conditional gamble, where X is a gamble and B is a non-impossible event ($B \neq \varnothing$). It is understood here that $X : \mathbb{P} \to \mathbb{R}$ is defined on an underlying partition $\mathbb{P}$ of atomic events ω, and that B belongs to the powerset of $\mathbb{P}$. Therefore, any $\omega \in \mathbb{P}$ implies either B or its negation $\neg B$ (in words, knowing that ω is true determines the truth value of B, i.e. B is known to be either true or false). Given B, the conditional partition $\mathbb{P}|B$ is formed by the conditional events $\omega|B$, such that ω implies B (implies that B is true) and $X|B : \mathbb{P}|B \to \mathbb{R}$ is such that $X|B(\omega|B) = X(\omega)$, $\forall \omega|B \in \mathbb{P}|B$. Because of this equality, several computations regarding $X|B$ can be performed by means of the restriction of X on B. In particular, it is useful for the sequel to recall that $\sup(X|B) = \sup_B X$, and $\inf(X|B) = \inf_B X$.

As special cases, we have that $X|\Omega = X$ is an unconditional gamble, $A|B$ a conditional event if A is an event (or its indicator I_A - we shall generally employ the same notation A for both).

As customary, a lower prevision $\underline{P}$ is, without further qualifications, a map from $\mathcal{D}$ into the real line, $\underline{P} : \mathcal{D} \to \mathbb{R}$. However, a lower prevision is often interpreted as a supremum buying price [16]. For instance, if a subject assigns $\underline{P}(X|B)$ to $X|B$, he is willing to buy X, conditional on B occurring, at any price lower than $\underline{P}(X|B)$. Under this behavioural interpretation, Definitions 1, 3, 5 check the consistency of $\underline{P}$, depending on whether it avoids losses bounded away from 0, according to different buying and selling constraints.

Definition 1. *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ be given.*

a) $\underline{P}$ *is a* coherent *conditional lower prevision on $\mathcal{D}$ iff, for all $m \in \mathbb{N}_0$, $\forall X_0|B_0, \ldots, X_m|B_m \in \mathcal{D}$, $\forall s_0, \ldots, s_m$ real and non-negative, defining $S(\underline{s}) = \bigvee\{B_i : s_i \neq 0, i = 0, \ldots, m\}$ and $\underline{G} = \sum_{i=1}^{m} s_i B_i(X_i - \underline{P}(X_i|B_i)) - s_0 B_0(X_0 - \underline{P}(X_0|B_0))$, it holds, whenever $S(\underline{s}) \neq \varnothing$, that $\sup\{\underline{G}|S(\underline{s})\} \geq 0$.*

b) $\underline{P}$ *is a* convex *conditional lower prevision on $\mathcal{D}$ iff, for all $m \in \mathbb{N}^+$, $\forall X_0|B_0, \ldots, X_m|B_m \in \mathcal{D}$, $\forall s_1, \ldots, s_m$ real and non-negative such that $\sum_{i=1}^{m} s_i = 1$ (convexity constraint), defining $\underline{G}_c = \sum_{i=1}^{m} s_i B_i(X_i - \underline{P}(X_i|B_i)) - B_0(X_0 - \underline{P}(X_0|B_0))$, $S(\underline{s}) = \bigvee\{B_i : s_i \neq 0, i = 1, \ldots, m\}$, it holds that $\sup\{\underline{G}_c|S(\underline{s}) \vee B_0\} \geq 0$.*

b1) $\underline{P}$ *is centered convex or* C-convex *on $\mathcal{D}$ iff it is convex and, $\forall X|B \in \mathcal{D}$, it is $0|B \in \mathcal{D}$ and $\underline{P}(0|B) = 0$.*

In the behavioural interpretation recalled above, Definition 1a) considers buying at most m conditional gambles $X_1|B_1, \ldots, X_m|B_m$ (also no one, when $m = 0$) at prices $\underline{P}(X_1|B_1), \ldots, \underline{P}(X_m|B_m)$, respectively, and selling at most one gamble $X_0|B_0$ at a supremum buying price $\underline{P}(X_0|B_0)$. The gain $\underline{G}$ is a linear combination with stakes $s_0, \ldots, s_m$ of the gains from these transactions. It is conditioned on $S(\underline{s})$, to rule out both trivial transactions ($\underline{G} = 0$, since $s_1 = \ldots = s_m = 0$) and the case that $\underline{G} = 0$ because no transaction takes place (when $B_0, \ldots, B_m$ are all false). Then, coherence requires the non-negativity of the supremum of $\underline{G}$, conditional on at least one non-trivial transaction being effective. The interpretation of Definition 1b) is similar: what changes is the convexity constraint on the stakes $(s_0 = 1), s_1, \ldots, s_m$. This implies that $\underline{G}_c$ is the gain from one selling transaction and at least one buying transaction.

The definition of coherent lower prevision is a structure free version of Williams coherence, discussed in [11]. It is more general than Walley's coherence [16], in particular it always allows for a natural extension and is not necessarily conglomerable. The notion of convex lower prevision is still more general, and was introduced in [10], extending the unconditional convexity studied in [9]. Convex previsions can incorporate various uncertainty models, including convex risk measures, non-normalised possibility measures, and others. However, the special subclass of C-convex lower previsions guarantees better consistency properties. Among these, there always exists a convex natural extension of these measures, whose properties are analogous to those of the natural extension [10, Theorem 9].

Even though coherent and convex lower previsions can be defined on any set of conditional gambles, they are characterised by a few axioms on the special environment $\mathcal{D}_{LIN}$ defined next.

Definition 2. *Let $\mathcal{X}$ be a linear space of gambles and $\mathcal{B} \subset \mathcal{X}$ the set of all (indicators of) events in $\mathcal{X}$. Suppose $1 \in \mathcal{B}$ and $B X \in \mathcal{X}, \forall B \in \mathcal{B}, \forall X \in \mathcal{X}$. Setting $\mathcal{B}^{\varnothing} = \mathcal{B} - \{\varnothing\}$, define*

$$\mathcal{D}_{LIN} = \{X|B : X \in \mathcal{X}, B \in \mathcal{B}^{\varnothing}\}. \tag{1}$$

The sets $\mathcal{D}_{LIN}$ may be viewed as conditional generalisations of linear spaces of (unconditional) gambles. In fact, when $\mathcal{B} = \{\Omega, \varnothing\}$, $\mathcal{D}_{LIN}$ reduces to a linear space of unconditional gambles (including real constants). Not surprisingly then, characterisations on $\mathcal{D}_{LIN}$ have an unconditional counterpart on linear spaces.

Proposition 1. *Let $\underline{P} : \mathcal{D}_{LIN} \to \mathbb{R}$ be a conditional lower prevision.*

a) $\underline{P}$ is coherent on $\mathcal{D}_{LIN}$ if and only if [18]

(A1) $\underline{P}(X|B) - \underline{P}(Y|B) \leq \sup\{X - Y|B\}$, $\forall X|B, Y|B \in \mathcal{D}_{LIN}$.[1]

(A2) $\underline{P}(\lambda X|B) = \lambda \underline{P}(X|B)$, $\forall X|B \in \mathcal{D}_{LIN}, \forall \lambda \geq 0$.

(A3) $\underline{P}(X + Y|B) \geq \underline{P}(X|B) + \underline{P}(Y|B)$, $\forall X|B,\ Y|B \in \mathcal{D}_{LIN}$.

(A4) $\underline{P}(A(X - \underline{P}(X|A \wedge B))|B) = 0$, $\forall X \in \mathcal{X}, \forall A, B \in \mathcal{B}^{\varnothing} : A \wedge B \neq \varnothing$.

b) $\underline{P}$ is convex on $\mathcal{D}_{LIN}$ if and only if (A1), (A4) and the following axiom hold [10, Theorem 8]

(A5) $\underline{P}(\lambda X + (1 - \lambda)Y|B) \geq \lambda \underline{P}(X|B) + (1 - \lambda)\underline{P}(Y|B), \forall X|B, Y|B \in \mathcal{D}_{LIN}, \forall \lambda \in]0, 1[$.

[1] (A1) may be replaced by $\underline{P}(X|B) \geq \inf(X|B)$, $\forall X|B \in \mathcal{D}_{LIN}$, thus corresponding to the original version in [18].

Condition (A4) is the *Generalised Bayes Rule (GBR)*, introduced in [17, 18] and studied also in [16] in the special case $B = \Omega$.

Since our discussion will focus on minimal consistency properties for a conditional lower prevision, we have to mention a conditional generalisation of the implication (inclusion) relation between events, termed Goodman-Nguyen relation ($\leq_{GN}$). In fact, suppose $A \Rightarrow B$ (or $A \subseteq B$). Then, asking that $\mu(A) \leq \mu(B)$ is a really minimal rationality requirement for any μ aiming at measuring how likely an event is, given that, whenever event A will turn to be true, B will be true too. The following extension of the implication to conditional events was proposed in [8]:

$$A|B \leq_{GN} C|D \quad \begin{array}{l}\text{iff } A \wedge B \Rightarrow C \wedge D \\ \text{and } \neg C \wedge D \Rightarrow \neg A \wedge B.\end{array} \tag{2}$$

The Goodman-Nguyen relation $\leq_{GN}$ was extended to conditional gambles in [12]:

$$X|B \leq_{GN} Y|D \text{ iff}$$
$$I_B X + I_{\neg B \vee D} \sup(X|B) \leq I_D Y + I_{B \vee \neg D} \inf(Y|D)$$

showing that $X|B \leq_{GN} Y|D$ implies $\underline{P}(X|B) \leq \underline{P}(Y|D)$ for a C-convex or coherent $\underline{P}$ [12, Proposition 10].

3 2-Convex Lower Previsions

In Definition 1, a) and b), there is no upper bound to $m \in \mathbb{N}$. One may think of introducing it as a natural way of weakening coherence and convexity. More precisely, let us call *elementary gain* on $X_i|B_i$ any term $s_i B_i(X_i - \underline{P}(X_i|B_i))$, with the proviso that $-B_0(X_0 - \underline{P}(X_0|B_0))$ in Definition 1 b) is also an elementary gain, formally corresponding to $s_0 = -1$. Then, we may state that no more than n elementary gains are allowed in either $\underline{G}$ (Definition 1, a)) or $\underline{G}_c$ (Definition 1, b)). When doing so, we speak of *n-coherent* or *n-convex* lower previsions. This approach extends the notion of n-coherent (unconditional) prevision in [16, Appendix B].

Intuition suggests that the smaller n is, the more the corresponding consistency concept is looser. In the extreme cases n may be as small as 1 with coherence, 2 with convexity.

However, 1-*coherence* is too weak. In fact, $\underline{P}$ is 1-coherent on $\mathcal{D}$ iff, $\forall X_0|B_0 \in \mathcal{D}$, $\forall s_0 \in \mathbb{R}$, $\sup\{s_0 B_0(X_0 - \underline{P}(X_0|B_0))|B_0\} \geq 0$. It is easy to see that this is equivalent to *internality*, i.e. to requiring that $\underline{P}(X_0|B_0) \in [\inf(X_0|B_0), \sup(X_0|B_0)]$, $\forall X_0|B_0 \in \mathcal{D}$.

Since internality alone does not seem enough as a rationality requirement, we turn our attention in this

section to what seems to be the next weakest consistency notion, that is 2-convexity.[2]

Definition 3. *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a* 2-convex *conditional lower prevision on $\mathcal{D}$ iff, $\forall X_0|B_0, X_1|B_1 \in \mathcal{D}$, we have that*

$$\sup\{B_1(X_1 - \underline{P}(X_1|B_1)) - B_0(X_0 - \underline{P}(X_0|B_0))|B_0 \vee B_1)\} \geq 0. \quad (3)$$

We explore now some basic features of 2-convex previsions. Some critical aspects are discussed next, showing in Section 4 that they can be solved resorting to the subclass of centered 2-convex previsions.

A remarkable result in our framework is the characterisation of 2-convexity on a structured set $\mathcal{D}_{LIN}$.

Proposition 2. *A conditional lower prevision $\underline{P} : \mathcal{D}_{LIN} \to \mathbb{R}$ is 2-convex on $\mathcal{D}_{LIN}$ if and only if (A1) and (A4) hold.*

To point out an important consequence of Proposition 2, compare it with Proposition 1 b). It follows at once that the difference between 2-convexity and convexity, on $\mathcal{D}_{LIN}$, is due to axiom (A5). On the other hand, the proof that a convex prevision on $\mathcal{D}_{LIN}$ must satisfy (A5), given in [10, Theorem 8], only involves a gain $\underline{G}_c$ made up of 3 elementary gains, i.e. it does not fully exploit convexity, but only 3-convexity. This justifies the following conclusion:

On $\mathcal{D}_{LIN}$, n-convexity with $n \geq 3$ and convexity are equivalent concepts.

Hence, the very difference between convexity and n-convexity reduces to that between convexity and 2-convexity, at least on $\mathcal{D}_{LIN}$. Yet, if $\underline{P}$ is defined on a set $\mathcal{D}$ other than $\mathcal{D}_{LIN}$, we may think of extending it to some $\mathcal{D}_{LIN} \supset \mathcal{D}$. If $\underline{P}$ is n-convex on $\mathcal{D}$, $n \geq 3$, and has an n-convex extension to $\mathcal{D}_{LIN}$, then $\underline{P}$ is convex on $\mathcal{D}_{LIN}$ and therefore also on $\mathcal{D}$. It ensues that if $\underline{P}$ is n-convex ($n \geq 3$) but not convex on $\mathcal{D}$, $\underline{P}$ will have no n-convex extension on any sufficiently large superset of $\mathcal{D}$ (any $\mathcal{D}^*$ including some $\mathcal{D}_{LIN}$ containing $\mathcal{D}$) - see also the later Example 2. This is a negative aspect of n-convexity, when $n \geq 3$. More generally, the discussion above shows that n-convex previsions are not particularly significant as an autonomous concept, when $n \geq 3$.

Turning again to 2-convex previsions, let us define a special extension, the 2-convex natural extension.

Definition 4. *Given a lower prevision $\underline{P} : \mathcal{D} \to \mathbb{R}$ and an arbitrary conditional gamble $Z|B$, let*

$$L(Z|B) = \{\alpha : \sup\{A(X - \underline{P}(X|A)) - B(Z - \alpha)|A \vee B\} < 0, \text{ for some } X|A \in \mathcal{D}\}. \quad (4)$$

Then the 2-convex natural extension *$\underline{E}_{2c}$ of $\underline{P}$ on $Z|B$ is*

$$\underline{E}_{2c}(Z|B) = \sup L(Z|B). \quad (5)$$

In general, $\underline{E}_{2c}(Z|B)$ may not be real-valued (i.e. $+\infty$, or $-\infty$ when $L(Z|B) = \emptyset$). The results in the next proposition are helpful in hedging this occurrence.

Proposition 3. *a) $L(Z|B) \neq \emptyset$, if there exists $Y|C \in \mathcal{D}$ such that $C \Rightarrow B$.*

b) Let $\underline{P}$ be 2-convex and such that $0|B \in \mathcal{D}$ and $\underline{P}(0|B) = 0$, $\forall X|B \in \mathcal{D}$. Given $0|C \notin \mathcal{D}$, the extension of $\underline{P}$ on $\mathcal{D} \cup \{0|C\}$ such that $\underline{P}(0|C) = 0$ is 2-convex.

c) When $L(Z|B) \neq \emptyset$, $L(Z|B) =]-\infty, \underline{E}_{2c}(Z|B)[$.

d) If $L(Z|B) \neq \emptyset$ and $\sup(X|A) \geq \underline{P}(X|A)$, $\forall X|A \in \mathcal{D}$, then $\underline{E}_{2c}(Z|B) \leq \sup(Z|B)$, $\forall Z|B$.

e) Let $\underline{P}$ be 2-convex and $0|B \in \mathcal{D}$, $\forall X|B \in \mathcal{D}$. Then, $\forall X|B \in \mathcal{D}$, $\sup(X|B) \geq \underline{P}(X|B)$ iff $\underline{P}(0|B) \leq 0$.

Parts a) and b) of Proposition 3 suggest a simple way to ensure $\underline{E}_{2c}(Z|B) \neq -\infty$: just add the gamble $0|B$ to $\mathcal{D}$, putting $\underline{P}(0|B) = 0$. To guarantee $\underline{E}_{2c}(Z|B) \neq +\infty$, it is sufficient that any $0|C$ in $\mathcal{D}$ (or added to $\mathcal{D}$) is given a non-positive lower prevision, by d) and e). Clearly, the simplest and most obvious choice is to put $\underline{P}(0|C) = 0$, $\forall 0|C$. This would make $\underline{P}$ a centered 2-convex lower prevision; in the remainder of this section we do not however rule out the possibility that $\underline{P}(0|C) \neq 0$ for some $0|C$.

The properties of the 2-convex natural extension are very similar to those of the natural extension:

Proposition 4. *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ be a lower prevision, with $\mathcal{D} \subseteq \mathcal{D}_{LIN}$. If $\underline{E}_{2c}$ is finite on $\mathcal{D}_{LIN}$, then*

a) $\underline{E}_{2c}(X|B) \geq \underline{P}(X|B)$, $\forall X|B \in \mathcal{D}$.

b) $\underline{E}_{2c}$ is 2-convex on $\mathcal{D}_{LIN}$.

c) If $\underline{P}^*$ is 2-convex on $\mathcal{D}_{LIN}$ and $\underline{P}^*(X|B) \geq \underline{P}(X|B)$, $\forall X|B \in \mathcal{D}$, then $\underline{P}^*(X|B) \geq \underline{E}_{2c}(X|B)$, $\forall X|B \in \mathcal{D}_{LIN}$.

d) $\underline{P}$ is 2-convex on $\mathcal{D}$ if and only if $\underline{E}_{2c} = \underline{P}$ on $\mathcal{D}$.

e) If $\underline{P}$ is 2-convex on $\mathcal{D}$, $\underline{E}_{2c}$ is its smallest 2-convex extension on $\mathcal{D}_{LIN}$.

[2] 2-convex previsions were termed 1-convex in [1, 12]. Here we prefer the locution '2-convex' by analogy with the rule for fixing n in 'n-coherent' in [16].

In words, the 2-convex natural extension dominates $\underline{P}$ (by a)), characterises 2-convexity (by d)) and is the least-committal 2-convex extension of $\underline{P}$ (by b), c), e)).

Being rather weak a consistency concept, 2-convexity may not satisfy a number of properties which necessarily hold for coherent lower previsions. For instance, the *positive homogeneity* axiom (A2) of Proposition 1, $\underline{P}(\lambda X|B) = \lambda \underline{P}(X|B)$, with $\lambda \geq 0$, may not hold, not even weakening it to

$$\underline{P}(\lambda X|B) \geq \lambda \underline{P}(X|B), \forall \lambda \in [0,1]. \tag{6}$$

(Unconditional versions of (6) hold for centered convex previsions.)

It can instead be shown that

Proposition 5. *If, given $\lambda \in \mathbb{R}$, $\underline{P}$ is 2-convex on $\mathcal{D} \supseteq \{X|B, \lambda X|B\}$, then necessarily*

$$\begin{aligned} \inf\{(\lambda-1)X|B\} + \underline{P}(X|B) \leq \underline{P}(\lambda X|B) \\ \leq \sup\{(\lambda-1)X|B\} + \underline{P}(X|B). \end{aligned} \tag{7}$$

Condition (7) seems rather mild, as the next example points out.

Example 1. *Given $\mathcal{D} = \{X|B, 2X|B\}$ ($\lambda = 2$), where the image of $X|B$ is $[-1,1]$ and $\underline{P}(X|B) = 0.2$, equation (7) gives the bounds $\underline{P}(2X|B) \in [-0.8, 1.2]$. It is easy to check that $\underline{P}$ is 2-convex on $\mathcal{D}$ whatever is the choice for $\underline{P}(2X|B)$ in the interval $[-0.8, 1.2]$. According to the value for $\underline{P}(2X|B)$ selected in this interval, it may be $\underline{P}(2X|B) \gtreqless 2\underline{P}(X|B)$.*

An annoying feature of 2-convexity is that *internality* may fail, i.e. $\underline{P}(X|B)$ need not belong to the closed interval $[\inf(X|B), \sup(X|B)]$. Thus, 2-convex previsions may not satisfy a property holding even for 1-coherent previsions.

It has to be noticed that 2-convexity permits no complete freedom in departing from internality. There are two issues to be emphasized with respect to this question. The first tells us that lack of internality cannot be two-sided, because of the following result.

Proposition 6. *If $\underline{P} : \mathcal{D} \to \mathbb{R}$ is 2-convex on $\mathcal{D}$ and $\underline{P}(Y|D) < \inf(Y|D)$ for some $Y|D \in \mathcal{D}$, then $\underline{P}(X|B) \leq \sup(X|B)$, $\forall X|B \in \mathcal{D}$. Similarly, $\underline{P}(Y|D) > \sup(Y|D)$ for some $Y|D \in \mathcal{D}$ implies $\underline{P}(X|B) \geq \inf(X|B)$, $\forall X|B \in \mathcal{D}$.*

The second is the observation that 2-convexity imposes a sort of, so to say, two-component internality. To see this, note that

Lemma 1. *If $\underline{P} : \mathcal{D} \to \mathbb{R}$ is 2-convex on $\mathcal{D}$, and $X|B$, $Y|B \in \mathcal{D}$, then*

$$\begin{aligned} \inf\{X - Y|B\} \leq \underline{P}(X|B) - \underline{P}(Y|B) \\ \leq \sup\{X - Y|B\}. \end{aligned} \tag{8}$$

Recall now that $\underline{P}(X|B)$ is interpreted as a supremum buying price for $X|B$, and that Definition 3 ensures that buying $X|B$ for $\underline{P}(X|B)$ and selling $Y|B$ at its supremum buying price $\underline{P}(Y|B)$ would be (marginally) acceptable for 2-convexity. Then, equation (8) tells us that the profit $\underline{P}(X|B) - \underline{P}(Y|B)$ from this two-component exchange ($X|B$ vs. $Y|B$) guarantees no arbitrage. For instance, it cannot exceed $\sup\{X - Y|B\}$.

As a further critical issue with 2-convexity, we have that the Goodman-Nguyen relation may not induce an agreeing ordering on a 2-convex prevision. This is tantamount to saying that the partial ordering of some 2-convex conditional previsions may conflict with the ordering of the extended implication (inclusion) relation $\leq_{GN}$.

For instance, from (2), if $B \Rightarrow C$ then $0|C \leq_{GN} 0|B$. Agreement with the Goodman-Nguyen relation requires $\underline{P}(0|C) \leq \underline{P}(0|B)$ to hold, but it can be proven that if $\underline{P}(0|B) < 0$ and $B \Rightarrow C$, then 2-convexity asks instead that $\underline{P}(0|C) \geq \underline{P}(0|B)$ (the inequality may be strict).

4 Centered 2-Convex Lower Previsions

The critical issues on 2-convexity discussed in the preceding section can be solved or softened requiring the additional property

$$\forall X|B \in \mathcal{D}, 0|B \in \mathcal{D} \text{ and } \underline{P}(0|B) = 0, \tag{9}$$

i.e. restricting our attention to centered 2-convex conditional lower previsions. This is shown in the following proposition.

Proposition 7. *Let $\underline{P} : \mathcal{D} \to \mathbb{R}$ be a centered 2-convex lower prevision on $\mathcal{D}$. Then,*

a) $\forall X|B \in \mathcal{D}$, $\underline{P}(X|B) \in [\inf X|B, \sup X|B]$.

b) $\underline{P}$ has a finite 2-convex natural extension $\underline{E}_{2c}$ on any superset of $\mathcal{D}$.

c) $X|B \leq_{GN} Y|D$ implies $\underline{P}(X|B) \leq \underline{P}(Y|D)$.

Comment. The condition $\underline{P}(0|B) = 0$ appears as obvious, and in fact guarantees more satisfactory properties to 2-convexity. In our view, the main reason for considering the alternative $\underline{P}(0|B) \neq 0$ is to encompass additional uncertainty models. This is patent already in the unconditional framework: convex risk measures, as introduced in [6, 7], correspond to convex, not necessarily centered previsions [9].

Note that by Proposition 7 a) centered 2-convexity implies 1-coherence, while being obviously implied by

2-coherence. Hence, the centering condition $\underline{P}(0|B) = 0$ appears as a technical instrument to guarantee that the lower prevision $\underline{P}$ satisfies more properties than a generic 2-convex prevision, without having to assume the more demanding properties of 2-coherence.

5 2-Coherent Lower Previsions

Our next step is a discussion of which additional properties are achieved by 2-coherent lower prevision.

Definition 5. *$\underline{P} : \mathcal{D} \to \mathbb{R}$ is a 2-coherent lower prevision on $\mathcal{D}$ iff $\forall X_0|B_0, X_1|B_1 \in \mathcal{D}$, $\forall s_1 \geq 0$, $\forall s_0 \in \mathbb{R}$, defining $S(\underline{s}) = \bigvee\{B_i : s_i \neq 0, i = 0, 1\}$ we have that, whenever $S(\underline{s}) \neq \varnothing$,*

$$\sup\{s_1 B_1(X_1 - \underline{P}(X_1|B_1)) - s_0 B_0(X_0 - \underline{P}(X_0|B_0))|S(\underline{s})\} \geq 0. \quad (10)$$

2-coherent lower previsions are characterized on $\mathcal{D}_{LIN}$ as follows:

Proposition 8. *Let $\underline{P} : \mathcal{D}_{LIN} \to \mathbb{R}$ be a conditional lower prevision. $\underline{P}$ is 2-coherent on $\mathcal{D}_{LIN}$ if and only if (A1), (A2), (A4) and the following axiom hold:*

(A6) $\underline{P}(X|B) \leq -\underline{P}(-X|B)$.

Remark 1. *Proposition 8 can be equivalently restated replacing axiom (A1) with*

(A7) If $X|B, Y|B \in \mathcal{D}_{LIN}$, $\mu \in \mathbb{R}$ are such that $X|B \geq Y|B + \mu$, then $\underline{P}(X|B) \geq \underline{P}(Y|B) + \mu$.

In fact, it can be easily verified that (A1) and (A7) are equivalent.

Comment A comparison of Propositions 1 and 8 is useful in detecting at once two major differences between (centered) 2-convex and 2-coherent previsions.

One is positive homogeneity (axiom (A2)), a condition which, on any set $\mathcal{D}$, is necessary for 2-coherence, but not for 2-convexity. The need for positive homogeneity depends on the specific model we wish to consider. We might be willing to reject it in some instance, typically because of *liquidity risk* considerations. Basically, this means that for a large positive λ difficulties might be encountered at exchanging $\lambda X|B$ at a price $\underline{P}(\lambda X|B) = \lambda \underline{P}(X|B)$, because of lack of market liquidity at some degree.

The second difference is pointed out by axiom (A6). To fix its meaning, recall that given $\underline{P}(X|B)$, its *conjugate* upper prevision $\overline{P}(X|B)$ is defined by

$$\overline{P}(X|B) = -\underline{P}(-X|B). \quad (11)$$

Hence, by (11) axiom (A6) ensures that $\overline{P}(X|B) \geq \underline{P}(X|B)$, $\forall X|B \in \mathcal{D}_{LIN}$.

Therefore, 2-coherence is preferable to 2-convexity whenever we fix an upper ($\overline{P}$) and a lower ($\underline{P}$) bound for the uncertainty evaluation of $X|B$, while keeping positive homogeneity.

As an aside to the above discussion, we note that 2-coherence requires a weak form of homogeneity when $\lambda < 0$:

Proposition 9. *Given $\lambda < 0$, if $\underline{P}$ is 2-coherent on $D \supseteq \{\lambda X|B, X|B\}$, then necessarily $\underline{P}(\lambda X|B) \leq \lambda \underline{P}(X|B)$.*

Compare Propositions 8 and 1, a). Recalling that any 2-coherent lower prevision satisfies internality (being 1-coherent too), while (A6) is a necessary condition for coherence, only the superlinearity axiom (A3) distinguishes 2-coherence and coherence on $\mathcal{D}_{LIN}$. From this, deductions on the role of n-coherence, $n \geq 3$, can be made which are quite analogue to those on n-convexity in Section 3. This time, it can be shown that any n-coherent lower prevision, $n \geq 3$, must satisfy (A3), and hence that:

On $\mathcal{D}_{LIN}$, n-coherence with $n \geq 3$ and coherence are equivalent concepts.

And again, we may in general argue that n-coherence has no special relevance, compared to coherence, when $n \geq 3$. In particular, n-coherent extensions of an n-coherent $\underline{P}$ exist on sufficiently large sets if and only if $\underline{P}$ is coherent.

The latter concept is illustrated in the next example, elaborating on Example 2.7.6 in [16].

Example 2. *Let $\mathbb{P} = \{a, b, c, d\}$ be a partition of the sure event Ω. Define $\underline{P}$ on the powerset of $\mathbb{P}$ as follows:*

- $\underline{P}(\Omega) = 1$
- $\underline{P}(E) = \frac{1}{2}$ *if E is made up of 2 or 3 elements of $\mathbb{P}$, one of which is a.*
- $\underline{P}(E) = 0$ *otherwise.*

It is shown in [16] that $\underline{P}$ is not coherent, while being 3-coherent, and hence also 3-convex. We show now that $\underline{P}$ has no 3-convex extension to the linear space $\mathcal{L}(\mathbb{P})$ of all gambles defined on $\mathbb{P}$.

In fact, suppose a 3-convex extension, also termed $\underline{P}$, exists, and define $A = a$, $B = a \vee b$, $C = a \vee c$, $D = a \vee d$. Note that, by applying (7) with $\lambda = \frac{1}{2}$, $X = A$ and $B = \Omega$, we get $\underline{P}(\frac{1}{2}A) \leq \underline{P}(A) = 0$. Therefore, also the 3-convex extension of $\underline{P}$ to $\frac{1}{4}(B + C + D - 1) = \frac{1}{2}A$ should be non-positive. However, by applying axiom

(A5) as a necessary condition of 3*-convexity and noting that* (7) *(with* $\lambda = -1$*,* $X = 1$ *and* $B = \Omega$*) ensures also that* $\underline{P}(-1) = -1$*, we obtain* $\underline{P}(\frac{1}{4}(B + C + D - 1)) = \underline{P}(\frac{1}{2}(\frac{1}{2}B + \frac{1}{2}C) + \frac{1}{2}(\frac{1}{2}D - \frac{1}{2})) \geq \frac{1}{2}\underline{P}(\frac{1}{2}B + \frac{1}{2}C) + \frac{1}{2}\underline{P}(\frac{1}{2}D - \frac{1}{2}) \geq \frac{1}{4}\underline{P}(B) + \frac{1}{4}\underline{P}(C) + \frac{1}{4}\underline{P}(D) + \frac{1}{4}\underline{P}(-1) \geq 3 \cdot \frac{1}{4} \cdot \frac{1}{2} - \frac{1}{4} = \frac{1}{8} > 0$*, a contradiction.*

From what we have just proven, we may conclude that:

a) the given $\underline{P}$ *on the powerset of* $\mathbb{P}$ *has no* 3*-convex extension to* $\mathcal{L}(\mathbb{P})$*;*

b) $\underline{P}$ *(viewed now as* 3*-coherent on the powerset of* $\mathbb{P}$*) has no* 3*-coherent extension on* $\mathcal{L}(\mathbb{P})$ *either: if it had one, this extension would be* 3*-convex too, contradicting a).*

We may thus conclude that centered 2-convexity and 2-coherence appear to be the most significant weakenings of (centered) convexity and coherence.

6 Weak Consistency in a Desirability Approach

In this section we examine centered 2-convexity and 2-coherence from the viewpoint of desirability. This is an alternative approach to rationality concepts for uncertainty measures going back to [17] in the case of conditional imprecise previsions. It has been recently applied to a variety of other situations, see e.g. the discussion in [13] and the results in [14].

Roughly speaking, a set $\mathcal{A}$ of gambles is considered.[3] It is such that its gambles are regarded as *desirable* or *acceptable*. We may in general be willing to establish some *rationality criteria*, requiring that certain gambles do, or do not, belong to $\mathcal{A}$. The basic problem we shall consider here is: which is the correspondence between the rationality criteria we adopt and the consistency concepts of centered 2-convexity or alternatively 2-coherence? More specifically, the following two questions arise:

Q1) Which rationality criteria should be required to the elements of a set $\mathcal{A}$, so that a conditional lower prevision $\underline{P}$ may be obtained from $\mathcal{A}$ that is 2-coherent (alternatively, 2-convex)?

Q2) Conversely, given a 2-coherent (alternatively, 2-convex) $\underline{P}$, does it determine a set $\mathcal{A}'$ with certain rationality properties?

In the case that $\underline{P}$ is coherent, the answer to Q1) and Q2) was given by Williams in [17]. Our approach to solving Q1) and Q2) was largely influenced by his work. Preliminarily, some notation must be introduced.

Definition 6. *Let* $\mathcal{X}$ *be a linear space of gambles,* $\mathcal{B} \subset \mathcal{X}$ *a set of (indicators of) events,* $\mathcal{B}^{\varnothing} = \mathcal{B} - \{\varnothing\}$*. We suppose* $\Omega \in \mathcal{B}$ *and* $BX \in \mathcal{X}$*,* $\forall B \in \mathcal{B}$*,* $\forall X \in \mathcal{X}$*.*[4] *Define then*

$$\begin{aligned} \mathcal{X}^{\succeq} &= \{X \in \mathcal{X} : \inf X \geq 0\}, \\ \mathcal{X}^{\preceq} &= \{X \in \mathcal{X} : \sup X \leq 0\}, \end{aligned} \tag{12}$$

and, $\forall B \in \mathcal{B}$*,*

$$\begin{aligned} \mathcal{R}(B) &= \{X \in \mathcal{X} : BX = X\}, \\ \mathcal{R}(B)^{\succ} &= \{X \in \mathcal{R}(B) : \inf\{X|B\} > 0\}, \\ \mathcal{R}(B)^{\prec} &= \{X \in \mathcal{R}(B) : \sup\{X|B\} < 0\}. \end{aligned} \tag{13}$$

If $\mathcal{S}$ *and* $\mathcal{T}$ *are subsets of* $\mathcal{X}$*, their* Minkowski sum *is*

$$\mathcal{S} + \mathcal{T} = \{X + Y : X \in \mathcal{S}, Y \in \mathcal{T}\}.$$

We shall use similar compact notation later. For instance, $\lambda\mathcal{S} + \mu\mathcal{T} \subseteq \mathcal{U}$*,* $\forall \lambda, \mu \geq 0$*, means:* $\forall X \in \mathcal{S}$*,* $\forall Y \in \mathcal{T}$*,* $\forall \lambda, \mu \geq 0$*,* $\lambda X + \mu Y \in \mathcal{U}$*.*

The following proposition answers question Q1) completely for 2-coherence:

Proposition 10. *Let* $\mathcal{A} \subseteq \mathcal{X}$ *be such that*

a) $\lambda\mathcal{A} + \mathcal{R}(B)^{\succ} \subseteq \mathcal{A}$*,* $\forall \lambda \geq 0$*,* $\forall B \in \mathcal{B}$*;*

b) $\mathcal{R}(B)^{\prec} \cap \mathcal{A} = \emptyset$*,* $\forall B \in \mathcal{B}$*.*

c) $(\mathcal{R}(B_1) \cap \mathcal{A}) + (\mathcal{R}(B_2) \cap \mathcal{A}) \subseteq \mathcal{R}(B_1 \vee B_2) \setminus \mathcal{R}(B_1 \vee B_2)^{\prec}, \forall B_1, B_2 \in \mathcal{B}$*.*

Define, $\forall X|B \in \mathcal{D}_{LIN}$*,*

$$\underline{P}(X|B) = \sup\{x : B(X - x) \in \mathcal{A}\}. \tag{14}$$

Then, $\underline{P}$ *is* 2*-coherent on* $\mathcal{D}_{LIN}$*.*

Unlike the case of coherent conditional lower previsions examined in [17, Section 3.1], $\mathcal{A}$ does not need to be a cone in Proposition 10: given $X, Y \in \mathcal{A}$, $\lambda \geq 0$, neither $X + Y$ nor λX are guaranteed to belong to $\mathcal{A}$. Actually, condition a) represents a weakening of the cone axioms: if $X \in \mathcal{A}$, $Y \in \mathcal{R}(B)^{\succ}$ and $\lambda \geq 0$, then $\lambda X + Y \in \mathcal{A}$. This implies also $\mathcal{R}(B)^{\succ} \subseteq \mathcal{A}$ $\forall B \in \mathcal{B}$, a condition that, like also b), is required for coherence as well (see (C1'), (C2') in [17, Section 3.1]).

The interpretation of b) is that of an *avoiding partial loss* condition: we can expect no gain from owning a gamble in $\mathcal{R}(B)^{\prec}$, when B is true, therefore such gambles cannot be included into $\mathcal{A}$.

[3] As will appear later, $\mathcal{A}$ is included into some fixed linear space of gambles.

[4] Note that if $X \in \mathcal{X}$ and $B \in \mathcal{B}^{\varnothing}$, $X|B \in \mathcal{D}_{LIN}$ in the notation of the preceding sections.

As for c), writing it in an extended form, it tells us that: if $X_1, X_2 \in \mathcal{A}$, $B_1X_1 = X_1$, $B_2X_2 = X_2$, then $(B_1 \vee B_2)(X_1+X_2) = X_1+X_2$ and $\sup(X_1+X_2|B_1\vee B_2) \geq 0$. Note that if $X_1 \in \mathcal{R}(B_1)$ and $X_2 \in \mathcal{R}(B_2)$, it always holds that $X_1 + X_2 \in \mathcal{R}(B_1 \vee B_2)$, without having to impose it by means of axiom c). In fact, we have that $(B_1 \vee B_2)(X_1 + X_2) = (B_1 \vee B_2)(B_1X_1 + B_2X_2) = (B_1 \vee B_2)B_1X_1 + (B_1 \vee B_2)B_2X_2 = B_1X_1 + B_2X_2 = X_1 + X_2$, so that $X_1 + X_2 \in \mathcal{R}(B_1 \vee B_2)$.

Therefore, the essential condition in axiom c) is that if X_1, X_2 are desirable (belonging to $\mathcal{A}$), this does not imply that $X_1 + X_2 \in \mathcal{A}$ (which is required for coherence in [17, 18]), but only that $X_1 + X_2$ is not necessarily discarded by resorting to b). To illustrate this concept, let for instance $B_1 = B_2 = \Omega$ in c), so that $\mathcal{R}(B_1) = \mathcal{R}(B_2) = \mathcal{R}(B_1 \vee B_2) = \mathcal{R}(\Omega) = \mathcal{X}$. Then, c) implies $X_1+X_2 \notin \mathcal{R}(\Omega)^{\prec}$, making impossible to apply b) in order to discard $X_1 + X_2$ from $\mathcal{A}$.

As for question Q2), an answer is given by the following proposition, when $\underline{P}$ is 2-coherent.

Proposition 11. *Let $\underline{P} : \mathcal{D}_{LIN} \to \mathbb{R}$ be 2-coherent. Define*

$$\mathcal{A}' = \{\lambda B(X - x) + Y : X|B \in \mathcal{D}_{LIN}, \quad x < \underline{P}(X|B), Y \in \mathcal{X}^{\succeq}, \lambda \geq 0\}. \tag{15}$$

Then the set $\mathcal{A}'$ is such that:

a') $a\mathcal{A}' + \mathcal{X}^{\succeq} \subseteq \mathcal{A}'$, $\forall a \geq 0$;

b') $\mathcal{X}^{\preceq} \cap \mathcal{A}' = \{0\}$;

c') $(\mathcal{A}' + \mathcal{A}') \setminus \{0\} \subseteq \mathcal{X} \setminus \mathcal{X}^{\preceq}$;

d') $\underline{P}(X|B) = \sup\{x : B(X - x) \in \mathcal{A}'\}$, $\forall X|B \in \mathcal{D}_{LIN}$.

Proposition 11 states the existence of a set of desirable gambles $\mathcal{A}'$, in accordance with a given 2-coherent conditional lower prevision $\underline{P}$ and satisfying the rationality criteria a'), b'), c'). Comparing a'), b') with a), b) in Proposition 10, a clear similarity comes evident: essentially, the sets $\mathcal{R}(B)^{\succ}$, $\mathcal{R}(B)^{\prec}$, $B \in \mathcal{B}$, have been replaced with $\mathcal{X}^{\succeq}$, $\mathcal{X}^{\preceq}$ respectively. As a consequence, note that $0 \in \mathcal{A}'$.

The interpretation of c') is similar to c) in Proposition 10. It tells that: if $X_1, X_2 \in \mathcal{A}'$, $X_1 + X_2 \neq 0$, then $\sup(X_1 + X_2) > 0$. Again, coherence would allow the stronger implication $X_1, X_2 \in \mathcal{A}' \to X_1+X_2 \in \mathcal{A}'$, while 2-coherence only ensures that $X_1 + X_2$ does not belong to the (near) rejection set $\mathcal{X}^{\preceq}$.

Actually, a'), b') c') prove to be stronger than a), b), c). This means that any 2-coherent conditional prevision can be represented through a set of desirable gambles $\mathcal{A}'$ satisfying the necessary axioms a'), b'), c'), but also that, at the same time, $\mathcal{A}'$ satisfies the weaker axioms a), b), c) in Proposition 10.

A comparison between (3) in Definition 3 and (10) in Definition 5 intuitively suggests that we can get an answer to Q1) for 2-convexity from a reduced form of Proposition 10, with $\lambda = 1$. More precisely, the following proposition holds:

Proposition 12. *Let $\mathcal{A} \subseteq \mathcal{X}$ be such that*

a) $\mathcal{A} + \mathcal{R}(B)^{\succ} \subseteq \mathcal{A}$, $\forall B \in \mathcal{B}$;

b) $\mathcal{R}(B)^{\prec} \cap \mathcal{A} = \emptyset$, $\forall B \in \mathcal{B}$.

Define, $\forall X|B \in \mathcal{D}_{LIN}$,

$$\underline{P}(X|B) = \quad \sup\{x : B(X - x) \in \mathcal{A}\}. \tag{16}$$

Then, $\underline{P}$ is 2-convex on $\mathcal{D}_{LIN}$. Moreover, $\underline{P}$ is centered iff $\mathcal{R}(B)^{\succ} \subseteq \mathcal{A}$ $\forall B \in \mathcal{B}$.

An analogously reduced form of Proposition 11 allows us to answer question Q2) for 2-convexity.

Proposition 13. *Let $\underline{P} : \mathcal{D}_{LIN} \to \mathbb{R}$ be 2-convex. Define*

$$\mathcal{A}' = \{B(X - x) + Y : X|B \in \mathcal{D}_{LIN}, \quad x < \underline{P}(X|B), Y \in \mathcal{X}^{\succeq}\}. \tag{17}$$

Then the set $\mathcal{A}'$ is such that:

a) $\mathcal{A}' + \mathcal{X}^{\succeq} \subseteq \mathcal{A}'$;

b) $\mathcal{X}^{\preceq} \cap \mathcal{A}' = \emptyset$ iff $\underline{P}(0|B) \leq 0$, $\forall B \in \mathcal{B}^{\varnothing}$;

c) $\underline{P}(X|B) = \sup\{x : B(X - x) \in \mathcal{A}'\}$, $\forall X|B \in \mathcal{D}_{LIN}$.

Further, $\underline{P}$ is centered iff $\mathcal{R}(B)^{\succ} \subseteq \mathcal{A}'$ $\forall B \in \mathcal{B}$.

Comparing Propositions 10 and 11 with, respectively, Propositions 12 and 13, we note that, in addition to the constraint $\lambda = 1$, 2-convexity requires no condition like c) and c') in Propositions 10 and 11 respectively. Referring, for instance, to c'), this means that, given $X, Y \in \mathcal{A}'$ with $X + Y \neq 0$, 2-convexity does not guarantee $\sup(X + Y) > 0$: summing up two individually desirable gambles could therefore give rise to a partial or even to a sure loss. Moreover, a non-centered 2-convex $\underline{P}$ suffers from a more serious shortcoming: if $\mathcal{R}(B)^{\succ} \subseteq \mathcal{A}'$ does not necessarily hold, then a non-negative gamble $X = BX$ $(X \neq 0)$ exists, that is considered non-desirable. The main drawbacks of 2-convexity relative to 2-coherence are therefore clearly pointed out by a comparison through desirability axioms.

7 Weakly Consistent Uncertainty Models

As mentioned in the Introduction, a motivation for studying the loose forms of consistency introduced in this paper is their capability of encompassing or extending uncertainty models already investigated in the literature. Even though these models may depart also considerably from coherence and convexity, they can nevertheless be accommodated into a unifying betting scheme, ranging from 2-convex to coherent lower previsions.

Focusing on 2-convexity, we first recall a few definitions and some results concerning unconditional 2-convex lower previsions.

Definition 7. *Given a finite partition $\mathbb{P}$, and denoting with $2^{\mathbb{P}}$ its powerset, a mapping $c : 2^{\mathbb{P}} \to [0,1]$ is a (normalised)* capacity *whenever $c(\varnothing) = 0$, $c(\Omega) = 1$ (normalisation) and $\forall A_1, A_2 \in 2^{\mathbb{P}}$ such that $A_1 \Rightarrow A_2$, $c(A_1) \leq c(A_2)$ (1-monotonicity).*

Definition 8. *Given a linear space $\mathcal{L}$ of random variables, a* niveloid *[2, 5] is a functional $N : \mathcal{L} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ which is translation invariant and monotone, i.e. such that*

$$\begin{array}{l} N(X+\mu) = N(X) + \mu, \forall X \in \mathcal{L}, \forall \mu \in \mathbb{R}; \\ X \geq Y \text{ implies } N(X) \geq N(Y), \forall X, Y \in \mathcal{L}. \end{array} \tag{18}$$

As well-known, capacities are uncertainty measures with really minimal quantitative requirements. Niveloids can be viewed as a generalisation of theirs to linear spaces of random variables which preserves their minimality properties. Strictly speaking, this is true for centered niveloids, i.e. such that $N(0) = 0$. In fact, the centering condition $N(0) = 0$ does not ensue from the definition of niveloid. Note also that niveloids apply to random variables which may be unbounded too.

It has been proven in [1, Section 4.1][5] that:

Proposition 14. *a) Let $\underline{P}$ be defined on $2^{\mathbb{P}}$. Then $\underline{P}$ is a centered 2-convex lower prevision if and only if it is a capacity.*

b) Let $\underline{P}$ be defined on a linear space $\mathcal{L}$ of bounded random variables (gambles). Then $\underline{P}$ is a 2-convex lower prevision if and only if it is a (finite-valued) niveloid.

Hence, an unconditional 2-convex lower prevision is equivalent to a capacity or a niveloid, on structured sets ($2^{\mathbb{P}}$ or $\mathcal{L}$ respectively). On non-structured sets,ù it extends these concepts.

[5] See Footnote 2.

2-convex conditional lower previsions are natural candidates to define conditional capacities and niveloids on *arbitrary* sets of, respectively, conditional events or gambles. To the best of our knowledge, such conditional versions have not been considered yet in this general conditional environment, but rather in more specific cases. For instance, [3] focuses on updating rules for 'convex' capacities, which means for 2-monotone lower probabilities, while considering a single conditioning event.

Thus 2-convex previsions may provide an appropriate framework for such extensions, guaranteeing some minimal properties like the existence of a 2-convex natural extension (when being centered). Take for instance centered 2-convex conditional lower probabilities. They satisfy the properties one would require to a conditional capacity: $\underline{P}(0|B) = 0$, $\underline{P}(\Omega|B) = 1$ (this follows from Proposition 7, a)), and $A|B \leq_{GN} C|D$ implies $\underline{P}(A|B) \leq \underline{P}(C|D)$ (Proposition 7, c)). Similarly, centered 2-convex lower previsions ensure generalisations of properties (18) (see especially Proposition 2 and Remark 1 for the first property, Proposition 7, c) for the second).

8 Conclusions

N-convex and n-coherent conditional lower previsions broaden the spectrum of uncertainty measures that can be accommodated into a behavioural approach to imprecision, including, for instance, conditional extensions of capacities and niveloids when $n = 2$. This choice for n is the most neatly distinguished from coherence, the other extreme in the spectrum, and that retaining more interesting properties. Among these the GBR must still hold. Centered 2-convex and 2-coherent previsions also have a clear meaning in terms of desirability. Further work is necessary to investigate additional properties, like the possible existence of envelope theorems, or properties of already defined notions. In particular, we conjecture that the 2-convex natural extension may simplify computing the convex natural extension. As a further generalisation of this work, the consistency notions defined here could be extended to the case of unbounded conditional random variables. This has been done in [15] for coherent conditional lower previsions, while, to the best of our knowledge, a similar investigation for convex conditional previsions is still missing.

9 Acknowledgments

We wish to thank the referees for their helpful suggestions. We acknowledge partial support by the FRA2013 grant 'Models for Risk Evaluation, Uncer-

tainty Measurement and Non-Life Insurance Applications'.

References

[1] P. Baroni, R. Pelessoni and P. Vicig. Generalizing Dutch risk measures through imprecise previsions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 17(2):153–177, 2009.

[2] S. Cerreia-Vioglio, F. Maccheroni, M. Marinacci and A. Rustichini. Niveloids and their extensions: Risk measures on small domains. *Journal of Mathematical Analysis and Applications*, 413(1):343–360, 2014.

[3] A. Chateauneuf, R. Kast and A. Lapied. Conditioning capacities and Choquet integrals: the role of comonotony. *Theory and Decision*, 51(2–4):367–386, 2001.

[4] B. de Finetti. *Theory of Probability*. Wiley, 1974.

[5] S. Dolecki and G.H. Greco. Niveloids. *Topological Methods in Nonlinear Analysis*, 5:1–22, 1995.

[6] H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447, 2002.

[7] H. Föllmer and A. Schied. Robust preferences and convex measures of risk. In: *Advances in Finance and Stochastics*, edited by K. Sandmann, K. and P. J. Schönbucher, 39–56, Springer-Verlag, 2002.

[8] I.R. Goodman and H.T. Nguyen. Conditional objects and the modeling of uncertainties. In: *Fuzzy Computing*, edited by M. Gupta and T. Yamakawa, 119–138, Elsevier (North-Holland) 1988.

[9] R. Pelessoni and P. Vicig. Convex imprecise previsions. *Reliable Computing*, 9:465–485, 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. Williams coherence and beyond. *International Journal of Approximate Reasoning*, 50(4):612–626, 2009.

[12] R. Pelessoni and P. Vicig. The Goodman–Nguyen relation within imprecise probability theory. *International Journal of Approximate Reasoning*, 55(8):1694–1707, 2014.

[13] E. Quaeghebeur. Desirability. In: *Introduction to Imprecise Probabilities*, edited by T. Augustin, F.P.A. Coolen, G. de Cooman, M.C.M. Troffaes, 1–27, J. Wiley and Sons, 2014.

[14] E. Quaeghebeur, G. de Cooman and F. Hermans. Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning*, 57:69–102, 2015.

[15] M. C .M. Troffaes and G. de Cooman. *Lower Previsions*. Wiley, 2014.

[16] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.

[17] P. M. Williams. *Notes on conditional previsions.* Research Report, School of Math. and Phys. Science, University of Sussex, 1975.

[18] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383,2007.

Statistical Modelling under Epistemic Data Imprecision: Some Results on Estimating Multinomial Distributions and Logistic Regression for Coarse Categorical Data

Julia Plass
Department of Statistics, LMU Munich
julia.plass@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

Marco E. G. V. Cattaneo
Department of Mathematics, University of Hull
m.cattaneo@hull.ac.uk

Georg Schollmeyer
Department of Statistics, LMU Munich
georg.schollmeyer@stat.uni-muenchen.de

Abstract

The paper deals with parameter estimation for categorical data under epistemic data imprecision, where for a part of the data only coarse(ned) versions of the true values are observable. For different observation models formalizing the information available on the coarsening process, we derive the (typically set-valued) maximum likelihood estimators of the underlying distributions. We discuss the homogeneous case of independent and identically distributed variables as well as logistic regression under a categorical covariate. We start with the imprecise point estimator under an observation model describing the coarsening process without any further assumptions. Then we determine several sensitivity parameters that allow the refinement of the estimators in the presence of auxiliary information.

Keywords. Coarse data, missing data, epistemic data imprecision, sensitivity analysis, partial identification, categorical data, multinomial logit model, coarsening at random (CAR), likelihood.

1 The Problem and its Background

A frequent challenge in statistical modelling is *data imprecision*, where some data are *coarse*, i.e. they are not observed in the resolution originally intended in the subject matter context. Throughout this paper, we focus on the case where the coarse observations are data under *epistemic data imprecision*. For categorical data as considered here this means that there exists a true precise value y of a generic variable Y taking values in a finite sample space $\Omega_Y = \{1, \ldots, K\}$, but we may only observe a non-singleton set $\mathscr{y}$ containing y. It is important to distinguish epistemic from *ontic* data imprecision, where data are coarse by nature and thus have to be interpreted as indivisible entities of their own (see, in particular, [7, 8]; [24] for an application in a multinomial logit model and classification.)

Epistemic data imprecision emerges most naturally in a huge variety of applications. Missing data, interpreted as the prominent special case where the whole sample space is observed only, arise, for instance directly by design in observational studies on treatment effects, see, e.g., [27], and unit non-response is quite frequent in surveys, in particular as refusals to answer sensitive questions. Typical instances of not missing but still coarse data include the numerous data sets where coarsening is deliberately applied as an anonymization technique (see, e.g., [10]), matched data sets with not completely identical categories, secondary data where the originally coded categories turn out to be not fine enough and, as a technical example, reliability analysis of a system whose components are tested separately prior to assembly [30].

Trapped in the framework of precise probabilities, traditional statistical methods are forced to neglect data imprecision or to impose quite strong, empirically untestable assumptions on the underlying coarsening process. Thus, except the very rare cases where the external information on the subject matter problem is rich enough to justify such an extent of precision of the modelling of the coarsening process, the price of the (seemingly) precise result is a substantial debilitation of the reliability of the conclusions drawn.

Against this background, set-valued approaches, aiming at a proper reflection of the available information, have been gathering momentum, also becoming a popular topic at the ISIPTA symposia ([5, 26, 17, 32, 33], to name just a few contributions). In different areas of application concepts of cautious data completion emerged, where a classical procedure is extended by considering the set of all virtual precise observations in accordance with the coarse data (see, e.g., the exposition in [2], and the references therein). General investigations of coarse data from an imprecise-probability-based Bayesian point of view include [6, 36]; random set-based perspectives are developed for instance in

[8, 21]. Linear regression under metrical coarse data (interval data) is vividly discussed in the partial identification literature in the spirit of [19] (see also, e.g, [26], and the references therein). Mainly focusing on missing data, [34] suggests a framework for a systematic sensitivity analysis for statistical modelling under epistemic data imprecision. [5] introduces a profile likelihood approach for coarse data (for missing data see also [37]) and derive from it a uniform framework for robust regression analysis with imprecise data.

This paper will develop another likelihood-based (see, e.g., [4, § 6.3, 7.2.2] for a general introduction) approach and we will in addition briefly sketch Bayesian approaches in Section 3. Our work is strongly influenced by the methodology of partial identification, dealing with the trade-off between information and credibility by first using the empirical evidence only, i.e. using information implied by the data and including only those assumptions about which there exists a common consensus concerning their validity (e.g., [19, 28, 20]). Sensitivity analysis pursues the same goal, but proceeds in a different direction. While partial identification starts from total uncertainty and gradually adds assumptions, in the framework of sensitivity analysis the collection of all precise results from successively relaxed assumptions is considered. Thereby, the analysis is framed by a sensitivity parameter, which is not identified but suffices to identify the parameter of interest, (e.g., [34]).

Our paper is structured as follows. In the next section we fix the notation and formulate the problem setting more exactly for the cases considered in this paper: independent and identically distributed (i.i.d.) variables and logistic regression with a categorical covariate. The crucial technical argument underlying our paper (developed in general terms in Section 3) is to introduce an observation model and utilize invariance properties of the likelihood. In Section 4 we derive and discuss the set-valued estimators arising from a fully non-committal observation model, and we then turn to settings where this interval is narrowed when we benefit from the presence of additional auxiliary information. For technically handling this by sensitivity parameters, it is helpful to go to the other extreme, investigating point identifying additional assumptions in some special cases. For the homogeneous situation, after studying known coarsening in Section 5.1, we focus on the coarsening at random (CAR) assumption and illustrate the disastrous behaviour of the resulting point estimator when CAR is inappropriate (Section 5.2). Then in Section 5.3 we consider an extension of CAR and determine the corresponding ratio of coarsening probabilities as a sensitivity parameter. For the logistic regression case in Section 5.4 we work out that there is, as an alternative to CAR and its extensions, a further assumption refining the initial set of estimators to a precise result. This assumption is called subgroup independent coarsening and its generalization again can serve as a sensitivity parameter (Section 5.5). These sensitivity parameters frame a systematic sensitivity analysis, resulting in imprecise point estimators reflecting justifiable auxiliary information.

2 The Basic Setting

Let $Y_1, \ldots, Y_n$ be a random sample of a categorical response variable of interest Y with realizations $y_1, \ldots, y_n$ in sample space $\Omega_Y = \{1, \ldots, j, \ldots, K\}$. Problematically, some of those realizations are not known in a precise form, and thus only realizations $\mathscr{y}_1, \ldots, \mathscr{y}_n$ of a sample $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ of a random variable $\mathcal{Y}$ within sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \emptyset$ can be observed, where $\mathcal{P}$ denotes the power set. The possible categories of $\mathcal{Y}$ constitute the singletons of $(\Omega_{\mathcal{Y}},\ \mathcal{P}(\Omega_{\mathcal{Y}}))$, with corresponding probability mass functions $p_{\mathscr{y}_i} = P(\mathcal{Y}_i = \mathscr{y}_i)$ $(i = 1, \ldots, n)$. But as we are interested in the random variables $Y_1, \ldots, Y_n$, our basic goal consists of gathering information about the individual probabilities $\pi_{i1} = P(Y_i = 1), \ldots, \pi_{iK} = P(Y_i = K)$. Thereby, we assume throughout the paper that the coarsening process is error-free, in the sense that $\mathscr{y}_i \ni y_i$, $i = 1, \ldots, n$.

We discuss the homogeneous case (i.i.d. case), in biometrical terms *prevalence* estimation, as well as situations with one precise categorical covariate X, in biometrical terms called *treatment*, with sample space Ω_X, being available. Both situations will be illustrated by means of the following example.

Running Example: *We refer to the data from the German panel study "Labor Market and Social Security" (PASS, wave 1, 2006/2007, [29]). As asking for the income may be regarded as a sensitive question and thus the response rate is expected to be low, in this study non-responders are required to report their income in classes starting from rather large classes that are narrowed by following questions. By proceeding in this way, anonymization is guaranteed in the level that is requested by the respondents and answers of different degrees of coarseness are obtained. Keeping things simple, here we refer to the data from question "HEK0700", where respondents are asked to report if their income Y is < 1000€ or ≥ 1000€ ($y_i \in \{<, \geq\}$; "$<$" and "$\geq$" abbreviating these classes, respectively) and our main goal is the estimation of $\pi_<$. As some respondents gave no suitable answer ("na") and cannot be allocated to one of the classes, partly only coarsened values of the variable $\mathcal{Y}$ are observed ($\mathscr{y}_i \in \{<, \geq, na\}$).*

Example, version 1: In order to illustrate the i.i.d. case, we only consider the reported answers of the income question, where 238, 835 and 338 respondents reported "<", "≥" and "na", respectively ($n_< = 238,\ n_\geq = 835,\ n_{\text{na}} = 338$).

In the case with categorical covariates, we here confine ourselves to one categorical covariate only, as this is technically equivalent to any finite set of categorical covariates. While in the i.i.d. case probabilities $\pi_{i1} = \pi_1, \ldots,\ \pi_{iK} = \pi_K$ are assumed to be independent of individual i, in the case with one covariate the probabilities $\pi_{i1} = P(Y_i = 1|X_i = x_i) = \pi_{x_i 1}, \ldots,\ \pi_{iK} = P(Y_i = K|X_i = x_i) = \pi_{x_i K}$ are influenced by individual i through the corresponding value of the covariate X_i. One of most generally applied models is the *multinomial logit model*. It describes the dependence of a categorical dependent variable Y of nominal scale on covariates X by

$$\pi_{ij} = P(Y_i = j|\mathbf{x}_i) = \frac{\exp(\beta_{j0} + \mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} \tag{1}$$

$i = 1, \ldots, n$ for categories $j = 1, \ldots, K-1$ and by

$$\pi_{iK} = \left(1 + \sum_{s=1}^{K-1} \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)\right)^{-1} \tag{2}$$

with category specific regression coefficients, that is $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$ referring to m covariates and intercept β_{j0}. As we here address the case of one categorical covariate $X_i \in \{1, \ldots, c\}$, dummy coded variables $X_{i1}, \ldots, X_{im}$ with $m = c - 1$ are included into the model.[1]

It is common to summarize categorical data in contingency tables by reporting the counts for possible outcomes, where the covariates X are supposed to be in the rows (e.g., [31]). Thus, in our case the contingency table in Table 1 will be addressed. The number of observations with $\mathcal{Y} = \mathscr{y}$ and treatment group $X = x$ is denoted by $n_{x\mathscr{y}}$, where $n_0 = n_{0A} + n_{0B} + n_{0AB}$, $n_1 = n_{1A} + n_{1B} + n_{1AB}$, $n_A = n_{0A} + n_{1A}$, $n_B = n_{0B} + n_{1B}$ and $n_{AB} = n_{0AB} + n_{1AB}$.

Example, version 2: Illustrating the case with a categorical covariate, apart from the partial income knowledge, the receipt of the so-called Unemployment Benefit II (variable alg2abez; here denoted by UBII) is considered and serves in the model in Expressions (1) and (2) as covariate X_i, $i, \ldots, n$. The data are summarized in Table 2.

[1] Dummy variable X_{il} $(l = 1, \cdots, m)$ attains value 1 if the l-th category is chosen by individual i, otherwise it is 0. In this way, reference category c is represented by all dummy variables being 0.

		$\mathcal{Y}$			
		A	B	AB	total
X	0	n_{0A}	n_{0B}	n_{0AB}	n_0
	1	n_{1A}	n_{1B}	n_{1AB}	n_1
	total	n_A	n_B	n_{AB}	n

Table 1: Contingency table that introduces used notation.

		income			
		<	≥	na	total
UBII	yes (0)	130	114	75	319
	no (1)	108	721	263	1092
	total	238	835	338	1411

Table 2: Contingency table to illustrate some results by means of the PASS data.

3 Sketch of the Basic Argument

This paper, similarly to [5, 37], relies on the likelihood as the fundamental concept to derive parameter estimators under epistemic data imprecision, but looks at it from a different angle. In order to support the appropriate incorporation of the available information provided by the data and the background knowledge, we explicitly formulate, and utilize, an *observation model* relating the observable level and the ideal level. The observation model is a set $\mathcal{Q}$ of (precise) coarsening probabilities,[2] and thus the medium to specify carefully and flexibly the available information about the coarsening process.

By virtue of the theorem of total probability, the elements of $\mathcal{Q}$ relate the probability distribution of the imprecise observation $\mathcal{Y}$ to the distribution of the underlying latent variable Y (and, if present, certain covariates).

Parametrizing the distributions, again possibly after splitting with respect to certain covariate values, let ϑ (the various p's in the following sections) and η (the various π's below) be the parameters determining the distribution of $\mathcal{Y}$ and Y, respectively, and let ζ be the parameter characterising the elements of $\mathcal{Q}$ (the various q's, possibly constrained by the specified constraints: $\left(q_{\mathscr{y}|y} := P(\mathcal{Y} = \mathscr{y}|Y = y)\right)_{(\mathscr{y} \in \Omega_{\mathcal{Y}}, y \in \Omega_Y)}$ in the i.i.d. case, while in the regression context the coarsening mechanisms generally also depend on the values of X_i, i.e., $(q_{\mathscr{y}|xy} := P(\mathcal{Y} = \mathscr{y}|X = x, Y = y)_{(\mathscr{y} \in \Omega_{\mathcal{Y}}, y \in \Omega_Y, x \in \Omega_X)}$ has to be considered).

Then we can describe the relationship between $\gamma := (\eta^T, \zeta^T)^T \in \Gamma$ and $\vartheta \in \Theta$ via the mapping $\Phi : \Gamma \to \Theta$, $\gamma \mapsto \vartheta$. Figure 1 and the running example illustrate

[2] More precisely, $\mathcal{Q}$ is a generalized transition kernel, consisting of credal sets indexed by the values of Y.

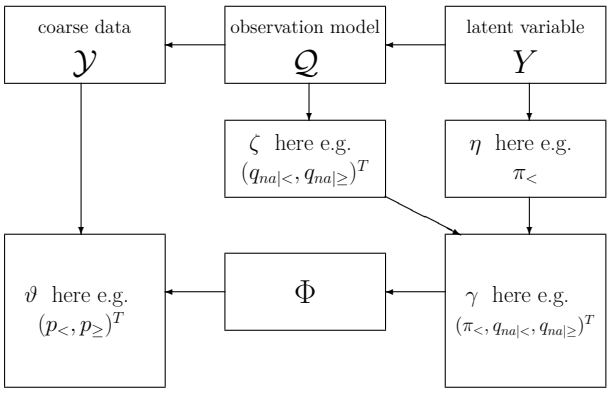


Figure 1: Observable and latent variable and the corresponding parameters.

this mapping $\Phi(\cdot)$ and all parameters involved.

Example, version 1 (cont.): The mapping $\Phi(\cdot)$ with arguments $\zeta = (q_{\text{na}|<}, q_{\text{na}|\geq})^T$ and $\eta = \pi_<$ establishes a connection to the parameters determining the probabilities of the observable income variable $\mathcal{Y}$, namely $\vartheta = (p_<, p_\geq)^T$.

In a first step (Section 4), we will only assume that the coarsening process is error-free and therefore take $\mathcal{Q}$ as the set of all coarsening mechanisms compatible with error-freeness. Then (Section 5), by using auxiliary information, we sharpen this set $\mathcal{Q}$. Note that we do neither assume anything about the plausibility of different elements ζ of $\mathcal{Q}$ nor do we treat different $y \in \mathscr{Y}$ as differently plausible. To derive the estimators, the invariance of the likelihood under parameter transformations is crucial: evaluating the likelihood in terms of γ and in terms of $\vartheta = \Phi(\gamma)$ is equivalent here. Our random set modelling will allow us to determine the ML-estimator $\hat{\vartheta}$ of ϑ, which moreover, apart from trivial extreme cases, can be shown to be single-valued. Then the possibly set-valued maximum-likelihood estimator for γ is obtained as

$$\hat{\Gamma} = \left\{ \gamma \,\middle|\, \Phi(\gamma) = \hat{\vartheta} \right\} \tag{3}$$

(see also [5, Section 2]). Thus, adapting the concept of maximum likelihood (ML) estimators to a persistent set-based perspective and to random set-based situations, we achieve a general and powerful framework for handling coarse categorical data via the mapping $\Phi(\cdot)$. If $\Phi(\cdot)$ is injective, then $\hat{\Gamma}$ is a singleton as well, and γ so-to-say empirically point identified; otherwise $\hat{\Gamma}$ is set-valued in the literal sense and γ empirically partially identified.

This compares to other approaches: A classical Bayesian analysis would put some prior on ζ and on η (cf., e.g., [23, 14]) while a generalized Bayesian analysis would replace one or both priors by a set of priors. This can be seen as imposing imprecise priors on ζ and on η. The non-committal analysis would start with a near-ignorance prior, for instance based on Dirichlet distributions adapting [35]'s imprecise Dirichlet model, and auxiliary information can be expressed by smaller credal sets; compare also the general Bayesian treatment of incomplete information in [6, 36]. Partially differently, in [3, Section 4.4.] an approach is presented that puts a precise prior on η and no prior on ζ and models the coarsening process with a multivalued mapping. This may be seen as imposing a vacuous imprecise probability on ζ. In another direction, one could impose some prior knowledge w.r.t. the imprecise data point $\mathscr{Y}$ by assuming different $y \in \mathscr{Y}$ as differently plausible. This can be done for example by imposing a possibility distribution on y (cf., e.g., [9, Section 3.2.]) or constructing observations directly by data augmentation (cf., e.g., [18]).

The dimension of the parameter vectors η and ζ increases substantially with the cardinality of Ω_Y and Ω_X. In the i.i.d. case $m = (\sum_{z=1}^{|\Omega_Y|} \binom{|\Omega_Y|}{z} \cdot z) - 1$ or equivalently $m = K \cdot 2^{K-1} - 1$ parameters have to be estimated, where in the case with one covariate this number even increases to $|\Omega_X| \cdot m$. Thus, for reasons of conciseness of presentation, we confine detailed explanations and derivations on the special, yet still representative cases of a binary response variable Y with sample space $\Omega_Y = \{A,\ B\}$ and observations within $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, as well as a binary precise categorical covariate X with values 0 and 1. Then the underlying model expressed in Expression (1) and (2) is called *logit model*. As the inclusion of more than one dummy variable simply leads to an increase of the number of subgroups, all results can be transferred straightforwardly to more general cases, namely cases with more than one non-binary covariates. Furthermore, the main results not only will be shown for the situation of a binary Y, where coarsening corresponds to missingness, but also in a general way.

4 Maximum Likelihood Estimation without Additional Information

In this section we derive the maximum likelihood estimators for the case where no additional information on the coarsening process is available, i.e. there are no constraints on the elements of $\mathcal{Q}$. A crucial step is to rely on the random set view that treats data imprecision as a change of the sample space with corresponding random variables $\mathcal{Y}_i$, $i = 1, \ldots, n$, which then lead to multinomially distributed variables with parameter ϑ for the counts based on the new sample space. According to the argumentation in Section 3, the resulting likelihood in ϑ, and the estimator derived

from maximizing it, will then be related to the parameters of the distribution of the latent variable (and the observation model). As just discussed, we explain the construction in some detail for the representative special cases with $\Omega_Y = \{A,\ B\}$ (and $\Omega_X = \{0,1\}$) and then report the general results.

4.1 Estimation in the i.i.d. Case

Considering categorical i.i.d. random variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ with realizations $\mathscr{y}_1, \ldots, \mathscr{y}_n$ in the sample space $\Omega_{\mathcal{Y}} = \{A,\ B,\ AB\}$, we obtain the following likelihood function for the parameter $\vartheta = (p_A, p_B)^T$ given the data, summarized by the counts n_A, n_B and n_{AB} (with $p_{AB} = 1 - p_A - p_B$):[3]

$$\begin{aligned} L(\vartheta) = L(p_A,\ p_B) &= L(p_A,\ p_B || \mathscr{y}_1, \ldots, \mathscr{y}_n) \qquad (4) \\ &= P(\mathscr{y}_1, \ldots, \mathscr{y}_n || p_A,\ p_B) \propto p_A^{n_A} \cdot p_B^{n_B} \cdot p_{AB}^{n_{AB}} \,. \end{aligned}$$

For $n = n_A + n_B + n_{AB} > 0$ this likelihood is uniquely maximized by the relative frequencies (see [25]),

$$\hat{p}_A^{(MLE)} = \frac{n_A}{n}, \qquad \hat{p}_B^{(MLE)} = \frac{n_B}{n}\,, \qquad (5)$$

and thus $\hat{p}_{AB}^{(MLE)} = 1 - \hat{p}_A^{(MLE)} - \hat{p}_B^{(MLE)} = \frac{n_{AB}}{n}$.

Essentially, we are interested in the parameter $\eta = \pi_A$ determining the probabilities of the true, but unobserved variable Y being equal to particular categories and the associated maximum likelihood estimator. Those probabilities of interest, in our case π_A and $\pi_B = 1 - \pi_A$, can be related with probabilities p_A, p_B and p_{AB} corresponding to the observable variables by

$$\begin{aligned} p_A &= (1 - q_{AB|A}) \cdot \pi_A \,, \qquad (6) \\ p_B &= (1 - q_{AB|B}) \cdot (1 - \pi_A) \,, \end{aligned}$$

where $p_{AB} = q_{AB|A} \cdot \pi_A + q_{AB|B} \cdot (1 - \pi_A)$ results from the law of total probability.

This means that the likelihood in terms of $\vartheta = (p_A, p_B)^T$ in Expression (4) and in terms of $\gamma = (\pi_A, q_{AB|A}, q_{AB|B})^T$, coincide, indeed.

By the invariance of the likelihood under parameter transformations, Expressions (5) and (6) can be combined, resulting in the following system of equations:

$$\begin{aligned} (1 - \hat{q}_{AB|A}) \cdot \hat{\pi}_A &= \frac{n_A}{n} = \hat{p}_A^{(MLE)}, \\ (1 - \hat{q}_{AB|B}) \cdot (1 - \hat{\pi}_A) &= \frac{n_B}{n} = \hat{p}_B^{(MLE)}, \qquad (7) \\ \hat{q}_{AB|A} \cdot \hat{\pi}_A + \hat{q}_{AB|B} \cdot (1 - \hat{\pi}_A) &= \frac{n_{AB}}{n} = \hat{p}_{AB}^{(MLE)}. \end{aligned}$$

For reasons of redundancy we can leave the third equation out of consideration. As there typically are multiple triples $\hat{\gamma} = (\hat{\pi}_A,\ \hat{q}_{AB|A},\ \hat{q}_{AB|B})^T$ that lead to the same values of $\hat{\vartheta} = (\hat{p}_A^{(MLE)},\ \hat{p}_B^{(MLE)})^T$, the mapping $\Phi:\ [0,\,1]^3 \to [0,\,1]^2$ with

$$\Phi \begin{pmatrix} \pi_A \\ q_{AB|A} \\ q_{AB|B} \end{pmatrix} = \begin{pmatrix} \pi_A \cdot (1 - q_{AB|A}) \\ (1 - \pi_A) \cdot (1 - q_{AB|B}) \end{pmatrix} = \begin{pmatrix} p_A \\ p_B \end{pmatrix} \quad (8)$$

(cf. Figure 1 for the case of the running example) connecting both parametrizations in general is not injective. Thus the maximum likelihood estimate $\hat{\Gamma}$ from Expression (3) is set-valued in the literal sense. Points in this set are constrained through the relationships in (7), and thus $\hat{\Gamma}$ is not a cuboid in $[0,\,1]^3$. Building the one dimensional projections, set-valued estimators of the single components of γ are obtained via

$$\begin{aligned} \hat{\pi}_A &\in \left[\frac{n_A}{n},\ \frac{n_A + n_{AB}}{n}\right], \qquad (9) \\ \hat{q}_{AB|A} &\in \left[0,\ \frac{n_{AB}}{n_A + n_{AB}}\right], \end{aligned}$$

and analogously for $\hat{q}_{AB|B}$, where $\frac{0}{0} := 1$.

Extending the discussion here to the general case of $\Omega_Y = \{1, \ldots, K\}$ and the corresponding $\Omega_{\mathcal{Y}}$, the estimators in Expression (9) generalize to

$$\hat{\pi}_y \in \left[\frac{n_{\{y\}}}{n},\ \frac{\sum_{\mathscr{y} \ni y} n_{\mathscr{y}}}{n}\right] \quad \hat{q}_{\mathscr{y}|y} \in \left[0,\ \frac{n_{\mathscr{y}}}{n_{\{y\}} + n_{\mathscr{y}}}\right], \qquad (10)$$

(where as above $\frac{0}{0} := 1$) for all $y \in \Omega_y = \{1, \ldots, K\}$ and all $\mathscr{y} \in \Omega_{\mathcal{Y}}$ such that $\{y\} \subset \mathscr{y}$.[4]

Example, version 1 (cont.): Applying Expression (10) to our example, one obtains

$$\hat{\pi}_< \in \left[\frac{238}{1411},\ \frac{238 + 338}{1411}\right] = [0.17,\ 0.41] \,.$$

4.2 Logistic Regression with a Categorical Covariate

Now we consider the heterogeneous situation expressed by a discrete covariate X, which also has been depicted in Table 1. Again we can derive set-valued estimators of the parameters of interest $\eta = (\pi_{0A}, \pi_{1A})^T$ (and the auxiliary parameter ζ characterizing the coarsening mechanisms) by taking the random set perspective, setting up the corresponding likelihood function and

[3] In the following, we will use the abbreviated notation of the likelihood without referring to the data.

[4] The estimators of the probability components of the distribution of Y_i prove to be the same as arising from a belief functions like construction of empirical probabilities and also coincide with the estimator obtained from cautious data completion, plugging in all potential precise sample outcome compatible with the observations $\mathscr{y}_1, \ldots, \mathscr{y}_n$ (see, e.g., [2])

applying the appropriate parameter transformations. Proceeding in this way, for fixed treatment group x the cell counts $(n_{xA}, n_{xB},\ n_{xAB})$ follow a multinomial distribution, i.e. $(n_{xA}, n_{xB}, n_{xAB}) \sim M(n_x, (p_{xA}, p_{xB}, p_{xAB}))$ with conditional probabilities $p_{x\mathscr{y}} = P(\mathcal{Y} = \mathscr{y}|X = x)$ (see [31, 1]).[5] Therefore, the corresponding likelihood function is given by

$$\begin{aligned} L(\vartheta) &= L(p_{0A},\ p_{1A},\ p_{0B},\ p_{1B}) \\ &\propto p_{0A}^{n_{0A}} \cdot p_{0B}^{n_{0B}} \cdot p_{0AB}^{n_{0AB}} \cdot p_{1A}^{n_{1A}} \cdot p_{1B}^{n_{1B}} \cdot p_{1AB}^{n_{1AB}}. \end{aligned} \tag{11}$$

For $n_x > 0$ the maximum likelihood estimators for the parameters are unique and given by (see [25])

$$\hat{p}_{x\mathscr{y}}^{(MLE)} = \frac{n_{x\mathscr{y}}}{n_x},\ \text{for } x \in \{0, 1\}.$$

Analogously to Section 4.1, we consider the mapping, which connects both parametrizations, $\Phi:\ [0,\ 1]^6 \to [0,\ 1]^4$ with

$$\Phi \begin{pmatrix} \pi_{0A} \\ \pi_{1A} \\ q_{AB|0A} \\ q_{AB|1A} \\ q_{AB|0B} \\ q_{AB|1B} \end{pmatrix} = \begin{pmatrix} \pi_{0A} \cdot (1 - q_{AB|0A}) \\ \pi_{1A} \cdot (1 - q_{AB|1A}) \\ (1 - \pi_{0A}) \cdot (1 - q_{AB|0B}) \\ (1 - \pi_{1A}) \cdot (1 - q_{AB|1B}) \end{pmatrix} = \begin{pmatrix} p_{0A} \\ p_{1A} \\ p_{0B} \\ p_{1B} \end{pmatrix} \tag{12}$$

(cf. Figure 1) and observe that in this case it is also not injective and thus $\hat{\Gamma}$, constructed along the line of (3), is strictly set-valued, too. Illustrating $\hat{\Gamma}$ again by the corresponding projections along the axes, we obtain for given value $x \in \{0, 1\}$ in the general case with more than two categories in Y, i.e. $y \in \Omega_Y = \{1, \ldots, K\}$ and $\mathscr{y} \in \Omega_{\mathcal{Y}}$ with $\{y\} \subset \mathscr{y}$,

$$\hat{\pi}_{xy} \in \left[\frac{n_{x\{y\}}}{n_x},\ \frac{\sum\limits_{\mathscr{y} \ni y} n_{x\mathscr{y}}}{n_x} \right],\ \hat{q}_{\mathscr{y}|xy} \in \left[0,\ \frac{n_{x\mathscr{y}}}{n_{x\{y\}} + n_{x\mathscr{y}}} \right], \tag{13}$$

where again $\frac{0}{0} := 1$.[6]

Example, version 2 (cont.): Applying Expression (13) to our example, one obtains

$$\begin{aligned} \hat{\pi}_{0<} &\in \left[\frac{130}{319},\ \frac{130 + 75}{319} \right] = [0.41,\ 0.64], \\ \hat{\pi}_{1<} &\in \left[\frac{108}{1092},\ \frac{108 + 263}{1092} \right] = [0.10,\ 0.34]. \end{aligned}$$

By recurring on the relation defined in Expression (1) and (2), and utilizing the injectivity of the logistic function, the likelihood function considered here can also be uniquely expressed in terms of the regression coefficients. In this way, instead of the estimators $\hat{\pi}_{0A}$ and $\hat{\pi}_{1A}$ determined by Expression (13), equivalently one can consider the estimators

$$\begin{aligned} \hat{\beta}_{A0} &\in \left[\log\left(\frac{n_{0A}}{n_{0B} + n_{0AB}} \right), \log\left(\frac{n_{0A} + n_{0AB}}{n_{0B}} \right) \right] \\ \hat{\beta}_{A} &\in \left[\log\left(\frac{n_{1A} \cdot (n_{0B} + n_{0AB})}{n_{0A} \cdot (n_{1B} + n_{1AB})} \right),\right. \\ &\qquad \left. \log\left(\frac{n_{0B} \cdot (n_{1A} + n_{1AB})}{n_{1B} \cdot (n_{0A} + n_{0AB})} \right) \right], \end{aligned} \tag{14}$$

assuming all expressions to be well-defined.

Example, version 2 (cont.): In terms of the regression coefficients, we obtain the estimates $\hat{\beta}_{<0} \in [-0.37,\ 0.59]$ and $\hat{\beta}_{<} \in [-1.83,\ -1.25]$.

Interpreting the indeterminate sign of intercept $\beta_{<0}$, one notes that for the group of persons that receives UBII (i.e. $X = 0$) the chance of being in the lower income group (< 1000€) in comparison to being in the higher income group (≥ 1000€) varies between $\exp(-0.37) = 0.69$ and $\exp(0.59) = 1.89$. In this way, one cannot judge the impact of the UBII on the dependent variable income without implying further assumptions about the coarsening. Unjustifiably ignoring the coarsening (see Section 5.2) pretends a particular sign of the regression coefficients. This corroborates the importance of including all imaginable coarsening mechanisms for obtaining a trustworthy result, which will be discussed now more in detail.

5 Reliable Incorporation of Auxiliary Information: Sensitivity Parameters and Partial Identification

The set-valued estimators from Expression (9) (and analogously from Expression (13)) are a typical application of the methodology of partial identification, emphasizing that only justified assumptions should be made which do not have to induce point identified parameters, but at least identify the parameter of interest in parts compared to the set of parameters that seemed to be possible in the beginning of the analysis (e.g., [19]). In this way, the trivial bounds [0, 1] on the probabilities have been refined substantially. In the spirit of partial identification and sensitivity analysis we can further refine the analysis if, and also only if, auxiliary information beyond the empirical evidence is available. Vansteelandt et al. [34] suggests to determine a sensitivity parameter δ in some range Δ under which the problem is identified and then to calculate the parameter of interest η for different values of the sensitivity parameter, where the whole region of the

[5] This corresponds to a product-multinomial sampling scheme (e.g. [31, 1]).

[6] Reminiscing about the derivation given here, we see that the categorical covariate case for the logistic model – in strict contrast to the continuous case (see Section 6) – in essence consists of a subgroup-specific consideration of the i.i.d. case.

resulting parameters of interest is called Ignorance Region $ir(\eta, \Delta)$ and the corresponding region of estimates Honestly Estimated Ignorance Region (HEIR) $\hat{ir}_n(\eta, \Delta)$. In order to account for statistical uncertainty due to finite sample size as well, in context of sensitivity analysis uncertainty regions are addressed that either can be constructed as covering the parameter of interest or the whole ignorance region with a probability of at least $(1 - \alpha)$ [13, 34].

To handle the inclusion of reliable information technically, we start with distinguishing and investigating point identifying additional assumptions, in order to utilize them as a technical means to derive sensitivity parameters, governing the incorporation of additional information.

Due to the fact that the imprecise point estimators in Expression (13) directly result from considering Expression (9) in a subgroup specific way, in Section 5.1 to Section 5.3 the detailed presentation is confined on the i.i.d. case. In Section 5.4, considering explicitly the regression model, another point-identifying assumption is suggested, where again the corresponding generalization may be used as a sensitivity parameter which allows the inclusion of partial knowledge.

5.1 Known Coarsening

If one or both coarsening parameters $q_{AB|A}$ and $q_{AB|B}$ are known (and different from 1), one can conclude directly that the corresponding mapping $\Phi(\cdot)$ from (8) is injective as in this case the parameter π_A can be uniquely related to the parameter p_A. Therefore, the set-valued estimator for π_A specified in Expression (9) can be shrunk to a single-valued estimator. The exact values of the coarsening parameters are most often unknown, but in case there is material information available that allows to bound them in non-trivial intervals, the consideration here gives a first way to perform a systematic sensitivity analysis. In most situations however such direct bounds will not be available. Therefore we look for alternative ways to introduce auxiliary knowledge.

5.2 Coarsening at Random (CAR)

If the coarsening is non-stochastic, the underlying degree of coarsening is predetermined and known. For instance, if respondents are requested to give their answer in a grouped way and we assume that all respondents answer correctly, then the coarsening is predefined in the sense that there is a unique coarsened outcome for every true answer. In the context of distinguishing between non-stochastic and stochastic coarsening mechanisms, Heitjan and Rubin [12] investigated under which properties the corresponding likelihood can be simplified to the so-called grouped likelihood and introduced the concept of *coarsening at random (CAR)*. This is a simplifying property requesting that the probability $q_{\mathscr{Y}|y}$ is constant, no matter which true value y is underlying as long as it fits to the observed value $\mathscr{y}$. Illustrated by the running example, CAR postulates that the probability of giving no suitable answer should not depend on the true income category, which contradicts practical experiences (e.g., [16]). In the dichotomous situation of this example we are then actually concerned with the assumption of missing at random (MAR) [18], which can be regarded as a special case of CAR.

Focusing again on the i.i.d. case, incorporating the CAR assumption of $q_{AB|A} = q_{AB|B}$ into the likelihood and in the observation model specifying $\Phi(\cdot)$, the situation simplifies substantially. Indeed, Φ is (almost) injective now, and we get the empirically point identified estimators, corresponding to having simply ignored the units with coarse values:

$$\begin{aligned} \hat{\pi}_A &= \frac{n_A}{n_A + n_B} \\ \hat{q}_{AB|A} &= \hat{q}_{AB|B} = \frac{n_{AB}}{n_A + n_B + n_{AB}} \,. \end{aligned}$$

There are ideal-type situations in which CAR can be justified indeed.[7] Nevertheless, this assumption must be treated with greatest care. Deviating from such an ideal-type situation and wrongly assuming CAR can lead to a bias of an extent that for sure destroys the relevance of the analysis, as is also illustrated in Figure 2. There the estimation of π_A under obstinately assumed CAR but varying coarsening probabilities is evaluated by the median relative empirical bias $\frac{\hat{\pi}_A - \pi_A}{\pi_A}$ based on 100 simulated datasets (here with $\pi_A = 0.6$).[8] The absolute value of the relative median bias increases the more one deviates from the case of CAR, indeed, up to a median relative bias of almost 80%.

5.3 Ratio of Coarsening Parameters

In our context the paper by Nordheim [22] obtains new importance. He considers the ratio between different mechanisms in the context of non-randomly missing and misclassified data. By fixing the ratio between the coarsening probabilities the corresponding maximum likelihood problem leads to quadratic equations, where

[7] For instance, rounding, type I censoring, which is present if the censoring times are fixed, and progressive type II censoring, which investigates censoring after the fixed d-th failure, in their pure form are CAR [15, 11].

[8] Thereby, in all addressed situations characterized by different true underlying coarsening mechanisms ($q_{AB|A}$ and $q_{AB|B}$ varying between 0.1 and 0.9 in equidistant breaks of 0.1, respectively), the assumption of CAR is involved into the estimation by plugging $q_{AB|A} = q_{AB|B}$ into the likelihood that is maximized.

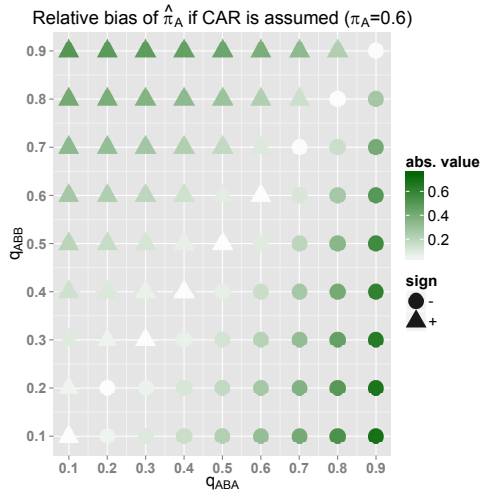


Figure 2: Consequences for the median relative bias of $\hat{\pi}_A$ if there is a deviation from assumed CAR.

one solution is contained in the interval of $\hat{\pi}_A$ from Expression (9), while the other solution lies outside of [0, 1] (cf. [22, p. 774]). Here we set $R = \frac{q_{B|B}}{q_{A|A}} = \frac{1-q_{AB|B}}{1-q_{AB|A}}$, slightly modifying the ratio of Nordheim by referring to the probabilities of the complementary events. Treating this ratio between the probabilities of precise observation fixed and including it into the likelihood in Section 4.1, unique, empirically point identified estimators are obtained as

$$\begin{aligned} \hat{\pi}_A &= \frac{n_A \cdot R}{n_B + n_A \cdot R}, \\ \hat{q}_{AB|A} &= \frac{n_B \cdot (R-1) + n_{AB} \cdot R}{n \cdot R} \end{aligned} \tag{15}$$

containing CAR as the special case $R = 1$. As in the case of CAR, the impact of assuming a wrong value of R has been investigated (results are available on request, see also [22]), where again a substantial bias can occur. The fact that there a similar variance of the estimators is obtained independently of the amount of deviation from the true value of R shows drastically that such deviations do not increase statistical uncertainty in the traditional sense and thus cannot be discovered by a traditional statistical analysis.

Because the parameter of interest π_A is identified given the typically unknown value of R, the ratio R can be used as a sensitivity parameter. In many cases it might be difficult to gain information about the exact value of R, but it seems quite realistic that a rough evaluation of the magnitude of R can be derived from material considerations, former studies or experiments. Thus, it is interesting to investigate the gain of information resulting from implying a factor R that is roughly known only, compared to the situation without any additional assumptions.[9] Considering the ratio R as a sensitivity parameter leads to the HEIRs.[10]

5.4 Subgroup Independent Coarsening

In the situation with covariates, there is apart from CAR, i.e. $\hat{q}_{AB|xA} = \hat{q}_{AB|xB}$, an alternative kind of uninformative coarsening, namely the independence of the underlying covariate value. Illustrated by the running example, imposing this kind of assumption means that answering in a coarse form, i.e., giving no suitable answer, does not depend on the receipt of unemployment benefit. As the receipt of unemployment benefit depends on the income, and the value of the income may influence the non-response to the income question (cf. Section 5.2), this assumption should be treated with particular caution here.

We will establish injectivity of the corresponding mapping $\Phi(\cdot)$ under an intuitive regularity condition and then, analogously to the procedure in Sections 5.2 and 5.3, this idea will be generalized in Section 5.5 by again considering the corresponding fraction as a sensitivity parameter. Imposing such *subgroup independent coarsening*

$$\begin{aligned} q_{AB|0A} &= q_{AB|1A} =: q_{AB|A} \\ q_{AB|0B} &= q_{AB|1B} =: q_{AB|B}, \end{aligned} \tag{16}$$

in the estimation problem of Section 4.2, the mapping $\Phi(\cdot)$ from Expression (12) is now injective[11] if restricted to the arguments $(\pi_{0A}, \pi_{1A}, q_{AB|A}, q_{AB|B})^T \in (0,1)^4$ such that

$$\pi_{0A} \notin \{0,1\},\ \pi_{1A} \notin \{0,1\} \ \underline{\text{and}}\ \pi_{0A} \neq \pi_{1A}. \tag{17}$$

One obtains the following unique estimators

$$\begin{aligned} \hat{\pi}_{0A} &= \frac{n_{0A}}{n_0} \frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}}, \\ \hat{\pi}_{1A} &= \frac{n_{1A}}{n_1} \frac{n_{1B}n_0 - n_1 n_{0B}}{n_{0A}n_{1B} - n_{0B}n_{1A}}, \\ \hat{q}_{AB|A} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{1B}n_0 - n_1 n_{0B}}, \\ \hat{q}_{AB|B} &= 1 - \frac{n_{0A}n_{1B} - n_{0B}n_{1A}}{n_{0A}n_1 - n_{1A}n_0}, \end{aligned} \tag{18}$$

[9] An example is given in the preliminary version of a technical report available at `http://www.statistik.lmu.de/~jplass/forschung.html`

[10] In more general cases of $|\Omega_Y| > 2$, the relations between the precise observation probabilities are not sufficient and relations concerning different coarsening mechanisms have to be known in order to obtain point identified estimators. More detailed information can be found in the preliminary version of a technical report cited in footnote 9.

[11] A proof of the injectivity of Φ in this situation is given in the preliminary version of a technical report cited in footnote 9. The case of $\pi_{0A} = \pi_{1A}$ reproduces the i.i.d. case, where there are multiple solutions.

when these are well-defined and inside the interval [0, 1]. Otherwise the maximum likelihood estimation is more challenging, but it can be shown that asymptotically ($n \to \infty$) the estimators of Expression (18) typically for all cases satisfying Expression (17) will be in [0, 1]. It has to be re-emphasized that in practical applications one must carefully reflect the plausibility of the subgroup independent coarsening assumption of Expression (16). In addition, the restrictions

$$p_{0A} \leq \frac{P(X=0) \cdot p_{1B} - p_{0B} \cdot P(X=1)}{p_{1B} - p_{0B} \cdot \frac{p_{1A}}{p_{0A}}} \leq 1 - p_{0B}$$

offer, at least under large sample sizes, a possibility to check whether the subgroup independent coarsening is appropriate at all.

5.5 A Generalization of Subgroup Independent Coarsening

There are situations in which one might have an idea about the relative magnitude of the probabilities of precise observations in both subgroups. For instance, knowledge from former studies could be available concerning the question whether respondents who do receive Unemployment Benefit II rather report their income class in a precise or a coarse way compared to the respondents that do not receive this benefit.

Analogously to the generalization of CAR in Section 5.3, we now generalize the assumption of subgroup independent coarsening by considering the ratio between the subgroup specific probabilities of precise observation, i.e., $R_1 = \frac{q_{A|1A}}{q_{A|0A}}$ and $R_2 = \frac{q_{B|1B}}{q_{B|0B}}$, where the case of $R_1 = R_2 = 1$ corresponds to assuming subgroup independent coarsening. As in Section 5.4, the mapping $\Phi(\cdot)$ from Expression (12) is injective for all cases in Expression (17) and thus unique estimators result.[12] Again, inclusion of partial knowledge is possible by regarding R_1 and R_2 as sensitivity parameters and considering all estimators resulting from incorporating a region of plausible values R_1 and R_2.

6 Concluding Remarks

We presented a maximum likelihood analysis of categorical data under epistemic data imprecision. Our approach working with possibly set-valued maximum likelihood estimators overcomes the dilemma of the precise probability based approaches, often damned to debilitate conclusions by the need to incorporate unjustified formal assumptions to ensure identifiability of parameters. The explicit reliance on an observation model specifying the coarsening process allows us to incorporate properly auxiliary information whenever it is present, in order to refine appropriately estimates derived from the empirical evidence alone.

The crucial arguments were developed, mutatis mutandis, for the i.i.d. case as well as a logistic regression based on one (or more) categorical covariates. From the applied point of view, an extension to metrical covariates is highly desirable. Although then a subgroup specific investigation is not possible any more, appropriate generalizations seem achievable in further work, especially when sensitivity parameters can be determined. However, to allow estimation of the underlying distribution from the data and to maintain the metric character, (partially) parametric modelling is needed. This implicitly restricts the set of distributions considered and in particular raises further issues in the understanding of statistical models as discussed, e.g., in [26, Sec. 3.1] for linear regression modelling.

In addition to this, the invariance property of the likelihood under different parametrizations, which is the technical basis of our results, offers two further directions of generalization. Further work may utilize these relationships beyond maximum likelihood estimation, in order to derive likelihood-based hypotheses tests and regions taking finite sample variability into account explicitly. These estimators also should be compared to confidence intervals derived along the lines of [34] when an appropriate sensitivity parameter could be determined.

Other areas of further research include a deeper investigation of the alternative generalized Bayesian (and possibilistic) approaches briefly mentioned in Section 3 as well as the consideration of other "deficiency" processes, most notably misclassification, which can be formalized in a very similar way. Our methodology thus also offers an alternative to, and a generalization to logistic regression of, recent work on misclassification from a partial identification perspective [20, 17].

Acknowledgements We are very grateful for the helpful remarks of three anonymous reviewers and appreciate their open sharing of stimulating ideas very much.

[12] They are given in the preliminary version of the technical report cited in footnote 9.

References

[1] A. Agresti. *Categorical Data Analysis*. 3rd edn., Wiley, 2013.

[2] T. Augustin, G. Walter, F. Coolen. Statistical inference. In: T. Augustin, F. Coolen, G. de Cooman, M. Troffaes (eds.): *Introduction to Imprecise Probabilities*, Wiley, 2014, pp. 135–189.

[3] A. Benavoli. Belief function and multivalued mapping robustness in statistical estimation. *Int. J. Approx. Reasoning*, 55:311–329, 2014.

[4] G. Casella, R. Berger. *Statistical Inference.* 2nd edn., Duxbury, 2002.

[5] M. Cattaneo, A. Wiencierz. Likelihood-based imprecise regression. *Int. J. Approx. Reasoning*, 53:1137–1154, 2012. [based on an ISIPTA '11 paper]

[6] G. de Cooman, M. Zaffalon. Updating beliefs with incomplete observations. *Artif. Intell.*, 159:75–125, 2004.

[7] I. Couso, D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reasoning*, 55:1502–1518, 2014.

[8] I. Couso, D. Dubois, L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Springer, Cham, 2014.

[9] T. Denoeux. Likelihood-based belief function: justification and some extensions to low-quality data. *Int. J. Approx. Reasoning*, 55:1535–1547, 2014.

[10] A. Dobra, S. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *P. Natl. Acad. Sci. USA*, 97: 11885–11892, 2000.

[11] D. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49:1099–1109, 1993.

[12] D. Heitjan, D. Rubin. Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253, 1991.

[13] G. Imbens, C. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857, 2004.

[14] T. Jiang, J. Dickey, Bayesian methods for categorical data under informative censoring, *Bayesian Anal.*, 3:541–553, 2008.

[15] J. Kalbfleisch, R. Prentice. *The Statistical Analysis of Failure Time Data.* 2nd edn., Wiley, 2002.

[16] A. Korinek, J. Mistiaen, M. Ravallion. Survey nonresponse and the distribution of income. *J. Econ. Inequal.*, 4:33–55, 2006.

[17] H. Küchenhoff, T. Augustin, A. Kunz. Partially identified prevalence estimation under misclassification using the kappa coefficient. *Int. J. Approx. Reasoning* 53:1168–1182, 2012. [based on an ISIPTA '11 paper]

[18] R. Little, D. Rubin, *Statistical Analysis with Missing Data.* 2nd edn., Wiley, 2002.

[19] C. Manski. *Partial Identification of Probability Distributions.* Springer, 2003.

[20] F. Molinari. Partial identification of probability distributions with misclassified data. *J. Econom.*, 144:81–117, 2008.

[21] H. Nguyen, B. Wu. Random and fuzzy sets in coarse data analysis. *Comput. Stat. Data. An.*, 51:70–85, 2006.

[22] E. Nordheim. Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Stat. Assoc.*, 79:772–780, 1984.

[23] D. Paulino, C. De B. Pereira. Bayesian analysis of categorical data informatively censored. *Commun. Stat., Theory Methods*, 21:2689–2705, 1992.

[24] J. Plass, P. Fink, N. Schöning, T. Augustin. Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision. *under revision for ISIPTA '15.* See also: *Techn. Rep., 179, Dep. Statistics, LMU Munich,* 2015 (url: www.epub.ub.uni-muenchen.de/23816).

[25] C. Rao. Maximum likelihood estimation for the multinomial distribution. *Indian J. Stat.*, 18:139–148, 1957.

[26] G. Schollmeyer, T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning*, 56:224–248, 2015. [based on an ISIPTA '13 paper]

[27] J. Stoye. Partial identification and robust treatment choice: An application to young offenders. *J. Statistical Theory and Practice*, 3:239–254, 2009.

[28] E. Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2:167–195, 2010.

[29] M. Trappmann, S. Gundert, C. Wenzig, D. Gebhardt. PASS: a household panel survey for research on unemployment and poverty. *Schmollers Jahrbuch*, 130:609–623, 2010.

[30] M. Troffaes, F. Coolen. Applying the imprecise Dirichlet model in cases with partial observations and dependencies in failure data. *Int. J. Approx. Reasoning*, 50:257–268, 2009.

[31] G. Tutz. *Regression for Categorical Data.* Cambridge University Press, 2011.

[32] L. Utkin, T. Augustin. Decision making under imperfect measurement using the imprecise Dirichlet model. *Int. J. Approx. Reasoning*, 44: 322–338, 2007. [based on an ISIPTA '05 paper]

[33] L. Utkin, F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (eds.), *ISIPTA '11*, pp. 371–380, 2011.

[34] S. Vansteelandt, E. Goetghebeur, M. Kenward, G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, 16:953–979, 2006.

[35] P. Walley. Inferences from multinomial data: Learning about a bag of marbles (with discussion). *J. R. Stat. Soc. B*, 58:3–57, 1996.

[36] M. Zaffalon, E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *J. Artif. Intell. Res.*, 34:757–821, 2009.

[37] Z. Zhang. Profile likelihood and incomplete data. *Int. Stat. Rev.*, 78:102–116, 2010.

Statistical Modelling in Surveys without Neglecting *The Undecided*: Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision

Julia Plass
Department of Statistics, LMU Munich
julia.plass@stat.uni-muenchen.de

Paul Fink
Department of Statistics, LMU Munich
paul.fink@stat.uni-muenchen.de

Norbert Schöning
Geschwister Scholl Institute of
Political Science, LMU Munich
norbert.schoening@gsi.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

Abstract

In surveys, and most notably in election polls, undecided participants frequently constitute subgroups of their own with specific individual characteristics. While traditional survey methods and corresponding statistical models are inherently damned to neglect this valuable information, an ontic random set view provides us with the full power of the whole statistical modelling framework. We elaborate this idea for a multinomial logistic regression model (which can be derived as a discrete choice model for voting behaviour) and an imprecise classification tree, and apply them as a prototypic illustration to the German Longitudinal Election Study 2013. Our results corroborate the importance of a sophisticated, random set-based modelling. Furthermore, by reinterpreting the undecided respondents' answers as disjunctive random sets, general forecasts based on interval-valued point estimators are calculated.

Keywords. Ontic data imprecision, survey methodology, election polls, multinomial logistic models, discrete choice models, imprecise classification trees, conjunctive random sets, disjunctive random sets, epistemic prediction, German Longitudinal Election Study 2013 (GLES 2013)

1 Introduction

Although pondering between several options is characteristic for human beings, indecisiveness of respondents is not reflected in most surveys. Instead it is common to force a precise answer, and at best to provide an additional category "Don't know" for those that are not decided. Frequently, in the framework of the analysis respondents reporting this "Don't know" category are no longer taken into consideration as those answers are understood as unusable. In many cases indecisive respondents are able to definitely exclude some options, which is not expressed by category "Don't know", and additionally characteristics of indecisive and decisive respondents may systematically differ. Consequently, the common proceeding leads to a substantial loss of information in data collection and biased results in the analysis of data.

In order to deal with this problem, it is necessary that questionnaire designers allow for multiple answers as "option A or option B" or at least provide ways to construct them. Hence, the preferences of the indecisive respondents are reflected in the most informative way and we are able to distinguish between different types of indecisive respondents. In this sense, we explicitly account for the heterogeneity within the group of indecisive respondents.

In order to embed this idea into a proper statistical modelling framework, we mainly will rely on the notion of *ontic sets* in the sense of Dubois and Prade ([15, 16]) as well as Dubois and Couso ([11]). They stressed the importance of differentiating between two views of a set, one representing precise collections of elements (*ontic view*) and the other reflecting incomplete knowledge about a particular precise value (*epistemic view*) ([12]). As answers of indecisive respondents are interpreted as ontic sets, we will call data that are coarse induced by indecision like "A or B" *data under ontic imprecision*.

Our paper is structured as follows. In Section 2 we will recapitulate some notions mainly based on random set theory ([19]) that have already been investigated in the framework of ontic sets ([11, 12]). In this context, we will emphasize the applicability of ontic sets to the general analysis in the presence of answers of indecisive respondents, where the focus will be on incorporating the idea of the ontic view into multinomial logistic regression analysis and classification trees in order to

model heterogeneity of respondents by their covariates. By briefly digressing into the epistemic view, in Section 3 interval-valued forecasts will be constructed. The aforementioned techniques are used in an illustrative analysis based on the German Longitudinal Election Study that is briefly presented in Section 4. Corresponding results are shown and compared to those obtained from classical statistical analyses in Section 5.

For sake of simplicity, we focus on categorical data of nominal scale, yet adaptation to ordinal scale for other applications may be derived only with little additional effort. Moreover, an extension to coarse categorical covariates under ontic data imprecision may be achieved with similar arguments.

2 Data under Ontic Imprecision: Basic Idea and Extending some Statistical Approaches

As argued in the introduction, it is crucial to distinguish between the ontic and epistemic view and thus between *random conjunctive sets* and *ill-known random variables* ([11, 12]). In this section we focus on *random conjunctive sets*, underlying the ontic view.

2.1 General Analysis

As we regard the case of categorical data with a finite state space, it is sufficient to focus on the definition of *finite random sets*, which can be considered as a simplification of the more general definition of random closed sets. A finite random set is a mapping $Z^* : \Omega \to \mathcal{P}(S)$ such that for any $A \subseteq S$ holds: $Z^{*-1}(\{A\}) = \{\omega \in \Omega : Z^*(\omega) = A\} \in \mathcal{A}$, where S denotes the state space, $\mathcal{P}$ the power set and $(\Omega, \mathcal{A})$ the underlying measurable space, equipped later with a probability measure P (e.g. [20]). In other words, a finite random set is characterized by a measurable mapping on the power set. Couso and Dubois call this notion *random conjunctive set* or *(ontic) set* ([11, 12]).

The important characteristic of an ontic set is that it represents a precise collection of elements in the sense that there is no true element of S underlying, but the set itself constitutes an entity of its own ([11]). Answers like "A or B" may be regarded as an ontic set $\{A,\ B\}$ as there is no unique preference. Therefore, the nature of coarse data under ontic imprecision is well represented by the ontic view. Consequently, this leads to a power set based view, meaning an extension of the classical precise state space S to $S^* = \mathcal{P}(S) \setminus \emptyset$, with the asterisk stressing ontic imprecision. Thus, basing the analysis on S^*, and therefore regarding coarse categories as own entities, provides the main idea of dealing with ontic imprecision. The one and only difference compared to the classical case is the adapted state space S^*.

Hence, by reinterpreting the random conjunctive set as precise random variable, classical probability theory and all statistical methods based on it are applicable. In other words, the idea of the adapted state space is independent of the statistical method and exploiting this idea further for formulating regression models and classification trees in the next sections should be regarded as an example.

A short example shall be given already here. It consists of calculating the probability of respondents, who are at least indecisive between particular options C_0, by the probability of the family of corresponding supersets $\mathcal{C} = \{T \subseteq S : C_0 \subseteq T\}$ to

$$P_{Z^*}(\mathcal{C}) = \sum_{C \in \mathcal{C}} P_{Z^*}(C)\ , \tag{1}$$

which is essentially a summation over singletons of the space S^* (cf. [11, p. 8]).

2.2 Regression Analysis

Generally, the main goal of regression analysis consists of modelling the relation between several covariates X and a dependent variable Y, without claiming to describe necessarily the causal impact of variables. In our case the dependent variable is assumed to be coarse under ontic imprecision, whereas we address precise covariates. As we restrict ourselves to a coarse categorical variable of nominal scale, a multinomial logit model is an appropriate statistical model.

2.2.1 Multinomial Logit Model

In this section it is mainly referred to [17, pp. 329-331]. A more thorough treatment of discrete choice models can be found for instance in [29]. We denote by $Y_i \in S = \{1, \ldots, c\}$ the random variable describing the response of individual $i = 1, \ldots, n$. Assuming a multinomial logit model, the probability of occurrence of category $s \in \{1, \ldots, c-1\}$ for i with given covariate values $\mathbf{x}_i$ is set to be

$$P(Y_i = s \,|\, \mathbf{x}_i) \ = \ \pi_{is} \ = \ \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s)}{1 + \sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)}\ , \tag{2}$$

with $\tilde{\mathbf{x}}_i^T = (1, \mathbf{x}_i^T)$ and category specific regression coefficients $\boldsymbol{\beta}_s = (\beta_{s0}, \beta_{s1}, \ldots, \beta_{sp})^T$ referring to p covariates. Because of the redundancy resulting from the fact that all probabilities add up to one, the corresponding probability for the so-called reference category c can

be determined by

$$P(Y_i = c \,|\, \mathbf{x}_i) \;=\; \pi_{ic} \;=\; 1 - \pi_{i1} - \ldots - \pi_{ic-1} \\ =\; \left(1 + \textstyle\sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)\right)^{-1} .$$

This corresponds to the side constraint that the regression coefficients of category c are set to zero.[1]

Expressing Equation (2) in terms of the linear predictor $\eta_{is} = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s$, one obtains the logarithmised chances and the relative risks of category $s \in \{1, \ldots, c-1\}$ and reference category c by

$$\log\left(\frac{\pi_{is}}{\pi_{ic}}\right) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s \quad \text{and} \quad \frac{\pi_{is}}{\pi_{ic}} = \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s) \; . \quad (3)$$

Accordingly, the exponential of β_{sj} $(j = 1, \ldots, p)$ expresses how the chance for category s compared to the reference category c changes if the value of a certain covariate x_j is increased by one unit in the case of metric covariates or if x_j is taken instead of reference category x_J in the case of categorical covariates.

2.2.2 A Multinomial Logit Model Based Approach under Ontic Imprecision

The redefinition of the original precise state space $S = \{1, \ldots, c\}$ of Y to the state space $S^* = \mathcal{P}(S) \setminus \emptyset$ of Y^* is crucial for adapting the multinomial logit model to account for ontic imprecision, treating answers of indecisive respondents as own categories, as already pointed out in Section 2.1.

Consequently, the number of categories of the dependent variable Y^* amounts to the cardinality of the new state space S^* $(m = |S^*| = |\mathcal{P}(S) \setminus \emptyset| = 2^{|S|} - 1)$. It formalizes the idea that no longer for each $Y \in \{1, \ldots, c\}$ but for each $Y_i^* \subseteq \{1, \ldots, c\}$ probabilities $\pi_{i1}^*, \ldots, \pi_{im}^*$ are modeled and coefficients $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{m-1}^*$ are estimated. Hence, the probability of occurrence of category $s \in \{1, \ldots, m-1\}$ for i with given covariate values $\mathbf{x}_i$ is determined by

$$P^*(Y_i^* = s \,|\, \mathbf{x}_i) \;=\; \pi_{is}^* \;=\; \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s^*)}{1 + \sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)}$$

and for reference category m by

$$P^*(Y_i^* = m \,|\, \mathbf{x}_i) \;=\; \pi_{im}^* \;=\; 1 - \pi_{i1}^* - \ldots - \pi_{im-1}^* \\ =\; \left(1 + \textstyle\sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)\right)^{-1} .$$

[1] In order to ensure identifiability it is important to include a side constraint for the regression coefficients into the basic model. Alternatively, any other category may be chosen as reference category or a symmetric type of constraint like $\sum_{r=1}^{c} \beta_r^T = (0, \ldots, 0)^T$ can be applied (e.g. [30]).

In this way, one obtains own regression coefficients for each coarse category, which exactly reflects the underlying idea that different types of indecisive respondents are regarded as own group.

In summary, one can account for ontic imprecision within categorical variable Y of nominal scale by incorporating coarse answers as own categories into a multinomial logit model. Apart from the up to exponential increase in the number of categories nothing changes: All statistical methods refining and extending the classical multinomial logit model, like penalization approaches, flexible covariate modelling or random effects under repeated measurements (e.g. [30]), and their fundamental statistical properties, like consistency and asymptotic normality of estimators, can be transferred. In this way, the here addressed adaptation of the multinomial logit model serves as an example for incorporating the power set based idea into categorical regression models.

2.3 Classification Trees

Whereas in regression we are mainly interested in the estimation of the regression coefficients, which provide a structural interpretation of the data, in the framework of classification trees one major goal is to predict the value(s) of a dependent variable (called *class variable* Y later on) of a future observation, based on values of some independent, so-called feature, variables. Learning a classification tree involves recursively partitioning the full data space as it is available in the beginning, into disjoint subspaces by splitting with respect to some (in-)homogeneity criterion. A most favourable property of a single classification tree from a statistical modelling point of view is that it still allows a structural interpretation, while such is lacking in the even more prediction orientated ensemble of trees, so-called bags or forests.

In the framework of classification trees there are numerous algorithms available that are able to deal both with nominal and numerical variables, some even account for missingness at random, for instance Quinlan's ID3 [23] and Breiman's CART [9] and their successors. They share the concept of selecting splitting feature variables performing the partitioning by a similarity measure, in our context the entropy. For sake of simplicity we confine ourselves to class and feature variables of nominal scale.

In order to calculate the entropy and decide on a splitting feature variable, it is required to estimate the class' probabilities, classically achieved by the corresponding relative frequency. Abellan and Moral [4] introduced *imprecise classification trees* by changing the estimation to involve imprecise probability mod-

els. As a split criterion they favoured a maximum entropy approach and presented in [4] an adaptation of Quinlan's ID3 algorithm, both of which for sake of simplicity we employ.

Yet there are more general approaches, where for instance the full entropy range is taken into account, as in [18] or [13], the latter naturally growing a forest. Further improvements of the initial imprecise algorithm also include the concept of bagging [2, 3].

In our analyses in Section 5.3 we grow classification trees accordingly to [4] but relying on a Nonparametric Predictive Inference (NPI) model for estimation of the class probability distribution within a node instead, yet an Imprecise Dirichlet Model would have been also applicable; see [10] for a more detailed introduction to NPI for categorical data and [4] or [18] for a description on how an imprecise classification tree based on it is actually constructed. Yet, we briefly recall the estimation with NPI within a tree's node.

Each node of the tree consists of a collection of observations. They are assigned to nodes in such a way that they form the aforementioned disjoint subspaces in an optimal way with respect to the splitting criterion. In the context of an entropy based splitting criterion the probability distribution of the class variable is required. In [4] the assumption of a precise probability distribution is relaxed to a credal set leading to a maximum entropy split criterion approach. According to NPI the predictive probability that for a virtual next observation the class variable attains a value y_i of its state space is within the following interval

$$P(Y = y_i) \in \left[\max\left(0, \frac{n_i - 1}{n}\right), \min\left(\frac{n_i + 1}{n}, 1\right) \right], \quad (4)$$

with n_i the number of observations having a class value of y_i and n the overall number of observations, both with respect to the node under consideration.

In the situation where the class variable is only observable under ontic imprecision, we embed ontic sets into the framework of classification trees properly by a redefinition of the class variable as a finite random set, thus basing the analysis on the power set of the class variable space, similarly to the regression analysis. This is a direct implementation of the crucial idea, allowing us to reinterpret the ontic sets as a new precise class variable, i.e. an answer "A or B" is interpreted now as the precise class "AB". Therefore, any classification tree technique might be applied that is able to deal with a precise classification variable, regardless of the underlying probability model(s). This power set based technique is frequently applied in the framework of multi-label classification (e.g. MODEL–n in [8]). Due to the increased number of classes the concept of entropy correction ([27]) becomes more important, besides substituting Y by Y^* in (4).

Furthermore, basically any classification technique may be applied, after the state space of the variables under ontic uncertainty is substituted by its power set. The classification trees serve as a feasible example.

3 Interval-valued Forecast

We consider the same data situation, but change our perspective and the aim of our analyses. Instead of modelling the underlying structure of voting (in)decisions, we now turn to forecasts based on an epistemic reinterpretation of our data.

Let's assume that our main interest lies now in forecasting certain events by enforcing a final decision expressed by a variable Y_{final}. In the context of voting behaviour such a situation arises when a forecast on the election result is required. Under the assumption that the final decision is precise and consistent with the data collected now, this means a precise true value is underlying the set-valued response.

In this way, set-valued elements A^* of S^* are no longer interpreted as own entities, but are regarded as incomplete knowledge, which for every event B from the space $(S, \mathcal{P}(S))$ is given by (cf. [7, p.185])

$$P(Y_{\text{final}} \in B \,|\, Y^* = A^*) \in \begin{cases} \{0\}, & \text{if } B \cap A^* = \emptyset \\ \{1\}, & \text{if } B \supseteq A^* \} \\ [0,1], & \text{otherwise} \end{cases},$$

postulating that the final answer is compatible with the initial information from the ontic view.

This corresponds to an epistemic view of modelling[2]. However, models should be cautiously interpreted as the data were originally obtained under ontic imprecision, yet it may be justified for modelling purpose.

In the context of the epistemic view Couso and Dubois ([11]) consider *ill-known random variables* Y_{epist} with precise, but incomplete realizations y_{epist}. An *ill-known random variable* Y_{epist} is a multiple-valued mapping $Y_{\text{epist}} : \Omega \to \mathcal{P}(S)$ described by the disjunctive set of mappings

$$\left\{ Y_{\text{precise}} : \; Y_{\text{precise}}(\omega) \in Y_{\text{epist}}(\omega) \;, \forall\, \omega \in \Omega \right\},$$

where $Y_{\text{precise}} : \Omega \to S$ is a precise random variable. Thus, Y_{epist} is interpreted as the collection of several precise models that can be deduced from incomplete knowledge.

[2] First steps towards statistical modelling under epistemic data imprecision can be found in ([21]).

Taking the reinterpretation as disjunctive sets seriously, the range covering the true probability of a certain event of interest E can be expressed by Dempster's lower and upper probabilities ([14]) that are

$$\underline{P}_{Y_{\text{epist}}}(E) = \sum_{Y_{\text{epist}}(\omega)\subseteq E} p(\omega) ,$$
$$\overline{P}_{Y_{\text{epist}}}(E) = \sum_{Y_{\text{epist}}(\omega)\cap E\neq\emptyset} p(\omega) ,$$

where p is the probability mass function of P ([11]).

Thus, the proportion of an option E can be forecasted by the sample counterparts $\widehat{I}(E)$ of the interval

$$I(E) = \left[\underline{P}_{Y_{\text{epist}}}(E) , \ \overline{P}_{Y_{\text{epist}}}(E) \right] . \qquad (5)$$

As the difference between the values of the lower and the upper probability represents the lack of knowledge induced by indecisive answers, it is apparent that the length of this interval can be interpreted as the extent of the underlying epistemic imprecision.

In order to account additionally for statistical uncertainty due to finite sampling, confidence intervals for $I(E)$ may be calculated. This leads to so-called uncertainty regions aiming to cover both: imprecision due to incompleteness and statistical uncertainty ([31]).

4 Data

Until now the German Longitudinal Election Study (GLES) ([25]) is the most elaborated German electoral poll and currently focuses on three federal elections (2009, 2013, 2017). The sampling method of the initial data set of the *GLES* 2013 is a (3-step) random sample, which is treated here in our illustrative analysis as a simple random sample. As voting intentions before the election day are of main interest, we consider the preliminary study of *GLES 2013*, which is a face-to-face interview two months prior to the election day.

To our present knowledge there is not any pre-election study allowing indecisive respondents to express their voting intention by multiple answers. The main advantage of *GLES 2013* is that respondents are also explicitly required to report their voting intention's certainty ("certainty")[3] along with the assessments of several parties (*q21a-q21h*[4]). Those and the respondent's current voting intention[5], collected in a precise answer, allow us the construction of a variable "ontic", reflecting the respondent's indecision by multiple answers. The procedure for our construction of the variable "ontic" is as follows: While for all "very certain" respondents the reported party of the variable "vote" is taken, the party or parties with maximal assessment are chosen for the respondents that are "fairly certain" explicitly allowing by construction indecision between the corresponding parties. For the respondents that decide for "neither/nor" or "not certain at all" parties with maximal and second highest assessments are taken.The chosen way of construction of the variable "ontic" is to some extent arbitrary, but at least it accounts reasonably for ontic imprecision. In the following we focus on the second vote, as similar steps and explanations hold for the first vote as well.

	case 13	case 126	case 1515
certainty	very certain	fairly certain	neither/ nor
vote	GREEN	SPD	CD
assessCD	-1	-1	+3
assessSPD	+2	+1	+3
assessFDP	-4	0	0
assessLEFT	-4	+1	-5
assessGREEN	+4	-3	+2
	⇓	⇓	⇓
ontic	GREEN	LEFT:SPD	CD:GREEN:SPD

Table 1: Construction of variable "ontic" (example)

The examples in Table 1 illustrate the way of construction by means of three randomly chosen respondents.[6]As our goal consists of demonstrating the difference in results from an analysis including ontic imprecision and a classical analysis, such a constructed variable is required.

Partly due to the construction of variable "ontic" several respondents had to be excluded[7]. All conducted filtering steps (e.g. excluding voters of smaller parties or non-voters) that reduced the sample of initially 2003 to 1196 respondents can be found in [22]. The associated loss of information caused by the reduced

[3] *q13* with categories "very certain", "fairly certain", "neither/nor" and "not certain at all"

[4] Each measured on a scale from "-5" ("a very negative view of this political party") to "+5" ("a very positive view of this political party")

[5] The German election system mixes elements of election by proportionality and by majority. The voters have two votes (*q11ab*: second vote, *q11aa*: first vote). The second vote is generally considered as more important, because the proportion of seats in the German Bundestag mainly is allocated according to the second vote. The first vote determines the direct representative of an election district in the Bundestag.

[6] Translations of German abbreviations of political parties are used here. Considered parties are: *Christlich Demokratische Union Deutschlands* (CDU) and *Christlich-Soziale Union in Bayern* (CSU) representing throughout Germany one option only (here denoted by CD), *Sozialdemokratische Partei Deutschlands* (SPD), *Die Linke* (LEFT), *Bündnis 90/Die Grünen* (GREEN), *Freie Demokratische Partei* (FDP).

[7] In voting studies sample loss is rather common. Usually empirical analyses are reduced to those parties, who entered the German Bundestag finally (e.g. [28]).

CD	SPD	GREEN
495	271	125
LEFT	FDP	GREEN:SPD
106	39	36
CD:SPD	CD:FDP	GREEN:LEFT
35	18	15
LEFT:SPD	CD:GREEN:SPD	GREEN:LEFT:SPD
14	17	13
CD:FDP:SPD		
12		

Table 2: Absolute frequencies of constructed variable "ontic" (second vote)

sample size is undesirable, but unavoidable for an ontic analysis illustrated by this data set. Because of the underrepresentation of indecisive persons induced by the current design of the questionnaire, which implicitly excludes indecisive respondents by the preceding filtering of the "certainty" item (cf. [22]), we expect less marked differences between an ontic and a classical analysis, described in the following sections.

The resulting illustrative data set containing variable "ontic", whose absolute frequencies are given in Table 2, forms the basis of the following analysis.[8]

5 Data Analysis

The principal goal consists of comparing the results obtained by an analysis using the constructed variable "ontic" (cf. Section 4 and [22]) to a classical analysis excluding all uncertain respondents. This issue will be considered in this section with regard to the findings from Section 2. Hereby, we focus on the second vote, only where mentioned explicitly the first vote is considered. All analyses are based on complete cases, dependent on the variables effectively under consideration. We performed our analyses with the open-source statistical software R [24]. The code is available on request from the authors.

5.1 General Analysis

The analysis incorporating ontic imprecision is based on $S^* = \mathcal{P}(S) \setminus \emptyset$, where

$$S = \{\text{CD}, \text{SPD}, \text{GREEN}, \text{LEFT}, \text{FDP}\}$$

is the state space. Since only 13 elements of S^* are attained in the addressed data set, we adapted S^* to cover those values of variable "ontic" only (see Table 2).

If for instance the probability of respondents is of interest that are (at least) indecisive between party "SPD" and "GREEN", according to Equation (1) all probabilities referring to respondents that are (at least) indecisive between both parties have to be summed up, which can be estimated by associated relative frequencies to

$$\begin{aligned}
&\widehat{P}_{Z^*}\big(Z^* \supseteq \{\text{GREEN, SPD}\}\big) \\
&= \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{GREEN, SPD}\}\big\}\Big) \\
&\quad + \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{CD, GREEN, SPD}\}\big\}\Big) \\
&\quad + \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{GREEN, LEFT, SPD}\}\big\}\Big) \\
&= \frac{36}{1196} + \frac{17}{1196} + \frac{13}{1196} \;\approx\; 0.06\,.
\end{aligned}$$

The estimated proportion of indecisive respondents is 0.13, calculated analogously. Consequently, if just decisive respondents are considered an amount of 13% of respondents are not taken into account. As respondents are excluded because of the value of the variable of interest itself, we are concerned with a *not missing at random* situation and thus ignoring the indecisive respondents may lead to biased results. This is particularly fatal for a theoretical understanding of voting decisions as well as from a practical campaigners' view, because this percentage covers those respondents that are of particular interest.

5.2 Regression Analysis

In order to analyse the heterogeneity within the coarse dependent variable Y under ontic data imprecision, the models presented in Section 2.2 are applied. The multinomial logit model has a longstanding tradition in the context of modelling voting behaviour[9].

In our analysis the variable "ontic" represents the coarse dependent variable, where "SPD" is chosen as reference category. Generally, it is important to choose all reference categories in such a way that interpretations enable answering the question of interest. For our illustrative purpose we use a very simple voting model with only two covariates[10], namely socio-demographical variable "religious denomination" (*q228*) as well as variable "most important source of information" (*q97*). In both variables certain categories were aggregated. Thus, variable "religious denomination" here only takes values "Christian" and "non-Christian", where the categories of "most important

[8] Absolute frequencies of singletons differ from those of variable "vote" due to the construction of variable "ontic".

[9] Actually, the multinomial logit model is the simplest model of the discrete choice family. Although it has several disadvantages for the modelling of voting behaviour as discussed by [6], for the sake of our illustrative application yet the multinomial logit model is appropriate, because it shows the basic concept in handling data under ontic imprecision, which can be extended analogously to more tailored models.

[10] Recent models of voting behaviour use policy distance, party identification and socio-demographical variables and yield a remarkable fit and prognostic validity (cf. [5])

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.37	−1.47***	0.13
rel.christ	0.32*	−0.05	0.49***
info.tv	0.01	−0.29	−0.01
info.np	−0.05	−1.67**	−0.01

Table 3: Comparison of results (second vote).

source of information" are translated to "television", "newspaper" and "other source", the latter also covering "radio", "internet" and "talking to other people". Every reclassification is subject to avoid categories with only few observations in order to decrease statistical uncertainty. By including "most important source of information" as a covariate into the model, we assume that the way how voters inform themselves of the federal election influences their voting intention. Nevertheless, one cannot exclude an opposite (causal) direction as respondents who vote for particular parties potentially avoid or prefer certain information sources because of the way this party is represented in it. This needs to be kept in mind when interpreting the model's results.

For reasons of conciseness estimated regression coefficients are shown just for category "CD" and "GREEN:SPD" (G:S) here.[11] With $n_{\text{CD}} = 508$ and $n_{\text{G:S}} = 36$ they form the largest groups of decisive and indecisive respondents, respectively, such that the interpretation of corresponding regression coefficients is comparably trustworthy. Especially in the context of estimators for indecisive groups, we remark that some of the regression coefficients' calculations are based on few observations, and thus corresponding interpretations have to be treated cautiously.

Furthermore, in context of interpretation one should check by taking the statistical significance[12] into account whether the regression coefficients vary just randomly. The small sample size within several groups of variable "ontic" may be responsible for non-significant estimators. Thus, from an increase in sample size statistical uncertainty is reduced and potentially significant results can be obtained.

Considering the results of the second vote analysis presented in Table 3 (ontic)[13], for Christian respondents the probability of electing "CD" instead of "SPD" is increased by the multiplicative factor $\exp(0.32) = 1.38$ compared to non-Christian respondents under the ceteris paribus assumption of unchanged other covariates.[14] Furthermore, regression coefficients closely to zero indicate that no influence of covariate "most important information source" on the probability of electing "CD" in comparison to the reference category "SPD" may be verified.

The crucial property of the multinomial regression under ontic imprecision consists of estimating own coefficients for the different indecisive groups. For instance, for respondents reporting "newspaper" as their most important information source in comparison to those naming another information source the probability of being indecisive between the two parties "GREEN" and "SPD" instead of voting for "SPD" is decreased by the factor $\exp(-1.67) = 0.19$ on the ceteris paribus premise. Likewise investigations are important for election campaigners to adjust their strategies adequately, as they show how potential voters differ from the core voters of a party (as here "SPD") in the choice of their favourable information source.

Results from a classical analysis that chooses variable "vote" as response variable and takes only those respondents into consideration that are "very certain" or "certain" may be found in Table 3 as well, again just displaying coefficients for "CD".

Comparing results from both analyses, estimators of similar magnitude are obtained throughout. In this way, the classical and the generalized approach reflecting ontic imprecision do not contradict each other.

The importance of our ontic set based modelling is corroborated even stronger when we consider the first vote instead. Now the analyses reveal remarkable differences partly associated with a change in sign. Thus, some covariates have an amplifying effect on the dependent variable in one analysis, while in the other analysis a weakening effect is underlying (cf. Table 4), yet those are not statistically significant.

Although the complete case analysis and the carried out filtering steps mainly induced by the questionnaire design led to a further decrease in the number of indecisive respondents, this illustrative analysis already shows striking differences between both analyses. Because of the here provided proof of concept for an ontic analysis, it is strongly suggested to include the option of reporting multiple answers such that those can be

[11] Estimated regression coefficients for the other categories may be found in [22]

[12] "***", "**" and "*" denotes statistical significance of level $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$, respectively.

[13] Covariates "religious denomination" and "most important information source" are dummy coded with "non-Christian" and "other source" as reference category, respectively. The estimates quantify the difference between the group under consideration and the reference category (rel.christ: "religious denomination" is "Christian"; info.tv, info.np: "most important information source" is television, newspaper, respectively).

[14] Despite the name "CD" and the above results indicating a strong Christian relation, nowadays the "CD" parties understand themselves as a general conservative party with members and supporters regardless their religious affiliation.

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.33	−1.41 **	−0.12
rel.christ	0.37 **	−0.25	0.52 ***
info.tv	−0.02	−0.32	0.25
info.np	−0.12	−1.69 **	0.13

Table 4: Comparison of results (first vote).

	ontic	vote	classical
Scenario 1	0.407 (0.040)	0.425 (0.050)	0.446 (0.041)
Scenario 2	0.704 (0.026)	0.796 (0.031)	0.817 (0.042)

Table 5: Correct classification rate (standard deviation) for second vote based on 10-fold cross-validation

included into the analysis in an appropriate way. In cases of large data sets with numerous indecisive respondents, we even expect increased differences in the estimation of regression coefficients.

5.3 Classification Trees Analysis

In a first scenario the settings are the same as we explored in the regression analysis, thus considering "ontic" coarse class variable and "religious denomination" and "most important source of information" as split feature variables, in the same scaling as previously in section 5.2 (Scenario 1). We are considering this setting to retain direct comparability with the regression analysis, yet we are aware that a classification tree's ability lies in reducing the sample space by discovering few favourable independent variables out of a potentially huge number of candidates. Therefore, we are not expecting an outstanding performance in this scenario. As discussed above we decided in favour of a Nonparametric Predictive Inference model as underlying (imprecise) model of the classification tree. We choose the most frequent class as prediction rule in the leaves, thus enforcing a precise result. Furthermore, we grew imprecise classification trees on the data set neglecting the undecided, but in this case we chose "vote" as the dependent variable as a counter part to the classical regression analysis. In order to assess the predictive ability of the trees a 10-fold cross-validation each was performed.

The results are to be found in the first row of Table 5, with respect to the second vote. For a fair comparison we measure the accuracy for both data situations by the correct classification rate (columns *ontic* and *classical*), and furthermore in case of the ontic data sets we checked the prediction result of "ontic" against "vote" (column *vote*). Any value of "vote" which was contained in the predicted coarse category was considered correctly classified. Furthermore the standard deviation is reported.

As it is clearly visible the predictive ability of the imprecise trees is unsurprisingly poor, and an inspection of the underlying trees reveals the culprits. The selection of the independent variables only allows growing of 13 different trees, which only in case of a strong dependency between the independent and depend variables leads to reasonable accuracy results. Furthermore when looking at the relative class frequencies in the root nodes, the category of "CD" is with over 40% by far the most observed one. While the construction of most trees involved at least one split, category "CD" is still predicted in a vast majority of the tree's leaves, in few cases even in all.

In further analyses, we incorporated more independent variables, allowing a higher variation in potential trees (Scenario 2). Further splitting candidate variables were the party identification (*q119*), the person's social stratum (*q192*), the sex (*q1*), general political interest (*q3*) and the personal economic situation (*q17*). With those and the previous variables the same analysing steps were repeated, but now with the accuracy nearly doubling in either scenario as the second row of Table 5 indicates. Especially the party identification has a high influence.

Similar prediction results as above are obtained when considering the first vote, instead of the second, displayed in [22]. Quite interestingly, the correct classification rate is lower when we are predicting the "ontic" variable than in the case when predicting "vote". In the second scenario there is a notable gap of around 10%, which is mainly caused by an ontic coarse class prediction, whereas vote is (naturally) precise.

In both scenarios the classical procedure of omitting the undecided persons leads to better results, when just considering the predictive ability, yet with the help of our ontic view we are able to identify hard to classify respondents.

A major reason for the small differences between the classical and ontic analyses is the comparably little percentage of undecided persons (less than 10% within the data under consideration). As mentioned in the discussion in the regression analyses, this is partly due to the conducted complete case analysis and the construction of variable "ontic", but more gravely imposed by the design of the questionnaire. When allowing for multiple answers directly in variable "vote", we expect an increase in the accuracy of the ontic prediction, as the number of hard to precisely classify, indecisive persons raises.

5.4 Interval-valued Forecast

In Section 3 the epistemic view has been used in order to calculate interval-valued forecast $I(E)$, which will be illustrated in this section.

For instance, if one is interested in the forecasted proportion of respondents electing "CD", by referring to the absolute frequencies of variable "ontic" in Table 2 and to Equation (5), the interval-valued forecast

$$\widehat{I}\big(\{\text{CD}\}\big) \;=\; \left[\, \frac{495}{1196} \,,\; \frac{495+35+18+17+12}{1196} \,\right]$$

is obtained. All fractions that are included in the lower bound refer to respondents who vote for the "CD" party for sure while all fractions that are used within the calculation of the upper bound concern respondents who generally could imagine to vote for it. Political studies gradually proceed to calculate the fraction of "potential voters" which corresponds to the upper bound of interval $\widehat{I}(E)$ (cf. [1]).

Nevertheless, forecasts are commonly based on respondents that are characterized by a high degree of certainty concerning their voting intention only. In our data example there are $n = 1096$ respondents that are "very certain" or "fairly certain" according to their voting intention, where 490 of those intend to vote for "CD" and thus the naive estimated forecasting probability results in

$$\widehat{P}_{\text{naive}}\big(\{\text{CD}\}\big) \;=\; \frac{490}{1096}\,.$$

As indecisive voters may systematically differ from respondents that are sure of their voting intention, the proportion in terms of interval $\widehat{I}(E)$ contains valuable information that is not expressed by $\widehat{P}_{\text{naive}}(E)$. Because of the difference between these groups it is important to treat results ignoring indecisive respondents with caution.

In practice forecasting the proportion of a set containing more than one element is of considerable relevance: Frequently, for instance in Germany, the main interest is the voters' percentage not just for a particular single party, but for a coalition. In this context the interval-valued forecast $\widehat{I}(E)$ becomes of particular interest, as respondents that are indecisive between the parties contained in the coalition of interest E are incorporated for sure. Thus, these coarse observations constitute a precise vote for the coalition (e.g. [22]).

6 Concluding Remarks

While currently data under ontic imprecision are still neglected in statistical analysis, they could prove a valuable source of information. Especially in context of election studies incorporating the different types of "The Undecided" into statistical analyses becomes increasingly important as more and more voters decide shortly before the election day (cf., e.g. [26]). Once the practitioner changes the state space, the statistical methods remain the same, as we could demonstrate. Even as the group was comparably small and we were forced to assess indecisiveness indirectly by constructing an ontic variable, we corroborated in our data example that including the undecided respondents did make a difference. Therefore, as now appropriate statistical methodology has been proven to be available, we strongly recommend allowing for multiple answers directly within questionnaires.

As the underlying idea is somewhat generic, the in here presented analyses by a multinomial regression model and imprecise classification trees are just the tip of the iceberg. One may think of more complex methods to study the data set, mutatis mutandis. For simplicity we restricted ourselves to the case of a nominal scale of the variable under ontic imprecision, yet the adaptation to an ordinal scale is achievable with little additional effort as well. In further studies it is worth considering not only the dependent variable under ontic imprecision but also the covariates. In principle, this is achievable by involving the power-set based idea again.

Acknowledgements

We are grateful to two of three anonymous reviewers for their very helpful remarks, also stimulating further research.

References

[1] Großteil der Wähler würde sich noch umstimmen lassen. *Süddeutsche Zeitung*, 16 August 2013. Accessed 24 January 2015, `http://www.sueddeutsche.de/politik/umfrage-zur-bundestagswahl-die-meisten-waehler-wuerden-sich-noch-umstimmen-lassen-1.1747539`.

[2] J. Abellán and A. Masegosa. Bagging decision trees on data sets with classification noise. In S. Link and H. Prade, editors, *Foundations of Information and Knowledge Systems*, pages 248–265. Springer Berlin Heidelberg, 2010.

[3] J. Abellán and A. Masegosa. An ensemble method of using credal decision trees. *European Journal of Operations Research*, 205(1):218–226, 2010.

[4] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[5] J. Adams, S. Merrill, and B. Grofman. *A Unified Theory of Party Competition: A Cross-National Analysis Integrating Spatial and Behavioral Factors*. Cambridge University Press, Cambridge, 2005.

[6] R. Alvarez and J. Nagler. When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42(1):55–96, 1998.

[7] T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014.

[8] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Books, Monterey, CA, 1984.

[10] F. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.

[11] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.

[12] I. Couso, D. Dubois, and L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Springer, Cham, 2014.

[13] R. Crossman, J. Abellán, T. Augustin, and F. Coolen. Building imprecise classification trees with entropy ranges. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 129–138, Innsbruck, 2011. SIPTA.

[14] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[15] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

[16] D. Dubois and H. Prade. Formal representations of uncertainty. In D. Bouyssou, D. Dubois, M. Pirlot, and H. Prade, editors, *Decision-Making Process: Concepts and Methods*, pages 85–156. ISTE & Wiley, London, 2009.

[17] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer, Berlin, 2013.

[18] P. Fink and R. Crossman. Entropy based classification trees. In F. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pages 139–147, Compiègne, 2013. SIPTA.

[19] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[20] H. Nguyen. *An Introduction to Random Sets*. CRC Press, Boca Raton, Florida, 2006.

[21] J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Towards statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression under coarse categorical data. Under revision for ISIPTA '15, preprint temporary available at `http://www.statistik.lmu.de/~jplass/forschung.html` (20.03.2015).

[22] J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical Modelling in Surveys without Neglecting "The Undecided": Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision - extended version. Technical Report 179, University of Munich, Department of Statistics, 2015. `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-23816-6`.

[23] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[25] H. Rattinger, S. Roßteutscher, R. Schmitt-Beck, B. Weßels, and C. Wolf. Vorwahl-Querschnitt (GLES 2013), 2014. GESIS Datenarchiv, Köln. ZA5700 Datenfile Version 2.0.0, Accessible from `https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5700&tab=3&ll=10&notabs=&af=&nf=1&search=gles&search2=&db=e`.

[26] O. Schirg. Wahlforscher: Jeder Dritte ist noch unentschlossen. *Die Welt*, 10 August 2001. Accessed 22 January 2015, `http://www.welt.de/print-welt/article467015/Wahlforscher-Jeder-Dritte-ist-noch-unentschlossen.html`.

[27] C. Strobl. Variable selection in classification trees based on imprecise probabilities. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 339–348, Carnegie Mellon University, Pittsburgh, 2005. SIPTA.

[28] P. Thurner. The empirical application of the spatial theory of voting in multiparty systems with random utility models. *Electoral Studies*, 19(4):493–517, 2000.

[29] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[30] G. Tutz. *Regression for Categorical Data*. Cambridge University Press, 2011.

[31] S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953–979, 2006.

A Logic with Upper and Lower Probability Operators

Nenad Savić
Faculty of Technical Sciences
University of Novi Sad, Serbia
nsavic@uns.ac.rs

Dragan Doder
Computer Science and Communications
University of Luxembourg
dragan.doder@uni.lu

Zoran Ognjanović
Mathematical Institute
Serbian Academy of Sciences and Arts
zorano@mi.sanu.ac.rs

Abstract

We present a propositional logic with unary operators that speak about upper and lower probabilities. We describe the corresponding class of models and discuss decidability issues. We provide an infinitary axiomatization for the logic and we prove that the axiomatization is sound and strongly complete. For some restrictions of the logic we provide finitary axiomatic systems.

Keywords. Probabilistic Logic, Upper and Lower Probabilities, Axiomatization, Completeness theorem.

1 Introduction

During the last few decades, uncertain reasoning has emerged as one of main fields in computer science and artificial intelligence. Many different tools are developed for representing, and reasoning with, uncertain knowledge. One particular line of research concerns the formalization in terms of probabilistic logic. After Nilsson [24] gave a procedure for probabilistic entailment which, given probabilities of premises, calculates bounds on the probabilities of the derived sentences, researchers from the field started investigation about formal systems for probabilistic reasoning [6, 7, 8, 9, 10, 13, 21, 25, 26].

However, in many applications, sharp numerical probabilities appear too simple for modelling uncertainty. This calls for developing different imprecise probability models [4, 5, 19, 22, 28, 29, 30, 32]. In order to model some situations of interest, some approaches use sets of probability measures instead of one fixed measure, and the uncertainty is represented by two boundaries – lower and upper probabilities [12, 18]. Consider the following example, essentially taken from [11].

Example 1 *Suppose that a bag contains 10 marbles and we know that 4 of them are red, and the remaining 6 are either black or green, but we do not know the exact proportion (for example, it is possible that there are no green marbles at all). The goal is to model a situation where the person picks a marble from the bag randomly. The cases when person picks up a red marble (red event), when person picks up a black marble (black event) and when person picks up a green marble (green event) will be denoted by R, B and G, respectively. Clearly, the probability of the red event is* 0.4*, but we cannot assign strict probability to black or green event. Therefore, we use the set of probability measures* $P = \{\mu_\alpha \mid \alpha \in [0, 0.6]\}$*, where* μ_α *assigns* 0.4 *probability to red event,* α *to black event, and* $0.6 - \alpha$ *to green event. We assign two functions to arbitrary set of probability measures ,P, first one is* $P^\star(X) = sup\{\mu(X) \mid \mu \in P\}$ *and the second one is* $P_\star(X) = inf\{\mu(X) \mid \mu \in P\}$ *which will be used to define a range of probabilities, i.e. they will be an upper and a lower probability, respectively.*

Halpern and Pucella [11] provided a finitary axiomatization for reasoning about linear combinations of upper probabilities, but they proved only weak completeness (every consistent formula is satisfiable). Their formulas are Boolean combinations of the expressions of the form $r_1\ell(\alpha_1) + \cdots + r_n\ell(\alpha_n) \geq r_{n+1}$, where ℓ is the upper probability operator and r_i are real numbers[1], for $i \in \{1, 2, \ldots, n+1\}$. Since nonrestricted real-valued formalisms are rich enough to express the type of a proper infinitesimal $\{0 < x < \frac{1}{n} \mid n = 1, 2, 3, \ldots\}$ (see Example 3), the logic from [11] is not compact. As an unpleasant logical consequence, for any finitary axiomatic system, there are consistent sets of formulas which are unsatisfiable [31].

In this paper, we propose sound and strongly complete (every consistent set of formulas is satisfiable) propositional logic for reasoning about lower and up-

[1] In [11], Halpern and Pucella define the rich language with formulas with all the reals as coefficients. But, in order to obtain decidability result, they have to restrict their language and allow only integer coefficients, i.e. $r_i \in \mathbb{Z}$.

per probabilities ($LUPP$[2]), whose syntax is simpler than the one in [11]. We extend propositional calculus with modal-like unary operators of the form $U_{\geq s}$ and $L_{\geq s}$, where s ranges over the unit interval of rational numbers. The intended meanings of $U_{\geq s}\alpha$ and $L_{\geq s}\alpha$ are "the upper/lower probability of α is at least s". The corresponding semantics consists of special types of Kripke models (possible worlds), with addition of sets of probability measures defined over the worlds. In order to obtain strong completeness, we use infinitary inference rules. Thus our languages are countable and formulas are finite, while only proofs are allowed to be infinite. We also propose the restricted logics $LUPP^{Fr(n)}$ (for each n in $\mathbb{N}$). For those logics, we achieve compactness using only a finite set of probability values, which is still enough for many practical applications. We propose finitary axiomatization for $LUPP^{Fr(n)}$.

From the technical point of view, we have modified some of our earlier developed completion methods presented in [14, 16, 17, 25, 27]. The complete axiomatic system for the logic is the key issue in formalizing the reasoning about lower and upper probabilities, since having a completeness theorem is the only formal way to prove the correctness of the hardware and software.

The contents of this paper are as follows. In Section 2 we recall the notions of lower and upper probability, as well as the representation theorem we use in our axiomatization. In Section 3 we present the syntax and semantics of $LUPP$ and discuss its decidability. In Section 4 we propose an axiomatization for the logic, and we prove some auxiliary propositions. We prove the soundness and completeness of the axiomatization in Section 5. In Section 6 we present the logics $LUPP^{Fr(n)}$, where the probabilities are restricted to a finite set. We conclude in Section 7.

2 Preliminaries

Let $W \neq \emptyset$ and let H be an algebra of subsets of W, i.e., a set of subsets of W such that:

- $W \in H$,
- if $A, B \in H$, then $W \setminus A \in H$ and $A \cup B \in H$.

A function $\mu : H \longrightarrow [0,1]$ is a finitely additive probability measure, if the following conditions hold:

- $\mu(W) = 1$,
- $\mu(A \cup B) = \mu(A) + \mu(B)$, whenever $A \cap B = \emptyset$.

For a set P of probability measures defined on H, the lower probability measure $P_\star$ and the upper probability measure $P^\star$ are defined by

- $P_\star(X) = \inf\{\mu(X) \mid \mu \in P\}$
- $P^\star(X) = \sup\{\mu(X) \mid \mu \in P\}$

for every $X \in H$. In the proof of soundness and completeness, we will use the following basic properties of $P_\star$ and $P^\star$:

- $P_\star(X) \leq P^\star(X)$,
- $P_\star(X) = 1 - P^\star(X^c)$,
- $P^\star(X \cup Y) \leq P^\star(X) + P^\star(Y)$, whenever $X \cap Y = \emptyset$.

In order to axiomatize upper and lower probabilities, we need to completely characterize $P_\star$ and $P^\star$ with a finite number of properties. Many complete characterizations are proposed in the literature, the earliest appears to be by Lorentz [20]. We will use the characterization by Anger and Lembcke [2] (also used by Halpern and Pucella [11, Theorem 2.3]). We start with the definition of (n,k)-cover.

Definition 1 ((n,k)-cover) . *A set A is said to be covered n times by a multiset $\{\{A_1, \ldots, A_m\}\}$ of sets if every element of A appears in at least n sets from $A_1, \ldots, A_m$, i.e., for all $x \in A$, there exists $i_1, \ldots, i_n$ in $\{1, \ldots, m\}$ such that for all $j \leq n$, $x \in A_{i_j}$. An (n,k)-cover of (A, W) is a multiset $\{\{A_1, \ldots, A_m\}\}$ that covers W k times and covers A $n+k$ times.*

Theorem 1 (Anger and Lembcke [2]) *Let W be a set, H an algebra of subsets of W, and f a function $f : H \longrightarrow [0,1]$. There exists a set P of probability measures such that $f = P^\star$ iff f satisfies the following three properties:*

(1) $f(\emptyset) = 0$,

(2) $f(W) = 1$,

(3) for all natural numbers m, n, k and all subsets $A_1, \ldots, A_m$ in H, if $\{\{A_1, \ldots, A_m\}\}$ is an (n,k)-cover of (A, W), then $k + nf(A) \leq \sum_{i=1}^{m} f(A_i)$.

3 The Logic $LUPP$

In this section we will describe the syntax and semantics of the logic $LUPP$, and we discuss the decidability problem of satisfiability of $LUPP$-formulas.

3.1 Syntax

Let S be the set of rational numbers from $[0,1]$ and let $\mathcal{L} = \{p, q, r, \ldots\}$ be a countable set of propositional letters. The language of logic $LUPP$ consists of the elements of set $\mathcal{L}$, classical propositional connectives $\neg$ and $\wedge$ and the lists of upper probability operators $U_{\geq s}$ and $L_{\geq s}$, for every $s \in S$. The set of all classical propositional formulas over $\mathcal{L}$ is defined as usual,

[2] LUP stands for "lower and upper probability", while the second P indicates that the logic is propositional.

and we will denote it by For_C. We will denote the propositional formulas by α, β and γ.

Definition 2 ***(Lower and upper probabilistic formulas)*** *If $\alpha \in For_C$ and $s \in S$, then a basic lower probability formula is any formula of the form $L_{\geq s}\alpha$, and a basic upper probability formula is any formula of the form $U_{\geq s}\alpha$. The set of all lower and upper probabilistic formulas, denoted by For_P, is the smallest set containing all basic lower and upper probability formulas which is closed under Boolean connectives.*

We denote the lower and upper probabilistic formulas by ϕ and ψ, possibly indexed. Let

$$For_{LUPP} = For_C \cup For_P.$$

The formulas from the set For_{LUPP} will be denoted by ρ and σ, possibly with subscripts.

We use the following abbreviations to introduce other types of inequalities: $U_{<s}\alpha$ is $\neg U_{\geq s}\alpha$, $L_{<s}\alpha$ is $\neg L_{\geq s}\alpha$, $U_{\leq s}\alpha$ is $L_{\geq 1-s}\neg\alpha$, $L_{\leq s}\alpha$ is $U_{\geq 1-s}\neg\alpha$, $U_{=s}\alpha$ is $U_{\leq s}\alpha \wedge U_{\geq s}\alpha$, $L_{=s}\alpha$ is $L_{\leq s}\alpha \wedge L_{\geq s}\alpha$, $U_{>s}\alpha$ is $\neg U_{\leq s}\alpha$, $L_{>s}\alpha$ is $\neg L_{\leq s}\alpha$. We also denote both $\alpha \wedge \neg\alpha$ and $\phi \wedge \neg\phi$ by $\bot$ (and similarly for $\top$).

Note that formulas are defined in the same style as in [3, 26], i.e. neither mixing of pure propositional formulas and lower and upper probabilistic formulas, nor nested lower and upper probability operators is allowed.

Example 2 *Continuing Example 1, it is clear that upper and lower probability, for the case that picked marble is green or black, are equal to* 0.6*. If there are no green marbles at all, then we obtain that lower probability for the case that picked marble is not green equals to* 1*. We can express that by the following formula of our language:*

$$U_{=0.6}(G \vee B) \wedge L_{=0.6}(G \vee B) \Rightarrow L_{=1}\neg G.$$

Another example of a lower and upper probabilistic formula is

$$U_{<\frac{1}{3}}\alpha \rightarrow L_{\geq\frac{1}{2}}(\alpha \wedge \beta),$$

where α, $\beta \in For_C$.

Next we state two formulas that are not well defined lower and upper probabilistic formulas of the logic $LUPP$:

$$\alpha \wedge U_{=1}\beta, \qquad U_{\geq s}U_{\geq r}\alpha.$$

The first formula is not well defined since it is a Boolean combination of pure propositional formula and an upper probabilistic formula, while the second formula is not well defined lower and upper probabilistic formula because it contains nested operators.

3.2 Semantics

The semantics for $LUPP$ is based on the possible-world approach.

Definition 3 ($LUPP$-structure) *An $LUPP$-structure is a tuple $\langle W, H, P, \upsilon\rangle$, where:*

- *W is a nonempty set of worlds.*
- *H is an algebra of subsets of W. The elements of H are called measurable worlds.*
- *P is a set of finitely additive probability measures defined on H.*
- *$\upsilon : W \times \mathcal{L} \longrightarrow \{true, false\}$ provides for each world $w \in W$ a two-valued evaluation of the primitive propositions, which is extended to classical propositional formulas as usual.*

For given $\alpha \in For_C$ and $LUPP$-structure M, let $[\alpha]_M = \{w \in W \mid \upsilon(w)(\alpha) = true\}$. We will not write the subscript M when it's clear from context.

Definition 4 (Measurable structure) *The structure M is measurable if $[\alpha]_M \in H$ for every $\alpha \in For_C$. The class of a measurable structures of the logic $LUPP$ will be denoted by $LUPP_{Meas}$.*

Definition 5 (Satisfiability relation) *The satisfiability relation $\models \subseteq LUPP_{Meas} \times For_{LUPP}$ is defined in the following way:*

- *$M \models \alpha$ iff $\upsilon(w)(\alpha) = true$, for all $w \in W$,*
- *$M \models U_{\geq s}\alpha$ iff $P^\star([\alpha]) \geq s$,*
- *$M \models L_{\geq s}\alpha$ iff $P_\star([\alpha]) \geq s$,*
- *$M \models \neg\phi$ iff it is not the case that $M \models \phi$,*
- *$M \models \phi \wedge \psi$ iff $M \models \phi$ and $M \models \psi$.*

Definition 6 (Satisfiability of a formula) *A formula $\rho \in For_{LUPP}$ is satisfiable if there is an $LUPP_{Meas}$-model M such that $M \models \rho$; ρ is valid if for every $LUPP_{Meas}$-model M, $M \models \rho$. A set of formulas T is satisfiable if there is an $LUPP_{Meas}$-model M such that $M \models \rho$ for every $\rho \in T$.*

Example 3 *Consider the set $T = \{\neg U_{=0}\alpha\} \cup \{U_{<\frac{1}{n}}\alpha \mid n$ is a positive integer$\}$. Every finite subset of T is $LUPP_{Meas}$-satisfiable, but the set T itself is not. Therefore, the compactness theorem which states that "if every finite subset of T is satisfiable, then T is satisfiable" does not hold for $LUPP$.*

3.3 Decidability

Recall that Halpern and Pucella [11] provide decidability result for the formulas which are Boolean combinations of the expressions of the form

$$r_1\ell(\alpha_1) + \cdots + r_n\ell(\alpha_n) \geq r_{n+1},$$

where ℓ is the upper probability operator and r_i are integers, for $i \in \{1, 2, \ldots, n+1\}$. First of all, note that using only integers as coefficients has the same expressive power as using all rational numbers. For example, if we want to express $\frac{3}{7}\ell(\alpha) \geq 1$ by using only integers, we can reformulate the formula as $3\ell(\alpha) \geq 7$, etc. Also note that our formula $U_{\geq s}\alpha$ is satisfiable iff the formula $\ell(\alpha) \geq s$ is satisfiable in the logic from [11]. Similarly, $L_{\geq s}\alpha$ is satisfiable iff the formula $-\ell(\neg\alpha) \geq -(1-s)$ is satisfiable. Then decidability of our logic is a consequence of decidability of the logic from [11]. Moreover, since the problem of deciding whether a formula of their language is satisfiable is NP-complete [11, Theorem 5.2], we have an upper bound of the decidability problem for $LUPP$. The lower bound follows from the fact that the complexity of decision problem for classical propositional logic is NP-complete. Thus, the satisfiability problem for $LUPP$-formulas is NP-complete as well.

4 The Axiomatization Ax_{LUPP}

We will introduce an axiomatic system for the logic $LUPP$ which will be denoted by Ax_{LUPP}.

Axiom schemes

(1) all instances of the classical propositional tautologies

(2) $U_{\leq 1}\alpha \wedge L_{\leq 1}\alpha$

(3) $U_{\leq r}\alpha \rightarrow U_{<s}\alpha$, $s > r$

(4) $U_{<s}\alpha \rightarrow U_{\leq s}\alpha$

(5) $(U_{\leq r_1}\alpha_1 \wedge \cdots \wedge U_{\leq r_m}\alpha_m) \rightarrow U_{\leq r}\alpha$, if $\alpha \rightarrow \bigvee_{J\subseteq\{1,\ldots,m\},|J|=k+n}\bigwedge_{j\in J}\alpha_j$ and $\bigvee_{J\subseteq\{1,\ldots,m\},|J|=k}\bigwedge_{j\in J}\alpha_j$ are propositional tautologies, where $r = \frac{\sum_{i=1}^{m} r_i - k}{n}$, $n \neq 0$

(6) $\neg(U_{\leq r_1}\alpha_1 \wedge \cdots \wedge U_{\leq r_m}\alpha_m)$, if $\bigvee_{J\subseteq\{1,\ldots,m\},|J|=k}\bigwedge_{j\in J}\alpha_j$ is a propositional tautology and $\sum_{i=1}^{m} r_i < k$

(7) $L_{=1}(\alpha \rightarrow \beta) \rightarrow (U_{\geq s}\alpha \rightarrow U_{\geq s}\beta)$

Inference Rules

(1) From ρ and $\rho \rightarrow \sigma$ infer σ

(2) From α infer $L_{\geq 1}\alpha$

(3) From the set of premises

$$\{\phi \rightarrow U_{\geq s-\frac{1}{k}}\alpha \mid k \geq \frac{1}{s}\}$$

infer $\phi \rightarrow U_{\geq s}\alpha$

(4) From the set of premises

$$\{\phi \rightarrow L_{\geq s-\frac{1}{k}}\alpha \mid k \geq \frac{1}{s}\}$$

infer $\phi \rightarrow L_{\geq s}\alpha$.

We have, by Axiom 1, that the classical propositional logic is sublogic of $LUPP$. Axiom 2 announce that the upper bound for upper and lower probabilities is 1. Axioms 5 and 6 are the logical analogue of the third condition from Theorem 1. To see that, note that equivalent way to say that $\{\{A_1, \ldots, A_m\}\}$ covers a set A n times is that

$$A \subseteq \bigcup_{J\subseteq\{1,\ldots,m\},|J|=n}\bigcap_{j\in J} A_j.$$

Therefore, the condition that the formula $\alpha \rightarrow \bigvee_{J\subseteq\{1,\ldots,m\},|J|=k+n}\bigwedge_{j\in J}\alpha_j$ is a tautology gives us that $[\alpha]$ is covered $n+k$ times by a multiset $\{\{[\alpha_1], \ldots, [\alpha_m]\}\}$, while the condition that $\bigvee_{J\subseteq\{1,\ldots,m\},|J|=k}\bigwedge_{j\in J}\alpha_j$ is a propositional tautology ensures that $W = [\top]$ is covered k times by a multiset $\{\{[\alpha_1], \ldots, [\alpha_m]\}\}$. Axiom 7 is crucial for proving that equivalent formulas have equal lower and upper probabilities.

Rule 1 is modus ponens, Rule 2 is the lower probability necessitation. Both Rule 3 and Rule 4 are infinitary rules of inference and Rule 3 intuitively says that if upper probability is arbitrary close to s then it is at least s, while Rule 4 intuitively says that if lower probability is arbitrary close to s then it is at least s.

Definition 7 (Inference relation)

- *$T \vdash \rho$ (ρ is derivable from T) if there is an at most denumerable sequence of formulas $\rho_1, \rho_2, \ldots, \rho$, such that every ρ_i is an axiom or a formula from the set T, or it is derived from the preceding formulas by an inference rule;*

- *$\vdash \rho$ (ρ is a theorem) iff $\emptyset \vdash \rho$;*

- *T is consistent if there is at least a formula $\alpha \in For_C$ and a formula $\phi \in For_P$ that are not deducible from T, otherwise T is inconsistent;*

- *T is maximally consistent set if it is consistent and:*

(1) for every $\alpha \in For_C$, if $T \vdash \alpha$, then $\alpha \in T$ and $L_{\geq 1}\alpha \in T$

(2) for every $\phi \in For_P$, either $\phi \in T$ or $\neg\phi \in T$.

- *T is deductively closed if for every $\rho \in For_{LUPP}$, if $T \vdash \rho$, then $\rho \in T$.*

The equivalent way to say that T is inconsistent is that $T \vdash \bot$. Note that it is not required that for every $\alpha \in For_C$, either α or $\neg\alpha$ belongs to a maximal consistent set (as it is done for formulas from For_P). Otherwise, by Rule 2, for each α we would have $L_{\geq 1}\alpha$ or $L_{\geq 1}\neg\alpha$.

Theorem 2 (Deduction theorem) *Let T be a set of formulas. Then $T \cup \{\phi\} \vdash \psi$ iff $T \vdash \phi \to \psi$.*

Proof. The only interesting case is when ϕ,$\psi \in For_P$.
($\Leftarrow$) Direct consequence of Rule 1.
($\Rightarrow$) Suppose that $T \cup \{\phi\} \vdash \psi$. We will use the induction on the length of the inference.
The cases when either $\vdash \psi$ or $\phi = \psi$ or ψ is obtained by application of modus ponens are the same as in the classical propositional case. Thus, let us consider the case where $\psi = L_{\geq 1}\alpha$ is obtained from $T \cup \{\phi\}$ by an application of Rule 2. In that case:

- $T, \phi \vdash \alpha$
- $T, \phi \vdash L_{\geq 1}\alpha$ by Rule 2

However, since $\alpha \in For_C$ and $\phi \in For_P$, ϕ cannot affect the proof of α from $T \cup \{\phi\}$, and we have:

(1) $T \vdash \alpha$

(2) $T \vdash L_{\geq 1}\alpha$ by Rule 2

(3) $T \vdash L_{\geq 1}\alpha \to (\phi \to L_{\geq 1}\alpha)$

(4) $T \vdash \phi \to L_{\geq 1}\alpha$ by Rule 1.

Next, let us consider the case where $\psi = \psi_1 \to U_{\geq s}\alpha$ is obtained from $T \cup \{\phi\}$ by an application of Rule 3. Then:

(1) $T, \phi \vdash \psi_1 \to U_{\geq s-\frac{1}{k}}\alpha$, for all $k \geq \frac{1}{s}$

(2) $T \vdash \phi \to (\psi_1 \to U_{\geq s-\frac{1}{k}}\alpha)$, by the induction hypothesis

(3) $T \vdash (\phi \wedge \psi_1) \to U_{\geq s-\frac{1}{k}}\alpha$

(4) $T \vdash (\phi \wedge \psi_1) \to U_{\geq s}\alpha$, by Rule 3

(5) $T \vdash \phi \to \psi$.

If the formula is obtained by an application of Rule 4, the proof is similar. □

We will not always explicitly emphasize moments in proofs where we use Deduction theorem.

Lemma 1 $\vdash U_{\leq r}\alpha \to L_{\leq r}\alpha$.

Proof. We consider two cases.

(1) $r \neq 1$. From Axiom (6) we obtain that $\neg(U_{\leq r}\alpha \wedge U_{\leq s}\neg\alpha)$, whenever $r + s < 1$. Therefore $U_{\leq r}\alpha \to U_{>s}\neg\alpha$, and because that holds for every $s < 1-r$, by inference rule (3) we have $U_{\leq r}\alpha \to U_{\geq 1-r}\neg\alpha$, i.e. $U_{\leq r}\alpha \to L_{\leq r}\alpha$.

(2) $r = 1$. Direct consequence of Axiom (2). □

Consequently, we obtain that $\vdash L_{\geq r}\alpha \to U_{\geq r}\alpha$, for each $r \in S$.

Lemma 2

(a) $\vdash U_{\geq 0}\alpha$

(b) $\alpha \vdash U_{=1}\alpha$

(c) $\vdash U_{=1}\top$

(d) $\vdash U_{=0}\bot$

(e) $\vdash U_{\geq s}\alpha \to U_{>r}\alpha$, $s > r$

(f) $\vdash U_{>s}\alpha \to U_{\geq s}\alpha$

(g) If $T \vdash \alpha \leftrightarrow \beta$ then $T \vdash U_{\geq s}\alpha \leftrightarrow U_{\geq s}\beta$

Proof.
(a) From Axiom (2), considering $\neg\alpha$, we have that $\vdash L_{\geq 0}\alpha$, and therefore, by Lemma 1 we have that $\vdash U_{\geq 0}\alpha$.
(b) Direct consequence of Inference Rule 2 and Lemma 1. The proofs of (c) and (d) are straightforward, (e) and (f) are obtained from Axioms (3) and (4) and contraposition, and (g) is direct consequence of Rule (2) and Axiom (7). □

5 Soundness and Completeness

5.1 Soundness

Theorem 3 (Soundness) *The axiomatic system Ax_{LUPP} is sound with respect to the class of $LUPP_{Meas}$-models.*

Proof. Our goal is to show that every instance of an axiom schemata holds in every model

and that the inference rules preserve the validity. For example, let us consider Axiom 5. Suppose that $\alpha \rightarrow \bigvee_{J\subseteq\{1,\ldots,m\},|J|=k+n} \bigwedge_{j\in J} \alpha_j$ and $\bigvee_{J\subseteq\{1,\ldots,m\},|J|=k} \bigwedge_{j\in J} \alpha_j$ are propositional tautologies, and suppose that $(U_{\leq r_1}\alpha_1 \wedge \cdots \wedge U_{\leq r_m}\alpha_m)$ holds in a model $M = \langle W, H, P, v\rangle$. We already explained that this means that a multiset $\{\{[\alpha_1], \ldots, [\alpha_m]\}\}$ is an (n, k)-cover of $([\alpha], [\top])$. Also, the inequalities $P^\star([\alpha_1]) \leq r_1, \ldots, P^\star([\alpha_m]) \leq r_m$ hold, by assumption. Since $P^\star$ is an upper probability measure, by Theorem 1, we know that $k + nP^\star([\alpha]) \leq \sum_{i=1}^{m} P^\star([\alpha_i])$, so we obtain that $P^\star([\alpha]) \leq \frac{\sum_{i=1}^{m} r_i - k}{n}$, $n \neq 0$ therefore $P^\star([\alpha]) \leq r$, where $r = \frac{\sum_{i=1}^{m} r_i - k}{n}$, i.e. $M \models U_{\leq r}\alpha$ as well. Consider now the Axiom (7). If $M \models L_{=1}(\alpha \rightarrow \beta)$, we have that $P_\star([\alpha \rightarrow \beta]) = 1$, and therefore $P^\star([\alpha \wedge \neg\beta]) = 1 - P_\star([\alpha \rightarrow \beta]) = 0$. Therefore $P^\star([\alpha]) = P^\star([\alpha \wedge \beta] \cup [\alpha \wedge \neg\beta]) \leq P^\star([\alpha \wedge \beta]) + P^\star([\alpha \wedge \neg\beta]) \leq P^\star([\beta])$. Hence, if $P^\star([\alpha]) \geq s$, then $P^\star([\beta]) \geq s$, so $M \models U_{\geq s}\alpha \rightarrow U_{\geq s}\beta$. The other axioms can be proved to be valid in a similar way and the proof is easier.

Rule (1) is validity-preserving for the same reason as in classical logic. Rule (2): if α holds in $M = \langle W, H, P, v\rangle$, then $[\alpha] = W$, and therefore $\mu([\alpha]) = 1$ for every $\mu \in P$. Then $P_\star([\alpha]) = 1$, so $M \models L_{\geq 1}\alpha$. Rule (3): Suppose that $M \models \phi \rightarrow U_{\geq s-\frac{1}{k}}\alpha$ whenever $k \geq \frac{1}{s}$. If $M \not\models \phi$, then obviously $M \models \phi \rightarrow U_{\geq s}\alpha$. Otherwise $M \models U_{\geq s-\frac{1}{k}}\alpha$ for every $k \geq \frac{1}{s}$, so $M \models U_{\geq s}\alpha$ because of the properties of the set of reals. Rule (4) is validity-preserving for the same reason as Rule (3). □

5.2 Completeness

In order to prove the completeness theorem we start with some auxiliary statements. After that, we show how to extend a consistent set of formulas T to a maximal consistent set of formulas $T^\star$. Finally, we construct the canonical model using the set $T^\star$ such that $M_{T^\star} \models \rho$ iff $\rho \in T^\star$.

Lemma 3 *Let T be a consistent set of formulas.*

(1) For any formula $\phi \in For_P$, either $T \cup \{\phi\}$ is consistent or $T \cup \{\neg\phi\}$ is consistent.

(2) If $\neg(\phi \rightarrow U_{\geq s}\alpha) \in T$, then there is some $n > \frac{1}{s}$ such that $T \cup \{\phi \rightarrow \neg U_{\geq s-\frac{1}{n}}\alpha\}$ is consistent.

(3) If $\neg(\phi \rightarrow L_{\geq s}\alpha) \in T$, then there is some $n > \frac{1}{s}$ such that $T \cup \{\phi \rightarrow \neg L_{\geq s-\frac{1}{n}}\alpha\}$ is consistent.

Proof.

(1) If $T \cup \{\phi\} \vdash \bot$, and $T \cup \{\neg\phi\} \vdash \bot$, then by Deduction theorem we have $T \vdash \neg\phi$ and $T \vdash \phi$. Contradiction.

(2) Suppose that for all $n > \frac{1}{s}$:

$$T, \phi \rightarrow \neg U_{\geq s-\frac{1}{n}}\alpha \vdash \bot.$$

Therefore, by Deduction theorem and propositional reasoning, we have

$$T \vdash \phi \rightarrow U_{\geq s-\frac{1}{n}}\alpha,$$

and by application of Rule 3 we obtain $T \vdash \phi \rightarrow U_{\geq s}\alpha$. Contradiction with the fact that $\neg(\phi \rightarrow U_{\geq s}\alpha) \in T$.

(3) can be proved in a similar way. □

Theorem 4 *Every consistent set can be extended to a maximal consistent set.*

Proof. Consider a consistent set T. By $Cn_C(T)$ we will denote the set of all classical formulas that are consequences of T. Let $\phi_0, \phi_1, \ldots$ be an enumeration of all formulas from For_P. We define a sequence of sets T_i, $i = 0, 1, 2, \ldots$ as follows:

(1) $T_0 = T \cup Cn_C(T) \cup \{L_{\geq 1}\alpha \mid \alpha \in Cn_C(T)\}$

(2) for every $i \geq 0$,

  (a) if $T_i \cup \{\phi_i\}$ is consistent, then $T_{i+1} = T_i \cup \{\phi_i\}$, otherwise

  (b) if ϕ_i is of the form $\psi \rightarrow U_{\geq s}\beta$, then $T_{i+1} = T_i \cup \{\neg\phi_i, \psi \rightarrow \neg U_{\geq s-\frac{1}{n}}\beta\}$, for some positive integer n, so that T_{i+1} is consistent, otherwise

  (c) if ϕ_i is of the form $\psi \rightarrow L_{\geq s}\beta$, then $T_{i+1} = T_i \cup \{\neg\phi_i, \psi \rightarrow \neg L_{\geq s-\frac{1}{n}}\beta\}$, for some positive integer n, so that T_{i+1} is consistent, otherwise

  (d) $T_{i+1} = T_i \cup \{\neg\phi_i\}$.

(3) $T^\star = \bigcup_{i=0}^{\infty} T_i$.

The set T_0 is obviously consistent because it contains consequences of an consistent set. Note that existence of the natural numbers (n) from the steps 2(b) and 2(c) of the construction is provided by Lemma 3, and each T_i is consistent.

It still remains to show that $T^\star$ is maximal consistent set. The steps (1) and (2) of the above construction ensure that $T^\star$ is maximal.

$T^\star$ obviously doesn't contain all formulas. If $\alpha \in For_C$, by the construction of T_0, α and $\neg\alpha$ can not be both

in T_0. For a formula $\phi \in For_P$, the set $T^\star$ does not contain both $\phi = \phi_i$ and $\neg\phi = \phi_j$, because the set $T_{\max\{i,j\}+1}$ is consistent.

Let us prove that $T^\star$ is deductively closed. If a formula $\alpha \in For_C$ and $T \vdash \alpha$, then by the construction of T_0, $\alpha \in T^\star$ and $L_{\geq 1}\alpha \in T^\star$. Let $\phi \in For_P$. It can be easily proved (induction on the length of the inference) that if $T^\star \vdash \phi$, then $\phi \in T^\star$. Note the fact that if $\phi = \phi_j$ and $T_i \vdash \phi$ it has to be $\phi \in T^\star$ because $T_{\max\{i,j\}+1}$ is consistent. Suppose that the sequence $\phi_1, \phi_2, \ldots, \phi$ is the proof of ϕ from $T^\star$. If mentioned sequence is finite, there must be some set T_i such that $T_i \vdash \phi$, and $\phi \in T^\star$. Therefore, suppose that the sequence is countably infinite. We can show that, for every i, if ϕ_i is obtained by an application of an arbitrary inference rule, and all the premises belong to $T^\star$, then, also $\phi_i \in T^\star$. If the inference rule is finitary one, then there must be a set T_j which contains all the premises and $T_j \vdash \phi_i$. So, we conclude that $\phi_i \in T^\star$. Now, consider the infinitary Rule 3. Let $\phi_i = \psi \to U_{\geq s}\alpha$ be obtained from the set of premises $\{\phi_i^k = \psi \to U_{\geq s_k}\alpha \mid s_k \in S\}$. By the induction hypothesis, we have that $\phi_i^k \in T^\star$, for every k. If $\phi_i \notin T^\star$, by step (2)(b) of the construction, there are some l and j so that $\neg(\psi \to U_{\geq s}\alpha)$, $\psi \to \neg U_{\geq s-\frac{1}{l}}\alpha \in T_j$. Thus, we have that for some $j' \geq j$:
- $\psi \wedge \neg U_{\geq s}\alpha \in T_{j'}$,
- $\psi \in T_{j'}$,
- $\neg U_{\geq s-\frac{1}{l}}\alpha$, $U_{\geq s-\frac{1}{l}}\alpha \in T_{j'}$.

Contradiction with the consistency of a set $T_{j'}$.
If we consider the infinitary Rule 4, the proof is similar.

Thus, $T^\star$ is deductively closed set which does not contain all formulas, so it is consistent. □

Definition 8 *If $T^\star$ is the maximally consistent set of formulas, then a tuple $M_{T^\star} = \langle W, H, P, \upsilon\rangle$ is defined:*

- *$W = \{w \mid w \models Cn_C(T)\}$ contains all classical propositional interpretations that satisfy the set $Cn_C(T)$,*
- *$H = \{[\alpha] \mid \alpha \in For_C\}$, where $[\alpha] = \{w \in W \mid w \models \alpha\}$,*
- *P is any set of probability measures such that $P^\star([\alpha]) = \sup\{s \mid U_{\geq s}\alpha \in T^\star\}$,*
- *for every world w and every propositional letter p, $\upsilon(w,p) = true$ iff $w \models p$.*

Lemma 4 *$M_{T^\star}$ is well defined.*

Proof. The prove that H is an algebra is straightforward.

First, $P^\star([\alpha]) := \sup\{s \mid U_{\geq s}\alpha \in T^\star\}$ is well defined because the value of the supremum does not depend on the choice of element from $[\alpha]$, by Lemma 2(g). Let's prove that $P^\star$ is an upper probability measure for some set of probability measures P. It is sufficient to prove the three conditions from Theorem 1. The conditions $P^\star(\emptyset) = 0$ and $P^\star(W) = 1$ are trivial. The only thing left to prove is that if $\{\{[\alpha_1], \ldots, [\alpha_m]\}\}$ is (n,k)-cover of $([\alpha], W)$, then $k + nP^\star([\alpha]) \leq \sum_{i=1}^m P^\star([\alpha_i])$.

Let $P^\star([\alpha_i]) = a_i$, i.e. $\sup\{r \mid U_{\geq r}\alpha_i \in T^\star\} = a_i$, $i = 1, \ldots, m$. For arbitrary $\varepsilon > 0$ there exists rational numbers $q_i \in (a_i, a_i + \varepsilon)$ such that $U_{\leq q_i}\alpha_i \in T^\star$ (otherwise $U_{>q_i}\alpha_i \in T^\star$ which is contradiction with the fact that a_i is supremum). Hence, we have $T^\star \vdash U_{\leq q_1}\alpha_1 \wedge \cdots \wedge U_{\leq q_m}\alpha_m$, and by Axiom 5, we have $T^\star \vdash U_{\leq q}\alpha$, where $q = \frac{\sum_{i=1}^m q_i - k}{n}$, $n \neq 0$ i.e., $\sup\{r \mid U_{\geq r}\alpha_i \in T^\star\} \leq q$ or $P^\star([\alpha]) \leq q$. Therefore, we have $P^\star([\alpha]) \leq \frac{\sum_{i=1}^m q_i - k}{n} = \frac{\sum_{i=1}^m a_i + m\varepsilon - k}{n}$, and because this holds for every $\varepsilon > 0$ we obtain $k + nP^\star([\alpha]) \leq \sum_{i=1}^m P^\star([\alpha_i])$.

If $n = 0$, we need to show that $k \leq \sum_{i=1}^m P^\star([\alpha_i])$. Reasoning as above, we have that $T^\star \vdash U_{\leq q_1}\alpha_1 \wedge \cdots \wedge U_{\leq q_m}\alpha_m$, for some $q_i \in (a_i, a_i + \varepsilon)$, and because of Axiom (6), how $\bigvee_{J \subseteq \{1,\ldots,m\}, |J|=k} \bigwedge_{j \in J} \alpha_j$ are propositional tautologies, we have that $\sum_{i=1}^m q_i \geq k$. Since that holds for every $\varepsilon > 0$, we obtain $\sum_{i=1}^m a_i \geq k$. □

Lemma 5 *Let $T^\star$ be a maximal consistent set of formulas. Then, $M_{T^\star} \in LUPP_{Meas}$.*

Proof. Directly from the construction of $M_{T^\star}$. □

Now we are ready to prove the main result of this paper.

Theorem 5 (Strong completeness) . *A set of formulas T is consistent iff it is $LUPP_{Meas}$ − satisfiable.*

Proof. Direction from right to left follows from the Soundness Theorem. For the proof of the other direction we construct $LUPP_{Meas}$-model $M_{T^\star}$ and show that for every $\rho \in For_{LUPP}$, $M_{T^\star} \models \rho$ iff $\rho \in T^\star$. We use the induction on the complexity of the formula.

- $\rho = \alpha \in For_C$. If $\alpha \in Cn_C(T)$, then by definition of $M_{T^\star}$ we have $M_{T^\star} \models \alpha$. Conversely, if $M_{T^\star} \models \alpha$, by the completeness of classical propositional logic we have that $\alpha \in Cn_C(T)$.
- Consider the case when $\rho = U_{\geq s}\alpha$. If $U_{\geq s}\alpha \in T^\star$, then $\sup\{r \mid U_{\geq r}\alpha \in T^\star\} = P^\star([\alpha]) \geq s$, and so $M_{T^\star} \models U_{\geq s}\alpha$. Now, suppose that $M_{T^\star} \models U_{\geq s}\alpha$, i.e. $\sup\{r \mid U_{\geq r}\alpha \in T^\star\} \geq s$. If $P^\star([\alpha]) > s$, then by the properties of supremum and monotonicity of $P^\star$, we have $U_{\geq s}\alpha \in T^\star$. If $P^\star([\alpha]) = s$, then, as a direct consequence of inference Rule 3, we have that $U_{\geq s}\alpha \in T^\star$.

- Next, let $\rho = L_{\geq s}\alpha$, i.e. $\rho = U_{\leq 1-s}\neg\alpha$. First, suppose that $U_{\leq 1-s}\neg\alpha \in T^\star$. We want to show that $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} \leq 1-s$, so suppose that $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} > 1-s$. Then, there exist a rational number $q \in (1-s, 1-s+\epsilon]$, for some $\epsilon > 0$, such that $U_{\geq q}\alpha \in T^\star$. Hence, $U_{>1-s}\neg\alpha \in T^\star$ which leads us to contradiction. So, $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} \leq 1-s$, i.e. $P^\star([\neg\alpha]) \leq 1-s$ and thus we obtain $M_{T^\star} \models L_{\geq s}\alpha$. Now, for the other direction, suppose that $M_{T^\star} \models U_{\leq 1-s}\neg\alpha$, i.e. $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} \leq 1-s$. Consider the following two cases:

  (1) $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} < 1-s$. Then, if $U_{>1-s}\neg\alpha \in T^\star$, then also $U_{\geq 1-s}\neg\alpha \in T^\star$, so $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} \geq 1-s$. Contradiction.

  (2) $\sup\{r \mid U_{\geq r}\neg\alpha \in T^\star\} = 1-s$. We want to show that then must be $\inf\{r \mid U_{\leq r}\neg\alpha \in T^\star\} = 1-s$ as well. First, suppose that $\inf\{r \mid U_{\leq r}\neg\alpha \in T^\star\} < 1-s$. Hence, there exist a rational number $q_1 \in [1-s-\epsilon, 1-s)$ such that $U_{\leq q_1}\neg\alpha \in T^\star$, and so $U_{<1-s}\neg\alpha \in T^\star$, contradiction with the fact that $U_{\geq 1-s}\neg\alpha \in T^\star$ (direct consequence of inference rule (3)). Now, suppose that $\inf\{r \mid U_{\leq r}\neg\alpha \in T^\star\} > 1-s$, i.e. $\inf\{r \mid U_{\leq r}\neg\alpha \in T^\star\} = 1-s+\varepsilon$. Take an arbitrary rational number $q_2 \in (1-s, 1-s+\varepsilon)$ and then both $U_{\leq q_2}\neg\alpha \in T^\star$ and $U_{\geq q_2}\neg\alpha \in T^\star$ leads us to contradiction (because of the properties of infimum and supremum), which is impossible. Therefore, $\inf\{r \mid U_{\leq r}\neg\alpha \in T^\star\} = 1-s$, or equivalently $\inf\{r \mid L_{\geq 1-r}\alpha \in T^\star\} = 1-s$ and then, by the inference Rule 4, we obtain that $L_{\geq s}\alpha \in T^\star$.

- Now, let $\rho = \neg\psi \in For_P$. Then $M_{T^\star} \models \neg\psi$ iff it is not the case that $M_{T^\star} \models \psi$ iff $\psi \notin T^\star$ iff $\neg\psi \in T^\star$.

- Finally, let $\rho = \phi\wedge\psi \in For_P$. Then, $M_{T^\star} \models \phi\wedge\psi$ iff $M_{T^\star} \models \phi$ and $M_{T^\star} \models \psi$ iff $\phi, \psi \in T^\star$ iff $\phi \wedge \psi \in T^\star$. □

6 The Logic $LUPP^{Fr(n)}$

In this section we introduce the Logic $LUPP^{Fr(n)}$ which is similar to $LUPP$. The main difference is that the finitely additive measures map H to $N = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$, for a fixed positive integer n. Therefore, we obtain countably many different logics, one for each n. Considering the semantics, a model is the tuple $\langle W, H, P, v\rangle$ defined as above but the set P consists of finitely additive measures with restricted range N, i.e., for each $\mu \in P$, $\mu : H \longrightarrow N$. Hence, for every $X \in H$, $P^\star(X)$ also belongs to N, because N is finite and therefore $\sup\{\mu(x) \mid \mu \in P\} = \max\{\mu(x) \mid \mu \in P\}$.

We want to show that there are finitary axiomatizations of these logics and to prove that they are sound and complete with respect to the considered classes of models.

For $s \in [0,1)$, let $s^+ = \min\{r \in N \mid s < r\}$, and if $s \in (0,1]$, let $s^- = \max\{r \in N \mid s > r\}$.
The axiomatization of the logic $LUPP^{Fr(n)}$ includes all the axioms from Section 4, plus one more axiom:

(8) $U_{>s}\alpha \rightarrow U_{\geq s^+}\alpha$.

The inference rules of the axiomatization are rules (1) and (2) from Section 4. Consequently, our axiomatization is finite, and the proofs are finite sequences of formulas.

Lemma 6 *(a)* $\vdash U_{>s}\alpha \leftrightarrow U_{\geq s^+}\alpha$,

(b) $\vdash U_{<s}\alpha \leftrightarrow U_{\leq s^-}\alpha$,

(c) $\vdash \bigvee_{s\in N} U_{=s}\alpha$,

(d) $\vdash \underline{\bigvee}_{s\in N} U_{=s}\alpha$.

Proof. Proofs for (a) and (b) are trivial (direct consequences of Axiom 8 including contrapositive).
(c) Clearly $\vdash (U_{\geq 1}\alpha \vee \neg U_{\geq 1}\alpha) \wedge \neg U_{>1}\alpha$. Therefore

$$\vdash (U_{\geq 1}\alpha \wedge \neg U_{>1}\alpha) \vee (\neg U_{\geq 1}\alpha \wedge \neg U_{>1}\alpha).$$

Since $U_{\geq 1}\alpha \wedge \neg U_{>1}\alpha = U_{=1}\alpha$ and $\vdash U_{<1}\alpha \rightarrow U_{\leq 1}\alpha$ we have $\vdash U_{=1}\alpha \vee U_{<1}\alpha$. Furthermore, from $\vdash U_{<1}\alpha \leftrightarrow ((U_{\geq 1^-}\alpha \vee \neg U_{\geq 1^-}\alpha) \wedge U_{<1}\alpha)$ and $\vdash (U_{\geq s}\alpha \rightarrow U_{\geq s^-}\alpha) \leftrightarrow (U_{<s^-}\alpha \rightarrow U_{<s}\alpha)$ we obtain that $\vdash U_{<1}\alpha \leftrightarrow ((U_{\geq 1^-}\alpha \wedge \neg U_{>1^-}\alpha) \vee (U_{<1^-}\alpha \wedge U_{<1}\alpha))$, and $\vdash U_{=1}\alpha \vee U_{=1^-}\alpha \vee U_{<1^-}\alpha$. Finally, we have that $\vdash (\bigvee_{s\in N} U_{=s}\alpha) \vee U_{<0}\alpha$, so $\vdash (\bigvee_{s\in N} U_{=s}\alpha)$.
(d) $U_{=r}\alpha = U_{\geq r}\alpha \wedge \neg U_{>r}\alpha$, so $\vdash U_{=r}\alpha \rightarrow \neg U_{=s}\alpha$, for every $s > r$. Similarly, we can prove that $\vdash U_{=r}\alpha \rightarrow \neg U_{=s}\alpha$, for every $s < r$. As a consequence, we obtain $\vdash \underline{\bigvee}_{s\in N} U_{=s}\alpha$. □

The proof of the strong completeness theorem is very similar to one presented in Section 5. We will only explain the idea of the proof without going into the details. First, we can prove the soundness theorem and the deduction theorem in a straightforward way. After that, while proving that every consistent set can be extended to a maximal consistent set, we skip the steps where we use infinitary inference rules, i.e. steps 2(b) and 2(c). One more fact needs some explanation. In the proof of the strong completeness theorem we use that if $\sup\{r \mid U_{\geq r}\alpha \in T^\star\} = s$, and $s \in S$, then $U_{\geq s}\alpha \in T^\star$. Now, we have that s must be in a set N, because if $s \notin N$ then there is some $r < s$ such that $r^+ = s^+$, so $T^\star \vdash U_{\geq s^+}\alpha$, but $s < s^+$. Contradiction.

Furthermore, $U_{\geq s}\alpha \in T^{\star}$ because of Lemma 6(d). The rest of the proof of strong completeness theorem is identical as in Section 5.

7 Conclusion

In this paper, we introduced the logic $LUPP$, whose language is obtained by adding the operators for upper and lower probabilities to propositional logic. We proposed an axiomatization for the logic and proved strong completeness. Since the logic is not compact, the axiomatization contains infinitary rules of inference. Then we simplified the semantics and we achieved compactness using finite sets of probability values for logics $LUPP^{Fr(n)}$. For those logics we provide finitary axiomatizations.

As a topic for further research, we propose developing a first order extension of the logics $LUPP$ and $LUPP^{Fr(n)}$. To the best of our knowledge, there is no axiomatization for first order logics for reasoning about lower and upper probabilities. Note that such a logic would extend classical first order logic, so the set of all valid formulas is not recursively enumerable [1] and no complete finitary axiomatization is possible in this undecidable framework. On the other hand, our completion techniques are already applied to some first order probabilistic logics [15, 23, 26].

Acknowledgements

This work was supported by the National Research Fund (FNR) of Luxembourg through project PRIMAT, and by the Serbian Ministry of Education and Science through projects ON174026 and III44006.
We thank the anonymous ISIPTA reviewers for their valuable remarks and suggestions.

References

[1] M. Abadi J. Y. Halpern. Decidability and expressiveness for first-order logics of probability. *Information and Computation*, 112: 1–36. 1994.

[2] B. Anger, J. Lembcke. Infnitely subadditive capacities as upper envelopes of measures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 68: 403–414. 1985.

[3] P. Cintula, C. Noguera. Modal Logics of Uncertainty with Two-Layer Syntax: A General Completeness Theorem. *Volume 8652 of Lecture Notes in Computer Science*, pages 124-136. Springer 2014.

[4] G. de Cooman, F Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence* 172(11): 1400–1427. 2008.

[5] D. Dubois, H. Prade. Possibility Theory. Plenum Press, New York, 1988.

[6] R. Fagin, J. Halpern, N. Megiddo. A logic for reasoning about probabilities. *Information and Computation* 87(1-2):78–128. 1990.

[7] R. Fagin, J. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2): 340–367, 1994.

[8] M. Fattorosi-Barnaba, G. Amati. Modal operators with probabilistic interpretations I. *Studia Logica* 46(4): 383–393. 1989.

[9] A. Frish, P. Haddawy. Anytime deduction for probabilistic logic. *Artificial Intelligence* 69: 93–122. 1994.

[10] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence* 46: 311–350. 1990.

[11] J. Y. Halpern, R. Pucella. A Logic for Reasoning about Upper Probabilities. *Journal of Artificial Intelligence Research* 17: 57–81. 2002.

[12] P.J. Huber. Robust Statistics. Wiley, New York, 1981.

[13] A. Heifetz, P. Mongin. Probability logic for type spaces. *Games and economic behavior* 35: 31–53. 2001.

[14] N. Ikodinović, Z. Ognjanović, M. Rašković, A. Perović. Hierarchies of probabilistic logics. *International Journal of Approximate Reasoning*, 55(9): 1830–1842. 2014.

[15] N. Ikodinović, M. Rašković, Z. Marković, Z. Ognjanović. A first-order probabilistic logic with approximate conditional probabilities. *Logic Journal of the IGPL* vol. 22, no. 4: 539–564. 2014.

[16] A. Ilić-Stepić, Z. Ognjanović. Complex valued probability logics. *Publications de l'Institut Mathematique, N.s. tome 95 (109) (2014)* 73-.86. 2014.

[17] A. Ilić-Stepić, Z. Ognjanović, N. Ikodinović. Conditional p-adic probability logic. *International Journal of Approximate Reasoning* vol. 55, no. 9: 1843–1865. 2014.

[18] H.E. Kyburg. Probability and the Logic of Rational Belief. Wesleyan University Press, Middletown, Connecticut, 1961.

[19] I. Levi. The Enterprise of Knowledge. MIT Press, London, 1980.

[20] G.G. Lorentz. Multiply subadditive functions. *Canadian Journal of Mathematics* 4(4): 455–462. 1952.

[21] M. Meier. An infinitary probability logic for type spaces. *Israel Journal of Mathematics* 192(1): 1–58. 2012.

[22] E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning* vol. 48, no. 2: 628–658. 2008.

[23] M. Milošević, Z. Ognjanović. A first-order conditional probability logic. *Logic Journal of the IGPL* vol. 20, no. 1: 235–253. 2012.

[24] N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28: 71–87. 1986.

[25] Z. Ognjanović, M. Rašković. Some probability logics with new types of probability operators. *Journal of Logic and Computation* 9(2): 181–195. 1999.

[26] Z. Ognjanović, M. Rašković. Some first-order probability logics. *Theoretical Computer Science* 247(1-2): 191–212. 2000.

[27] M. Rašković, Z. Marković, Z. Ognjanović. A logic with approximate conditional probabilities that can model default reasoning. *International Journal of Approximate Reasoning* vol. 49, no. 1: 52–66. 2008.

[28] G. Shafer. A Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ, 1976.

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

[30] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning* vol. 24, no. 2–3: 125–148. 2000.

[31] W. van der Hoeck. Some consideration on the logics $P_F D$. *Journal of Applied Non-Classical Logics*, vol. 7, no. 3, 287 – 307. 1997.

[32] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1: 3–28. 1978.

On the Number and Characterization of the Extreme Points of the Core of Necessity Measures on Finite Spaces

Georg Schollmeyer
Department of Statistics, LMU Munich
georg.schollmeyer@stat.uni-muenchen.de

Abstract

This paper develops a combinatorial description of the extreme points of the core of a necessity measure on a finite space. We use the ingredients of Dempster-Shafer theory to characterize a necessity measure and the extreme points of its core in terms of the Möbius inverse, as well as an interpretation of the elements of the core as obtained through a transfer of probability mass from non-elementary events to singletons. With this understanding we derive an exact formula for the number of extreme points of the core of a necessity measure and obtain a constructive combinatorial insight into how the extreme points are obtained in terms of mass transfers. Our result sharpens the bounds for the number of extreme points given in [15] or [14, 13]. Furthermore, we determine the number of edges of the core of a necessity measure and additionally show how our results could be used to enumerate the extreme points of the core of arbitrary belief functions in a not too inefficient way.

Keywords. necessity measure, core, extreme point, enumeration, belief function, Möbius inverse, mass transfer, possibility measure, credal set, focal set.

1 Introduction

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite space and let $N : 2^\Omega \longrightarrow [0,1]$ be a necessity measure.[1] The core $\mathscr{M}(N)$ of a necessity measure N is defined as the set of probability measures dominating N:

$$\mathscr{M}(N) := \{P \in \mathscr{P}_n | \forall A \in 2^\Omega : P(A) \geq N(A)\},$$

where $\mathscr{P}_n$ denotes the set of all probability measures[2] on Ω. If one identifies a probability measure P with its characterizing vector $(P(\{\omega_1\}), \ldots, P(\{\omega_n\}))$ then the core of N is a convex polytope[3] with finite many extreme points.

[1] A necessity measure $N : 2^\Omega \longrightarrow [0,1]$ is a map satisfying $N(\emptyset) = 0, N(\Omega) = 1$ and $N(A \cap B) = min\{N(A), N(B)\}$ for all $A, B \in 2^\Omega$. For a general introduction to necessity measures, see, e.g., [6].

[2] Since Ω is finite here, it does not make a difference if we take finitely additive or σ-additive probabilities.

[3] For basics of polytopes, see, e.g., [12].

The aim of this paper is to give a formula for the number as well as a constructive description of these extreme points. Since we will derive an exact formula for the number of extreme points in this paper, we are in fact able to improve the bounds for the number of extreme points given in [14, 13] that are not tight.

Studying the geometry of the core and describing the extreme points of the core is interesting for its own, not only in the context of necessity measures. Furthermore, for different applications of imprecise probability theory it is helpful to efficiently describe and compute the extreme points of the core to make different computational tasks tractable.

For example in decision making under partial prior information, for one approach for computing optimal decisions given in [19, Section 4], one needs to compute all extreme points of the underlying imprecise probability model. Also for statistical hypothesis testing under imprecise probabilistic models one can use the extreme points of the cores of the underlying models for the construction of Niveau-α-Maximin-Tests tests, cf., [1, Section 4,5].

In the field of game theory, where more general set functions (games) are treated, the core is an object of interest as well, cf., e.g. [10, 18, 3]. There, for example in the context of convex games the so-called Shapley value appears as the center of gravity of the extreme points of the core (cf., [18]).

The idea of studying complex set functions (here, necessity measures, or more generally, belief functions) via a characterizing set of more easy to handle set functions (here, classical probability measures) is also present in the context of qualitative capacities (cf., e.g., [11]), where the so-called possibilistic core consisting of all (qualitative) possibility measures dominating a given qualitative capacity was introduced in [7]. There, results similar to Theorem 2 of our paper and an enumerating procedure for the "extreme points" of this possibilistic core (which are defined differently in an order theoretic manner) are given.

To describe the core of necessity measures we use

Dempster-Shafer theory[4] and treat necessity measures as special kinds of belief functions. A belief function $\text{Bel}: 2^\Omega \longrightarrow [0,1]$ is a function that is induced by a so-called basic probability assignment $m: 2^\Omega \longrightarrow [0,1]$ via

$$\forall A \in 2^\Omega : \text{Bel}(A) = \sum_{B \subseteq A} m(A).$$

The basic probability assignment m generating *Bel* can be interpreted as a generalization of a probability measure that assigns probability mass not only to elementary events but also to any arbitrary event in $2^\Omega \backslash \{\emptyset\}$. Since m is thought of as a probability measure, it is assumed that $\sum_{A \in 2^\Omega} m(A) = 1$ and furthermore $m(\emptyset) = 0$. Events $A \subseteq \Omega$ with $m(A) > 0$ are called focal sets and the set of all focal sets of a belief function Bel is denoted with $\mathscr{F}(\text{Bel})$. The motivation for introducing the basic probability assignment is the modeling of some kind of uncertainty that cannot be associated with exactly one state $\omega \in \Omega$, but only with a non-elementary event $A \subseteq \Omega$. The belief function *Bel* induced by the basic probability assignment is then of interest if one wants to know for some set A, which portion of the whole probability mass can overall be associated to the states of A. For a given belief function Bel the basic probability assignment m generating Bel can be recovered from Bel by applying the so-called Möbius inversion, thus m is also called the Möbius inverse of Bel.

Now, a necessity measure N (on a finite space) can be characterized[5] as a special belief function[6] where all focal sets are nested, i.e.: $\forall A, B \in \mathscr{F}(N) : A \subseteq B$ or $B \subseteq A$. The core of a belief function can be understood as the set of all probability measures that are consistent with the belief function in the sense that every $P \in \mathscr{M}(Bel)$ can be obtained via a "transfer" of probability mass of the basic probability assignment m from non-elementary events $A \subseteq \Omega$ to singletons $\{\omega\} \subseteq A$. To make this more precise, we state the following definition and theorem:

Definition 1 *Let Bel be a belief function with corresponding basic probability assignment m. A selection* $\lambda : \mathscr{F}(\text{Bel}) \longrightarrow \mathscr{P}_n : A \mapsto \lambda_A$ *is a mapping that assigns to every focal set A a probability measure* λ_A *whose support is in A. The set of all selections associated to a belief function on a space* 2^Ω *with* $|\Omega| = n$ *is denoted with* Λ_n*. A selection* λ *could be understood as specifying for every focal set A and for every state* $\omega \in A$*, how much of mass assigned to A should be transferred from A to* ω*. More precisely, for a belief function* Bel *and a selection* λ *there is an induced probability measure* P_λ *via*

$$P_\lambda(\{\omega_i\}) = \sum_{A \in \mathscr{F}(\text{Bel})} m(A) \cdot \lambda_A(\{\omega_i\}).$$

[4] For an introduction, see, e.g., [17].

[5] For a proof, see, e.g., [17, p.220].

[6] Note that the interpretation of a necessity measure is not necessarily identical to that of a belief function, in this paper we analyze only purely mathematical properties of necessity measures in the framework of Dempster-Shafer theory.

Theorem 1 *For a belief function Bel we have*

$$\mathscr{M}(Bel) = \{P_\lambda \mid \lambda \in \Lambda_n\}.$$

The proof can be found in [4, Corollary 3, p.273] or, in the context of game theory, in [5, Theorem 2]. In the context of game theory, the set $\{P_\lambda \mid \lambda \in \Lambda_n\}$ is called selectope and the set $\mathscr{M}(\nu)$, where ν is a game, is called core and both sets coincide iff the Möbius inverse of the game ν is non-negative, as is also shown in [5, Theorem 2]. Since selections are simply mappings, we can introduce convex combinations. For selections $\lambda, \lambda' \in \Lambda_n$ and $c \in [0,1]$ define

$$c \cdot \lambda + (1-c) \cdot \lambda' : \mathscr{F}(\text{Bel}) \longrightarrow \mathscr{P}_n : A \mapsto c \cdot \lambda_A + (1-c) \cdot \lambda'_A.$$

Note that the probability measure associated to a convex combination of two selections equals the convex combination of the probability measures associated to the two selections: For $\lambda, \lambda' \in \Lambda_n$ and $c \in [0,1]$ we have

$$P_{c\lambda + (1-c)\lambda'} = cP_\lambda + (1-c)P_{\lambda'}.$$

This suggests that it is possible to characterize the extreme points of $\mathscr{M}(Bel)$ in terms of the corresponding selections in Λ_n.

Lemma 1 *For an extreme point* $P = P_\lambda \in \mathscr{M}(Bel)$ *we have:* $\forall A \in \mathscr{F}(Bel) : \exists! \omega \in A : \lambda_A(\{\omega\}) = 1$.

Proof: Let $A \in \mathscr{F}(Bel)$. If for all $\omega \in A : \lambda_A(\{\omega\}) \neq 1$ then there would exist $\omega_i, \omega_j \in A$ with $\lambda_A(\{\omega_i\}) > 0$ and $\lambda_A(\{\omega_j\}) > 0$. Now set $\varepsilon := \min\{\lambda_A(\{\omega_i\}), \lambda_A(\{\omega_j\})\} > 0$ and define the selections μ and ν via

$$\mu_B(\{\omega\}) = \begin{cases} \lambda_B(\{\omega\}) & \text{if } B \neq A \\ \lambda_A(\{\omega\}) & \text{if } B = A, \omega \notin \{\omega_i, \omega_j\} \\ \lambda_A(\{\omega\}) + \varepsilon & \text{if } B = A, \omega = \omega_i \\ \lambda_A(\{\omega\}) - \varepsilon & \text{if } B = A, \omega = \omega_j \end{cases};$$

$$\nu_B(\{\omega\}) = \begin{cases} \lambda_B(\{\omega\}) & \text{if } B \neq A \\ \lambda_A(\{\omega\}) & \text{if } B = A, \omega \notin \{\omega_i, \omega_j\} \\ \lambda_A(\{\omega\}) - \varepsilon & \text{if } B = A, \omega = \omega_i \\ \lambda_A(\{\omega\}) + \varepsilon & \text{if } B = A, \omega = \omega_j \end{cases}.$$

Then $P_\lambda = \frac{1}{2}P_\mu + \frac{1}{2}P_\nu$ and $P_\mu \neq P_\nu$ because

$$\begin{aligned} P_\mu(\{\omega_i\}) - P_\nu(\{\omega_i\}) &= \sum_{B \neq A} m(B) \cdot \mu_B(\{\omega_i\}) + m(A) \cdot \mu_A(\{\omega_i\}) \\ &\quad - \sum_{B \neq A} m(B) \cdot \nu_B(\{\omega_i\}) - m(A) \cdot \nu_A(\{\omega_i\}) \\ &= 2\varepsilon \cdot m(A) \neq 0. \end{aligned}$$

This is a contradiction to the assumption that P_λ is an extreme point of $\mathscr{M}(\text{Bel})$, so there exists an ω with $\lambda_A(\{\omega\}) = 1$. Because λ_A is a probability measure, there could be only one ω with $\lambda_A(\omega) = 1$. ∎

Lemma 1 suggests the following definition:

Definition 2 *Let* $\mathscr{D}_n := \{P \in \mathscr{P}_n \mid \exists!\omega \in \Omega : P(\{\omega\}) = 1\}$ *and let* Bel *be a belief function. Let furthermore* λ *be a selection and* $A \in \mathscr{F}(\text{Bel})$. *If* $\lambda_A \in \mathscr{D}_n$ *we denote by* $\omega_\lambda(A)$ *the unique* ω *with* $\lambda_A(\{\omega\}) = 1$.

Theorem 2 *Let* Bel *be a belief function and let* P_λ *be an extreme point of the core of* Bel. *For focal sets* $A, A' \in \mathscr{F}(\text{Bel})$ *with* $\{\omega_\lambda(A), \omega_\lambda(A')\} \subseteq A \cap A'$ *we have*

$$\omega_\lambda(A) = \omega_\lambda(A').$$

Proof: Assume that $\omega_\lambda(A) \neq \omega_\lambda(A')$. We now show that if this would be the case then we could construct two different elements P_μ and P_ν of the core of Bel such that $P_\lambda = cP_\mu + (1-c)P_\nu$ for some appropriate chosen $c \in [0,1]$ and thus P_λ could not be an extreme point, so $\omega_\lambda(A) = \omega_\lambda(A')$: Define the selections μ and ν as

$$\mu_B(\omega) = \begin{cases} \lambda_B(\omega) & \text{if } B \neq A' \\ 1 & \text{if } B = A', \omega = \omega_\lambda(A) \\ 0 & \text{else} \end{cases}$$

$$\nu_B(\omega) = \begin{cases} \lambda_B(\omega) & \text{if } B \neq A \\ 1 & \text{if } B = A, \omega = \omega_\lambda(A') \\ 0 & \text{else} \end{cases}.$$

These selections lead in fact to two different probability measures P_μ and P_ν. Now, with $c = \frac{m(A)}{m(A)+m(A')}$ we have $P^* := c \cdot P_\mu + (1-c) \cdot P_\nu = P_\lambda$. To see this, look at the three different cases $\omega = \omega_\lambda(A), \omega = \omega_\lambda(A')$ and $\omega \notin \{\omega_\lambda(A), \omega_\lambda(A')\}$:

$$\begin{aligned} P^*(\{\omega_\lambda(A)\}) &= c \sum_{\substack{B \neq A', \\ \omega_\lambda(B) = \omega_\lambda(A)}} m(B) + m(A') + (1-c) \sum_{\substack{B \neq A, \\ \omega_\lambda(B) = \omega_\lambda(A)}} m(B) \\ &= \sum_{\substack{B \notin \{A,A'\} \\ \omega_\lambda(B) = \omega_\lambda(A)}} m(B) + c \cdot (m(A) + m(A')) \\ &= \sum_{\substack{B \notin \{A,A'\} \\ \omega_\lambda(B) = \omega_\lambda(A)}} m(B) + m(A) \\ &= P_\lambda(\{\omega_\lambda(A)\}). \end{aligned}$$

Here, the first sum in the first equation is valid because of Lemma 1 and because all mass $m(A')$ is assigned by μ to $\omega_\lambda(A)$ and the second sum does not contain $m(A)$ and $m(A')$ because the mass $m(A)$ and $m(A')$ is assigned by ν to $\omega_\lambda(A') \neq \omega_\lambda(A)$.

$$\begin{aligned} P^*(\{\omega_\lambda(A')\}) &= c \sum_{\substack{B \neq A', \\ \omega_\lambda(B) = \omega_\lambda(A')}} m(B) + (1-c) \sum_{\substack{B \neq A, \\ \omega_\lambda(B) = \omega_\lambda(A')}} m(B) + m(A) \\ &= \sum_{\substack{B \notin \{A,A'\} \\ \omega_\lambda(B) = \omega_\lambda(A')}} m(B) + (1-c)(m(A') + m(A)) \\ &= \sum_{\substack{B \notin \{A,A'\} \\ \omega_\lambda(B) = \omega_\lambda(A')}} m(B) + m(A') \\ &= P_\lambda(\{\omega_\lambda(A')\}). \end{aligned}$$

Analogously, here, the first sum in the first equation does not contain $m(A)$ and $m(A')$ because these masses are assigned by μ to $\omega_\lambda(A) \neq \omega_\lambda(A')$ and in the second sum the mass $m(A)$ is assigned by ν to $\omega_\lambda(A')$. For $\omega \notin \{\omega_\lambda(A), \omega_\lambda(A')\}$ we have

$$\begin{aligned} P^*(\{\omega\}) &= c \sum_{\substack{B \neq A' \\ \omega_\lambda(B) = \omega}} m(B) + (1-c) \sum_{\substack{B \neq A \\ \omega_\lambda(B) = \omega}} m(B) \\ &= \sum_{\substack{B \notin \{A,A'\} \\ \omega_\lambda(B) = \omega}} m(B) = P_\lambda(\{\omega\}). \end{aligned}$$

Here, the masses $m(A)$ and $m(A')$ essentially play no role, because they are not assigned to ω by neither μ nor ν. ∎

2 Description of the Core of a Necessity Measure

Now we are prepared to describe the extreme points of the core of a necessity measure. As already mentioned, a necessity measure N is a belief function where the focal sets are nested. This enables a concise description of the extreme points of the core:

Theorem 3 *Let* N *be a necessity measure with focal sets* $\mathscr{F}(N) = \{A_1 \subset A_2 \subset \ldots \subset A_k\}$. *The number of extreme points of the core* $\mathscr{M}(N)$ *is given by*

$$|\operatorname{ext}(\mathscr{M}(N))| = |A_1| \cdot \prod_{i=2}^{k} (|A_i \backslash A_{i-1}| + 1). \qquad (1)$$

Furthermore, the set of extreme points can be described as

$$\operatorname{ext}(\mathscr{M}(N)) = \{P_\lambda \mid \lambda \in \Lambda_n^{\text{ext}}\}$$

with $\Lambda_n^{\text{ext}} = \{\lambda \in \Lambda_n \mid \forall A_l \in \mathscr{F}(N) : \lambda_{A_l} \in \mathscr{D}_n \;\&\; \omega_\lambda(A_l) \in A_{l-1} \Rightarrow \omega_\lambda(A_l) = \omega_\lambda(A_{l-1})\}$.

Proof: We firstly show that the number of extreme points is lower or equal to $|A_1| \cdot \prod_{i=2}^{k} (|A_i \backslash A_{i-1}| + 1)$. For this we only have to observe that we could inductively look at the focal sets of N starting from the smallest focal set A_1. For a given extreme point P_λ, the mass assigned to A_1 can be assigned to any $\omega \in A_1$, for which one has $|A_1|$ possibilities. Then, for the second focal set A_2 one has $|A_2 \backslash A_1|$ possibilities to assign the mass of A_2 outside of A_1 and only one possibility to assign the mass into A_1 because in this case, the element $\omega \in A_1$, to which the mass is assigned, is, because of Theorem 2, already determined as $\omega = \omega_\lambda(A_1)$, so for the assignment of the mass $m(A_2)$, we have maximal $|A_i \backslash A_{i-1}| + 1$ possibilities and so on. This gives maximal $|A_1| \cdot \prod_{i=2}^{k} (|A_i \backslash A_{i-1}| + 1)$ possibilities for constructing an extreme point.

Now we still have to show that the extreme points constructed in the above manner are all actually extreme points and that they

are all pairwise different. For this, we can analogously look at ascending focal elements. To see that any P_λ with $\lambda \in \Lambda_n^{\text{ext}}$ is in fact an extreme point we firstly assume that P_λ is the convex combination of r extreme points P_{μ_i} with μ_i in Λ_n^{ext} and show that then necessarily $P_\lambda = P_{\mu_1} = \ldots = P_{\mu_r}$ which shows that P_λ is an extreme point of $\mathscr{M}(N)$:

Since P_λ is such that $\lambda \in \Lambda_n^{\text{ext}}$, all mass of A_1 is assigned by λ to exactly one $\omega \in A_1$ and no other mass $m(B)$ is assigned by λ to some other $\omega \in A_1$, so $P_\lambda(A_1 \backslash \{\omega_\lambda(A_1)\}) = 0$. This implies that for all P_{μ_i} we also have $P_{\mu_i}(A_1 \backslash \{\omega_\lambda(A_1)\}) = 0$ and so $\lambda(A_1) = \mu_1(A_1) = \ldots = \mu_r(A_1)$. Now, look at A_2. If λ assigns the mass of A_2 somewhere into A_1 (namely to $\omega_\lambda(A_1)$), then no mass at all is assigned by λ to some $\omega \in A_2 \backslash A_1$ and thus necessarily all μ_i also have to assign all the mass into A_1 (namely to $\omega_\lambda(A_1)$), so, in this cases we have $\lambda(A_2) = \mu_1(A2) = \ldots = \mu_r(A_2)$. If λ assigns all mass of A_2 somewhere into $A_2 \backslash A_1$, then every μ_i also has to assign the mass of A_2 outside A_1 because if there was a P_{μ_i} that assigns the mass of A_2 into A_1 then we would have $P_{\mu_i}(A_1) > Bel(A_1) = P_\lambda(A_1)$ because if λ assigns the mass of A_2 not into A_1, then λ assigns also the mass of all further $A_3, \ldots, A_k$ not into A_1 and thus $P_\lambda(A_1) = Bel(A_1)$. But if $P_{\mu_i}(A_1) > P_\lambda(A_1)$ then because P_λ is assumed to be a convex combination of $P_{\mu_1}, \ldots, P_{\mu_r}$, there has to be a P_{μ_j} with $P_{\mu_j} < P_\lambda(A_1) = Bel(A_1)$. This is a contradiction to the fact that P_{μ_j} dominates Bel. So, in fact, in this case all μ's assign the mass of A_2 outside of A_1 and thus exactly to $\omega_\lambda(A_2)$ because $P_\lambda(A_2 \backslash \{\omega_\lambda(A_2)\}) = 0$. The same argumentation for all further $A_3, \ldots, A_k$ shows that altogether $\lambda(A_l) = \mu_1(A_l) = \ldots = \mu_r(A_l)$ for $l = 1, \ldots, k$ and so $P_\lambda = P_{\mu_1} = \ldots = P_{\mu_r}$.

To finally see that selections λ, λ' with at least one focal set A_l with $\omega_\lambda(A_l) \neq \omega_{\lambda'}(A_l)$ lead to different P_λ and $P_{\lambda'}$ look at the smallest focal set A_l with $\omega_\lambda(A_l) \neq \omega_{\lambda'}(A_l)$. If $l = 1$ then $P_\lambda(\{\omega_\lambda(A_l)\}) > 0$ and $P_{\lambda'}(\{\omega_\lambda(A_l)\}) = 0$ so P_λ and $P_{\lambda'}$ are different. If $l > 1$ then we have $\omega_\lambda(A_l) \notin A_{l-1}$ or $\omega_{\lambda'} A(l) \notin A_{l-1}$ because if both $\omega_\lambda(A_l)$ and $\omega_{\lambda'}(A_l)$ were in A_{l-1} then also $\omega_\lambda(A_{l-1})$ and $\omega_{\lambda'}(A_{l-1})$ would differ which would be a contradiction to the minimality of l. So assume without loss of generality $\omega_\lambda(A_l) \notin A_{l-1}$. Then $P_\lambda(\{\omega_\lambda(A_l)\}) > 0$ but $P_{\lambda'}(\{\omega_\lambda(A_l)\}) = 0$ because λ' assigns all mass of focal sets $A \supseteq A_l$ either outside of A_l or to $\omega_{\lambda'}(A_l)$ and all other focal sets $A \subseteq A_{l-1}$ do not contain $\omega_\lambda(A_l)$. ∎

With Theorem 3 we have a precise constructive description of the extreme points of the core of a necessity measure. It turns out that it is possible to give furthermore a formula for the number of edges of the core. For this purpose we can use the fact that if two extreme points P and P' are connected through an edge of the core, then they differ exactly at two states and thus the difference of P and P' is of the form $P - P' = (0, \ldots, 0, \varepsilon, 0, \ldots, -\varepsilon, 0 \ldots, 0)$ for some $\varepsilon \in \mathbb{R}$. This result is given in [20] that more generally treats capacities of order 2.

Definition 3 *Let* Bel *be a belief function with focal elements* $\mathscr{F}(\text{Bel}) = \{A_1, \ldots, A_k\}$ *and let* P_λ *be an extreme point of the core of* Bel *induced by a selection* λ. *The characteristic* χ *of* λ *is defined*[7] *as*

$$\chi : \mathscr{F}(Bel) \to \Omega \dot{\cup} \{0\} :$$
$$A_i \mapsto \begin{cases} 0 & \text{if } \exists j < i : \omega_\lambda(A_j) = \omega_\lambda(A_i) \\ \omega_\lambda(A_i) & \text{else} \end{cases}.$$

Lemma 2 *Let* N *be a necessity measure with focal sets* $\mathscr{F}(N) = \{A_1 \subset A_2 \subset \ldots \subset A_k\}$ *and* P_λ *and* $P_{\lambda'}$ *two different extreme points of* $\mathscr{M}(N)$ *induced by selections* λ *and* λ' *with corresponding characteristics* χ *and* χ'. *Then* P_λ *and* $P_{\lambda'}$ *are adjacent (meaning connected through an edge of* $\mathscr{M}(N)$*) if and only if there is exactly one focal set* A *with* $\chi(A) \neq \chi'(A)$.

Proof: Assume that P_λ and $P_{\lambda'}$ are adjacent and that there are two different focal sets where the characteristics χ and χ' differ. Look particularly at the smallest set A_l and some other set A_r where χ and χ' differ. Then P_λ and $P_{\lambda'}$ differ at the two different states $\omega_\lambda(A_l)$ and $\omega_{\lambda'}(A_l)$. Since furthermore either $\omega_\lambda(A_r)$ or $\omega_{\lambda'}(A_r)$ is not in A_l there exists a third state $\omega_\lambda(A_r)$ or $\omega_{\lambda'}(A_r)$ where P_λ and $P_{\lambda'}$ differ, so P_λ and $P_{\lambda'}$ could not be adjacent. This shows that in fact adjacent extreme points have characteristics that differ only on one focal set.

Let now λ and λ' be two selections with associated characteristics χ and χ' that differ only on one focal set A_l. For arbitrary $\omega \in \Omega$ let $i(\omega)$ denote the index of the smallest focal set that contains ω. Then for $\omega \notin \{\omega_\lambda(A_l), \omega_{\lambda'}(A_l)\}$ we have that $P_\lambda(\{\omega\}) = 0$ if $\chi(A_{i(\omega)}) \neq \omega$ and otherwise if $\chi(A_{i(\omega)}) = \omega$ that

$$P_\lambda(\{\omega\}) = \begin{cases} m(A_{i(\omega)}) + \sum_{\substack{B \in \{A_{i(\omega)}, \ldots, A_k\} \\ \chi(B) = 0}} m(B) & \text{if } i(\omega) > l, \\ m(A_{i(\omega)}) + \sum_{\substack{B \in \{A_{i(\omega)}, \ldots, A_{l-1}\} \\ \chi(B) = 0}} m(B) & \text{if } i(\omega) < l. \end{cases}$$

So $P_\lambda(\{\omega\}) = P_{\lambda'}(\{\omega\})$. This means that P_λ and $P_{\lambda'}$ differ at most at two states (namely $\omega_\lambda(A_l)$ and $\omega'_\lambda(A_l)$) and since they are different, they differ exactly at two states. Unfortunately, extreme points that differ only at two states need not to be adjacent (see for example the belief function of section 4) but in the case of necessity measures this is the case. In fact, one can show with the concepts of [20] that for extreme points with associated characteristics that differ only at one focal set (or equivalently, for extreme points that differ only at two states,) there exist permutations σ and μ such that $P_\lambda = p_\sigma$, $P_{\lambda'} = p_\mu$ and the associated equivalence classes $[p_\sigma]$ and $[p_\mu]$ are neighboured in the network and thus P_λ and $P_{\lambda'}$ are adjacent. Details about this can be given upon request. ∎

From Lemma 2 it follows that every extreme point P_λ has $|A_1| - 1 + \sum_{i=1}^{k}(|A_i \backslash A_{i-1}|)$ adjacent extreme points. With this we can count the number of edges of $\mathscr{M}(N)$:

[7] Note that this definition depends on the numbering of the focal sets. For the special case of necessity measures the focal sets are assumed to be numbered in increasing cardinality.

Theorem 4 *Let N be a necessity measure with focal sets $\mathscr{F}(N) = \{A_1 \subset A_2 \subset \ldots \subset A_k\}$. The number $|\text{edges}(\mathscr{M}(N))|$ of edges of the core $\mathscr{M}(N)$ is given by*

$$\frac{1}{2} \cdot |A_1| \cdot \prod_{i=2}^{k} (|A_i \backslash A_{i-1}| + 1) \cdot (|A_1| - 1 + \sum_{i=2}^{k} |A_i \backslash A_{i-1}|).$$

Proof: The statement about the number of edges follows simply by counting for all extreme points P_λ all adjacent extreme points $P_{\lambda'}$ that form an edge with P_λ and by taking into account that with this, every edge is counted two times. ∎

We now compare our result with results given in [14, 13]. There, the results are given in the language of possibility measures Π that are defined in a dual way as $\Pi : 2^\Omega \longrightarrow [0,1] : A \mapsto 1 - N(A^c)$ and are then join preserving mappings particularly satisfying $\Pi(A) = \max_{\omega \in A} \Pi(\{\omega\})$ and are thus uniquely defined through $\pi_i := \Pi(\{\omega_i\})$. Furthermore, in the sequel we assume $0 < \pi_1 \leq \pi_2 \leq \ldots \leq \pi_n = 1$ to simplify presentation. In [14, 13] the set $S := \{i \in \{1,\ldots,n-2\} \mid \pi_{i+1} > \pi_i\} \cup \{n-1\}$ and its cardinality $s := |S|$ play an important role in establishing bounds for the number of extreme points. In terms of the necessity measure N the set S writes as $S = \{i \in \{1,\ldots,n-2\} \mid \{\omega_{i+1},\ldots,n\} \in \mathscr{F}(N)\} \cup \{n-1\}$ and s equals the number of non-elementary focal sets.

Theorem 5 ([14, 13]) [8] *Let N be a necessity measure with associated possibility measure Π satisfying $0 < \pi_1 \leq \ldots \leq \pi_n = 1$. Let s denote the number of non-elementary focal sets of N (or equivalently the cardinality of the set $S = \{i \in \{1,\ldots,n-2\} \mid \pi_{i+1} > \pi_i\} \cup \{n-1\}$). Then the core $\mathscr{M}(N)$ is a $n-1$ dimensional simple polytope*[9] *with $n-1+s$ facets. The number of extreme points is bounded by*

$$|\text{ext}(\mathscr{M}(N))| \geq s(n-2)+2, \tag{2}$$

$$|\text{ext}(\mathscr{M}(N))| \leq \begin{pmatrix} n-2+s-\lfloor \frac{n-2}{2} \rfloor \\ \lfloor \frac{n-1}{2} \rfloor \end{pmatrix} + \begin{pmatrix} n-2+s-\lfloor \frac{n-1}{2} \rfloor \\ \lfloor \frac{n-2}{2} \rfloor \end{pmatrix} \tag{3}$$

and by

$$|\text{ext}(\mathscr{M}(N)| \leq 2^s \prod_{j=1}^{s} (i_j - i_{j-1}) \tag{4}$$

where $i_0 = 0$ and $i_1, i_2, \ldots, i_s$ denote the increasingly ordered indices of the set S.

3 Illustration of the Results

We can now illustrate our results via an example taken from [14, Example 2, p.242]. There, $\Omega = \{\omega_1,\ldots,\omega_5\}$ and a possibility measure Π is given by $\pi_1 = 0$, $\pi_2 = \pi_3 = 0.5$, $\pi_4 = 0.75$, $\pi_5 = 1$. This translates to an associated necessity measure N with focal elements $A_1 = \{\omega_5\}$, $A_2 = \{\omega_4, \omega_5\}$ and $A_3 = \{\omega_2, \omega_3, \omega_4, \omega_5\}$ and masses $m(A_1) = 0.25$, $m(A_2) = 0.25$, $m(A_3) = 0.5$. Because of $\pi_1 = 0$ we have $P(\{\omega_1\}) = 0$ for all $P \in \mathscr{M}(N)$ and thus state ω_1 plays essentially no role and the core $\mathscr{M}(N)$ is a 3-dimensional polytope that is uniquely described by the second, third and fourth component of all probability vectors p of the core.

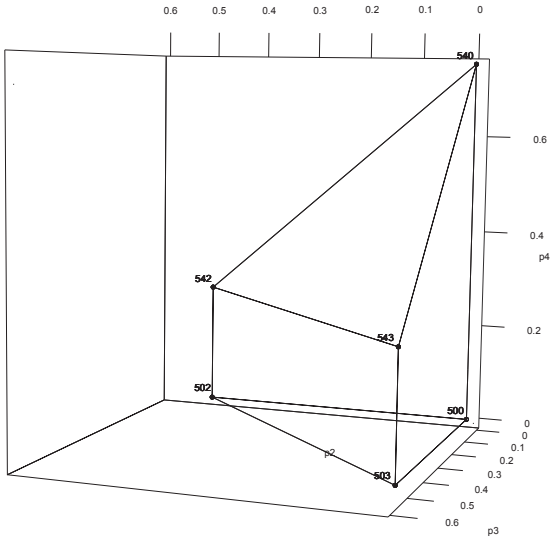


Figure 1: Illustration of the core of Example 2 given in [14].

Figure 1 depicts the core of N. One can see its 6 extreme points, its 9 edges and its 5 facets. This is in accordance with Theorem 3, Theorem 4 and Theorem 5:

$$|\text{ext}(\mathscr{M}(N))| = 1 \cdot 2 \cdot 3 = 6$$
$$|\text{edges}(\mathscr{M}(N))| = \frac{1}{2} \cdot 1 \cdot 2 \cdot 3 \cdot (0+1+2) = 9$$
$$|\text{fac}(\mathscr{M}(N))| = 5 - 2 + 2 = 5.$$

Furthermore, exactly $0+1+2 = 3$ edges meet at every extreme point as argued in the leader of Theorem 4. The digit sequence at every extreme point in Figure 1 indicates the characteristic of the corresponding selection. For example the sequence 503 at the extreme point in the foreground means that the mass of A_1 is assigned to ω_5, the mass of A_2 is assigned to the same ω as the mass of A_1 (thus to ω_5) and the mass of A_3 is assigned to ω_3. One can see that the characteristics of two different extreme points differ exactly at one position if and only if they are adjacent.

The extreme point with characteristic 500 is in a sense distinguished because it is obtained as all mass is assigned to one state ω_5. For every arbitrary necessity measure there exists (at least) one such degenerate extreme point $p \in \mathscr{D}_n$, namely $p = (0,\ldots,1)$. (If the smallest focal set contains k

[8] Note that unfortunately the bounds given in [14, Theorem 2] are misprinted, the correct bounds can be found in [13].

[9] A d-dimensional polytope is called simple, if all vertices are contained in exactly d facets.

$n-1$	s	2^{n-1}	l_1	u_1	u_2	l_3	u_3
2	2	4	4	4	4	4	4
2	1	4	3	3	4	3	3
3	3	8	8	8	8	8	8
3	2	8	6	6	6	6	6
3	1	8	4	4	4	4	4
4	4	16	14	20	16	16	16
4	3	16	11	14	16	12	12
4	2	16	8	9	16	8	9
8	8	256	58	660	256	256	256
8	7	256	51	450	256	192	192
8	6	256	44	294	256	128	144
8	5	256	37	182	256	80	108
9	9	512	74	1430	512	512	512
9	8	512	66	990	512	384	384
9	7	512	58	660	512	256	288
9	6	512	50	420	512	160	216
10	5	1024	47	378	1024	112	243
15	5	32768	72	1584	7776	192	1024
20	5	1048576	97	5005	32768	272	3125
20	10	1048576	192	277134	1048576	6144	59049
$m\cdot s$	s	$2^{m\cdot s}$	$s(m\cdot s-1)+2$	$\binom{(m+1)s+1-\lfloor\frac{m\cdot s+1}{2}\rfloor}{\lfloor\frac{m\cdot s}{2}\rfloor}+\binom{(m+1)s+1-\lfloor\frac{m\cdot s}{2}\rfloor}{\lfloor\frac{m\cdot s+1}{2}\rfloor}$	$2^s\cdot m^s$	$2^{s-1}\cdot((m-1)s+2)$	$(m+1)^s$

Table 1: Different bounds for the number of extreme points of the core of a necessity measure for different sizes of $n-1$ and s.

elements then there are even k degenerate extreme points). This extreme point p is adjacent to extreme points of the form $(0,\ldots,\pi_k,\ldots,1-\pi_k)$ obtained bay assigning all mass $m(A)$ to ω_k if $\omega_k \in A$ and to ω_n else.

Additionally, we can investigate the behaviour of the different bounds for the number of extreme points for different sizes of $n-1$ and s. Table 1 shows the exponential bound 2^{n-1} given in [15], the lower bound l_1 and the upper bound u_1 of [14] (these are here the inequalities (2) and (3)) and the upper bound u_2 of [13] (here inequality (4)) obtained by maximizing (4) under fixed sizes of $n-1$ and s). Additionally, the herein established bounds l_3 and u_3 obtained via minimizing/maximizing (1) for fixed $n-1$ and s are given in the last columns. The last row shows the general situation when $n-1$ is a multiple of s. The sharp upper bound u_3 is obtained by choosing s focal sets $A_1,\ldots,A_s$ with cardinality $|A_l| = l\cdot m+1$ where $m=(n-1)/s$. One can see that for fixed m this bound is exponential in s and in the special case of $m=1$ we get the bound $2^s = 2^{1\cdot s} = 2^{n-1}$ of [15]. For higher m the expansion rate of the exponential growth of the extreme points in dependence on s is greater. If the "density" $\frac{1}{m}$ of focal sets decreases and n is fixed, then the number $(m+1)^s = (m+1)^{\frac{n-1}{m}}$ decreases. For a fixed number of focal sets the number of extreme points is polynomial in the reciprocal m of the density of focal elements.

Our result on the description of the extreme points suggests that it is possible to enumerate all extreme points in a time proportional to $(m+1)^s\cdot s$ because for every extreme point one needs to add s mass values $m(A)$ to some state $\omega^* \in A$ as $p(\omega^*) = p(\omega^*)+m(A)$ to obtain this extreme point.

To get an impression about the possible gain in efficiency, we compare the term $(m+1)^s\cdot s$ with the time two standard enumeration procedures need to enumerate the extreme points. We used implementations of firstly the Double Description Method (cf., [8, 16]) and secondly the Reverse Search Method (cf., [2]) to enumerate the extreme points for different values of m and s and necessity measures that maximize the number of extreme points for given values of m and s.

Figure 2 shows the logarithm of the execution time[10] t in seconds in dependence of s (or m respectively) for the Double Description Method[11] where the value of m (or s respectively) was fixed at different levels. The term $\ln((m+1)^s\cdot s) = s\ln(m+1)+\ln(s)$ is approximately linearly increasing in s (with slope roughly $\ln(m+1)$) and logarithmically increasing in m. Compared to this, the log of computation time increases seemingly linearly in s, but with higher slopes. For example for $m=7$ the slope of $\ln(t)$ is around 4 whereas the slope of $\ln((m+1)^s\cdot s)$ is

[10] We used a personal computer (64 bit) with an Intel(R) Xeon(R) CPU (E5-2650v2, 2.60 Ghz, 2 cores).

[11] We used the r-package rcdd (cf., [9]) which is an interface to the C++ implementation [8] of the Double Description Method.

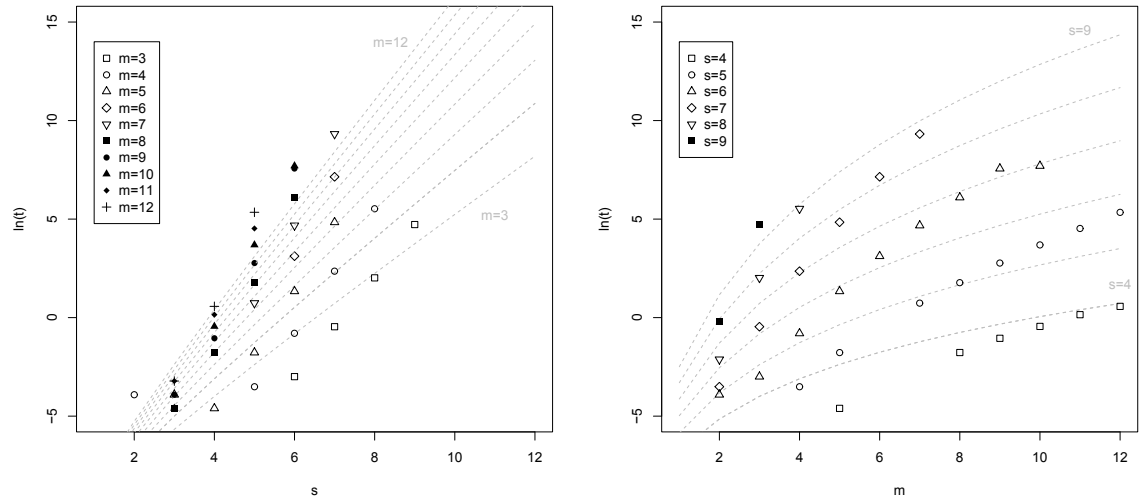


Figure 2: Different execution times of the Double Description Method together with the logarithm of a multiple of the term $(m+1)^s \cdot s$ (grey dashed lines) expected for an efficient enumerating procedure that uses our result.

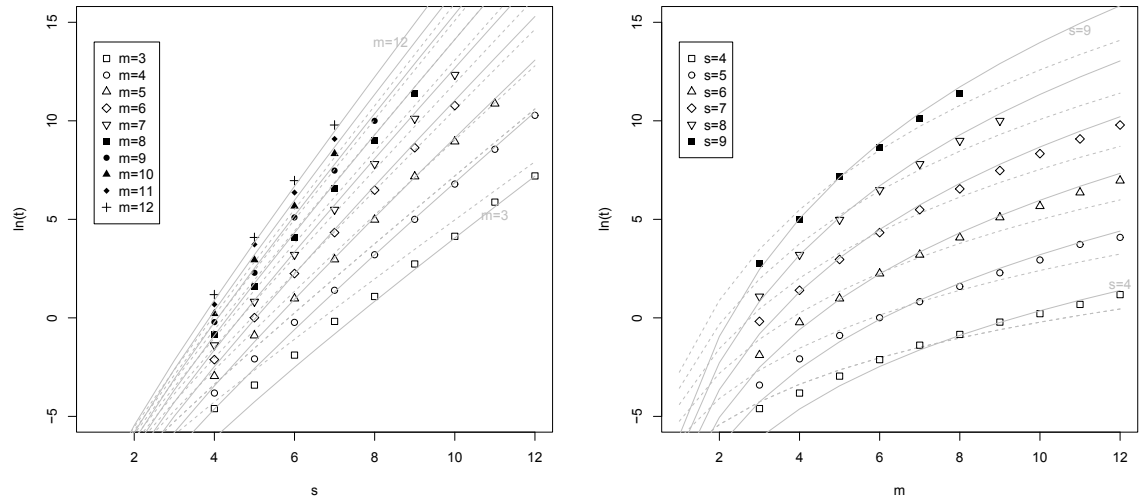


Figure 3: Different execution times of the Reverse Search Method together with the logarithm of a multiple of the term $(m+1)^s \cdot s$ (grey dashed lines) and the logarithm of a multiple of the term $(m+1)^s \cdot (ms)^2$ (grey lines).

somewhere around 2.3, so the expansion rate of the seemingly exponential growth of computation time is larger than the computation time expected for an ideal enumeration procedure. Also the growing of $\ln(t)$ in dependence on m seems to be linearly, so computation time seems to grow also exponentially in m and not polynomially as would be the case with an ideal enumeration procedure.

Figure 3 shows the results for the Reverse Search Method.[12] For this method it is known (cf., [2, Theorem 6.2]) that for simple polytopes the time complexity for enumerating the extreme points is $O(kn)$ per vertex, where n is the number of variables and k is the number of inequalities in the H-representation of the polytope. This would translate in our case to a time complexity of $O((m+1)^s \cdot (ms)^2)$ since we have $(m+1)^s$ vertices of a polytope of dimension $n-1 = m \cdot s$ that could be described by $O(n-1)$ inequalities (cf., [14, p.238]).

It turns out that the execution times are mostly smaller for the Reverse Search Method compared to the Double Description Method. In Figure 3 the grey dashed lines again display the logarithm of a multiple of the term $(m+1)^s \cdot s$, whereas the grey solid lines show the logarithm of a multiple of the term $(m+1)^s \cdot (ms)^2$. One can see that the theoretical time complexity of the Reverse Search Method is roughly in accordance with the actually obtained execution times and that one could still gain some improvement of performance if one uses our results to enumerate the extreme points instead of using the Reverse Search Method.

4 Extension to Belief Functions

With the insight of Theorem 3 and its proof we have not only an exact formula for the number of the extreme points of the core of a necessity measure but also a possibility to efficiently enumerate all extreme points. If we now extend our focus from necessity measures to arbitrary belief functions, then the analysis is more difficult, but Lemma 1 and Theorem 2 still hold. In the case of a necessity measure it was possible to look recursively at ascending focal sets and decide for every focal set if the corresponding mass should be assigned somewhere into the previous focal set (and then the previous focal set would already determine to which exact ω the mass should be assigned to actually obtain an extreme point) or if the mass should be assigned somewhere outside of the previous focal set and then every possible assignment would in fact lead to an extreme point.

If the focal sets are not nested then in the first place it is not clear with which focal set one should start some recursive procedure and how to proceed the recursion. But it is still possible to do a not too inefficient recursion that could generate a set of candidates of extreme points that actually includes all extreme points. One can (totally) order the focal sets in an arbitrary way that at least respects the order of set inclusion of the focal sets to make the recursion not unnecessarily ineffective. One possibility would be to order the focal sets according to their cardinality or another sort of rank function. (The linear ranking via cardinality is then not completely determined, so here comes some sort of arbitrariness into play). Then one could analogously go through ascending focal sets A_i and decide with the help of Theorem 2 to which state $\omega \in A_i$ the mass $m(A_i)$ should be assigned to actually obtain an extreme point. Then for a possible candidate of a selection λ that is already determined on the focal sets $A_1, \ldots, A_l$ one has to decide for the assignment of the mass $m(A_{l+1})$ to some $\omega^* \in A_{l+1}$ if this candidate ω^* is contained in some previous focal set $A \in \{A_1, \ldots, A_l\}$. If this is the case and if furthermore $\omega_\lambda(A) \in A_{l+1}$ and $\omega_\lambda(A) \neq \omega^*$ the assignment of the mass $m(A_{l+1})$ to this ω^* could be excluded, because it could not lead to an extreme point. (Note that in the case of a necessity measure it was enough to look only at the direct predecessor set A_l.)

We now shortly illustrate this recursive procedure via an example. Take $\Omega = \{\omega_1, \ldots, \omega_5\}$ and focal sets $A_1 = \{\omega_5\}$, $A_2 = \{\omega_4, \omega_5\}$, $A_3 = \{\omega_2, \omega_3, \omega_4\}$, $A_4 = \{\omega_2, \omega_3, \omega_4, \omega_5\}$. The indices indicate the ordering of the focal sets, here corresponding to the cardinality of the focal sets. In terms of focal sets this example is like the example above with the only exception that we added the focal set A_3 to make the focal sets not nested. As masses take for example $m(A_1) = 0.2$, $m(A_2) = 0.2$, $m(A_3) = 0.2$, $m(A_4) = 0.4$.

Table 2 shows all 8 selections obtained by the recursive procedure that could possibly lead to extreme points. The second column describes the corresponding selections. For example the digit sequence 5535 means that the masses of A_1,A_2 and A_4 are assigned to ω_5 and the mass of A_3 is assigned to ω_3. This is similar to the digit sequence describing the characteristics in Figure 1, but note that for example selections 2 and 3 have the same characteristic and this is the only reason for choosing this description. The third column shows the 5 components of the corresponding extreme point candidates.

no.	$\omega_\lambda(A_i)$				P_λ				
1	5	5	4	5	0	0	0	0.2	0.8
2	5	5	3	5	0	0	0.2	0	0.8
3	5	5	3	3	0	0	0.6	0	0.4
4	5	5	2	5	0	0.2	0	0	0.8
5	5	5	2	2	0	0.6	0	0	0.4
6	5	4	4	4	0	0	0	0.8	0.2
7	5	4	3	3	0	0	0.6	0.2	0.2
8	5	4	2	2	0	0.6	0	0.2	0.2

Table 2: Summary of altogether 8 candidates of selections that could lead to extreme points.

[12] We used the library lrslib, see http://cgm.cs.mcgill.ca/ avis/C/lrs.html.

Figure 4 shows the resulting core of the belief function for this example (black) together with the core of the necessity measure of the previous example (grey). One can see that compared to the necessity measure, the belief function has an extra facet and altogether 8 extreme points. In contrast to necessity measures, here for example the extreme points no. 1 and no. 8 differ only at two states but are not adjacent. Furthermore, for this example all 8 candidates of Table 2 are in fact extreme points, but this is generally not the case. A simple counterexample is $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and focal sets $A_1 = \{\omega_1, \omega_2\}$, $A_2 = \{\omega_1, \omega_3\}$, $A_3 = \{\omega_2, \omega_3\}$. Then Theorem 2 could not exclude any selection candidate. But for example with $m(A_1) = m(A_2) = m(A_3) = \frac{1}{3}$ and a selection with characteristic 132 we get an associated point $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. But this point is no extreme point of the core because it is a convex combination of the actual extreme points $p_1 = (0, \frac{2}{3}, \frac{1}{3})$ and $p_2 = (\frac{2}{3}, 0, \frac{1}{3})$ obtained by the selections with characteristics 232 and 113.

To exclude selections that do not lead to extreme points one can simultaneously consider the characterization of the extreme points given e.g. in [4, Proposition 9, p.274, Proposition 13, p.277]: Every extreme point of the core of a belief function (or even more generally a capacity of order 2) can be obtained via a total order $<$ on Ω and an associated selection λ that assigns all mass of a focal set A to the greatest element (w.r.t. $<$) of A. The selection with characteristic 132 of the above counterexample is obviously no λ associated to some total order $<$ because from $\omega_\lambda(\{\omega_1, \omega_2\}) = \omega_1$ it follows $\omega_2 < \omega_1$ and with $\omega_\lambda(\{\omega_1, \omega_3\}) = 3$ we have $\omega_1 < \omega_3$, so $\omega_2 < \omega_3$, but this is in contradiction with $\omega_\lambda(\{\omega_2, \omega_3\}) = \omega_2$. So with this "double description" of the extreme points one could exclude candidates of selections that do not lead to extreme points.[13] If we do this, then finally the question remains, if we possibly enumerate some of the extreme points more than one time with this modified procedure. Fortunately, we are able to show that this is not the case:

Theorem 6 *Let λ_1 and λ_2 be two different selections induced by some orderings $<_1$ and $<_2$ on Ω. Assume furthermore that for $i = 1, 2$ and for all focal sets A and A' the relation*

$$\{\omega_{\lambda_i}(A), \omega_{\lambda_i}(A')\} \subseteq A \cap A' \Longrightarrow \omega_{\lambda_i}(A) = \omega_{\lambda_i}(A')$$

of Theorem 2 is satisfied. Then the associated extreme points P_{λ_1} and P_{λ_2} are different.

Proof: Look at the (non-empty) system $D := \{A \in \mathscr{F}(\mathrm{Bel}) \mid \omega_{\lambda_1}(A) \neq \omega_{\lambda_2}(A)\}$. Then take that set $B \in D$ such that the associated $\omega_{\lambda_2}(B)$ is minimal w.r.t. $<_1$. Then the mass of B is transferred by λ_2 to $\omega := \omega_{\lambda_2}(B)$, so $P_{\lambda_2}(\{\omega\}) = \ldots + m(B) + \ldots$,

[13] Another way to exclude transportations that do not lead to extreme points would be to check, if the selection is consistent in the sense of [5, p.25], cf. also Lemma 2 therein. Furthermore, also in the context of qualitative capacities the situation is similar, cf., [7, p.13].

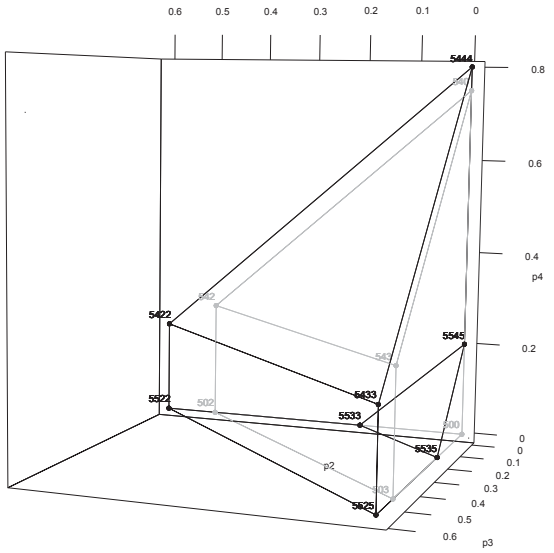


Figure 4: Comparison of the core of a necessity measure and a belief function.

but the mass of B is not transferred by λ_1 to ω. If $P_{\lambda_1}(\{\omega\}) = P_{\lambda_2}(\{\omega\})$ then there has to be another set $\tilde{B} \in D$ whose mass is transferred by λ_1 to ω but not by λ_2 to ω, so for the element $\tilde{\omega} := \omega_{\lambda_2}(\tilde{B}) \in \tilde{B}$ we have $\tilde{\omega} <_1 \omega$, but this is in contradiction to the minimality of $\omega_{\lambda_2}(B)$ w.r.t. $<_1$. So $P_{\lambda_1}(\{\omega\}) \neq P_{\lambda_2}(\{\omega\})$ and the two extreme points P_{λ_1} and P_{λ_2} are different. ∎

With this we can efficiently enumerate the extreme points of an arbitrary belief function (on a finite space).

If the only task is to compute all extreme points, then another nice option of preprocessing could be simplifying in some situations: One could firstly factorize the space Ω according to the equivalence relation $\sim$ of indistinguishability: Two states ω and ω' are indistinguishable if every focal set A either contains both ω and ω' or contains neither ω nor ω'. Especially if there are only few focal sets on a big space Ω then the quotient space $W := \Omega_{/\sim}$ could be much smaller. One can then look at the associated belief function $\mathrm{Bel}_{/\sim} : 2^W \longrightarrow [0,1] : A \mapsto \mathrm{Bel}(\bigcup A)$ and compute in a first step the extreme points of $\mathrm{Bel}_{/\sim}$. The extreme points of the original belief function Bel can then be obtained by deciding in a second step for every extreme point $P_{/\sim}$ of $\mathrm{Bel}_{/\sim}$ and every equivalence class $w = [\omega]$ with $P_{/\sim}(\{w\}) > 0$ to which $\omega \in w$ the mass $P_{/\sim}(\{w\})$ assigned to the equivalence class w should be further assigned.

5 Conclusion

In this paper we worked out a combinatorial description of the extreme points of a necessity measure on a finite space. We treated necessity measures as special kinds of belief functions and were thus able to apply parts of our results also to arbitrary belief functions. Based on this we gave

a possible procedure of seemingly efficiently enumerating the extreme points of belief functions.

For the case of arbitrary belief functions we did not explicitly analyze the complexity of enumeration procedures that use our results. This is a possible direction of further research.

Related to this there are a lot of further combinatorial questions. For instance: Is there a non-trivial bound for the number of extreme points in terms of the number of focal sets? Or: What is the maximal number of extreme points of a belief function where the set of focal elements builds an ordered set (w.r.t. set inclusion) that has a fixed width?[14]

Another direction of further research could be to analyze which parts of the given theorems and considerations of this paper still hold in the case of capacities of order 2 that are not belief functions.

Acknowledgements

The author would like to thank the anonymous reviewers and Thomas Augustin for their very helpful comments and suggestions.

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck und Ruprecht, Göttingen, 1998. In German, with an English summary on pages 247-249.

[2] D. Avis. A revised implementation of the reverse search vertex enumeration algorithm. In *Polytopes-combinatorics and computation*, pages 177–198. Birkhäuser, Basel, 2000.

[3] O. N. Bondareva. Some applications of linear programming methods to the theory of cooperative games. *Problemy kibernetiki*, 10:119–139, 1963. [in Russian]. English translatio in Selected Russian Papers in Game Theory 1959-1965. Princeton: Princeton University Press, 1968.

[4] A. Chateauneuf and J.-Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of möbius inversion. *Mathematical Social Sciences*, 17(3):263–283, 1989.

[5] J. Derks, H. Haller, and H. Peters. The selectope for cooperative games. *International Journal of Game Theory*, 29:23–38, 2000.

[6] D. Dubois and H. Prade. *Possibilty Theory*. Plenum Press, New York, 1988.

[7] D. Dubois, H. Prade, and A. Rico. Representing qualitative capacities as families of possibility measures. *International Journal of Approximate Reasoning*, 58:3–24, 2015.

[8] K. Fukuda and A. Prodon. Double description method revisited. In *Combinatorics and Computer Science*, pages 91–111. Springer, Berlin, 1996.

[9] C. Geyer and G. Meeden. R package rcdd (c double description for r), version 1.1, 2014.

[14] The width of a partially ordered set $(X, \leq)$ is the maximal cardinality of a subset of elements that are all pairwise incomparable.

[10] D. B. Gillies. Solutions to general non-zero-sum games. *Contributions to the Theory of Games*, 4(40):47–85, 1959.

[11] M. Grabisch. The Möbius transform on symmetric ordered structures and its application to capacities on finite sets. *Discrete Mathematics*, 287(1):17–34, 2004.

[12] B. Grunbaum. *Convex polytopes*. Interscience, New York, 1967.

[13] T. Kroupa. Geometry of cores of submodular coherent upper probabilities and possibility measures. In *Soft Methods for Handling Variability and Imprecision*, pages 306–312. Springer, Berlin, 2008.

[14] T. Kroupa. Geometry of possibility measures on finite sets. *International Journal of Approximate Reasoning*, 48(1):237–245, 2008.

[15] E. Miranda, I. Couso, and P. Gil. Extreme points of credal sets generated by 2-alternating capacities. *International Journal of Approximate Reasoning*, 33(1):95–115, 2003.

[16] T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall. The double description method. In *Contributions to the Theory of Games*, volume 2, pages 51–73. Princeton University Press, Princeton, 1953.

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

[18] L. S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1(1):11–26, 1971.

[19] L. V. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In *ISIPTA'05 (Pittsburg)*, volume 5, pages 349–358, 2005.

[20] J. F. Verdegay-López and S. Moral. Network of probabilities associated with a capacity of order-2. *Information Sciences*, 125(1):187–206, 2000.

Using Imprecise Continuous Time Markov Chains for Assessing the Reliability of Power Networks with Common Cause Failure and Non-Immediate Repair

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Jacob Gledhill
Durham University, UK
jacob.gledhill@durham.ac.uk

Damjan Skulj
University of Ljubljana
damjan.skulj@fdv.uni-lj.si

Simon Blake
Newcastle University, UK
simon.blake@ncl.ac.uk

Abstract

We explore how imprecise continuous time Markov chains can improve traditional reliability models based on precise continuous time Markov chains. Specifically, we analyse the reliability of power networks under very weak statistical assumptions, explicitly accounting for non-stationary failure and repair rates and the limited accuracy by which common cause failure rates can be estimated. Bounds on typical quantities of interest are derived, namely the expected time spent in system failure state, as well as the expected number of transitions to that state. A worked numerical example demonstrates the theoretical techniques described. Interestingly, the number of iterations required for convergence is observed to be much lower than current theoretical bounds.

1 Introduction

This paper is an initial exploration to apply recent advances in imprecise continuous time Markov chains to the reliability analysis of power networks.

A typical power network consists of multiple redundant power lines, and works as long as at least one of the power lines is working. A problem of interest occurs when single events can lead to the failure of multiple power lines, such as for instance a landslide causing collapse of a pylon carrying two power lines. Such events are called *common cause failures*. In this case, faults in different lines are not statistically independent, and require special care in modelling, estimation, and validation. In practice, a majority of power outages are due to common cause failure, and therefore modelling this type of failure is vital.

Because common cause failures are very hard to quantify statistically [4], methods from imprecise probability theory have been introduced that allow accurate yet robust prediction of behaviour under relatively weak statistical assumptions [7,9,10]. We model the power networks using imprecise continuous time Markov chains [5,6], which have not previously received much attention in the literature. We are particularly interested in the amount of time spent in the state where all power lines have failed, as well as the number of visits to this state. Whereas [7] considered immediate repair only, here we explicitly model repair as well.

Modelling repair requires much more sophisticated mathematical methods which have been only very recently developed, namely imprecise continuous time Markov chains [6]. Following [6], we will discretise our imprecise continuous time Markov chain and use lower and upper transition operators [2]. In this framework, practical calculations such as calculating lower and upper long run probabilities can be done via linear programming [6]. Throughout, we exploit the fact that repair times of power lines are much shorter than failure times. We use this fact to get a reasonable approximation for the expected number of times that the system visits the totally failed state, as well as the expected amount of time that it spends there, in a given time period. For the imprecise case, we derive simple bounds on these quantities.

The structure of the paper is as follows. Section 2 looks at how we can use continuous time Markov chains to model a power network with two components, accounting for common cause failure and non-immediate repair. Section 3 generalises this setting to imprecise continuous time Markov chains, and works through a detailed example. Section 4 concludes the paper.

2 Continuous Time Markov Chains

2.1 Definition

We start with reviewing the basic definition and properties of continuous time Markov chains.

Definition 1 *A* continuous time Markov chain *is a*

family $(X_t)_{t\in\mathbb{R}}$ *of random variables taking values in a finite state space* S*, such that for all* $s < t$ *and* $\delta t > 0$*,* $X_{t+\delta t}$ *is independent of* X_s *conditionally on* X_t*, and*

$$P(X_{t+\delta t} = j \mid X_t = i) = I_{ij} + \delta t\, Q_{ij} + o_{ij}(\delta t) \quad (1)$$

where $\lim_{\delta t\to 0+} o_{ij}(\delta t)/\delta t = 0$*,* I *is the identity matrix, and* Q *is called the* rate matrix*.*

In particular, the above process is stationary, that is, the transition probabilities $P(X_{t+\delta t} = j \mid X_t = i)$ do not depend on t. For $i \neq j$, the values Q_{ij} are non-negative and describe the rate at which the process switches from state i to state j. The rows of Q must sum to zero because, by Eq. (1),

$$\sum_{j\in S} Q_{ij} = -\sum_{j\in S} \frac{o_{ij}(\delta t)}{\delta t} \quad (2)$$

which tends to zero as $\delta t \to 0$, so all diagonal elements Q_{ii} will be non-positive.

The above definition implies that for any fixed time t there is a transition matrix T_t such that

$$P(X_{s+t} = j \mid X_s = i) = (T_t)_{ij}. \quad (3)$$

The transition matrix is a function of t and satisfies *Kolmogorov's forward* and *backward* equations:

$$\frac{d}{dt}T_t = T_t Q \quad (4)$$

and

$$\frac{d}{dt}T_t = Q T_t \quad (5)$$

respectively, with the initial condition $T_0 = I$. It is well known that in the stationary case, i.e. when Q is constant in time, the solution of the above equations is

$$T_t = e^{tQ}, \quad (6)$$

where e^{tQ} is the *matrix exponential* of tQ.

2.2 Inference

We briefly review the details of doing inference on precise continuous time Markov chains.

Typically, we are interested in the expectation of some function of the state at time t, conditional on some initial state at time 0. It follows from Eq. (6) that for any $f\colon S \to \mathbb{R}$

$$E(f(X_t) \mid X_0 = i) = \sum_{j\in S} P(X_t = j \mid X_0 = i) f(j) = [e^{tQ} f]_i \quad (7)$$

where f is interpreted as a column vector in the last expression.

Equation (7) lies at the basis of all practical calculations with continuous time Markov chains in this paper. For example,

$$P(X_t = j \mid X_0 = i) = E(I_j(X_t) \mid X_0 = i) \quad (8)$$
$$= \left[e^{tQ} I_j\right]_i = \left[e^{tQ}\right]_{ij}, \quad (9)$$

where I_j denotes the indicator function interpreted as a column vector:

$$I_j(k) \coloneqq \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

A wide variety of methods is available for calculating the matrix exponential; see [3] for a review and discussion. For small dimensions, the following method is slow but simple and sufficiently effective for the purpose of this paper. Equation (1) suggests that a continuous time Markov chain is a limit of discrete time Markov chains. Specifically,

$$T'_{\delta t} \coloneqq I + \delta t\, Q \quad (11)$$

maps the rate matrix Q to a discrete time Markov chain transition matrix $T'_{\delta t}$, provided that δt is small enough so that none of the diagonal entries of $T'_{\delta t}$ are negative. It can then be shown that

$$e^{tQ} = \lim_{n\to\infty} (T'_{t/n})^n \quad (12)$$

For practical calculations, we can take n to be a power of 2, so $(T'_{t/n})^n$ can be evaluated by repeated squaring, requiring only $\log_2 n$ matrix multiplications [12]. Although this method is conceptually and computationally simple, it may produce numerically unstable results. An improvement is to use Padé approximation, which also allows for error analysis [3, pp. 9–10]. Essentially, we calculate

$$e^{tQ} \simeq [R_{mm}(tQ/n)]^n \quad (13)$$

where R_{mm} is a known polynomial, and again we take n to be a power of 2 so we can use repeated squaring. Suitable values for m and n, as a function of the 2-norm of tQ, can be found in [3, p. 11, Table 1].

Concerning the limit behaviour for $t \to \infty$, the following definition and theorem are of importance.

Definition 2 *A probability mass function* π *on* S *is a* stationary distribution *for a continuous time Markov chain if*

$$\pi Q = 0. \quad (14)$$

Theorem 3 *If there is a unique stationary distribution π for a continuous time Markov chain, then*

$$\lim_{t\to\infty} \left[e^{tQ}\right]_{ij} = \pi_j. \tag{15}$$

In words, the limit behaviour does not depend on the initial state when $\pi Q = 0$ has a unique solution for π. In that case, π describes that unique limit behaviour.

For analysis and design of power systems, we are typically interested in the following quantities:

(i) the expected amount of time spent in a particular state i during a time period of length τ; it is easily shown that this expectation is simply equal to

$$\alpha_i := \tau\pi_i; \tag{16}$$

(ii) the expected number of transitions to state i during a time period of length τ; this can be shown to be equal to

$$\beta_i := -\tau\pi_i Q_{ii}. \tag{17}$$

2.3 Example

Although the methods described in this paper apply in principle to arbitrary power networks, for demonstrating the ideas of the paper, following [7], we will consider a simple network consisting of just two power lines, called A and B. We can set up a continuous time Markov chain to model this system as follows [1]. The state space is $S = \{AB, A, B, \emptyset\}$, where the labels of the states denote the non-faulty components (i.e. both A and B are non-faulty in AB, whereas both are faulty in $\emptyset$). Using the basic parameter model [4, 10], we can model common cause failures by assigning all failures to any one of the following three events:

- A_I: independent failure of A.
- B_I: independent failure of B.
- C_{AB}: common cause failure of both A and B.

Using standard notation from the literature on common cause failure modelling, denote by q_1^A the rate of A_I, q_1^B the rate of B_I and q_2 the rate of C_{AB}. Similarly, let r_A be the repair rate of A and r_B the repair rate of B—for simplicity we exclude simultaneous repair; extending the analysis to allow for this possibility is trivial. The rate matrix is then

$$Q = \begin{bmatrix} -q_1^A - q_1^B - q_2 & q_1^B & q_1^A & q_2 \\ r_B & -q_1^A - q_2 - r_B & 0 & q_1^A + q_2 \\ r_A & 0 & -q_1^B - q_2 - r_A & q_1^B + q_2 \\ 0 & r_A & r_B & -r_A - r_B \end{bmatrix} \tag{18}$$

The corresponding digraph of the continuous time Markov chain is depicted in Fig. 1.

To estimate the rate parameters q_1^A, q_1^B, q_2, we assume that the chain spends most of its time in state AB,

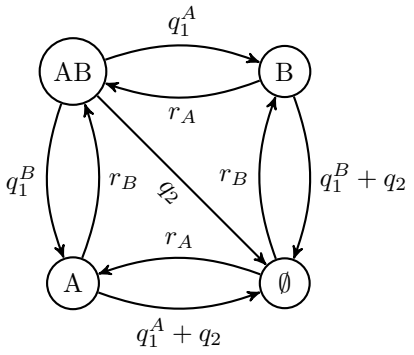


Figure 1: Markov chain for failure with non-instant repair. The nodes show non-faulty power lines.

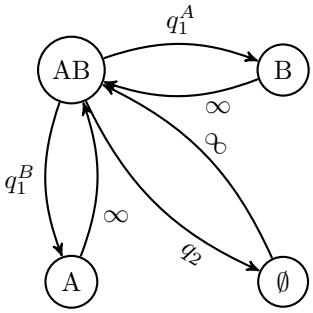


Figure 2: Markov chain for failure with instant repair.

which is reasonable, as repair times are much shorter than failure times. Therefore, from the point of view of AB, we can assume instant repair (see Fig. 2), leading us precisely to the situation discussed in [7]. We know from the theory of continuous time Markov chains that the number of transitions from each state are Poisson distributed. If we then make the simplifying assumption that all failures occur from AB, then the process reduces to three independent Poisson processes, each generating one of the events A_I, B_I and C_{AB}.

Let n_A be the number of single failures of A, n_B the number of single failures of B, and n_{AB} the number of double failures. Similarly, let T_{AB} denote the amount of time spent in state AB. We will use the data from the example in [7] where two circuits, A and B, have been observed for 12 years. A experienced 7 failures in this time, and B 4 failures, with 3 of these failures being double failures. So, using our notation, and under the approximate assumption of immediate repair, we have:

$$n_A = 4, \qquad n_B = 1, \qquad n_{AB} = 3, \tag{19}$$

$$T_{AB} = 12, \tag{20}$$

leading to the following maximum likelihood estimates:

$$\tilde{q}_1^A = \frac{n_A}{T_{AB}} = 1/3 \tag{21}$$

$$\tilde{q}_1^B = \frac{n_B}{T_{AB}} = 1/12 \tag{22}$$

$$\tilde{q}_2 = \frac{n_{AB}}{T_{AB}} = 1/4 \tag{23}$$

We have no repair time data, but a mean time to repair of 12 hours is not entirely unrealistic, so we take $r_A = r_B = 730$. The rate matrix is then:

$$Q = \begin{bmatrix} -\frac{2}{3} & \frac{1}{12} & \frac{1}{3} & \frac{1}{4} \\ 730 & -730 - \frac{7}{12} & 0 & \frac{7}{12} \\ 730 & 0 & -730 - \frac{1}{3} & \frac{1}{3} \\ 0 & 730 & 730 & -1460 \end{bmatrix} \tag{24}$$

The unique stationary distribution is

$$\pi = \begin{bmatrix} 9.989 \times 10^{-1} \\ 2.851 \times 10^{-4} \\ 6.271 \times 10^{-4} \\ 1.713 \times 10^{-4} \end{bmatrix} \tag{25}$$

The expected amount of time spent in the state $\emptyset$ in a period of 10 years, is

$$\alpha_\emptyset = 10 \text{ years} \times 1.713 \times 10^{-4} = 0.625 \text{ days}. \tag{26}$$

and the expected number of visits to $\emptyset$ in a 10 year period is

$$\beta_\emptyset = -10 \times \pi_\emptyset Q_{\emptyset\emptyset} = 2.501 \tag{27}$$

3 Continuous Time Imprecise Markov Chains

3.1 Motivation

The example of the previous section suffers from a number of issues:

- the Markov assumption of $X_{t+\delta t}$ being independent of X_s for $s < t$ conditionally on X_t may not be realistic, particularly for repair;
- the transition rates may not be constant in time, but are usually affected by a variety of factors; and
- estimation of the rates themselves is difficult, due to the lack of data, as extensively discussed in [7, 10].

Specifically, under constant transition rates, repair times are exponentially distributed, and are independent of the history of the system. But this is usually not the case. In some cases the repair may be virtually immediate, as a minor failure in a power line may be detected by a computer and then corrected immediately, but in other cases there may be need for an engineer to go out and work on the line, which obviously takes time. So, repairs times will often follow a bimodal distribution rather than an exponential distribution.

Similarly, failure rates often follow a so-called bathtub curve due to burn-in and wear-out effects, and can be affected in quite complex ways by the repair history of the system. A full modelling of these details requires a lot of data and expert knowledge.

It seems therefore convenient to consider our transition rates as not being fixed, but instead being bounded by an interval, to cover a range of distributions that is more likely to occur in reality, without having to be too precise about the details of this distribution, or on how this distribution depends on the history of the system.

As already mentioned, another source of severe uncertainty concerns the common cause failures, which are very hard to quantify. We will follow [7, 10] and use a robust Bayesian approach to bound our estimates, allowing robust prediction of behaviour under relatively weak statistical assumptions. Eventually, this leaves us with a set of rate matrices $\mathcal{Q}$ bounded by linear constraints. How can we interpret such a set as a statistical process?

3.2 Definitions

Consider a non-stationary non-Markovian continuous time process whose generator

$$Q_{ij}(t, t_n, x_n, \ldots, t_0, x_0) \coloneqq \lim_{\delta t \to 0+} \frac{P(X_{t+\delta t}{=}j | X_t{=}i, X_{t_n}{=}x_n, \ldots, X_{t_0}{=}x_0) - I_{ij}}{\delta t} \tag{28}$$

(where $t > t_n > \cdots > t_0$) is an arbitrary function of time and history which is only required to satisfy $Q(t, t_n, x_n, \ldots, t_0, x_0) \in \mathcal{Q}$ for all t, n, t_n, x_n, …, t_0, and x_0. Here, $\mathcal{Q}$ is a set of transition rate matrices—note that the set $\mathcal{Q}$ itself does not depend on time or history. A simple way to do our inference, which imposes very few assumptions about the additional structure of the process, is then to perform a sensitivity analysis over all these continuous time processes. Specifically, we are interested in the lower expectation of a function of the state at time t for a given initial state at time 0:

Definition 4 *Let $t > 0$. The* lower transition operator $\underline{T}_t \colon \mathbb{R}^S \to \mathbb{R}^S$ *is defined by*

$$[\underline{T}_t f]_i \coloneqq \underline{E}(f(X_t) \mid X_0 = i) \tag{29}$$

The upper transition operator is defined through conjugacy: $\overline{T}_t f = -\underline{T}_t(-f)$.

A clever way of calculating $\underline{T}_t$ goes via the so-called *lower rate operator*, provided that the set $\mathcal{Q}$ of rate matrices has a particular structure:

Definition 5 *We say that $\mathcal{Q}$ has* separately specified rows *if*

$$\mathcal{Q} = \left\{ \begin{bmatrix} Q_{1*} \\ Q_{2*} \\ \vdots \end{bmatrix} : Q_{i*} \in \mathcal{Q}_{i*} \right\} \tag{30}$$

where $\mathcal{Q}_{i*} \coloneqq \{Q_{i*} \colon Q \in \mathcal{Q}\}$, and Q_{i*} denotes the ith row of Q.

In other words, $\mathcal{Q}$ has separately specified rows if the set of matrices attained by forming matrices with any combination of rows from matrices in $\mathcal{Q}$ (where the first row can be chosen from any of the first rows of matrices in $\mathcal{Q}$ and so on) is again $\mathcal{Q}$. For example,

$$\mathcal{Q} \coloneqq \left\{ \begin{bmatrix} -a & a \\ a & -a \end{bmatrix} : a \in [0,1] \right\} \tag{31}$$

does not have separately specified rows, but

$$\mathcal{Q} \coloneqq \left\{ \begin{bmatrix} -a & a \\ b & -b \end{bmatrix} : a, b \in [0,1] \right\} \tag{32}$$

has separately specified rows.

Definition 6 *An* interval rate matrix *is a compact and convex set of rate matrices with separately specified rows.*

Definition 7 *Let $\mathcal{Q}$ be an interval rate matrix. The corresponding* lower rate operator $\underline{Q} \colon \mathbb{R}^S \to \mathbb{R}^S$ *is defined by*

$$\left[\underline{Q} f\right]_i \coloneqq \min_{Q \in \mathcal{Q}} [Qf]_i = \min_{Q_{i*} \in \mathcal{Q}_{i*}} Q_{i*} f \tag{33}$$

for any function $f \colon S \to \mathbb{R}$ on the state space S.

The upper rate operator $\overline{Q}$ is defined through conjugacy: $\overline{Q} f \coloneqq -\underline{Q}(-f)$. The properties of lower and upper rate operators are studied extensively in [6].

Clearly, it holds that

$$[\underline{Q} f]_i \leq [Qf]_i \leq [\overline{Q} f]_i \tag{34}$$

for every $i \in S$, $f \colon S \to \mathbb{R}$, and $Q \in \mathcal{Q}$. But we can make an even stronger statement. Because $\mathcal{Q}$ has separately specified rows, for any specific f, these bounds can be attained for the same Q independently of $i \in S$. Specifically, for every f, there is a $Q \in \mathcal{Q}$ such that for all $i \in S$ we have that $[Qf]_i = [\underline{Q} f]_i$. A similarly result holds for the upper bound. This property substantially simplifies calculations.

What makes $\underline{Q}$ so important is that it entirely determines $\underline{T}_t$, through the following generalisation of Kolmogorov's backward equation [6]:

$$\frac{d}{dt} \underline{T}_t = \underline{Q} \underline{T}_t. \tag{35}$$

Calculating $\underline{T}_t$ amounts to solving this non-linear differential equation with initial condition $\underline{T}_0 = I$. For a specific vector f, if we denote $\underline{T}_t f$ by $\underline{f}_t$, then we must simply solve the differential equation

$$\frac{d}{dt} \underline{f}_t = \underline{Q}\, \underline{f}_t. \tag{36}$$

subject to the initial condition $\underline{f}_0 = f$. This equation has been extensively studied in [6], where the existence of the solution is proved [6, Corollary 2] and numerical algorithms are proposed [6, Section 4].

Unfortunately, those algorithms provide no direct way to determine the limit distribution for $t \to \infty$, which is the main interest of this paper. In particular, the error bounds provided in [6] become too conservative in the long term limit.

Practical calculations of the solutions of Eq. (36) are done by approximations using some kind of discretisation. The simplest method is uniform grid discretisation, which approximates $\underline{T}_t$ by $\underline{T'}^n_{t/n}$, where $\underline{T}'_{\delta t} \colon \mathbb{R}^S \to \mathbb{R}^S$ is defined by

$$\left[\underline{T}'_{\delta t} f\right]_i \coloneqq \left[\left(I + \delta t\, \underline{Q}\right) f\right]_i. \tag{37}$$

It can now be shown that [5]:

$$\left[\underline{f}_t\right]_i = \lim_{n \to \infty} \left[\underline{T'}^n_{t/n} f\right]_i. \tag{38}$$

which generalises Eq. (12).

3.3 Inference

Equation (38) allows us, in principle, to calculate the limit behaviour for $t \to \infty$.

Definition 8 *The lower and upper stationary probability mass functions are defined by*

$$\underline{\pi}_i \coloneqq \lim_{t \to \infty} \underline{P}(X_t = i \mid X_0 = j) \tag{39}$$

$$\overline{\pi}_i \coloneqq \lim_{t \to \infty} \overline{P}(X_t = i \mid X_0 = j) \tag{40}$$

provided that the right hand side does not depend on j.

Clearly, we have that

$$\underline{\pi}_i = \lim_{t \to \infty} \lim_{n \to \infty} \left[\underline{T'}^n_{t/n} I_i\right]_j \tag{41}$$

with a similar equality for $\overline{\pi}_i$. Obviously, it would be much nicer to have a generalisation of the equality $\pi Q = 0$ for imprecise continuous time Markov chains; this is under investigation.

For our power system analysis, we are interested in bounds on the expected amount of time spent in state i during a time period of length τ. A simple heuristic bound is easily shown to be

$$\underline{\alpha}_i \coloneqq \tau \underline{\pi}_i \qquad \overline{\alpha}_i \coloneqq \tau \overline{\pi}_i \tag{42}$$

To see this, consider the problem for a discrete time Markov chain. The lower expected number of steps spent in state i during N time steps satisfies:

$$\underline{E}\left(\sum_{n=1}^{N} I_{X_{M+n}=i} \,\middle|\, X_0 = j\right) \geq \sum_{n=1}^{N} \underline{P}(X_{M+n} = i \mid X_0 = j) \tag{43}$$

for large M, where we used the superadditivity of the lower expectation operator [11, p. 76, §2.6.1(e)] [8]. Now apply this formula for the discretised chain with $M = t/\delta t$ and $N = \tau/\delta t$, note that the duration of each step is δt, and that $\underline{P}(X_{M+n} = i \mid X_0 = j) \simeq \underline{\pi}_i$ for large M.

Similarly, a simple heuristic bound on the expected number of transitions to state i during a time period of length τ is:

$$\underline{\beta}_i \coloneqq \tau \sum_{j \neq i} \underline{\pi}_j [\underline{Q} I_i]_j \qquad \overline{\beta}_i \coloneqq \tau \sum_{j \neq i} \overline{\pi}_j [\overline{Q} I_i]_j \tag{44}$$

To see this, again consider the problem for a discrete time Markov chain. The lower expected number of transitions to state i during N time steps satisfies:

$$\underline{E}\left(\sum_{n=1}^{N} I_{X_{M+n+1}=i \cap X_{M+n} \neq i} \,\middle|\, X_0 = k\right) \tag{45}$$

$$\geq \sum_{n=1}^{N} \underline{P}(X_{M+n+1} = i \cap X_{M+n} \neq i \mid X_0 = k) \tag{46}$$

$$\geq \sum_{n=1}^{N} \sum_{j \neq i} \underline{P}(X_{M+n+1} = i \mid X_{M+n} = j) \times \underline{P}(X_{M+n} = j \mid X_0 = k) \tag{47}$$

for large M, where we used the superadditivity [11, p. 76, §2.6.1(e)] [8] the multiplication rule [11, p. 296, §6.3.5(14)] [8] of the lower expectation operator, and the Markov property. Now apply this formula for the discretised chain with $M = t/\delta t$ and $N = \tau/\delta t$, and note that $\underline{P}(X_{M+n} = j \mid X_0 = k) \simeq \underline{\pi}_j$ for large M, and that

$$\underline{P}(X_{M+n+1} = i \mid X_{M+n} = j) = \delta t \, [\underline{Q} I_i]_j \tag{48}$$

for all $j \neq i$.

These discrete time analyses also say something about the continuous time process because, loosely speaking, the fraction of time that the continuous time process spends on jumping is zero, making the error in these bounds infinitesimally small, provided that δt is infinitesimally small as well.

3.4 Example

We now demonstrate how imprecise continuous time Markov chains can be used to model our power network. For q_1^A, q_1^B, and q_2, we use the data and intervals for failure rates derived in the example in [10], under the approximate assumption of immediate repair, which seems reasonable as the system will spend most of its time in state AB. In this data, A and B are two identical distribution lines, and the intervals for the expected failure rates are:

$$q_1^A \in [0.32, 0.37] \tag{49}$$

$$q_1^B \in [0.32, 0.37] \tag{50}$$

$$q_2 \in [0.19, 0.24] \tag{51}$$

expressed as failures per year. In this study, we did not have repair time data. Through expert elicitation, we judge repair rates between 6 and 12 hours to be reasonable:

$$r_A \in [730, 1460] \tag{52}$$

$$r_B \in [730, 1460] \tag{53}$$

expressed as number of repairs per year.

It may be worth noting that we are not assuming that repairs will happen at a fixed but unknown time between 6 and 12 hours. We are also not assuming that repair time has an exponential density

$$f(t) = \lambda \exp(-\lambda t) \tag{54}$$

with $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, where $\underline{\lambda} = 730$ (rate for a 12 hour mean repair time) and $\overline{\lambda} = 1460$ (rate for a 6 hour mean repair time). The exponential distribution is strongly skewed to the left, with a peak at 0. Although the parametric form of the actual distribution may deviate from the exponential, the feature of having a peak at 0 does reflect an important characteristic of network repairs, as many failures can be fixed remotely (such as for instance a circuit breaker tripping due to a power surge from lightning). Some repairs may also take much longer than 12 hours. An exponential shape is

judged to be a reasonable approximation for repair in the literature [1]. But in this paper, we actually allow a much more general class of distributions for repair, as we allow the rate to vary in time in an arbitrary way between 6 and 12 hours; intuitively, the corresponding set of densities is

$$f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(s)\, ds\right) \tag{55}$$

where $\lambda(t)$ is an arbitrary function of time satisfying $\lambda(t) \in [\underline{\lambda}, \overline{\lambda}]$, and which may also depend on the full system history—only the bounds are assumed to be independent of time and history. Our parametric assumptions are thus much weaker than what is usually assumed in the literature, thereby providing additional confidence in inferences, whilst at the same time making computations more efficient.

We let $\mathcal{Q}$ be an interval rate matrix defined through Eq. (18) and the above constraints. Specifically, with

$$Q_L = \begin{bmatrix} -0.98 & 0.32 & 0.32 & 0.19 \\ 730 & -1460.61 & 0 & 0.51 \\ 730 & 0 & -1460.61 & 0.51 \\ 0 & 730 & 730 & -2920 \end{bmatrix} \tag{56}$$

$$Q_U = \begin{bmatrix} -0.83 & 0.37 & 0.37 & 0.24 \\ 1460 & -730.51 & 0 & 0.61 \\ 1460 & 0 & -730.51 & 0.61 \\ 0 & 1460 & 1460 & -1460 \end{bmatrix} \tag{57}$$

we take

$$\mathcal{Q} \coloneqq [Q_L, Q_U] = \Big\{ Q \colon Q_L \leq Q \leq Q_U, \quad \forall i \in S, \sum_{j \in S} Q_{ij} = 0 \Big\} \tag{58}$$

which has separately specified rows, and therefore it is indeed an interval rate matrix. Note that simply taking the set of all rate matrices of Eq. (18) for all parameters in the above mentioned intervals leads to a set of rate matrices that does not have separately specified rows.

We can now evaluate the lower and upper stationary distributions via Eqs. (37) and (41), where $\underline{Q}$ and $\overline{Q}$ are evaluated through linear programming. To choose sufficiently large values for t and n, we increased the values until empirical convergence was observed. An interesting observation here is that the values required were much lower than some theoretical bounds derived in the literature (see for example [6]). We suspect that this is due to some additional structure of our problem (for instance, rows summing to zero), which in turn raises interesting theoretical questions concerning computation.

In our case, $t = 0.02$ (which roughly corresponds to one week) and $n = 80$ were found to be sufficiently large. For reference, the second largest eigenvalue of the transition matrix, for some extreme selections in $\mathcal{Q}$, was at most 0.817, and $0.817^{80} = 9.830 \times 10^{-8}$, so it seems intuitively reasonable to expect convergence to be of the order 9.830×10^{-8}. In any case, taking say $t = 0.04$ and $n = 320$ (this corresponds to a doubling of the time t and a halving of the time step t/n) leads to no further changes in the following results up to 4 significant digits, which empirically confirms convergence. For the stationary distribution, we find:

$$\underline{\pi} = \begin{bmatrix} 9.985 \times 10^{-1} \\ 2.623 \times 10^{-4} \\ 2.623 \times 10^{-4} \\ 6.513 \times 10^{-5} \end{bmatrix} \quad \overline{\pi} = \begin{bmatrix} 9.994 \times 10^{-1} \\ 7.252 \times 10^{-4} \\ 7.252 \times 10^{-4} \\ 1.647 \times 10^{-4} \end{bmatrix} \tag{59}$$

Concerning the time we expect to spend in state $\emptyset$, say for a period τ of 10 years, we immediately find

$$[\underline{\alpha}_\emptyset, \overline{\alpha}_\emptyset] = [6.513 \times 10^{-4}, 1.647 \times 10^{-3}] \text{ years} \tag{60}$$

$$= [5.705, 14.427] \text{ hours} \tag{61}$$

Similarly, the expected number of visits to $\emptyset$ in that same period is

$$[\underline{\beta}_\emptyset, \overline{\beta}_\emptyset] = [1.900, 2.407] \tag{62}$$

where we used:

$$\left[\underline{Q} I_\emptyset\right]_{AB} = 0.19 \qquad \left[\overline{Q} I_\emptyset\right]_{AB} = 0.24 \tag{63}$$

$$\left[\underline{Q} I_\emptyset\right]_{A} = 0.51 \qquad \left[\overline{Q} I_\emptyset\right]_{A} = 0.61 \tag{64}$$

$$\left[\underline{Q} I_\emptyset\right]_{B} = 0.51 \qquad \left[\overline{Q} I_\emptyset\right]_{B} = 0.61 \tag{65}$$

4 Conclusions

We have looked at a model for dealing with common cause failures in power networks with two power lines, where intervals for the failure and repair rates are used to allow us to make accurate yet robust prediction of behaviour under relatively weak statistical assumptions. Using imprecise Markov chains allows for the case where failure and repair rates are not constant in time, and allows us to properly capture the uncertainty regarding common cause failures which are very hard to quantify. For all these reasons, imprecise continuous time Markov chains have a lot of potential to improve traditional reliability models based on precise Markov chains.

We still assumed that the Markov property [1] holds which, while possibly an unrealistic assumption, is one that is still prevalent in the standard literature.

One disadvantage of the linear programming approach [6] for finding the limit behaviour for $t \to \infty$ is that it

is quite inefficient compared to the standard precise method of solving a linear system. An interesting piece of future research would be to see if we could find new algorithms that work much faster to identify bounds on the stationary distribution.

Another interesting follow up to this paper could be extending the model to apply it to a power network with more than two power lines. Similarly to what is detailed in [7], there would be difficulties in finding intervals for parameters relating to common cause events, because multiple failures can occur in many more ways when three or more power lines are involved.

Finally, we observed empirically that the number of steps required for convergence is much lower than current theoretical bounds. We suspect this is due to the specific structure of our rate matrices. This raises the question as to how current theoretical bounds can be improved for these cases.

Acknowledgements

The first two authors are grateful to BP for supporting the work reported in this paper.

References

[1] Roy Billinton and Ronald N. Allan. *Reliability Evaluation of Power Systems*. Plenum Press, 2nd edition, 1996.

[2] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, October 2009. arXiv:0801.0980, doi:10.1017/S0269964809990039.

[3] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003. doi:10.1137/S00361445024180.

[4] A. Mosleh, K. N. Fleming, G. W. Parry, H. M. Paula, D. H. Worledge, and D. M. Rasmuson. Procedures for treating common cause failures in safety and reliability studies: Procedural framework and examples. Technical Report NUREG/CR-4780, PLG Inc., Newport Beach, CA (USA), January 1988.

[5] Damjan Škulj. Interval matrix differential equations. arXiv:1204.0467v1 [math.CA], April 2012. arXiv:1204.0467.

[6] Damjan Škulj. Efficient computation of the bounds of continuous time imprecise Markov chains. *Applied Mathematics and Computation*, 250:165–180, 2015. doi:10.1016/j.amc.2014.10.092.

[7] Matthias C. M. Troffaes and Simon Blake. A robust data driven approach to quantifying common-cause failure in power networks. In F. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA'13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pages 311–317, Compiègne, France, July 2013. SIPTA. URL: http://www.sipta.org/isipta13/index.php?id=paper&paper=031.html.

[8] Matthias C. M. Troffaes and Gert de Cooman. *Lower Previsions*. Wiley Series in Probability and Statistics. Wiley, 2014. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470723777.html.

[9] Matthias C. M. Troffaes, Dana L. Kelly, and Gero Walter. Imprecise Dirichlet model for common-cause failure. In *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012 (PSAM11 ESREL 2012)*, pages 6722–6728, June 2012.

[10] Matthias C. M. Troffaes, Gero Walter, and Dana Kelly. A robust Bayesian approach to modelling epistemic uncertainty in common-cause failure models. *Reliability Engineering and System Safety*, 125:13–21, May 2014. Special issue of selected articles from ESREL 2012. arXiv:1301.0533, doi:10.1016/j.ress.2013.05.022.

[11] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[12] Guido Walz. Computing the matrix exponential and other matrix functions. *Journal of Computational and Applied Mathematics*, 21(1):119–123, 1988. doi:10.1016/0377-0427(88)90394-9.

Classification SVM Algorithms with Interval-Valued Training Data using Triangular and Epanechnikov Kernels

Lev V. Utkin
Saint Petersburg State Forest Technical University, Russia
lev.utkin@gmail.com

Anatoly I. Chekh
Saint Petersburg State Electrotechnical University, Russia
anatoly.chekh@gmail.com

Yulia A. Zhuk
Saint Petersburg State Forest Technical University, Russia
zhuk_yua@mail.ru

Abstract

Classification algorithms based on different forms of support vector machines (SVMs) for dealing with interval-valued training data are proposed in the paper. L_2-norm and L_∞-norm SVMs are used for constructing the algorithms. The main idea allowing us to represent the complex optimization problems as a set of simple linear or quadratic programming problems is to approximate the Gaussian kernel by the well-known triangular and Epanechnikov kernels. The minimax strategy is used to choose an optimal probability distribution from the set and to construct optimal separating functions.

Keywords. Classification, support vector machine, kernel, interval-valued data, minimax strategy, linear programming, quadratic programming, extreme points.

1 Introduction

The binary classification problem can be formally written as follows. Given n training data (examples, patterns) $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$, in which $\mathbf{x}_i \in \mathbb{R}^m$ represents a feature vector involving m features and $y_i \in \{-1, 1\}$ indices the class of the associated examples, the task of classification is to construct an accurate classifier $c : \mathbb{R}^m \to \{-1, 1\}$ that maximizes the probability that $c(\mathbf{x}) = y_i$ for $i = 1, ..., n$. Generally $\mathbf{x}_i$ may belong to an arbitrary set $\mathcal{X}$, but we consider the special case $\mathcal{X} = \mathbb{R}^m$ for simplicity. One of the ways for classification is to find a real valued separating function $f(\mathbf{x}, \mathbf{w}, b)$ having parameters $\mathbf{w}$ and b such that $\mathbf{w} = (w_1, ..., w_m) \in \mathbb{R}^m$ and $b \in \mathbb{R}$, for example, $f(\mathbf{x}, \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$. Here $\langle \mathbf{w}, \mathbf{x} \rangle$ denotes the dot product of two vectors $\mathbf{w}$ and $\mathbf{x}$. The sign of the function determines the class label prediction or $c(\mathbf{x})$. We also introduce the notation $x_i^{(k)}$ for the k-th element of the vector $\mathbf{x}_i$.

There are a lot of classification algorithms, but most of them are based on using a training set consisting of precise or point-valued data. However, training examples in many real applications can be obtained only in the interval form. Interval-valued data may result from imperfection of measurement tools or imprecision of expert information, from missing data. It should be noted that the interval-valued data can be regarded as a special case of a more general form of imprecise data. For example, we cannot observe some feature, but we know that the difference between values of the feature for data from different classes is less than some known value. In this case, we have imprecise training data.

Many classification algorithms have been presented for dealing with interval-valued data [11, 14, 17]. Most algorithms use an obvious approach when interval-valued observations are replaced by precise values based on some additional assumptions, for example, by taking middle points of intervals [12]. This approach is rather efficient when intervals are small and do not intersect each other. If intervals in training data are very large, then this approach may lead to incorrect classification.

One of the classification algorithms taking into account all points of intervals has been proposed by Utkin and Coolen [21]. However, this algorithm uses a weak assumption which restricts its usage. According to this assumption, the separating function f is monotone, for example, linear, because its lower and upper bounds in this case are determined only by the bounds of pattern intervals. However, in spite of the restricted application of the algorithm, it looks for "optimal" points to some extent of the expected classification risk, but not for points of intervals of training data. This is an important peculiarity of the algorithm. Similar approaches have been used by Hüllermeier [10], by Antonucci et al. [1] in their interesting classification algorithms under interval and fuzzy training data.

We propose a general approach for constructing robust classification algorithms dealing with imprecise training data which can be represented in the form of closed intervals or some compact convex sets of values

of training data. In contrast to the algorithms where intervals are replaced by points, the proposed algorithm searches for optimal precise points by applying the robust or maximin strategy of decision. In fact, we select a single probability distribution or a point in the interval of expected risk values in accordance with a certain decision strategy instead of points in intervals of training data.

We use the term robust in the sense defined by Xu et al. [26]. The robustness property means here reducing sensitivity of a classifier to incorrect replacement of intervals by point-valued analogues. There are different definitions of robustness. We use robustness stemmed from the robust optimization where a minimax optimization is performed over all possible values of intervals. This definition differs from robustness in statistics which studies how an estimator behaves under a small perturbation of the statistics model.

In order to construct new classification algorithms dealing with interval-valued training data, we propose to use the following three ideas:

1. Interval-valued observations produce a set of probability distributions such that the lower and upper expected classification risk measures can be determined in terms of the belief functions in a simple way.

2. There are many variants of SVMs. It is proposed to choose standard L_2-norm SVM. Moreover, it is proposed to use one of the L_∞-norm SVMs such that constraints in its dual form do not depend on vectors of observations $\mathbf{x}_i$, $i = 1, ..., n$. This allows us to solve the corresponding optimization problem by using extreme points of the polytope produced by the constraints.

3. It is proposed to replace the Gaussian kernel by the well-known triangular kernel and Epanechnikov kernel which can be regarded as two approximations of the Gaussian kernel. This replacement allows us to get a set of linear or quadratic optimization problems with variables $\mathbf{x}_i$ restricted by intervals $\mathbf{A}_i$, $i = 1, ..., n$.

It should be noted that the idea of approximating the Gaussian kernel by the triangular kernel in one-class classification problems has been studied by the authors [22]. This idea and other ones are exploited below for constructing new binary classification algorithms.

2 A Standard L_2-Norm SVM by Precise Data

Suppose we have training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}^m \times \{-1, +1\}$. Let ϕ be a feature map $\mathbb{R}^m \to G$ such that the data points are mapped into an alternative higher-dimensional feature space G. In other words, this is a map into an inner product space G such that the inner product in the image of ϕ can be computed by evaluating some simple kernel $K(\mathbf{x}, \mathbf{y}) = (\phi(\mathbf{x}), \phi(\mathbf{y}))$, such as the Gaussian kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\left\|\mathbf{x} - \mathbf{y}\right\|^2 / \sigma^2\right).$$

Here σ is the kernel parameter determining the geometrical structure of the mapped samples in the kernel space [24]. It is important to note that Gaussian kernels are very popular because they support many complex models and are rather flexible. Moreover, they show good features and strong learning capability [25].

The SVM minimizes the empirical risk measure

$$R = n^{-1} \sum_{i=1}^{n} l(\mathbf{x}_i),$$

as an approximation of the expected risk, which can be regarded as a bound depending on the so-called VC dimension introduced by Vapnik [23]. Here l is a loss function. The minimization of the above functional is an ill-posed problem because it admits an infinite number of solutions. In order to overcome this difficulty, regularization theory [19] provides a framework for solving the problem by adding appropriate constraints on the solution. This can be done by introducing a smoothness or penalty term $J(f)$ and a tuning "cost" parameter C which balances the tradeoff between the empirical risk measure and the penalty term. As a result, a general class of regularization problems has the form:

$$\min_f \left(C \sum l(\mathbf{x}_i) + J(f)\right).$$

Standard penalty terms are the L_s-norms such that $L_s = \|\mathbf{w}\|_s$, $s > 0$. In particular, the most popular penalty in the SVM classifier is $\|\mathbf{w}\|_2$. Hence, the SVM classifier can be represented in the form of the following convex optimization problem (the quadratic programming problem):

$$\min_{\xi, \mathbf{w}, b} R = \min_{\xi, \mathbf{w}, b} \left(\frac{1}{2} \|\mathbf{w}\|_2 + C \sum_{i=1}^{n} \xi_i\right), \qquad (1)$$

subject to

$$\xi_i \geq 0, \quad y_i \left(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b\right) \geq 1 - \xi_i, \ i = 1, ..., n. \qquad (2)$$

Here ξ_i, $i = 1, ..., n$, are the slack variables. The quantity $C\xi_i$ is the "penalty" for any data point $\mathbf{x}_i$ that either lies within the margin on the correct side of the hyperplane ($\xi_i \leq 1$) or on the wrong side of the hyperplane ($\xi_i > 1$). The above optimization problem is obtained under condition that the so-called hinge loss function is used, i.e., $l(\mathbf{x}) = \max(0, 1 - y_i f\left\langle \mathbf{w}, \phi(\mathbf{x}_i) \right\rangle)$.

Instead of minimizing the primary objective function (1), a dual objective function, the so-called Lagrangian, can be formed of which the saddle point is the optimum. The dual programming problem is of the form:

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \quad (3)$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, ..., n. \quad (4)$$

After substituting the obtained solution into the expression for the decision function f, we get the "dual" separating function

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

The above SVM is often called the L_2-norm SVM due to the definition of the regularization term. The parameter b is defined by using support vectors $\mathbf{x}_i$ from the following equation $b = y_j - \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j)$.

At the same time, there are other forms of the SVM defined by different L_s-norms of the penalty term. It turns out that the SVM with the L_∞-norm can be very useful when we deal with interval-valued data.

3 Interval-Valued Training Data and Belief Functions

Suppose we have training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$. We again have two classes, i.e., $y_i \in \{-1, 1\}$. However, in contrast to training data considered in the previous sections, $\mathbf{x}_i$ are interval-valued, i.e., $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$. Here $\mathbf{A}_i = [\underline{a}_i^{(1)}, \overline{a}_i^{(1)}] \times ... \times [\underline{a}_i^{(m)}, \overline{a}_i^{(m)}]$, i.e., $\underline{a}_i^{(k)} \leq x_i^{(k)} \leq \overline{a}_i^{(k)}$, $k = 1, ..., m$; $\underline{a}_i^{(k)}$, $\overline{a}_i^{(k)}$ are bounds for values of the k-th feature in the i-th training example.

There are several ways in which one could deal with interval-valued data. In this paper, we consider the expected risk by interval-valued data in the framework of belief functions or Dempster-Shafer theory. Below, we give some basic definitions in the framework of belief functions.

Let $\mathcal{X}$ be a universal set under interest, usually referred to in evidence theory as the frame of discernment. Suppose n observations were made of an element $u \in \mathcal{X}$, each of which resulted in an imprecise (non-specific) measurement given by a set $\mathbf{A}$ of values. Let c_i denote the number of occurrences of the set $\mathbf{A}_i \subseteq \mathcal{X}$, and $\mathcal{P}o(\mathcal{X})$ the set of all subsets of $\mathcal{X}$ (power set of $\mathcal{X}$). A frequency function m, called basic probability assignment (BPA), can be defined such that [6, 16]:

$$m : \mathcal{P}o(\mathcal{X}) \to [0, 1], \quad m(\emptyset) = 0, \quad \sum_{\mathbf{A} \in \mathcal{P}o(\mathcal{X})} m(\mathbf{A}) = 1.$$

According to [6], this function can be obtained as follows:

$$m(\mathbf{A}_i) = c_i / n.$$

According to [16], the belief $Bel(A)$ and plausibility $Pl(\mathbf{A})$ of an event $\mathbf{A} \subseteq \mathcal{X}$ can be defined as

$$\mathrm{Bel}(\mathbf{A}) = \sum_{\mathbf{A}_i : \mathbf{A}_i \subseteq \mathbf{A}} m(\mathbf{A}_i),$$

$$\mathrm{Pl}(\mathbf{A}) = \sum_{\mathbf{A}_i : \mathbf{A}_i \cap \mathbf{A} \neq \varnothing} m(\mathbf{A}_i).$$

As pointed out in [9], a belief function can formally be defined as a function satisfying axioms which can be viewed as a weakening of the Kolmogorov axioms that characterize probability functions. Therefore, it seems reasonable to understand a belief function as a generalized probability function [6] and the belief $Bel(\mathbf{A})$ and plausibility $Pl(\mathbf{A})$ measures can be regarded as lower and upper bounds for the probability of $\mathbf{A}$, i.e., $\mathrm{Bel}(\mathbf{A}) \leq \Pr(\mathbf{A}) \leq \mathrm{Pl}(\mathbf{A})$. This implies that for a function $l(\mathbf{x})$, we can define the lower expectation $\underline{R}$ and the upper expectation $\overline{R}$ of the function $l(\mathbf{x})$ in the framework of belief functions as follows [13, 18]:

$$\underline{R} = \sum_{i=1}^{n} m(\mathbf{A}_i) \inf_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i),$$

$$\overline{R} = \sum_{i=1}^{n} m(\mathbf{A}_i) \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i).$$

The above definition provides a simpler way for determining the bounds for the expected risk. By using the assumption accepted in the empirical expected risk, we can conclude that $m(\mathbf{A}_i) = 1/n$ for all $i = 1, ..., n$. Hence, we get

$$\underline{R} = \frac{1}{n} \sum_{i=1}^{n} \inf_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i), \quad \overline{R} = \frac{1}{n} \sum_{i=1}^{n} \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i).$$

It follows from the above that we have the interval $[\underline{R}, \overline{R}]$ of the expected risk measure instead of its precise value. In order to use this interval in solving the classification problem, we have to determine a strategy of decision making which selects one point within this interval for searching optimal classification parameters $\mathbf{w}$, ξ and b in (1)-(2) or $\alpha_1, ..., \alpha_n$ in (3)-(4).

One of the well-known and popular ways for dealing with the interval of the expected risk is to use the minimax (pessimistic or robust) strategy. According to the minimax strategy, we select a probability distribution from the set of distributions such that the expected risk R achieves its maximum for fixed values of parameters. It should be noted that the "optimal" probability distributions may be different for different values of parameters. If to return to the interval $[\underline{R}, \overline{R}]$, then the minimax strategy assumes the largest risk, i.e., the upper bound $\overline{R}$. The minimax strategy can be explained in a simple way. We do not know a precise probability distribution and every distribution from their predefined set can be selected. Therefore, we should take the "worst" distribution providing the largest value of the expected risk. The minimax criterion can be interpreted as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case [15]. This criterion of decision making can be regarded as the well-known Γ-minimax [4, 7].

Robust algorithms have been exploited in classification problems due to the opportunity to avoid some strong assumptions underlying the standard classification algorithms. As pointed out by Xu et al. [26], the use of robust optimization in classification is not new. One of the popular robust classification algorithms is based on the assumption that inputs are subject to an additive noise, i.e., $\mathbf{x}_i^* = \mathbf{x}_i + \triangle\mathbf{x}_i$, where noise $\triangle\mathbf{x}_i$ is governed by a certain distribution. The simplest way for dealing with noise is to assume that every "true" data point is only known to belong to the interior of an Euclidean ball centered at the "nominal" data point $\mathbf{x}_i$ and each point can move around within the Euclidean ball. This algorithm has a very clear intuitive geometric interpretation [3]. One can see that the algorithm with interval-valued data and the robust algorithms [3, 26] are very close.

Finally, we can write the optimization problem for computing the optimal classification parameters ($\mathbf{w}$, ξ, b or α, b) as follows:

$$\overline{R} = \sup_{\mathbf{x}_i \in \mathbf{A}_i, i=1,...,n} \min_{\xi, \mathbf{w}, b} \sum_{i=1}^{n} l(\mathbf{x}_i),$$

4 L_2-Norm SVM by Interval-Valued Data

4.1 A General Problem and a New Kernel

Let us rewrite the objective function of problem (3)-(4) by taking into account interval-valued elements of the training set

$$\sup_{\mathbf{x}_i \in \mathbf{A}_i, i=1,...,n} \max_{\alpha} \times \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right). \quad (5)$$

This is a nonlinear optimization problem whose solution is generally a hard problem. Therefore, we propose a method for its solution which can reduce this problem to a finite set of linear programming problems.

One of the ideas underlying the proposed algorithm is to approximate the Gaussian kernel $K(\mathbf{x}, \mathbf{y})$ by another kernel which could somehow simplify the optimization problem. It is proposed to introduce a new kernel function

$$K_1(\mathbf{x}, \mathbf{y}) = \max\{0, 1 - \|\mathbf{x} - \mathbf{y}\|^1 / \sigma^2\}, \quad (6)$$

This is the well-known triangular kernel. Its main peculiarity is that K_1 is linear. This peculiarity allows us to solve the above optimization problem.

Let us fix the values of α and write the dual optimization problem with the introduced kernel K_1 having optimization variables $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$:

$$\inf_{\mathbf{x}_i, i=1,...,n} \left(\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j G_{ij} - \sum_{i=1}^{n} \alpha_i \right), \quad (7)$$

subject to

$$G_{ij} = \max\left\{0, 1 - \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^1}{\sigma^2}\right\}, \ i, j = 1, ..., n, \quad (8)$$

$$\underline{a}_i^{(k)} \le x_i^{(k)} \le \overline{a}_i^{(k)}, \ k = 1, ..., m, \ i = 1, ..., n. \quad (9)$$

Here G_{ij} is a new variable such that $G_{ij} = K_1(\mathbf{x}_i, \mathbf{x}_j)$.

We do not add constraints (4) to the set of constraints (8)-(9) because the values of α are fixed, i.e., we consider the problem with variables $\mathbf{x}_i$, $i = 1, ..., n$. One can see from (7)-(9) that this problem is linear in case of the triangular kernel. According to some general results from linear programming theory, an optimal solution to the above problem is achieved at extreme

points or vertices of the polytope produced only by constraints (8)-(9). This is the first main feature of the proposed approach and the main reason for introducing the triangular kernels. Moreover, it can be seen from constraints (8)-(9) that they do not depend on variables α. This implies that the extreme points do not depend on α. This is the second feature which is used below. The linearity of the above problem and the independence of vertices of the polytope of variables α allow us to represent the initial optimization problem with objective function (5) as a finite set of standard quadratic programming problems which are formed by substituting extreme points $\mathbf{x}_i^*$ of the polytope produced by (8)-(9) into the kernel function $K_1(\mathbf{x}_i, \mathbf{x}_j)$ instead of $\mathbf{x}_i$.

We do not consider details of the optimization problem representation as a set of quadratic programming problems. However, we discuss about a set of extreme points $\mathbf{x}_i^*$, $i = 1, ..., n$. It is interesting to note that G_{ij} totally depends on $\mathbf{x}_i$, $i = 1, ..., n$. This implies that only constraints for $\mathbf{x}_i$ define the extreme points which are trivial and coincide with the bounds of intervals $\mathbf{A}_i, i = 1, ..., n$. Moreover, we do not need to represent constraints (8) in the form of standard inequalities. By enumerating the extreme points $\mathbf{x}_i^*$, we compute all values G_{ij} and substitute them into objective function (7). Finally, we have one of the standard quadratic programming problems corresponding to one combination of bounds of intervals $\mathbf{A}_i, i = 1, ..., n$, whose solution can be found, for example, by means of the packages "kernlab", "e1071", "wSVM" in the R-project.

The optimal values of α correspond to the *largest* value of objective function (7) over all extreme points $\mathbf{x}_i^*$. After substituting the obtained solution into the expression for the decision function f, we get

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{10}$$

If we have $n^* \leq n$ interval-valued observations such that all their features are interval-valued, then we have to solve m^{n^*} quadratic programming problems. Of course, when n^* is rather large or the training examples are characterized by many interval-valued features m, then the obtained algorithm leads to extremely hard computations. Therefore, we propose below another classification algorithm whose complexity does not depend on the number of features m.

5 L_∞-Norm SVM

5.1 The Primal Form

We aim to find such a form of the SVM that would separate classification parameters, for example, $\alpha_1, ..., \alpha_n$, and intervals of $\mathbf{x}_1, ..., \mathbf{x}_n$. The SVM whose dual form satisfies this condition was proposed by Zhou et al. in [27]. It is based on using the L_∞-norm for writing the regularization term $\|\mathbf{w}\|$. The L_∞-norm leads to one of the possible variants of the SVM. Suppose that we have fixed precise values $\mathbf{x}_1, ..., \mathbf{x}_n$ from $\mathbf{A}_1, ..., \mathbf{A}_n$, respectively. According to [27], the optimization problem for computing the separating function parameters is of the form:

$$\min R = \min \left(-r + C \sum_{i=1}^{n} \xi_i \right), \tag{11}$$

subject to

$$y_j \left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq r - \xi_j, \ j = 1, ..., n, \tag{12}$$

$$-1 \leq \alpha_i \leq 1, \ i = 1, ..., n, \tag{13}$$

$$r \geq 0, \ \xi_j \geq 0, \ j = 1, ..., n. \tag{14}$$

Here α_j, ξ_j, $j = 1, ..., n$, r, b are optimization variables; $C \geq 0$ is a constant. One can see that the separating function f is written in constraints in terms of Lagrange multipliers α_i (see (10)).

The authors of [27] show that the VC dimension in this case is bounded and the separating function f can be approached by minimizing the empirical expected risk measure. It is indicated in [27] that training SVMs is simpler than the L_2-norm SVMs, especially for large-scale problems.

5.2 The Dual Form

It should be noted that the SVM algorithm proposed by Zhou et al. in [27] is an interesting version of the SVM. However, its direct use does not help us in solving the classification problem with interval-valued data, which is viewed as a optimization problem with the objective function

$$R = \max_{\mathbf{x}_i} \min_{r, b, \alpha_j, \xi_j} \left(-r + C \sum_{i=1}^{n} \xi_i \right),$$

and constraints (12)-(14), $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$.

Another advantage of the above SVM is very important for us. This is a special form of the dual problem which allows us to get a simple way for dealing with interval-valued data. Therefore, let us write the dual form for the above problem *by fixed* $\mathbf{x}_i$, $i = 1, ..., n$.

First of all, we replace the variables α_j in (11)-(14) by non-negative variables $a_j \geq 0$ and $c_j \geq 0$ in order to have only non-negative variables, i.e., $\alpha_j = a_j - c_j$. By using the standard method for constructing the

dual form, we get the following linear programming problem:

$$\max \sum_{i=1}^{n} (-g_i - h_i),$$

subject to $g_i, h_i \geq 0$,

$$\sum_{i=1}^{n} z_i \geq 1,\ 0 \leq z_j \leq C,\ j = 1, ..., n,$$

$$\sum_{i=1}^{n} z_i y_i = 0,$$

$$g_j - h_j = y_j \left(\sum_{i=1}^{n} z_i y_i K(\mathbf{x}_j, \mathbf{x}_i) \right),\ j = 1, ..., n.$$

Here $z = (z_1, ..., z_n)$, g_i, h_i, $i = 1, ..., n$, are optimization variables. By substituting the last constraint into the objective function, we get another objective function

$$\max \sum_{i=1}^{n} \left(-2g_i - y_i \left(\sum_{j=1}^{n} z_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \right).$$

Note that the maximum of the objective function is achieved when variable g_i is as small as possible, i.e., $g_i = 0$ for all $i = 1, ..., n$. Hence, we get the following simplified optimization problem

$$\min_{z} \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{n} z_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right), \tag{15}$$

subject to

$$\sum_{i=1}^{n} z_i \geq 1,\ 0 \leq z_j \leq C,\ j = 1, ..., n, \tag{16}$$

$$\sum_{i=1}^{n} z_i y_i = 0. \tag{17}$$

At first glance, the above dual form of the optimization problem does not differ from the primal form from the viewpoint of its use. However, one can see that constraints of the dual form do not contain terms $K(\mathbf{x}_i, \mathbf{x}_j)$ and do not contain vectors $\mathbf{x}_i$. This is a very important feature of the dual form, which allows us to introduce interval-valued data into the SVM. It should be noted that the same property cannot be obtained by considering the standard SVM based on the L_1-norm or the L_2-norm. Therefore, problem (15)-(17) plays a key role in constructing the algorithm of classification with interval-valued data.

5.3 Extreme Points of the Polytope

If we assume that the values of $K(\mathbf{x}_i, \mathbf{x}_j)$ are precisely known, i.e., the values $\mathbf{x}_i$, $i = 1, ..., n$, are precise or fixed, then one of the ways for solving the linear programming problem (15)-(17) is to find the extreme points or vertices of the polytope produced by constraints (16)-(17) and denoted by $z^{(l)}$, $l = 1, ..., N$. Here N is the total number of extreme points. An optimal solution to the above problem is achieved at one of the extreme points.

Proposition 1 *Let n_- and n_+ be numbers of training examples in classes labelled $y = -1$ and $y = 1$, respectively. All extreme points of the polytope produced by constraints (16)-(17) can be divided into two subsets. The first subset consists of*

$$N_1 = \sum_{t=\lceil 1/2C \rceil}^{\min(n_-, n_+)} \binom{n_-}{t} \binom{n_+}{t}$$

extreme points such that every point contains t elements from every class equal to C and other elements are 0. Here t is an integer determined from the condition

$$\frac{1}{2C} < t \leq \min(n_-, n_+).$$

Let s be an integer determined from the condition

$$\frac{1}{2C} - 1 \leq s < \min\left(\frac{1}{2C}, n_-, n_+ \right).$$

If there exists $s \geq 0$, then the second subset consists of

$$N_2 = (n_- - s)(n_+ - s) \binom{n_-}{s} \binom{n_+}{s}$$

extreme points such that every point contains s (if there exists $s > 0$) elements from every class equal to C, one element from every class is $1/2 - sC$, other elements are 0.

Proposition 1 can be regarded as an extension of Proposition 5 in [20].

5.4 L_∞-Norm SVM by Interval-Valued Data

Let us rewrite the objective function of problem (15)-(17) by taking into account the interval-valued elements of the training set

$$\min_{l=1,...,N} \min_{\mathbf{x}_i \in \mathbf{A}_i, i=1,...,n} \sum_{i=1}^{n} \sum_{j=1}^{n} z_j^{(l)} y_i y_j K(\mathbf{x}_i, \mathbf{x}_j). \tag{18}$$

By having extreme points, we can replace the optimization problem (15)-(17) by a set of $N = N_1 + N_2$

(see Proposition 1) objective functions provided above. However, we cannot solve the obtained set of optimization problems with variables $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$, in a simple way because the function $K(\mathbf{x}_i, \mathbf{x}_j)$ is nonlinear. Therefore, we again apply the idea of replacement the Gaussian kernel by its approximations. According to this idea, the Gaussian kernel can be approximated by another kernel which could somehow simplify the optimization problem. It is proposed to introduce two kernel functions

$$K_1(\mathbf{x}, \mathbf{y}) = \max\{0, 1 - \|\mathbf{x} - \mathbf{y}\|^1 / \sigma^2\},$$
$$K_2(\mathbf{x}, \mathbf{y}) = \max\{0, 1 - \|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2\}.$$

Both the kernels can be regarded as approximations of the Gaussian kernel. The first one is the triangular kernel considered in the previous sections. The second kernel is known as the Epanechnikov kernel.

Let us fix the values of $z^{(l)} = (z_1^{(l)}, ..., z_n^{(l)})$ and write the dual optimization problem with the introduced kernels K_r, $r = 1, 2$, for the l-th extreme point $z^{(l)}$ of (16)-(17) as follows:

$$\min_{\mathbf{x}_i, i=1,...,n} \sum_{i=1}^{n} \sum_{j=1}^{n} z_j^{(l)} y_i y_j G_{ij}, \tag{19}$$

subject to

$$G_{ij} = \max\left\{0, 1 - \|\mathbf{x}_i - \mathbf{x}_j\|^r / \sigma^2\right\}, \ i, j = 1, ..., n, \tag{20}$$
$$\underline{a}_i^{(k)} \le x_i^{(k)} \le \overline{a}_i^{(k)}, \ k = 1, ..., m, \ i = 1, ..., n. \tag{21}$$

Here $x_i^{(j)}$ is the value of the j-th feature of the i-th example; G_{ij} is a new variable such that $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$; r is 1 or 2 if we use the triangular or Epanechnikov kernel, respectively.

Finally, we get the set of N linear programming problems in case of using the triangular kernel. In case of the Epanechnikov kernel, we have the same number of quadratically constrained linear programs (QCLPs). It can be numerically solved by means of several methods, for example, by using the sequential quadratic programming [5] which efficiently implemented by means of SNOP [8]. The optimal values of $\mathbf{x}_i$ correspond to the *smallest* value of objective function (19) over all extreme points $\mathbf{x}_i^*$.

It is interesting to note that the number N of optimization problems does not depend on the number of features m. This is an important peculiarity of the proposed algorithm, which allows us to apply the algorithm to application problems with many features.

The function $f(\mathbf{x})$ can be rewritten in terms of Lagrange multipliers as

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i^*, \mathbf{x}) + b.$$

However, we do not know the optimal values of α_i because we used the dual optimization problem. Here we have two ways for computing the separating function. The first way is based on the fact that, by knowing the optimal solution z^* of the dual problem, the optimal solution α^* of the primal problem can be found by well-known algorithms. In particular, if the algorithm is implemented by using R-project, then the function "solveLP" in the package "linprog" has the output variable "con$dual" which provides the dual solution.

The second way is simpler. If we know precise optimal values $\mathbf{x}_i^*$ of intervals $\mathbf{A}_i$, $i = 1, ..., n$, then we can return to the initial problem (11)-(14) or to its dual form (15)-(17) and solve them by given fixed $\mathbf{x}_i^*$.

5.5 Comments about Constraints with the Triangular and Epanechnikov Kernels

It should be noted that constraints (20) are written in the short form. In order to solve the corresponding optimization problems, they have to be represented by the standard linear or quadratic inequalities. We do not consider in detail the representation of (20) because it is trivial due to the following two tricks.

First, the "standard" representation of (20) depends on the sign of the product $y_i y_j$. If $y_i y_j \ge 0$, then we get two constraints of the form:

$$G_{ij} \ge 1 - \|\mathbf{x}_i - \mathbf{x}_j\|^r / \sigma^2, \ G_{ij} \ge 0.$$

If $y_i y_j < 0$, then we use the well-known equation $\max(0, w) = w/2 + |w/2|$.

Second, in order to represent the absolute values, we use interesting results proposed by Beaumont [2]. According to [2], if we know some interval of values $[\underline{w}, \overline{w}] \subset \mathbb{R}$ of a variable w, then we can write $\forall w \in [\underline{w}, \overline{w}]$, $|w| \le uw + v$, where

$$u = \frac{|\overline{w}| - |\underline{w}|}{\overline{w} - \underline{w}}, \quad v = \frac{\overline{w}|\underline{w}| - \underline{w}|\overline{w}|}{\overline{w} - \underline{w}}.$$

6 Conclusion Remarks

New classification algorithms dealing with interval-valued training data have been proposed in the paper. A part of proposed algorithms using the triangular kernel instead of the Gaussian kernel comes to a finite set of simple linear programming problems whose solution does not meet difficulties. Another part using the triangular kernel comes to a finite set of quadratic programming problems whose solution are implemented by many standard procedures. The third part of algorithms is based on quadratically constrained linear programs which can be solved by using

the package "cplexAPI" available in several programming languages, for instance, in R-project.

It is important to note that the proposed algorithms indirectly find "optimal" points of intervals corresponding to the robust or maximin decision strategy. However, they fundamentally differ from the algorithm using some point-valued counterpart of intervals. The obtained "optimal" points of intervals are optimal in the sense that they maximize the expected classification error or risk if we apply the robust or maximin strategy. These "optimal" points compose a single probability distribution among a set of distributions produced by intervals in the framework of Dempster-Shafer theory.

Of course, all algorithms have a bottle neck which is their complexity. However, the proposed algorithms should not be used when a training set is large and intervals are rather small. Moreover, the algorithms based on the L_2-norm SVM should be used when the number of features is small. At the same time, the algorithms based on the L_∞-norm SVM do not depend on the number of features. It does not mean that the value m does not impact on the complexity of these algorithms. One can see from constraints (21) that the number of constraints strongly depends on m.

Finally, we have to stress on the main idea allowing us to construct the above algorithms. This is the replacement of the Gaussian kernel by the triangular and Epanechnikov kernels. This idea can be also used for constructing the support vector regression algorithms when dependent as well as independent variables are interval-valued.

Acknowledgement

The reported study was partially supported by RFBR, research project No. 15-01-01414-a.

References

[1] A. Antonucci, R. de Rosa, A. Giusti, and F. Cuzzolin. Temporal data classification by imprecise dynamical models. In *Proc. of the 8th International Symposium on Imprecise Probability: Theories and Applications*, pages 13–22, Compiegne, France, 2013. SIPTA.

[2] O. Beaumont. Solving interval linear systems with linear programming techniques. *Linear Algebra and Its Applications*, 281:293–309, 1998.

[3] A. Ben-Tal, L.E. Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, Princeton, New Jersey, 2009.

[4] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[5] P.T. Boggs and J.W. Tolle. Sequential quadratic programming. *Acta numerica*, 4:1–51, 1995.

[6] A.P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annales of Mathematical Statistics*, 38(2):325–339, 1967.

[7] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

[8] P.E. Gill, W. Murray, and M.A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Journal on Optimization*, 12(4):979–1006, 2002.

[9] J.Y. Halpern and R. Fagin. Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54(3):275–317, 1992.

[10] E. Hullermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

[11] H. Ishibuchi, H. Tanaka, and N. Fukuoka. Discriminant analysis of multi-dimensional interval data and its application to chemical sensing. *International Journal of General Systems*, 16(4):311–329, 1990.

[12] E.A. Lima Neto and F.A.T. de Carvalho. Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis*, 52:1500–1515, 2008.

[13] H.T. Nguyen and E.A. Walker. On decision making using belief functions. In R.Y. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 311–330. Wiley, New York, 1994.

[14] P. Nivlet, F. Fournier, and J.-J. Royer. Interval discriminant analysis: An efficient method to integrate errors in supervised pattern recognition. In *Second International Symposium on Imprecise Probabilities and Their Applications*, pages 284–292, Ithaca, NY, USA, 2001.

[15] C.P. Robert. *The Bayesian Choice*. Springer, New York, 1994.

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

[17] A. Silva and P. Brito. Linear discriminant analysis for interval data. *Computational Statistics*, 21:289–308, 2006.

[18] T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4(5):391–418, 1990.

[19] A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems*. W.H. Winston, Washington DC, 1977.

[20] L.V. Utkin. A framework for imprecise robust one-class classification models. *International Journal of Machine Learning and Cybernetics*, 5(3): 379-393, 2014.

[21] L.V. Utkin and F.P.A. Coolen. Interval-valued regression and classification models in the framework of machine learning. In F. Coolen, G. de Cooman, Th. Fetz, and M. Oberguggenberger, editors, *Proc. of the Seventh Int. Symposium on Imprecise Probabilities: Theories and Applications, ISIPTA'11*, pages 371–380, Innsbruck, Austria, 2011. SIPTA.

[22] L.V. Utkin, Y.A. Zhuk, and A.I. Chekh. A robust one-class classification model with interval-valued data based on belief functions and minimax strategy. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 8556 of *Lecture Notes in Computer Science*, pages 107–118. Springer, 2014.

[23] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[24] J. Wang, H. Lu, K.N. Plataniotis, and J. Lu. Gaussian kernel optimization for pattern classification. *Pattern Recognition*, 42(7):1237 – 1247, 2009.

[25] W. Wang, Z. Xu, W. Lu, and X. Zhang. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, 55(3):643–663, 2003.

[26] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10(7):1485–1510, 2009.

[27] Weida Zhou, Li Zhang, and Licheng Jiao. Linear programming support vector machines. *Pattern Recognition*, 35(12):2927–2936, 2002.

Modelling Indifference with Choice Functions

Arthur Van Camp and **Gert de Cooman**
Ghent University
SYSTeMS Research Group
{Arthur.VanCamp,Gert.deCooman}@UGent.be

Enrique Miranda
University of Oviedo
Department of Statistics and Operations Research
mirandaenrique@uniovi.es

Erik Quaeghebeur
Centrum Wiskunde & Informatica
Algorithms & Complexity Group
Erik.Quaeghebeur@cwi.nl

Abstract

We investigate how to model indifference with choice functions. We take the coherence axioms for choice functions proposed by Seidenfeld, Schervisch and Kadane as a source of inspiration, but modify them to strengthen the connection with desirability. We discuss the properties of choice functions that are coherent under our modified set of axioms and the connection with desirability. Once this is in place, we present an axiomatisation of indifference in terms of desirability. On this we build our characterisation of indifference in terms of choice functions.

Keywords. Choice function, coherence, indifference, set of desirable gambles, maximality, E-admissibility.

1 Introduction

The language of classical probability—(probability) mass functions, say—is insufficiently versatile and powerful to describe certain aspects of beliefs, such as indecision. Imprecise probability uncertainty models, such as coherent lower previsions and coherent sets of desirable gambles, are often used to remedy this. Coherent sets of desirable gambles in particular play a crucial role in theories of conservative reasoning [16], predictive inference [10], credal networks [6], and so on. They have many advantages, such as mathematical elegance and the lack of problems for conditioning on an event with (lower) probability zero. However, they are not capable of modelling beliefs corresponding to 'or' statements, such as the belief that a coin has two equal sides of unknown type—twice heads or twice tails. It turns out such more general types of assessments can be modelled with choice functions.

To allow for incomparability, Seidenfeld, Schervisch and Kadane [23] introduce axioms for rational choice expressed by choice functions that are a weakened version of the ones suggested by Rubin [18]. We modify them slightly, in order to allow for Walley–Sen maximality [28, 26] to be coherent, and we drop their Archimedean continuity axiom to allow for a more direct connection with coherent sets of desirable gambles. We introduce our notion of coherence for choice functions in Section 2. We work with abstract vectors (called options), rather than horse lotteries or gambles: this will allow us to deal with indifference without too many mathematical difficulties, later on in this paper. Because we are interested in conservative reasoning with coherent choice functions, we introduce an 'is not more informative than' ordering, which allows us to consider the most conservative choice function compatible with an assessment as an infimum associated with this partial order.

In Section 3, we relate our theory of coherent choice functions to coherent sets of desirable options, and identify the most conservative coherent choice function compatible with a coherent set of desirable options as the one associated with Walley–Sen maximality: it selects the undominated options under the strict partial order generated by a coherent set of desirable options, and is therefore fully based on binary choice.

In Section 4, we show that there are other general classes of coherent choice functions not based on binary choice, and we relate them to each other.

An important aspect of any uncertainty theory is how it deals with indifference. Adding indifference to the picture typically reduces the complexity of the modelling effort. Also, knowing how to model indifference opens up a path towards modelling symmetry, which has many important practical applications. As an example of both aspects, the permutation symmetry that lies behind exchangeability has important applications in statistical modelling, and reduces the complexity of the modelling effort, as is exemplified by de Finetti's representation theorem [12]. Our treatment here lays the foundation for dealing with, say, exchangeability for choice functions.

In Section 5, we give an intuitive definition of indifference for choice functions that reduces to the existing account for sets of desirable gambles (options). We exhibit the power and simplicity of our definition of indifference in an interesting example.

2 Choice Functions on Option Sets

Consider a real vector space $\mathcal{V}$, provided with the vector addition + and scalar multiplication. We denote by 0 the additive identity, or null vector. For any subsets O_1 and O_2 of $\mathcal{V}$ and any λ in $\mathbb{R}$, we define $\lambda O_1 := \{\lambda u : u \in O_1\}$ and $O_1 + O_2 := \{u + v : u \in O_1, v \in O_2\}$. Elements u of $\mathcal{V}$ are intended as abstract representations of *options* amongst which a subject can express his preferences, by specifying, as we shall see below, choice functions. Mostly, options will be real-valued maps on the possibility space, also called *gambles*. We want to work with the more abstract notion of options—elements of some general vector space—because in Section 5, we will need choice functions defined on *equivalence classes* of options. These constitute a vector space—and hence are abstract options themselves—but can no longer be interpreted easily and directly as gambles.

We denote by $\mathcal{Q}(\mathcal{V})$ the set of all non-empty *finite* subsets of $\mathcal{V}$, a strict subset of the power set of $\mathcal{V}$. Elements O of $\mathcal{Q}(\mathcal{V})$ are the option sets amongst which a subject can choose his preferred options. When it is clear what vector space of options we are talking about, we will omit explicit mention of $\mathcal{V}$ and simply write $\mathcal{Q}$.

Definition 1. A *choice function* C on $\mathcal{Q}$ is a map

$$C : \mathcal{Q} \to \mathcal{Q} \cup \{\varnothing\} : O \mapsto C(O) \text{ such that } C(O) \subseteq O.$$

We collect all choice functions in the set $\mathcal{C}$.

The idea underlying this definition is that a choice function C selects the set $C(O)$ of 'best' options in the *option set* O. Our definition resembles the one commonly used in the literature [1, 23, 25], except for a not unusual restriction to *finite* option sets [13, 19, 24].

2.1 Rationality Axioms

Seidenfeld et al. [23, Section 3] call a choice function C *coherent* if there is a non-empty set of probability-utility pairs $\mathcal{S}$ such that $C(O)$ is the set of options in O that maximise expected utility for some probability-utility pair in $\mathcal{S}$. They also provide an axiomatisation for this type of coherence, based on the one for binary preferences [2]. One of their axioms is an 'Archimedean' continuity condition, and another one is a convexity condition, necessary for the connection with a set of probability-utility pairs.

We prefer to define coherence directly in terms of axioms, without reference to probabilities and utilities. In such a context, we see no compelling reason to adopt an Archimedean axiom, all the more so because we are interested in establishing the connection between choice functions and Walley's sets of desirable gambles Walley [29], which violate this axiom. Furthermore, the convexity condition does not allow for choice functions that select the undominated options under some partial ordering, which is something we find natural, and shall need later on.

We will weaken their axioms in Section 2.1.2 by dropping the Archimedean condition and by replacing their convexity condition with a weaker variant. On the other hand, our second axiom is a strengthened version of theirs, needed for the conditioning we intend to discuss in a later paper.

2.1.1 Some Useful Definitions

We call $\mathbb{N}$ the set of all (positive) integers, and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Also, we call $\mathbb{R}_{>0}$ the set of all (strictly) positive real numbers, and $\mathbb{R}_{\geq 0} := \mathbb{R}_{>0} \cup \{0\}$.

Given any subset O of $\mathcal{V}$, we define the *linear hull* $\operatorname{span}(O)$ as the set of all finite linear combinations of elements of O:

$$\operatorname{span}(O) := \left\{ \sum_{k=1}^{n} \lambda_k u_k : n \in \mathbb{N}, \lambda_k \in \mathbb{R}, u_k \in O \right\} \subseteq \mathcal{V},$$

the *positive hull* $\operatorname{posi}(O)$ as the set of all positive finite linear combinations of elements of O:

$$\operatorname{posi}(O) := \left\{ \sum_{k=1}^{n} \lambda_k u_k : n \in \mathbb{N}, \lambda_k \in \mathbb{R}_{>0}, u_k \in O \right\} \subseteq \mathcal{V},$$

and the *convex hull* $\operatorname{CH}(O)$ as the set of convex combinations of elements of O:

$$\operatorname{CH}(O) := \left\{ \sum_{k=1}^{n} \alpha_k u_k : n \in \mathbb{N}, \alpha_k \in \mathbb{R}_{\geq 0}, \sum_{k=1}^{n} \alpha_k = 1, u_k \in O \right\} \subseteq \mathcal{V}.$$

A subset O of $\mathcal{V}$ is called a *convex cone* if it is closed under positive finite linear combinations, i.e. if $\operatorname{posi}(O) = O$. A convex cone $\mathcal{K}$ is called *proper* if $\mathcal{K} \cap -\mathcal{K} = \{0\}$.

With any proper convex cone $\mathcal{K} \subseteq \mathcal{V}$ such that $0 \in \mathcal{K}$, we associate an ordering $\preceq$ on $\mathcal{V}$, defined for all u and v in $\mathcal{V}$ as follows:

$$u \preceq_{\mathcal{K}} v \Leftrightarrow v - u \in \mathcal{K} \Leftrightarrow 0 \preceq_{\mathcal{K}} v - u \Leftrightarrow u - v \preceq_{\mathcal{K}} 0.$$

We also write $u \succeq_{\mathcal{K}} v$ for $v \preceq_{\mathcal{K}} u$. The ordering $\preceq_{\mathcal{K}}$ is actually a *vector ordering*: it is a partial order (reflexive, antisymmetric and transitive) that satisfies the following two characteristic properties:

$$u_1 \preceq_{\mathcal{K}} u_2 \Leftrightarrow u_1 + v \preceq_{\mathcal{K}} u_2 + v; \tag{1}$$

$$u_1 \preceq_{\mathcal{K}} u_2 \Leftrightarrow \lambda u_1 \preceq_{\mathcal{K}} \lambda u_2, \tag{2}$$

for all u_1, u_2, v in $\mathcal{V}$ and λ in $\mathbb{R}_{>0}$. Conversely, given a vector ordering $\preceq$, the proper convex cone $\mathcal{K}$ from which it is derived can always be retrieved by $\mathcal{K} = \{u \in \mathcal{V} : u \succeq 0\}$. When the abstract options are gambles, $\mathcal{K}$ will usually be the non-negative orthant, and the ordering $\preceq$ is then pointwise. When the options are equivalence classes, as in Section 5.2, the ordering will be the induced ordering on equivalence classes, as defined in Eq. (10).

The vector space of options $\mathcal{V}$, ordered by the vector ordering $\preceq_{\mathcal{K}}$, is called an *ordered vector space* $\langle \mathcal{V}, \preceq_{\mathcal{K}} \rangle$. We

shall refrain from explicitly mentioning the actual proper convex cone $\mathcal{K}$ we are using, and simply write $\mathcal{V}$ to mean the ordered vector space, and $\preceq$ for the associated vector ordering.

Finally, with any vector ordering $\preceq$, we associate the strict partial ordering $\prec$ as follows:

$$u \prec v \Leftrightarrow (u \preceq v \text{ and } u \neq v) \Leftrightarrow v - u \in \mathcal{K} \setminus \{0\} \text{ for all } u, v \text{ in } \mathcal{V}.$$

We call u *positive* if $u \succ 0$, and collect all positive options in the convex cone $\mathcal{V}_{>0} := \mathcal{K} \setminus \{0\}$.

2.1.2 Rationality axioms for choice functions

Definition 2. We call a choice function C on $\mathcal{Q}(\mathcal{V})$ *coherent* if for all O, O_1, O_2 in $\mathcal{Q}$, u, v in $\mathcal{V}$ and λ in $\mathbb{R}_{>0}$:

C_1. $C(O) \neq \varnothing$;
C_2. if $u \prec v$ then $\{v\} = C(\{u, v\})$;
C_3. a. if $C(O_2) \subseteq O_2 \setminus O_1$ and $O_1 \subseteq O_2 \subseteq O$ then $C(O) \subseteq O \setminus O_1$;
b. if $C(O_2) \subseteq O_1$ and $O \subseteq O_2 \setminus O_1$ then $C(O_2 \setminus O) \subseteq O_1$;
C_4. a. if $O_1 \subseteq C(O_2)$ then $\lambda O_1 \subseteq C(\lambda O_2)$;
b. if $O_1 \subseteq C(O_2)$ then $O_1 + \{u\} \subseteq C(O_2 + \{u\})$;
C_5. if $O \subseteq \mathrm{CH}(\{u, v\})$ then $\{u, v\} \cap C(O \cup \{u, v\}) \neq \varnothing$.[1]

We collect all coherent choice functions on $\mathcal{V}$ in the set $\bar{\mathcal{C}}$.

Parts C_3a and C_3b of Axiom C_3 are respectively known as Sen's condition α and Aizerman's condition. They are more commonly written as, respectively:

$$(O_1 \cap C(O_2) = \varnothing \text{ and } O_1 \subseteq O_2 \subseteq O) \Rightarrow O_1 \cap C(O) = \varnothing \quad (3)$$

and

$$(O_1 \cap C(O_2) = \varnothing \text{ and } O \subseteq O_1) \Rightarrow O_1 \cap C(O_2 \setminus O) = \varnothing \quad (4)$$

for all O, O_1, O_2 in $\mathcal{Q}$.

Proposition 1. *The following statements hold for any coherent choice function C:*

(i) $\lambda C(O) + \{u\} = C(\lambda O + \{u\})$ *for all O in $\mathcal{Q}$, λ in $\mathbb{R}_{>0}$ and u in $\mathcal{V}$;*
(ii) *for all u_1, u_2 in $\mathcal{V}$ such that $u_1 \preceq u_2$, all O in $\mathcal{Q}$ and all v in $O \setminus \{u_1, u_2\}$:*
   a. *if $u_2 \in O$ and $v \notin C(O \cup \{u_1\})$ then $v \notin C(O)$;*
   b. *if $u_1 \in O$ and $v \notin C(O)$ then $v \notin C(\{u_2\} \cup O \setminus \{u_1\})$;*
(iii) *C is* insensitive to the omission of non-chosen options [9, Definition 11]*: $C(O') = C(O)$ for all O, O' in $\mathcal{Q}$ such that $C(O) \subseteq O' \subseteq O$;*
(iv) *$C(C(O)) = C(O)$ for all O in $\mathcal{Q}$.*

For Bradley [3], any choice function must at least satisfy property (iv). Seidenfeld *et al.* [23] impose the two properties (ii)a and (ii)b as rationality axioms [23, Axiom 4]. Our proofs for them rely quite heavily on, amongst other things, Axiom C_2, which is a strengthened version of another of their rationality axioms. This does not imply, however, that our rationality axioms are stronger than theirs, since we have dropped their Archimedean axiom [23, Axiom 3], and replaced their convexity axiom [23, Axiom 2b] by our strictly weaker variant C_5.

2.2 The 'Is Not More Informative Than' Relation

Because we are interested in *conservative reasoning* with choice functions, we look for the implications of a given assessment that are as 'uninformative' as possible. Therefore, we need some binary relation $\sqsubseteq$ on $\mathcal{C}$, having the specific interpretation of being 'not more informative than', or, in other words, 'at least as uninformative as'.

Definition 3. Given two choice functions C_1 and C_2 in $\mathcal{C}$, we call C_1 *not more informative than* C_2—and we write $C_1 \sqsubseteq C_2$—if $(\forall O \in \mathcal{Q}) C_1(O) \supseteq C_2(O)$.

This intuitive way of ordering choice functions is also used by Bradley [3], and in earlier work by the authors [27]. The underlying idea is that a choice function is more informative when it chooses more specifically, or restrictively, amongst the available options.

Since by definition $\sqsubseteq$ is a product ordering of set inclusions, the following result is immediate [5].

Proposition 2. *The structure $(\mathcal{C}; \sqsubseteq)$ is a complete lattice:*

(i) *it is a* partially ordered set*, or* poset*, meaning that the binary relation $\sqsubseteq$ on $\mathcal{C}$ is* reflexive*,* antisymmetric *and* transitive*;*
(ii) *for any subset $\mathcal{C}'$ of $\mathcal{C}$, its infimum* $\inf \mathcal{C}'$ *and its supremum* $\sup \mathcal{C}'$ *with respect to the ordering $\sqsubseteq$ exist in $\mathcal{C}$, and are given by* $\inf \mathcal{C}'(O) = \bigcup_{C \in \mathcal{C}'} C(O)$ *and* $\sup \mathcal{C}'(O) = \bigcap_{C \in \mathcal{C}'} C(O)$ *for all O in $\mathcal{Q}$.*

The idea is that $\inf \mathcal{C}'$ is the most informative model that is not more informative than any of the models in $\mathcal{C}'$, and $\sup \mathcal{C}'$ the least informative model that is not less informative than any of the models in $\mathcal{C}'$.

We also consider the poset $(\bar{\mathcal{C}}; \sqsubseteq)$, where $\bar{\mathcal{C}} \subseteq \mathcal{C}$ inherits the partial order $\sqsubseteq$ from $\mathcal{C}$.

Proposition 3. *$(\bar{\mathcal{C}}; \sqsubseteq)$ is complete infimum-semilattice: $\bar{\mathcal{C}}$ is closed under arbitrary non-empty infima, so* $\inf \mathcal{C}' \in \bar{\mathcal{C}}$ *for any non-empty subset $\mathcal{C}'$ of $\bar{\mathcal{C}}$.*

[1] This axiom is not needed to prove the results in this paper, and all results remain valid without it. We include it because it seems reasonable: the version with rational convex combinations can be derived from our other axioms, so C_5 amounts to requiring some very weak continuity. More importantly, this axiom is instrumental for the proofs of some results not included in this paper due to space limitations; because of this, we prefer to keep a unified set of axioms in all of our work in this topic.

3 Relation with Sets of Desirable Options

Choice functions cannot be characterised using pairwise comparison of options,[2] meaning that a binary relation on options does not uniquely determine a choice function. In this section, we study the ones that do correspond to a pairwise comparison of options.

3.1 Sets of Desirable Options

Sets of desirable options are a generalisation of *sets of desirable gambles*. Gambles are real-valued maps on a possibility space $\mathcal{X}$, interpreted as uncertain rewards. Such gambles can be seen as vectors in the vector space $\mathbb{R}^{\mathcal{X}}$. Here we generalise this notion by looking at a general (abstract) vector space $\mathcal{V}$ of (abstract) options, rather than gambles. We shall see that sets of desirable options amount to a pairwise comparison of options and therefore correspond to a special kind of choice functions.

A set of desirable options D is simply a subset of the vector space of options $\mathcal{V}$. We collect all sets of desirable options in the set $\mathcal{D}$. As we did for choice functions, we pay special attention to *coherent* sets of desirable options.

Definition 4. A set of desirable options D is called coherent if for all u and v in $\mathcal{V}$ and λ in $\mathbb{R}_{>0}$:

D_1. $0 \notin D$;
D_2. $\mathcal{V}_{>0} \subseteq D$;
D_3. if $u \in D$ then $\lambda u \in D$;
D_4. if $u, v \in D$ then $u + v \in D$.

We collect all coherent sets of desirable options in the set $\bar{\mathcal{D}}$.

Axioms D_3 and D_4 turn coherent sets of desirable options D into cones—$\operatorname{posi}(D) = D$. They include the positive options due to Axiom D_2, and do not contain the zero option due to Axiom D_1. As an immediate consequence, their intersection with $\mathcal{V}_{<0} := -\mathcal{V}_{>0}$ is empty. As usual, we may associate with the cone D a strict partial order $\prec$ on $\mathcal{V}$, by letting $u \prec v \Leftrightarrow 0 \prec v - u \Leftrightarrow v - u \in D$, so $D = \{u \in \mathcal{V} : 0 \prec u\}$ [8, 16].

3.2 The 'Is Not More Informative Than' Relation

As for choice functions, sets of desirable options can be ordered according to a 'not more informative than' relation.

Definition 5. Given two sets of desirable options D_1, D_2 in $\mathcal{D}$, we call D_1 *not more informative than* D_2 when $D_1 \subseteq D_2$.

Because the ordering of sets of desirable options $\subseteq$ is just set inclusion, it is a partial ordering on $\mathcal{D}$, and the poset $(\mathcal{D}; \subseteq)$ is a complete lattice, with supremum operator $\bigcup$, and infimum operator $\bigcap$.

[2] An equivalent representation of a coherent choice function C is a binary relation $\lhd$ on $\mathcal{Q}$—on *sets of options*—defined through $O_1 \lhd O_2 \Leftrightarrow O_1 \cap C(O_1 \cup O_2) = \varnothing$ for all O_1, O_2 in $\mathcal{Q}$. This binary relation $\lhd$ is a strict partial order on $\mathcal{Q}$ [14].

Proposition 4. *$(\bar{\mathcal{D}}; \subseteq)$ is a complete infimum-semilattice, or alternatively, $\bar{\mathcal{D}}$ is an intersection structure—closed under arbitrary non-empty intersections.*

Proposition 4 guarantees us that there is a unique least informative set of desirable options in $\bar{\mathcal{D}}$, called the *vacuous set of desirable options* D_{v}.

Proposition 5. *The least informative (smallest) set of desirable options D_{v} is given by $D_{\mathrm{v}} := \mathcal{V}_{>0}$.*

It will be useful to also consider the maximally informative, or *maximal* coherent sets of desirable options.[3] They are the undominated elements of the complete infimum-semilattice $(\bar{\mathcal{D}}; \subseteq)$; we collect them into a set $\hat{\mathcal{D}}$:

$$\hat{\mathcal{D}} := \{D \in \bar{\mathcal{D}} : (\forall D' \in \bar{\mathcal{D}})(D \subseteq D' \Rightarrow D = D')\}.$$

First we prove a useful proposition that will allow us to characterise these maximal elements very elegantly.

Proposition 6. *Given any coherent set of desirable options D and any non-zero option $u \notin D$, then $\operatorname{posi}(D \cup \{-u\})$ is a coherent set of desirable options.*

Proposition 7. *A coherent set of desirable options D is maximal if and only if*

$$(\forall u \in \mathcal{V} \setminus \{0\})(u \in D \text{ or } -u \in D). \tag{5}$$

Proposition 8. *For any coherent set of desirable options D, its set of dominating maximal coherent sets of desirable options $\hat{\mathcal{D}}_D := \{\hat{D} \in \hat{\mathcal{D}} : D \subseteq \hat{D}\}$ is non-empty.*

Proposition 9. *$(\bar{\mathcal{D}}; \subseteq)$ is dually atomic, meaning that any coherent set of desirable options D is the infimum of its non-empty set of dominating maximal coherent sets of desirable options $\hat{\mathcal{D}}_D$: $D = \inf \hat{\mathcal{D}}_D$.*

3.3 Connection Between Choice Functions and Sets of Desirable Options

In this section, we establish a connection between choice functions and sets of desirable options.

Definition 6. Given a choice functions C, we say that an option v is *chosen above* some option u whenever $u \notin C(\{u, v\})$, or equivalently whenever $v \neq u$ and $\{v\} = C(\{u, v\})$. Similarly, given a set of desirable options D, we say that an option v is *preferred to* some option u whenever $v - u \in D$, or equivalently, $u \prec v$. We call a choice function C and a set of desirable options D *compatible* when

$$u \notin C(\{u, v\}) \Leftrightarrow v - u \in D \Leftrightarrow u \prec v \text{ for all } u, v \in \mathcal{V}.$$

Compatibility means that the behaviour of a choice function *restricted to pairs of options* reflects the behaviour of a

[3] The discussion in the rest of this section is based on similar discussions about sets of desirable gambles [8, 4, 17]. We repeat the details here *mutatis mutandis* to make the paper more self-contained.

set of desirable options.[4] So, a choice function C will have at most one compatible set of desirable options, whereas conversely, a set of desirable options D may have many compatible choice functions: compatibility only directly influences the behaviour of a choice function on doubletons.

3.3.1 From Choice Functions to Desirability

We begin by studying the properties of the set of desirable options compatible with a given coherent choice function.

Proposition 10. *Given a coherent choice function C in $\bar{\mathcal{C}}$, there is a unique compatible coherent set of desirable options D_C, given by $D_C := \{u \in \mathcal{V} : 0 \notin C(\{0, u\})\}$.*

3.3.2 From Desirability to Choice Functions

We collect in $\bar{\mathcal{C}}_D$ all the compatible coherent choice functions with the given coherent set of desirable options D:

$$\begin{aligned}\bar{\mathcal{C}}_D :=& \{C \in \bar{\mathcal{C}} : (\forall u, v \in \mathcal{V})(v \notin C(\{u, v\}) \Leftrightarrow u - v \in D)\} \\ =& \{C \in \bar{\mathcal{C}} : D_C = D\}.\end{aligned}$$

Proposition 11. *Given a coherent set of desirable options D, the infimum—most uninformative element—$\inf \bar{\mathcal{C}}_D$ of its set of compatible coherent choice functions $\bar{\mathcal{C}}_D$ is the coherent choice function C_D, defined by*

$$\begin{aligned}C_D(O) :=& \{u \in O : (\forall v \in O) v - u \notin D\} \\ =& \{u \in O : (\forall v \in O) u \not\prec v\} \text{ for all } O \text{ in } \mathcal{Q}. \qquad (6)\end{aligned}$$

The coherent choice function C_D is the least informative choice function that is compatible with a coherent set of desirable options D: it is based on the binary ordering represented by D *and nothing else*. As we shall see in Proposition 17, there are other coherent choice functions C compatible with D, but they encode more information than just the binary ordering represented by D. Proposition 11 is especially interesting because it shows that the most conservative choice function based on a strict partial order of options, is the choice function based on *maximality*—the one that selects the *undominated* options under the strict partial order $\prec$ associated with a coherent set of desirable options D. Any choice function that is based on maximality under such a strict partial order is coherent.

Proposition 3 guarantees that there is a unique smallest—least informative—coherent choice function. We shall call it the *vacuous choice function*, and denoted it by C_{v}.

Proposition 12. *The vacuous choice function C_{v} is given by $C_{\mathrm{v}}(O) = C_{D_{\mathrm{v}}}(O) = \{u \in O : (\forall v \in O) u \not\prec v\}$ for all O in $\mathcal{Q}$. It selects from any set of options the ones that are undominated under the strict vector ordering $\prec$.*

[4] See Ref. [21] for an axiomatisation of imprecise preferences in the context of binary comparisons of horse lotteries.

Example 1. Consider, as a simple example, the case that the vector ordering is total, meaning that for any u, v in $\mathcal{V}$, either $u \prec v$, $v \prec u$ or $u = v$. It then follows from Proposition 12 that, for any coherent choice function C, $C(O) \subseteq C_{\mathrm{v}}(O) = \max O$ for all $O \in \mathcal{Q}$, where $\max O$ is the unique largest element of the finite option set O according to the strict total ordering $\prec$. But then Axiom C_1 guarantees that $C(O) = C_{\mathrm{v}}(O) = \max O$ for all $O \in \mathcal{Q}$, so C_{v} is the *only* coherent choice function. □

3.3.3 Properties of the Relation Between Choice Functions and Desirability

Since sets of desirable options represent only pairwise comparison, and are therefore generally less expressive than choice functions, we expect that going from a choice function to a compatible set of desirable options leads to a loss of information, whereas going the opposite route does not. This is confirmed by Propositions 13 and 14, but in particular by their Corollary 15. Example 2 in Section 4 further on shows that the inequalities in these results can be strict.

Proposition 13. *Consider any set of coherent choice functions $\mathcal{C}' \subseteq \bar{\mathcal{C}}$. Then $D_{\inf \mathcal{C}'} = \inf\{D_C : C \in \mathcal{C}'\}$ and $C_{\inf\{D_C : C \in \mathcal{C}'\}} \sqsubseteq \inf \mathcal{C}'$, and therefore also $C_{D_{\inf \mathcal{C}'}} \sqsubseteq \inf \mathcal{C}'$.*

Proposition 14. *Consider any set of coherent sets of desirable options $\mathcal{D}' \subseteq \bar{\mathcal{D}}$ and any coherent set of desirable options D'. Then $D_{\inf\{C_D : D \in \mathcal{D}'\}} = \inf \mathcal{D}'$ and therefore $D_{C_{D'}} = D'$. Moreover, $C_{\inf \mathcal{D}'} \sqsubseteq \inf\{C_D : D \in \mathcal{D}'\}$.*

Corollary 15. *Consider any coherent set of desirable options $D \in \bar{\mathcal{D}}$ and any coherent choice function $C \in \bar{\mathcal{C}}$. Then $D = D_{C_D}$ and $C_{D_C} \sqsubseteq C$.*

4 Other Types of Coherent Choice Functions

There are other types of coherent choice functions than the ones '*based on maximality*', derived from a coherent set of desirable options by selecting undominated elements as in Eq. (6). For instance, any infimum of such coherent choice functions is still coherent.

Definition 7. For any set of coherent sets of desirable options $\mathcal{D}' \subseteq \bar{\mathcal{D}}$, we define the *'infimum of maximality'* choice function as $C_{\mathcal{D}'} := \inf\{C_D : D \in \mathcal{D}'\}$.

Proposition 16. *Consider any set of coherent sets of desirable options $\mathcal{D}' \subseteq \bar{\mathcal{D}}$, then $C_{\mathcal{D}'}$ is a coherent choice function.*

We now consider two special cases of these infimum of maximality choice functions. In Definition 8, we focus only on sets of *maximal* coherent sets of desirable options.

Definition 8. If $\mathcal{D}' \subseteq \hat{\mathcal{D}}$ is a set of *maximal* coherent set of desirable options, the coherent choice function $C_{\mathcal{D}'}$ is called *M-admissible*. We shall also denote it by $C^M_{\mathcal{D}'}$ as a reminder that the infimum is taken over maximal sets.

In particular, we can consider the M-admissible choice functions for the set $\mathcal{D}' = \hat{\mathcal{D}}_D$ of all maximal coherent set of desirable options that include a coherent set of desirable options D. In order not to burden the notation, we let

$$C_D^M := C_{\hat{\mathcal{D}}_D}^M = \inf\{C_{\hat{D}} : \hat{D} \in \hat{\mathcal{D}} \text{ and } D \subseteq \hat{D}\}. \tag{7}$$

Proposition 17. *Consider any coherent set of desirable options $D' \in \bar{\mathcal{D}}$. Then $D' = D_{C_{D'}^M}$ and $C_{D'} \sqsubseteq C_{D'}^M$.*

The inequality in Proposition 17 can be strict—meaning that $C_{D'} \sqsubset C_{D'}^M$ for some coherent set of desirable options D'—as is shown in Example 3.

As another special case, we consider choice functions associated with Levi's [15, Chapter 5] notion of E-admissibility, as suggested by Seidenfeld *et al.* [23], and Troffaes [26]. They are based on a non-empty *set of mass functions*. Consider a *finite* possibility space $\mathcal{X}$, and maps from $\mathcal{X}$ to $\mathbb{R}$ (also called *gambles*), forming the vector space $\mathcal{V} = \mathbb{R}^{\mathcal{X}}$ of *finite* dimension $|\mathcal{X}|$. The vector ordering $\leq$ we associate with this vector space is the pointwise ordering of real numbers: $u \leq v \Leftrightarrow (\forall x \in \mathcal{X}) u_x \leq v_x$, where, for instance, u_x is the x-component of the option u. We call any map $p : \mathcal{V} \to \mathbb{R}$ with $(\forall x \in \mathcal{X}) p(x) \geq 0$ and $\sum_{x \in \mathcal{X}} p(x) = 1$ a (probability) *mass function*, and we associate an *expectation* E_p with p by letting $\mathrm{E}_p(u) := \sum_{x \in \mathcal{X}} p(x) u_x$ for all u in $\mathcal{V}$.

With a mass function p, we associate a set of desirable options

$$D_p := \mathcal{V}_{>0} \cup \{u \in \mathcal{V} : \mathrm{E}_p(u) > 0\} \tag{8}$$

and a choice function C_p defined for all O in $\mathcal{Q}$ by

$$C_p(O) := \{u \in O : (\forall v \in O)(\mathrm{E}_p(u) \geq \mathrm{E}_p(v) \text{ and } u \not\prec v)\}. \tag{9}$$

Proposition 18. *The set of desirable options D_p and the choice function C_p are coherent and compatible, and moreover $C_p = C_{D_p}$.*

This result allows us to introduce the following, second special case of 'infimum of maximality' choice functions.

Definition 9. With any non-empty set of mass functions K,[5] we associate the corresponding *E-admissible* choice function $C_K^E := \inf\{C_p : p \in K\} = C_{\{D_p : p \in K\}}$.

Proposition 19. *Given any non-empty set of mass functions K, we have for all O in $\mathcal{Q}$ that*

$$C_K^E(O) = \{u \in O : (\exists p \in K) u \in \arg\max_{v \in O} \mathrm{E}_p(v)\} \cap C_{\mathrm{v}}(O).$$

The following proposition establishes a connection between M-admissible and E-admissible choice functions.

[5] Although Levi's notion of E-admissibility was originally [15, Chapter 5] concerned with *convex closed* sets of mass functions, we impose no such requirement here on the set K.

Proposition 20. *For any non-empty set of mass functions K, $C_K^E \sqsubseteq C_{\hat{\mathcal{D}}_K}^M$, where $\hat{\mathcal{D}}_K := \bigcup_{p \in K} \hat{\mathcal{D}}_{D_p} \subseteq \hat{\mathcal{D}}$.*

The following examples show why choice functions are more powerful than sets of desirable options as uncertainty representations, and elucidates the difference between E-admissible and M-admissible choice functions.

Example 2. Consider the situation where you have a coin with two *identical* sides of unknown type: either both sides are heads (H), or both sides are tails (T). The random variable that represents the outcome of a coin flip assumes a value in the finite possibility space $\mathcal{X} := \{\mathrm{H}, \mathrm{T}\}$. The options we consider are gambles: real-valued functions on $\mathcal{X}$, which constitute the two-dimensional vector space $\mathbb{R}^{\mathcal{X}}$, ordered by the pointwise order. We model this situation using (a) coherent sets of desirable options, (b) M-admissible choice functions, and (c) E-admissible choice functions. In all three cases we start from two simple models: one that describes practical certainty of H and another that describes practical certainty of T, and we take their infimum—the most informative model that is still less informative than both—as a candidate model for the coin problem.

For (a), we use two coherent sets of desirable options D_{H} and D_{T}, expressing practical certainty of H and T, respectively, given by the *maximal* sets of desirable options $D_{\mathrm{H}} := \mathcal{V}_{>0} \cup \{u \in \mathcal{V} : u_{\mathrm{H}} > 0\}$ and $D_{\mathrm{T}} := \mathcal{V}_{>0} \cup \{u \in \mathcal{V} : u_{\mathrm{T}} > 0\}$, where u_{H} and u_{T} denote the values of the gamble u in H and T, respectively. The model for the coin with two identical sides is then $D_{\mathrm{H}} \cap D_{\mathrm{T}} = \mathcal{V}_{>0}$. This vacuous model D_{v} is incapable of distinguishing between this situation and the one where we are completely ignorant about the coin.

For an approach (b) that distinguishes between these two situations, we draw inspiration from Proposition 13: instead of working with the sets of desirable options themselves, we move to the corresponding choice functions $C_{\mathrm{H}} := C_{D_{\mathrm{H}}}$ and $C_{\mathrm{T}} := C_{D_{\mathrm{T}}}$, where

$$\begin{aligned} C_{\mathrm{H}}(O) &= \{u \in O : (\forall v \in O) v - u \notin D_{\mathrm{H}}\} \\ &= \arg\max\{u_{\mathrm{H}} : u \in O\} \cap C_{\mathrm{v}}(O) \text{ for all } O \text{ in } \mathcal{Q} \\ C_{\mathrm{T}}(O) &= \arg\max\{u_{\mathrm{T}} : u \in O\} \cap C_{\mathrm{v}}(O) \text{ for all } O \text{ in } \mathcal{Q}. \end{aligned}$$

We infer that $|C_{\mathrm{H}}(O)| = |C_{\mathrm{T}}(O)| = 1$ for every O in $\mathcal{Q}$. The M-admissible choice function we are looking for is $C_{\{D_{\mathrm{H}}, D_{\mathrm{T}}\}}^M = \inf\{C_{\mathrm{H}}, C_{\mathrm{T}}\}$, which selects at most two options from each option set. It is given by

$$\begin{aligned} &C_{\{D_{\mathrm{H}}, D_{\mathrm{T}}\}}^M(O) \\ &= (\arg\max\{u_{\mathrm{H}} : u \in O\} \cup \arg\max\{u_{\mathrm{T}} : u \in O\}) \cap C_{\mathrm{v}}(O) \end{aligned}$$

for all O in $\mathcal{Q}$, and differs from the vacuous choice function C_{v}. Indeed, consider the particular option set $O = \{u, v, w\}$, where $u = (1, 0)$, $v = (0, 1)$ and $w = (1/2, 1/2)$. Then $C_{\{D_{\mathrm{H}}, D_{\mathrm{T}}\}}^M(O) = \{u, v\} \neq O = C_{\mathrm{v}}(O)$.

For (c), the set of mass functions K consists of the two degenerate mass functions: $K = \{p_{\mathrm{H}}, p_{\mathrm{T}}\}$, where $p_{\mathrm{H}} = (1, 0)$

and $p_T = (0,1)$. The corresponding expectations $E_H := E_{p_H}$ and $E_T := E_{p_T}$ satisfy $E_H(u) = u_H$ and $E_T(u) = u_T$ for all u in $\mathcal{V}$. So we see that $C_{p_H} = C_H$ and $C_{p_T} = C_T$, and therefore this approach leads to the same choice function as the previous one: $C^E_{\{p_H,p_T\}} = C^M_{\{D_H,D_T\}} = \inf\{C_H, C_T\}$. □

Example 3. We consider the same finite possibility space $\mathcal{X} := \{H, T\}$ as in Example 2, with the same option space and vector ordering. Also consider the vacuous set of desirable options D_v and the option set $O := \{0, u, v\}$, where $u = (1, -1/4)$ and $v = (-1/4, 1)$. Because all options in O are pointwise undominated in O, we find that $C_{D_v}(O) = O$. On the other hand, it follows from the definition in Eq. (7) that

$$0 \in C^M_{D_v}(O) \Leftrightarrow (\exists \hat{D} \in \hat{\mathcal{D}}_{D_v})(u \notin \hat{D} \text{ and } v \notin \hat{D}),$$

also taking into account Axiom D_1. But $u \notin \hat{D}$ and $v \notin \hat{D}$ implies that $-u \in \hat{D}$ and $-v \in \hat{D}$ by Proposition 7, and therefore also $-u - v \in \hat{D}$ by Axiom D_4. But $-u - v = (-3/4, -3/4) < 0$, contradicting the coherence [Axiom D_1] of $\hat{D}$. This means that $0 \notin C^M_{D_v}(O)$, so $C_{D'} \sqsubset C^M_{D'}$.

This same example shows that $C_v = C_{\hat{\mathcal{D}}} \sqsubset C_{\hat{\mathcal{D}}} = C^M_{D_v}$. □

To conclude this section, we want to mention that there are other popular choice rules besides maximality and E-admissibility, such as, amongst others, Γ-*maximin*, Γ-*maximax* and *interval dominance* [26]. However, they are not coherent: none of them satisfies Axiom C_4b.

5 Indifference

5.1 Indifference and Desirability

For sets of desirable options, there is a systematic way of modelling indifference [8, 7, 17]. Let us recall what it means to express an assessment of indifference there.

In addition to a subject's set of desirable options D—the options he strictly prefers to the zero option—we can also consider the options that he considers to be *equivalent* to the zero option. We call these options *indifferent*. A set of indifferent options I is simply a subset of $\mathcal{V}$, but as before with desirable options, we pay special attention to *coherent* sets of indifferent options.

Definition 10. A set of indifferent options I is called coherent if for all u, v in $\mathcal{V}$ and λ in $\mathbb{R}$:

I_1. $0 \in I$;
I_2. if $u \in \mathcal{V}_{>0} \cup \mathcal{V}_{<0}$ then $u \notin I$;
I_3. if $u \in I$ then $\lambda u \in I$;
I_4. if $u, v \in I$ then $u + v \in I$.

Taken together, Axioms I_3 and I_4 are equivalent to imposing that $\text{span}(I) = I$, and due to Axiom I_1, I is non-empty and therefore a linear subspace of $\mathcal{V}$.

The interaction between indifferent and desirable options is subject to rationality criteria as well: they should be compatible with one another.

Definition 11. Given a set of desirable options D and a coherent set of indifferent options I, we call D *compatible* with I if $D + I \subseteq D$.

The idea behind Definition 11 is that adding an indifferent option to a desirable option does not make it non-desirable.

Since $D \subseteq D + I$ due to Axiom I_1, compatibility of D and I is equivalent to $D + I = D$. An immediate consequence of compatibility between a coherent set of desirable options D and a coherent set of indifferent options I is that $D \cap I = \varnothing$, meaning that no option can be assessed as desirable—strictly preferred to the zero option—and indifferent—equivalent to the zero option—at the same time.

5.2 Indifference and Quotient Spaces

In order to introduce indifference for choice functions, we shall build on a coherent set of indifferent options I, as defined in Definition 10. Two options u and v are considered to be indifferent, to a subject, whenever $v - u$ is indifferent to the zero option, or in other words $v - u \in I$. The idea behind indifference for choice functions will be that we identify indifferent options, and choose between equivalence classes of indifferent options, rather than between single options. We begin by formalising this idea.

We can collect all options that are indifferent to an option $u \in \mathcal{V}$ into the *equivalence class*

$$[u] := \{v \in \mathcal{V} : v - u \in I\} = \{u\} + I.$$

Of course, $[0] = \{0\} + I = I$ is a linear subspace, and the $[u] = \{u\} + I$ affine subspaces of $\mathcal{V}$. The set of all these equivalence classes is the *quotient space*

$$\mathcal{V}/I := \{[u] : u \in \mathcal{V}\} = \{\{u\} + I : u \in \mathcal{V}\}.$$

This quotient space is a vector space under the vector addition, given by

$$[u] + [v] = \{u\} + I + \{v\} + I = \{u + v\} + I = [u + v] \text{ for } u, v \in \mathcal{V},$$

and the scalar multiplication, given by

$$\lambda[u] = \lambda(\{u\} + I) = \{\lambda u\} + I = [\lambda u],$$

for $u \in \mathcal{V}$ and $\lambda \in \mathbb{R}$. $[0] = I$ is the additive identity of $\mathcal{V}/I$.

That we identify indifferent options, and therefore express preferences between equivalence classes of indifferent options, essentially means that we define choice functions on $\mathcal{Q}(\mathcal{V}/I)$. But in order to characterise coherence for such choice functions, we need to introduce a convenient vector ordering on $\mathcal{V}/I$, that is appropriately related to the vector ordering on $\mathcal{V}$; see Section 2.1. For two elements $[u]$ and $[v]$ of $\mathcal{V}/I$, we define

$$[u] \preceq [v] \Leftrightarrow (\exists w \in I) u \preceq v + w, \tag{10}$$

and as usual, the strict variant of the vector ordering on $\mathcal{V}/I$ is characterised by

$$[u] \prec [v] \Leftrightarrow ([u] \preceq [v] \text{ and } [u] \neq [v]).$$

Proposition 21. *The ordering $\preceq$ on $\mathcal{V}/I$ is a vector ordering, and $[u] \prec [v] \Leftrightarrow (\exists w \in I) u \prec v + w$ for any u, v in $\mathcal{V}$.*

We use the notation $O/I := \{[u] : u \in O\}$ for the option set of equivalence classes $[u]$ associated with the options u in an option set O in $\mathcal{Q}(\mathcal{V})$. $\cdot/I$ is an onto map from $\mathcal{Q}(\mathcal{V})$ to $\mathcal{Q}(\mathcal{V}/I)$ that preserves set inclusion.

Proposition 22. *Given any two option sets O_1 and O_2 in $\mathcal{Q}(\mathcal{V})$ such that $O_1 \subseteq O_2$, then $O_1/I \subseteq O_2/I$.*

5.3 Quotient Spaces and Sets of Desirable Options

We use this quotient space to prove interesting characterisations of indifference for sets of desirable options.

Proposition 23. *A set of desirable options $D \subseteq \mathcal{V}$ is compatible with a coherent set of indifferent options I if and only if there is some (*representing*) set of desirable options $D' \subseteq \mathcal{V}/I$ such that $D = \{u : [u] \in D'\} = \bigcup D'$. Moreover, the representing set of desirable options is unique and given by $D' = D/I := \{[u] : u \in D\}$.*

This, together with the definition of compatibility, shows that the correspondence between sets of desirable options on $\mathcal{V}$ and (their representing) sets of desirable options on $\mathcal{V}/I$ is one-to-one and onto. It also preserves coherence.

Proposition 24. *Consider any set of desirable options $D \subseteq \mathcal{V}$ that is compatible with a coherent set of indifferent options I, and its representing set of desirable options $D/I \subseteq \mathcal{V}/I$. Then D is coherent if and only if D/I is.*

5.4 Quotient Spaces and Choice Functions

The discussion above inspires us to combine indifference with choice functions in the following manner: a choice function expresses indifference if its behaviour is completely determined by a choice function on the equivalence classes of indifferent options.

Definition 12. We call a choice function C on $\mathcal{Q}(\mathcal{V})$ *compatible* with a coherent set of indifferent options I if there is some *representing* choice function C' on $\mathcal{Q}(\mathcal{V}/I)$ such that $C(O) = \{u \in O : [u] \in C'(O/I)\}$ for all O in $\mathcal{Q}(\mathcal{V})$.

This definition allows for characterisations that are similar to the ones for desirability in Propositions 23 and 24. If a choice function on $\mathcal{Q}(\mathcal{V})$ is compatible with I then the representing choice function on $\mathcal{Q}(\mathcal{V}/I)$ is necessarily unique, and we denote it by C/I:

Proposition 25. *For any choice function C on $\mathcal{Q}(\mathcal{V})$ that is compatible with some coherent set of indifferent options I, the unique representing choice function C/I on $\mathcal{Q}(\mathcal{V}/I)$ is given by $C/I(O/I) := C(O)/I$ for all O in $\mathcal{Q}(\mathcal{V})$. Hence also*

$$C(O) = O \cap \left(\bigcup C/I(O/I)\right) \text{ for all } O \text{ in } \mathcal{Q}(\mathcal{V}).$$

This, together with the definition of compatibility, shows that the correspondence between choice functions on $\mathcal{Q}(\mathcal{V})$ and (their representing) choice functions on $\mathcal{Q}(\mathcal{V}/I)$ is one-to-one and onto. It also preserves coherence.

Proposition 26. *Consider any choice function C on $\mathcal{Q}(\mathcal{V})$ that is compatible with a coherent set of indifferent options I, and its representing choice function C/I on $\mathcal{Q}(\mathcal{V}/I)$. Then C is coherent if and only if C/I is.*

To conclude this general discussion of indifference for choice functions, we mention that it is closed under arbitrary infima, which enables conservative inference under indifference: we can consider the least informative choice function that is compatible with some assessments and is still compatible with a coherent set of indifferent options.

Proposition 27. *Consider any coherent set of indifferent options I, and any non-empty collection of coherent choice functions $\{C_i : i \in \mathcal{I}\}$ that are compatible with I, then its coherent infimum $\inf\{C_i : i \in \mathcal{I}\}$ is compatible with I as well, and $C/I = \inf\{C_i/I : i \in \mathcal{I}\}$.*

5.5 Relation with Desirability

First, we consider a coherent choice function C compatible with some coherent set of indifferent options I, and check whether the corresponding coherent set of desirable options D_C is also compatible with I.

Proposition 28. *Consider any coherent set of indifferent options I, and any compatible coherent choice function C, then the corresponding coherent set of desirable options D_C is also compatible with I, and $D_C/I = D_{C/I}$.*

Next, and conversely, we consider a coherent set of desirable options D compatible with I, and check whether the corresponding coherent choice functions C_D is also compatible with I.

Proposition 29. *Consider any coherent set of indifferent options I, and any compatible coherent set of desirable options D, then the corresponding coherent choice function C_D is also compatible with I, and $C_D/I = C_{D/I}$.*

5.6 Example

To exhibit the power and simplicity of our definition of indifference, we reconsider the finite possibility space $\mathcal{X} := \{\mathrm{H}, \mathrm{T}\}$ of Example 2, where the vector space $\mathcal{V}$ is again the two-dimensional vector space $\mathbb{R}^{\mathcal{X}}$ of real-valued functions on $\mathcal{X}$, or gambles, and the vector ordering $\preceq$ is the usual pointwise ordering of gambles.

We want to express indifference between heads and tails, or in other words between $\mathbb{I}_{\mathrm{H}}$ and $\mathbb{I}_{\mathrm{T}}$, where $\mathbb{I}_{\mathrm{H}} := (1,0)$ and $\mathbb{I}_{\mathrm{T}} := (0,1)$. This means that $\mathbb{I}_{\mathrm{H}} - \mathbb{I}_{\mathrm{T}}$ is considered equivalent to the zero gamble, so the linear space of all gambles that are equivalent to zero—or in other words, the set of indifferent gambles (or options)—is then given by

$$I = \{\lambda(\mathbb{I}_{\mathrm{H}} - \mathbb{I}_{\mathrm{T}}) : \lambda \in \mathbb{R}\} = \{u \in \mathbb{R}^{\mathcal{X}} : \mathrm{E}_p(u) = 0\},$$

where E_p is the expectation associated with the uniform mass function $p = (1/2, 1/2)$ on $\{\mathrm{H}, \mathrm{T}\}$, associated with a fair coin: $\mathrm{E}_p(u) := \frac{1}{2}[u_{\mathrm{H}} + u_{\mathrm{T}}]$. So, for any option u in $\mathbb{R}^{\mathcal{X}}$—any real-valued function on $\mathcal{X}$:

$$[u] = \{u\} + I = \{v \in \mathbb{R}^{\mathcal{X}} : \mathrm{E}_p(v) = \mathrm{E}_p(u)\},$$

which tells us that the equivalence class $[u]$ can be characterised by the common uniform expectation $\mathrm{E}_p(u)$ of its elements. Therefore, $\mathbb{R}^{\mathcal{X}}/I$ has unit dimension, and we can identify it with the real line $\mathbb{R}$. The vector ordering between equivalence classes is given by, using Eq. (10):

$$\begin{aligned}
[u] \preceq [v] &\Leftrightarrow (\exists \lambda \in \mathbb{R}) u \preceq v + \lambda(\mathbb{I}_{\mathrm{H}} - \mathbb{I}_{\mathrm{T}}) \\
&\Leftrightarrow (\exists \lambda \in \mathbb{R})(u_{\mathrm{H}} \leq v_{\mathrm{H}} + \lambda \text{ and } u_{\mathrm{T}} \leq v_{\mathrm{T}} - \lambda) \\
&\Leftrightarrow (\exists \lambda \in \mathbb{R}) u_{\mathrm{H}} - v_{\mathrm{H}} \leq \lambda \leq -u_{\mathrm{T}} + v_{\mathrm{T}} \\
&\Leftrightarrow u_{\mathrm{H}} - v_{\mathrm{H}} \leq -u_{\mathrm{T}} + v_{\mathrm{T}} \Leftrightarrow \mathrm{E}_p(u) \leq \mathrm{E}_p(v),
\end{aligned}$$

and similarly $[u] \prec [v] \Leftrightarrow \mathrm{E}_p(u) < \mathrm{E}_p(v)$ for all u, v in $\mathbb{R}^{\mathcal{X}}$. Hence, the strict vector ordering $\prec$ on $\mathbb{R}^{\mathcal{X}}/I$ is total, so we infer from the argumentation in Example 1 that there is only one representing choice function, namely the vacuous one. Therefore, there is only one choice function C on $\mathcal{Q}(\mathbb{R}^{\mathcal{X}})$ that is compatible with I, namely, the one that has the vacuous choice function C_{v} on $\mathcal{Q}(\mathbb{R}^{\mathcal{X}}/I)$ as its representation C/I. Recall that for any O in $\mathcal{Q}(\mathbb{R}^{\mathcal{X}})$:

$$\begin{aligned}
C_{\mathrm{v}}(O/I) &= \{[u] : (\forall [v] \in O/I)[u] \not\prec [v]\} \\
&= \{[u] : (\forall [v] \in O/I)[v] \preceq [u]\} \\
&= \{[u] : (\forall [v] \in O/I)\mathrm{E}_p(v) \leq \mathrm{E}_p(u)\},
\end{aligned}$$

and therefore

$$C(O) := \{u \in O : (\forall v \in O)\mathrm{E}_p(v) \leq \mathrm{E}_p(u)\} = C^E_{\{p\}}(O).$$

The indifference assessment between heads and tails leaves us no choice but to use an E-admissible model for a probability mass function, associated with a fair coin.

The choice function C is therefore based on E-admissibility, but is not compatible with M-admissibility. To see this, consider the set of options $O := \{w, 0, -w\}$ with $w := (1, -1)$, so $w_{\mathrm{H}} + w_{\mathrm{T}} = 0$. Hence $C(O) = O$.

But no M-admissible choice function will select 0 in O: observe that $0 \notin C_{\hat{D}}(O)$ for all $\hat{D} \in \mathcal{D}'$, because $0 \in C_{\hat{D}}(O)$ would imply that $\{w, -w\} \cap \hat{D} = \varnothing$, contradicting that $\hat{D}$ is a maximal set of desirable options by Proposition 7.

6 Conclusion

We have developed a theory of conservative reasoning with choice functions, and related coherent choice functions to coherent sets of desirable options, showing that choice functions are indeed more informative than sets of desirable options as a tool for conservative reasoning. We have also provided an intuitive definition for indifference that subsumes the usual definition for sets of desirable options.

We still intend to address conditioning for choice functions, and look for an elegant conditioning rule that subsumes the one for sets of desirable options—and therefore also Bayes's rule. Another problem to tackle is related to indifference: Seidenfeld [20] (see also [3]) has given another elegant definition for indifference for choice functions, which he has also linked to sequential coherence. We know that our definition implies his, but the question whether the two approaches are equivalent is still open. The connection with sequential coherence is also an open issue, and we expect Axiom C_3 will play an important role in resolving it.

Acknowledgments

Gert de Cooman's research was partly funded through project number 3G012512 of the Research Foundation Flanders (FWO). Erik Quaeghebeur's contribution is part of the *Safe Statistics* project financed by the Netherlands Organisation for Scientific Research (NWO). Enrique Miranda acknowledges financial support by project TIN2014-59543-P. The authors would like to express their gratitude to anonymous reviewers for their helpful comments and suggestions, and to Jasper De Bock for the undominated background noise and stimulating discussion.

References

[1] Mark A. Aizerman. New problems in the general choice theory. *Social Choice and Welfare*, 2:235–282, 1984. doi:10.1007/BF00292690.

[2] Francis J. Anscombe and Robert J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34:199–205, 1963. URL http://www.jstor.org/stable/2991295.

[3] Seamus Bradley. Weak rationality and imprecise choice. in preparation. URL http://www.seamusbradley.net/Papers/imprecise-choice.pdf.

[4] Inés Couso and Serafín Moral. Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011. doi:10.1016/j.ijar.2011.04.004.

[5] Brian A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, 1990.

[6] Jasper De Bock. *Credal Networks under Epistemic Irrelevance: Theory and algorithms*. PhD thesis, Ghent University, 2015.

[7] Jasper De Bock, Arthur Van Camp, Márcio A. Diniz, and Gert de Cooman. Representation theorems for partially exchangeable random variables. *Fuzzy Sets and Systems*, In Press. doi:10.1016/j.fss.2014.10.027.

[8] Gert de Cooman and Erik Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53(3):363–395, 2012. doi:10.1016/j.ijar.2010.12.002. Precisely imprecise: A collection of papers dedicated to Henry E. Kyburg, Jr.

[9] Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal of Approximate Reasoning*, 39:257–278, 2005. doi:10.1016/j.ijar.2004.10.004.

[10] Gert de Cooman, Jasper De Bock, and Márcio Alves Diniz. Coherent predictive inference under exchangeability with imprecise probabilities. *Journal of Artificial Intelligence Research*, 52:1–95, 2015.

[11] Bruno de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[12] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester, 1974–1975. English translation of [11], two volumes.

[13] Junnan He. A generalized unification theorem for choice theoretic foundations: Avoiding the necessity of pairs and triplets. Economics Discussion Paper 2012-23, Kiel Institute for the World Economy, 2012. URL http://www.economics-ejournal.org/economics/discussionpapers/2012-23.

[14] Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. A Rubinesque theory of decision. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 45:45–55, 2004. URL http://www.jstor.org/stable/4356297.

[15] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[16] Erik Quaeghebeur. Desirability. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, chapter 1. John Wiley & Sons, 2014.

[17] Erik Quaeghebeur, Gert de Cooman, and Filip Hermans. Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning*, 57:69–102, February 2015. doi:10.1016/j.ijar.2014.12.003. URL http://arxiv.org/abs/1208.4462.

[18] Herman Rubin. A weak system of axioms for "rational" behavior and the nonseparability of utility from prior. *Statistics & Risk Modeling*, 5(1-2):47–58, 1987. URL http://EconPapers.repec.org/RePEc:bpj:strimo:v:5:y:1987:i:1-2:p:47-58:n:11.

[19] Thomas Schwartz. Rationality and the myth of the maximum. *Noûs*, 6(2):97–117, 1972.

[20] T. Seidenfeld. Decision without independence and without ordering: what is the difference? *Economics and Philosophy*, 4:267–290, 1988.

[21] Teddy Seidenfeld, Mark J. Schervish, and Jay B. Kadane. A representation of partially ordered preferences. *The Annals of Statistics*, 23:2168–2217, 1995. Reprinted in [22], pp. 69–129.

[22] Teddy Seidenfeld, Mark J. Schervish, and Jay B. Kadane. *Rethinking the Foundations of Statistics*. Cambridge University Press, Cambridge, 1999.

[23] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172 (1):157–176, 2010. doi:10.1007/s11229-009-9470-7.

[24] Amartya Sen. Choice functions and revealed preference. *The Review of Economic Studies*, 38(3):307–317, Jul. 1971.

[25] Amartya Sen. Social choice theory: A re-examination. *Econometrica*, 45:53–89, 1977. doi:10.2307/1913287.

[26] Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, January 2007. doi:10.1016/j.ijar.2006.06.001.

[27] Arthur Van Camp, Gert de Cooman, and Erik Quaeghebeur. Connecting choice functions and sets of desirable gambles. In *Imprecise Probabilities in Statistics and Philosophy, Proceedings*, 2014.

[28] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[29] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24 (2–3):125–148, 2000. doi:10.1016/S0888-613X(00)00031-1.

Credal Compositional Models and Credal Networks

Jiřina Vejnarová
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
vejnar@utia.cas.cz

Abstract

This paper studies the composition operator for credal sets introduced at the last ISIPTA conference in more detail. Our main attention is devoted to the relationship between a special type of compositional model, so-called perfect sequences of credal sets, and those of (precise) probability distributions, with the goal of finding the relationship between credal compositional models and credal networks. We prove that a perfect sequence of credal sets is a convex hull of perfect sequences of extreme points of these credal sets. Finally, we reveal the relationship among credal networks (in a general sense), perfect sequences of credal sets and separately specified credal networks.

Keywords. Credal sets, strong independence, credal networks, separate specification, compositional models.

1 Introduction

The most widely used models managing uncertainty and multidimensionality are, at present, the so-called *probabilistic graphical Markov models*. The problem of multidimensionality is solved in these models with the help of the concept of conditional independence, which enables factorisation of a multidimensional probability distribution into small parts (marginals, conditionals or just factors). Among them, the most popular are Bayesian networks. Therefore, it is not very surprising that analogous models have also been studied in several theories of imprecise probability [1, 2, 3].

Credal networks represent a generalisation of Bayesian networks capable of dealing with imprecision. Compositional models for credal sets, on the other hand, are intended to be a generalisation of compositional models for precise probabilities [6, 7, 8]. As the equivalence between Bayesian networks and precise compositional models is well known [9], it also seems quite natural to ask a similar question in this more general case.

Compositional models have also been introduced in possibility theory [13, 14] (where these models are parameterised by a continuous t-norm) and a few years ago in evidence theory [10, 11] as well. In all these frameworks the original idea is preserved but certain slight differences between them are present.

Although Bayesian networks and (precise) probabilistic compositional models represent the same class of distributions, they do not do it in the same way. Namely, Bayesian networks use *conditional distributions*, whereas compositional models consist of *unconditional distributions*. Naturally, both types of models contain the same information but, while some marginal distributions are explicitly expressed in compositional models, it may happen that their computation from the corresponding Bayesian network is rather computationally expensive.

Furthermore, the research concerning the relationship between compositional models in evidence theory and evidential networks [15] revealed an aspect that is probably even more important. Even though any evidential network (with a proper conditioning rule and conditional independence concept) can be expressed as a compositional model, if we do it in the opposite way and transform a compositional model into an evidential network, we may realise that the model is more imprecise than the original one. This is caused by the fact that conditioning increases imprecision.

In [16] we introduced a composition operator for credal sets, but due to the problem of discontinuity it needs a revision. This task seems to be rather difficult and has not been satisfactorily solved yet. Therefore, we decided to postpone its definition for the general case to the future and now we deal only with the case of projective credal sets, as this approach is sufficient for the topic of this paper.

The goal of this paper is to show that the composition operator for credal sets is worth developing, as compositional models seem to be a reasonable counterpart of

credal networks. We prove that the perfect sequence of credal sets is a convex hull of perfect sequences of extreme points of these credal sets. We prove that any separately specified credal network can be expressed in the form of a perfect sequence of credal sets, and any perfect sequence of credal sets can be expressed as a credal network (in a general sense). Finally, we present an algorithm for transforming a compositional model to a credal network.

This contribution is organized as follows. In Section 2 we summarise the basic concepts and notation. Definition of the operator of composition is recalled in Section 3, which is completely devoted to its basic properties and those of compositional models. Finally, in Section 4 the relationship between credal networks and compositional models is studied.

2 Basic Concepts and Notation

In this section we will recall the basic concepts and notation necessary for understanding the paper.

2.1 Variables and Distributions

For an index set $N = \{1, 2, \ldots, n\}$ let $\{X_i\}_{i\in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$, and $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$ be the Cartesian product of these sets.

In this paper we will deal with groups of variables on subspaces of the Cartesian product. Let us note that X_K will denote a group of variables $\{X_i\}_{i\in K}$ with values in

$$\mathbf{X}_K = \bigtimes_{i\in K}\mathbf{X}_i$$

throughout the paper.

Any group of variables X_K can be described by a *probability distribution* (sometimes also called *probability function*)

$$P : \mathbf{X}_K \longrightarrow [0, 1],$$

such that $\sum_{x_K\in\mathbf{X}_K} P(x_K) = 1$.

Having two probability distributions P_1 and P_2 of X_K, we say that P_1 is *absolutely continuous* with respect to P_2 (and denote $P_1 \ll P_2$) if for any $x_K \in \mathbf{X}_K$

$$P_2(x_K) = 0 \Longrightarrow P_1(x_K) = 0.$$

This concept plays an important role in the definition of the composition operator.

2.2 Credal Sets

A *credal set* $\mathcal{M}(X_K)$ describing a group of variables X_K is defined as a closed convex set of probability measures describing the values of these variables.[1]

[1]For $K = \emptyset$ let us set $\mathcal{M}(X_\emptyset) \equiv 1$.

In order to simplify the expression of operations with credal sets, it is often considered [12] that a credal set is the set of probability distributions associated with the probability measures in it. Under such consideration, a credal set can be expressed as a *convex hull* of its extreme distributions

$$\mathcal{M}(X_K) = \mathrm{CH}\{\mathrm{ext}(\mathcal{M}(X_K))\}.$$

Consider a credal set describing X_K, i.e., $\mathcal{M}(X_K)$. For each $L \subset K$ its *marginal credal set* $\mathcal{M}(X_L)$ is obtained by element-wise marginalisation, i.e.,

$$\mathcal{M}(X_L) = \mathrm{CH}\{P^{\downarrow L} : P \in \mathrm{ext}(\mathcal{M}(X_K))\}, \quad (1)$$

where $P^{\downarrow L}$ denotes the marginal distribution of P on $\mathbf{X}_L$.

Having two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing X_K and X_L, respectively (assuming that $K, L \subseteq N$), we say that these credal sets are *projective* if their marginals describing the common variables coincide, i.e., if

$$\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}). \quad (2)$$

Let us note that if K and L are disjoint, then $\mathcal{M}_1$ and $\mathcal{M}_2$ are always projective, as $\mathcal{M}_1(X_\emptyset) = \mathcal{M}_2(X_\emptyset) \equiv 1$.

Conditional credal sets are obtained from the joint ones by point-wise conditioning of the extreme points and subsequent linear combination of the resulting conditional distributions. More formally: Let $\mathcal{M}(X_{K\cup L})$ $(K \cap L = \emptyset)$ be a credal set describing (groups of) variables $X_{K\cup L}$. Then for any $x_L \in \mathbf{X}_L$

$$\mathcal{M}(X_K|x_L) = \mathrm{CH}\{P(X_K|x_L) : P \in \mathrm{ext}(\mathcal{M}(X_{K\cup L}))\}, \quad (3)$$

is a *conditional credal set* describing X_K given $X_L = x_L$.

2.3 Strong Independence

Among numerous definitions of independence for credal sets [4] we have chosen strong independence, as it seems to be the most appropriate for multidimensional models.

We say that (groups of) variables X_K and X_L (K and L disjoint) are *strongly independent* with respect to $\mathcal{M}(X_{K\cup L})$ iff (in terms of probability distributions)

$$\mathcal{M}(X_{K\cup L}) = \mathrm{CH}\{P_1 \cdot P_2 : P_1 \in \mathrm{ext}(\mathcal{M}(X_K)), P_2 \in \mathrm{ext}(\mathcal{M}(X_L))\}. \quad (4)$$

Again, several generalisations of this notion to conditional independence already exist, see, e.g., [12],

but since the following definition is suggested by the authors as the most appropriate for the marginal problem, it seems to be a suitable concept in our case as well, since the composition operator can also be used as a tool for solving the marginal problem, as shown (within the framework of possibility theory), e.g., in [14].

Given three groups of variables X_K, X_L and X_M (where K, L, M are mutually disjoint subsets of N such that K and L are nonempty), we say in a way analogous[2] to [12] that X_K and X_L are *conditionally strongly independent* given X_M under the global set $\mathcal{M}(X_{K\cup L\cup M})$ (we will denote this relationship by $K \perp\!\!\!\perp L|M$) iff

$$\begin{aligned}&\mathcal{M}(X_{K\cup L\cup M})\\ &\quad = \mathrm{CH}\{(P_1 \cdot P_2)/P_1^{\downarrow M} : P_1 \in \mathrm{ext}(\mathcal{M}(X_{K\cup M})),\\ &\qquad P_2 \in \mathrm{ext}(\mathcal{M}(X_{L\cup M})), P_1^{\downarrow M} = P_2^{\downarrow M}\}. \quad (5)\end{aligned}$$

This definition is a generalisation of stochastic conditional independence: if $\mathcal{M}(X_{K\cup L\cup M})$ is a singleton, then $\mathcal{M}(X_{K\cup M})$ and $\mathcal{M}(X_{L\cup M})$ are also (projective) singletons and the definition is reduced to the definition of stochastic conditional independence.

3 Compositional Models

In this section we will summarise the achieved results concerning compositional models for credal sets. Most of them are presented without proofs; missing proofs can be found in [16]. The concept of the composition operator is presented first in a precise probability framework, as it seems to be useful for better understanding to the concept.

3.1 Composition Operator and Its Properties

Now, let us recall the definition of composition of two credal sets. Consider two index sets $K, L \subset N$. We do not put any restrictions on K and L; they may be but need not be disjoint, and one may be a subset of the other.

In order to enable the reader to understand this concept, let us first present the definition of composition for precise probabilities [6]. Let P_1 and P_2 be two probability distributions of (groups of) variables X_K and X_L; then

$$(P_1 \triangleright P_2)(X_{K\cup L}) = \frac{P_1(X_K) \cdot P_2(X_L)}{P_2(X_{K\cap L})}, \quad (6)$$

whenever $P_1(X_{K\cap L}) \ll P_2((X_{K\cap L})$; otherwise, it remains undefined.

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be credal sets describing X_K and X_L, respectively. Our original goal in [16] was to define a new credal set, denoted by $\mathcal{M}_1 \triangleright \mathcal{M}_2$, which will be describing $X_{K\cup L}$ and will contain all of the information contained in $\mathcal{M}_1$ and, as much as possible, in $\mathcal{M}_2$.

The required properties are met by Definition 1 in [16][3]. However, the definition exhibits a kind of discontinuity and should be reconsidered. Therefore, we will only deal with the composition of projective credal sets in this paper.

Definition 1 For two projective credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing X_K and X_L, their *composition* $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is defined by the following expression:

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})\\ &\quad = \mathrm{CH}\{(P_1 \cdot P_2)/P_2^{\downarrow K\cap L} : P_1 \in \mathrm{ext}(\mathcal{M}_1(X_K)),\\ &\qquad P_2 \in \mathrm{ext}(\mathcal{M}_2(X_L)), P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}\}.\end{aligned}$$

The following lemma, proven in [16], contains basic properties possessed by this composition operator.

Lemma 1 *For two projective credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ describing X_K and X_L, respectively, the following properties hold true:*

(i) $\mathcal{M}_1 \triangleright \mathcal{M}_2$ *is a credal set describing* $X_{K\cup L}$.

(ii) $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K) = \mathcal{M}_1(X_K)$ *and* $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_L) = \mathcal{M}_2(X_L)$.

(iii) $\mathcal{M}_1 \triangleright \mathcal{M}_2 = \mathcal{M}_2 \triangleright \mathcal{M}_1$.

As the operator is, at present, defined only for projective sets, it is commutative, as suggested by *(iii)* of this lemma. Furthermore, it follows from *(ii)* that the operator keeps both marginals. Both of these properties are typical in other settings exactly for the case of projective marginals.

Despite these facts, it remains non-associative (in general), as can be seen from the following example.

Example 1 Let X_1 and X_2 be two binary variables and

$$\mathcal{M}_1(X_1) = \mathrm{CH}\{[0.2, 0, 8], [0.5, 0.5]\}$$

and

$$\mathcal{M}_2(X_2) = \mathrm{CH}\{[0.3, 0.7], [0.6, 0.4]\}$$

[2] Let us note that our definitions somehow differ from those presented in [12]: the authors there require point-wise satisfaction in (4) and (5), which leads to non-convexity. In [5], this type of independence is called *complete*.

[3] Let us note that the definition is based on Moral's concept of conditional independence with relaxing convexity.

be two credal sets describing X_1 and X_2, respectively; further let

$$\mathcal{M}_3(X_1X_2) = \mathrm{CH}\{[0.2, 0, 0.1, 0.7], [0.5, 0, 0.1, 0.4]\}$$

be another credal set describing both X_1 and X_2. Here $[a, b]$ means $P(x_1) = a$ and $P(\bar{x}_1) = b$, and similarly $[a, b, c, d]$ means $P(x_1x_2) = a$, $P(x_1\bar{x}_2) = b$, $P(\bar{x}_1x_2) = c$ and $P(\bar{x}_1\bar{x}_2) = d$.

Using (1) to $\mathcal{M}_3(X_1X_2)$, one can realise that both $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ are marginal to $\mathcal{M}_3(X_1X_2)$.

$\mathcal{M}_1 \triangleright \mathcal{M}_2$ is obtained via Definition 1:

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2)\\ &\quad = \mathrm{CH}\{[0.06, 0.14, 0.24, 0.56], [0.12, 0.48, 0.08, 0.32]\\ &\qquad [0.15, 0.35, 0.15, 0.35], [0.3, 0.2, 0.3, 0.2]\},\end{aligned}$$

but $\mathcal{M}_1 \triangleright \mathcal{M}_2$ cannot be composed with $\mathcal{M}_3$, as they are not projective. On the other hand

$$\begin{aligned}&(\mathcal{M}_2 \triangleright \mathcal{M}_3)(X_1X_2)\\ &\qquad = \mathrm{CH}\{[0.2, 0, 0.1, 0.7], [0.5, 0, 0.1, 0.4]\},\end{aligned}$$

as follows from (*ii*) of Lemma 1 and similarly, for the same reason,

$$\begin{aligned}&(\mathcal{M}_1 \triangleright (\mathcal{M}_2 \triangleright \mathcal{M}_3))(X_1X_2)\\ &\qquad = \mathrm{CH}\{[0.2, 0, 0.1, 0.7], [0.5, 0, 0.1, 0.4]\}.\end{aligned}$$ ♢

The following theorem, also proven in [16], expresses the relationship between strong independence and the operator of composition. It is, together with Lemma 1, the most important assertion enabling us to introduce multidimensional models.

Theorem 1 *Let $\mathcal{M}$ be a credal set describing $X_{K\cup L}$ with marginals $\mathcal{M}(X_K)$ and $\mathcal{M}(X_L)$. Then*

$$\mathcal{M}(X_{K\cup L}) = (\mathcal{M}^{\downarrow K} \triangleright \mathcal{M}^{\downarrow L})(X_{K\cup L})$$

iff

$$(K \setminus L) \perp\!\!\!\perp (L \setminus K) | (K \cap L).$$

3.2 Perfect Sequences of Credal Sets

In this subsection we will recall repetitive application of the composition operator with the goal to create a multidimensional model. Since the operator is not associative, as demonstrated in Example 1, we have to specify in which order the low-dimensional credal sets are composed together. To make the formulae more transparent, we will omit parentheses in the case the operator is to be applied from left to right, i.e., in what follows

$$\begin{aligned}&\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 \triangleright \cdots \triangleright \mathcal{M}_{m-1} \triangleright \mathcal{M}_m \qquad (7)\\ &\quad = (\cdots((\mathcal{M}_1 \triangleright \mathcal{M}_2) \triangleright \mathcal{M}_3) \triangleright \cdots \triangleright \mathcal{M}_{m-1}) \triangleright \mathcal{M}_m.\end{aligned}$$

Furthermore, we will always assume $\mathcal{M}_i$ to be a credal set describing X_{K_i} and call $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots, \mathcal{M}_m$ a *generating sequence* of model (7).

The reader familiar with some papers on probabilistic, possibilistic or evidential compositional models knows that one of the most important notions in this theory is that of a so-called *perfect sequence*, already introduced in [16] also for credal sets. Let us recall it here.

Definition 2 A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is called *perfect* if

$$\begin{aligned}\mathcal{M}_1 \triangleright \mathcal{M}_2 &= \mathcal{M}_2 \triangleright \mathcal{M}_1,\\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 &= \mathcal{M}_3 \triangleright (\mathcal{M}_1 \triangleright \mathcal{M}_2),\\ &\vdots\\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m &= \mathcal{M}_m \triangleright (\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{m-1}).\end{aligned}$$

Let us note that the concept of perfect sequence of probability distributions is a special case of this definition, in the case of all credal sets being singletons.

It is evident that the necessary condition for perfectness is pairwise projectivity (i.e., (2) holds for any pair of credal sets from the generating sequence in question) of low-dimensional credal sets. However, from Example 1 one can easily see that this condition need not be sufficient.

Therefore a stronger, necessary and sufficient condition, expressed by the following assertion, must be fulfilled.

Lemma 2 *A generating sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is perfect* iff *the pairs of credal sets $\mathcal{M}_j$ and $(\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{j-1})$ are projective, i.e., if*

$$\begin{aligned}&\mathcal{M}_j(X_{K_j \cap (K_1 \cup \cdots \cup K_{j-1})})\\ &\qquad = (\mathcal{M}_1 \triangleright \cdots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \cdots \cup K_{j-1})}),\end{aligned}$$

for all $j = 2, 3, \ldots, m$.

From Definition 2 one can hardly identify the properties of perfect sequences beyond the algebraic ones; the most important one is expressed by the following characterisation theorem, which also suggests why these sequences are called perfect.

Theorem 2 *A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is perfect* iff *all the credal sets from this sequence are marginal to the composed credal set $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m$:*

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m)(X_{K_j}) = \mathcal{M}_j(X_{K_j}),$$

for all $j = 1, \ldots, m$.

The following (almost trivial) assertion, which brings the sufficient condition for perfectness, resembles assertions concerning decomposable models.

Theorem 3 *Let a generating sequence of pairwise projective credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ be such that $K_1, K_2, \ldots, K_m$ satisfies the following running intersection property:*

$$\forall j = 2, 3, \ldots, m \quad \exists \ell (1 \leq \ell < j) \text{ such that } \quad K_j \cap (K_1 \cup \cdots \cup K_{j-1}) \subseteq K_\ell.$$

Then the sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is perfect.

It should be emphasised that the running intersection property of $K_1, K_2, \ldots, K_m$ is a sufficient condition to guarantee perfectness of a generating sequence of pairwise projective assignments. By no means is this condition necessary, as already demonstrated in [16].

Therefore, not only is perfectness of a sequence a structural property connected with the properties of $K_1, K_2, \ldots, K_m$ but it also depends on specific values of the respective basic assignments.

3.3 Perfect Sequence as Convex Hull

In this subsection we will study the relationship between perfect sequences of credal sets and those of a probability distribution. Before doing that, let us present a simple lemma necessary for the proof of the main theorem.

Lemma 3 *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two projective credal sets describing X_K and X_L, respectively. Then*

$$\begin{aligned} &\{\text{ext}((\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K \cup X_L))\} \\ &\quad \subseteq \{P_1 \triangleright P_2 : P_1 \in \text{ext}(\mathcal{M}_1(X_K)), \\ &\qquad P_2 \in \text{ext}(\mathcal{M}_2(X_L)), P_1^{\downarrow K \cap L} = P_2^{\downarrow K \cap L}\}. \end{aligned} \tag{8}$$

Proof. By Definition 1, $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K \cup L})$ is the convex hull of the set of probability distributions from the set on the right-hand side of (8), taking into account the definition of the composition operator for precise probabilities. Therefore its extreme points must also belong to this set. □

Equality need not hold in (8), as can be seen from the following simple example.

Example 2 Let

$$\mathcal{M}_1(X_1) = \text{CH}\{[0.2, 0.8], [0.5, 0.5]\}$$

and

$$\mathcal{M}_2(X_2) = \text{CH}\{[0.5, 0.5], [0.8, 0.2]\}$$

be two credal sets describing X_1 and X_2, respectively. Then, as mentioned above, $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ are projective, and therefore $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is obtained by Definition 1:

$$\begin{aligned} &(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1 X_2) \\ &\quad = \text{CH}\{[0.1, 0.4, 0.1, 0.4], [0.16, 0.04, 0.64, 0.16], \\ &\qquad [0.25, 0.25, 0.25, 0.25], [0.4, 0.1, 0.4, 0.1]\}, \end{aligned} \tag{9}$$

nevertheless $[0.25, 0.25, 0.25, 0.25]$ is not an extreme point of (9) because it can be obtained as a linear combination of $[0.1, 0.4, 0.1, 0.4]$ and $[0.4, 0.1, 0.4, 0.1]$. ♢

Theorem 4 *Let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ be a perfect sequence of credal sets such that each $\mathcal{M}_i, i = 1, \ldots m$, is the convex hull of its extreme points, i.e.,*

$$\mathcal{M}_i(X_{K_i}) = \text{CH}\{P_i : P_i \in \text{ext}(\mathcal{M}_i(X_{K_i}))\}.$$

Then

$$\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m$$

is a convex hull of all

$$P_1 \triangleright P_2 \triangleright \ldots \triangleright P_m$$

such that each $P_i \in \text{ext}(\mathcal{M}_i(X_{K_i}))$, and $P_1, P_2, \ldots, P_m$ form a perfect sequence.

Proof. Let us prove the assertion by induction. For $m = 2$ it is obvious as it follows directly from Definition 1. Let us suppose that

$$\begin{aligned} &\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_j \\ &\quad = \text{CH}\{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_j, P_i \in \text{ext}(\mathcal{M}_i), \\ &\qquad P_1, P_2, \ldots, P_j \text{ is perfect}\} \end{aligned}$$

for $2 \leq j < m$ and prove that

$$\begin{aligned} &\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_{j+1} \\ &\quad = \text{CH}\{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_{j+1}, P_i \in \text{ext}(\mathcal{M}_i), \\ &\qquad P_1, P_2, \ldots, P_{j+1} \text{ is perfect}\} \end{aligned} \tag{10}$$

holds as well.

By convention (7)

$$\begin{aligned} &\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_j \triangleright \mathcal{M}_{j+1} \\ &\qquad = (\cdots \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_j) \triangleright \mathcal{M}_{j+1} \end{aligned}$$

and since $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_j$ and $\mathcal{M}_{j+1}$ are projective, we can apply Definition 1 to these credal sets to obtain

$$\begin{aligned} &(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_j) \triangleright \mathcal{M}_{j+1} \\ &\quad = \text{CH}\{Q_j \cdot \frac{P_{j+1}}{P_{j+1}^{\downarrow (K_1 \cup \cdots \cup K_j) \cap K_{j+1}}}, \\ &\qquad Q_j \in \text{ext}(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_j), \\ &\qquad P_{j+1} \in \text{ext}(\mathcal{M}_{j+1}), \\ &\qquad Q_j^{\downarrow (K_1 \cup \cdots \cup K_j) \cap K_{j+1}} = P_{j+1}^{\downarrow (K_1 \cup \cdots \cup K_j) \cap K_{j+1}}\}. \end{aligned}$$

However, due to Lemma 3

$$Q_j \in \{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_j, P_i \in \mathrm{ext}(\mathcal{M}_i), \\ P_1, P_2, \ldots, P_j \text{ is perfect}\}.$$

Let us denote by $P_1^*, P_2^*, \ldots, P_j^*$ a perfect sequence such that

$$Q_j = P_1^* \triangleright P_2^* \triangleright \ldots \triangleright P_j^*.$$

Then, due to Lemma 2 (applied to precise probability distributions) $P_1^*, P_2^*, \ldots, P_j^*, P_{j+1}$ forms a perfect sequence. Therefore

$$\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_{j+1} \\ \subseteq \mathrm{CH}\{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_{j+1}, P_i \in \mathrm{ext}(\mathcal{M}_i), \\ P_1, P_2, \ldots, P_{j+1} \text{ is perfect}\}.$$

Let, on the other hand, $P_1, P_2, \ldots, P_{j+1}$ be a perfect sequence of distributions such that each $P_i \in \mathrm{ext}(\mathcal{M}_i)$. Then

$$P_1 \triangleright P_2 \triangleright \ldots \triangleright P_{j+1} \in \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_{j+1},$$

and therefore also

$$\mathrm{CH}\{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_{j+1}, P_1, P_2, \ldots, P_{j+1} \text{ is perfect}\} \\ \subseteq \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_{j+1}.$$

Therefore (10) is satisfied. □

Example 3 Let $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ be the two credal sets from Example 2,

$$\mathcal{M}_3(X_1X_2X_3) \\ = \mathrm{CH}\{[0.1, 0, 0.3, 0.1, 0.05, 0.05, 0.1, 0.3], \\ [0.16, 0, 0.03, 0.01, 0.32, 0.32, 0.04, 0.12], \\ [0.4, 0, 0.075, 0.025, 0.2, 0.2, 0.025, 0.075]\}$$

and

$$\mathcal{M}_4(X_3X_4) \\ = \mathrm{CH}\{[0.44, 0.11, 0.18, 0.27], [0.56, 0.14, 0.12, 0.18], \\ [0.33, 0.22, 0.09, 0.36], [0.42, 0.28, 0.06, 0.24]\}.$$

These credal sets form a perfect sequence $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$, since $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is marginal to $\mathcal{M}_3$, and $\mathcal{M}_3$ and $\mathcal{M}_4$ are projective, as from (1) one gets

$$\mathcal{M}_3(X_3) = \mathrm{CH}\{[0.55, 0.45], [0.7, 0.3]\} = \mathcal{M}_4(\mathrm{X}_3).$$

The credal set $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 \triangleright \mathcal{M}_4(X_1, X_2, X_3, X_4)$ is then expressed as

$$\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 \triangleright \mathcal{M}_4 \qquad (11) \\ = \mathrm{CH}\{[0.08, 0.02, 0, 0, 0.24, 0.06, 0.04, 0.06, 0.04, \\ 0.01, 0.02, 0.03, 0.08, 0.02, 0.12, 0.18], \\ [0.06, 0.04, 0, 0, 0.18, 0.12, 0.02, 0.08, 0.03, \\ 0.02, 0.01, 0.04, 0.06, 0.04, 0.06, 0.24], \\ [0.128, 0.032, 0, 0, 0.024, 0.006, 0.004, \\ 0.006, 0.256, 0.064, 0.128, 0.192, \\ 0.032, 0.008, 0.048, 0.072], \\ [0.096, 0.064, 0, 0, 0.018, 0.012, 0.002, \\ 0.008, 0.192, 0.128, 0.064, 0.256, \\ 0.024, 0.016, 0.024, 0.096], \\ [0.32, 0.08, 0, 0, 0.06, 0.015, 0.015, 0.01, 0.16, \\ 0.04, 0.08, 0.12, 0.02, 0.005, 0.03, 0.015], \\ [0.24, 0.16, 0, 0, 0.045, 0.03, 0.005, 0.02, 0.12, \\ 0.08, 0.04, 0.16, 0.015, 0.01, 0.015, 0.06]\}.$$

This credal set can be obtained either directly by successive application of Definition 1 or as a convex hull of $P_1^{i_1} \triangleright P_2^{i_2} \triangleright P_3^{i_3} \triangleright P_4^{i_4}$, where any $P_1^{i_1}, P_2^{i_2}, P_3^{i_3}, P_4^{i_4}$ forms a perfect sequence, and any $P_j^{i_j} \in \mathrm{ext}(\mathcal{M}_j)$. In this example we have six perfect sequences, namely

$$\begin{array}{ll} P_1^1, P_2^1, P_3^1, P_4^1; & P_1^1, P_2^1, P_3^1, P_4^3; \\ P_1^1, P_2^2, P_3^2, P_4^1; & P_1^1, P_2^2, P_3^2, P_4^3; \\ P_1^2, P_2^2, P_3^3, P_4^2; & P_1^2, P_2^2, P_3^3, P_4^4, \end{array} \qquad (12)$$

where

$$P_1^1 = [0.2, 0.8], \qquad P_1^2 = [0.5, 0.5], \\ P_2^1 = [0.5, 0.5], \qquad P_2^2 = [0.8, 0.2], \\ P_3^1 = [0.1, 0, 0.3, 0.1, 0.05, 0.05, 0.1, 0.3], \\ P_3^2 = [0.16, 0, 0.03, 0.01, 0.32, 0.32, 0.04, 0.12], \\ P_3^3 = [0.4, 0, 0.075, 0.025, 0.2, 0.2, 0.025, 0.075], \\ P_4^1 = [0.44, 0.11, 0.18, 0.27], \\ P_4^2 = [0.56, 0.14, 0.12, 0.18], \\ P_4^3 = [0.33, 0.22, 0.09, 0.36], \\ P_4^4 = [0.42, 0.28, 0.06, 0.24]. \qquad \diamondsuit$$

As we stated in the Introduction, in the precise probability framework any multidimensional distribution representable by a Bayesian network can also be represented in the form of a perfect sequence, and vice versa. An analogous result, although somewhat weaker, for perfect sequences of credal sets will be presented in the next section.

4 Credal Networks

In this section we will deal with credal networks and their relationship to credal compositional models.

4.1 Basic Concepts

A *credal network* [1] over X_N is (in analogy to Bayesian networks) a pair $(\mathcal{G}, \{\mathbf{P}^1, \ldots, \mathbf{P}^k\})$ such that, for any

$i = 1, \ldots, k$, $(\mathcal{G}, \mathbf{P}^i)$, is a Bayesian network over X_N, i.e., each $\mathbf{P}^i$ is a system of conditional probability distribution forming the joint distribution of X_N, $P^i(X_N)$.

The resulting model is a credal set, which is the convex hull of the Bayesian networks, i.e.,

$$\mathrm{CH}\{P^1(X_N), \ldots, P^k(X_N)\}.$$

It is evident that this definition loses the attractiveness of Bayesian networks, where the overall information is computed from local pieces of information. Let us denote by $\mathcal{CN}(X_N)$ the class of all credal networks over X_N.

The most popular (and also most effective) type of credal networks is represented by those called separately specified. A *separately specified credal network* over X_N is a pair $(\mathcal{G}, \mathbf{M})$, where $\mathbf{M}$ is a set of conditional credal sets $\mathcal{M}(X_i|pa(X_i))$ for each $X_i \in X_N$, and $pa(X_i)$ denotes the *set of parent variables* of X_i. Here the overall model is, in analogy to Bayesian networks, obtained as a strong extension of the $\mathcal{M}(X_i|pa(X_i)), i \in N$. Analogous to the previous paragraph, let us denote by $\mathcal{SCN}(X_N)$ the class of all separately specified credal networks over X_N.

Nevertheless, a lot of situations exist in which separately specified credal networks either cannot be used or their use leads to less specific models. For more details, the reader is referred to [1]; one extremely simple example can be found in the next subsection (Example 5).

4.2 Credal Networks and Perfect Sequences of Credal Sets

In this subsection we will prove, using the preceding results, a relationship between credal networks and perfect sequences of credal sets. For this purpose, let us denote by $\mathcal{CM}(X_N)$ the class of compositional models over X_N.

Theorem 5 *For any X_N*

$$\mathcal{SCN}(X_N) \subset \mathcal{CM}(X_N) \subset \mathcal{CN}(X_N). \qquad (13)$$

Proof. Let

$$(\mathcal{G}, \mathcal{M}(X_i|pa(X_i)), i \in N) \qquad (14)$$

be a separately specified credal network over X_N and N be ordered in such a way that $i > j \in pa(i)$ for each $i \in N$. The overall model (joint credal set describing X_N) is then obtained as a strong extension of credal sets from (14).

Let us define $\mathcal{M}_i(X_i \cup pa(X_i))$ as a strong extension of $\mathcal{M}(X_i|pa(X_i))$ and $\mathcal{M}(pa(X_i))$, where $\mathcal{M}(pa(X_i))$ is a marginal of the strong extension of $\mathcal{M}(X_j|pa(X_j)), j = 1, \ldots, i-1$. Now it easily follows that any $\mathcal{M}_i(X_i \cup pa(X_i))$ is a marginal of the strong extension of (14). Therefore, credal sets $\mathcal{M}_1(X_1), \ldots, \mathcal{M}_n(X_i \cup pa(X_n))$ form a perfect sequence defining the same joint model as (14).

If $\mathcal{M}_1(X_{K_1}), \ldots, \mathcal{M}_m(X_{K_m})$ is perfect, then according to Theorem 4

$$\begin{aligned}\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m \\ = \mathrm{CH}\{P_1 \triangleright P_2 \triangleright \ldots \triangleright P_m, P_i \in \mathrm{ext}(\mathcal{M}_i), \\ P_1, P_2, \ldots, P_m \text{ is perfect}\}.\end{aligned}$$

For any perfect sequence $P_1, P_2, \ldots, P_m$ a Bayesian network exists representing the distribution

$$P_1 \triangleright \ldots \triangleright P_m$$

such that, for each variable X_j, $\ell \in \{1, \ldots, m\}$ exists such that $(\{X_j\} \cup pa(X_j)) \subset \{X_i\}_{i \in K_\ell}$. Therefore,

$$\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \cdots \triangleright \mathcal{M}_m = \mathrm{CH}\{(\mathcal{G}^{\mathrm{i}}, \mathbf{P}^{\mathrm{i}}), 1, \ldots, \mathrm{k}\}.$$

As any perfect sequence represents the same system of conditional independences, it is evident that any Bayesian network can be defined on the same graph $\mathcal{G}$, which concludes the proof. □

For the description of an algorithm reconstructing a credal network from a perfect sequence of credal sets the reader is referred to the following subsection.

The following simple examples demonstrate that the inclusions in (13) are proper.

Example 4 Let X_1 and X_2 be two binary variables and P_1 and P_2 be defined as follows

$$\begin{array}{lll} P_1(x_1) = 0.4 & P_1(x_2|x_1) = 0.25 & P_1(x_2|\bar{x}_1) = 0.5, \\ P_2(x_1) = 0.6 & P_2(x_2|x_1) = 0.5 & P_2(x_2|\bar{x}_1) = 0.25. \end{array}$$

They form, together with the graph $X_1 \longrightarrow X_2$, two Bayesian networks. The corresponding credal network is

$$\mathrm{CH}\{[0.1, 0.3, 0.3, 0.3], [0.3, 0.3, 0.1, 0.3]\}. \qquad (15)$$

From these distributions one can get the following credal sets forming a perfect sequence

$$\begin{aligned}\mathcal{M}_1(X_1) &= \mathrm{CH}\{[0.4, 0.6], [0.6, 0.4]\}, \\ \mathcal{M}_2(X_1X_2) &= \mathrm{CH}\{[0.1, 0.3, 0.3, 0.3], [0.2, 0.2, 0.15, 0.45]\} \\ &\quad \{[0.15, 0.45, 0.2, 0.2], [0.3, 0.3, 0.1, 0.3]\}.\end{aligned}$$

It is evident that $\mathcal{M}_1 \triangleright \mathcal{M}_2(X_1X_2) = \mathcal{M}_2(X_1X_2)$, which also contains other Bayesian networks not contained in (15). ♢

Example 5 Let

$$\begin{aligned}&\mathcal{M}_1(X_1X_2)\\ &\quad = \mathrm{CH}\{[0.2,0.2,0,0.6],[0.1,0.4,0.1,0.4],\\ &\qquad [0.25,0.25,0.25,0.25],[0.2,0.3,0.3,0.2]\}.\end{aligned}$$

be a credal set describing variables X_1 and X_2 with values in $\mathbf{X}_1$ and $\mathbf{X}_2$ ($\mathbf{X}_i = \{x_i, \bar{x}_i\}$), respectively.

From its extreme points we obtain the following distributions:

$$\begin{array}{lll} P_1(x_2) = 0.2 & P_1(x_1|x_2) = 1 & P_1(x_1|\bar{x}_2) = 0.25\\ P_2(x_2) = 0.2 & P_2(x_1|x_2) = 0.5 & P_2(x_1|\bar{x}_2) = 0.5\\ P_3(x_2) = 0.5 & P_3(x_1|x_2) = 0.5 & P_3(x_1|\bar{x}_2) = 0.5\\ P_4(x_2) = 0.5 & P_4(x_1|x_2) = 0.4 & P_2(x_1|\bar{x}_2) = 0.6.\end{array}$$

These are, together with the graph $X_2 \longrightarrow X_1$, four Bayesian networks. Their convex hull is exactly the set $\mathcal{M}_1(X_1X_2)$. Nevertheless, it is not a separately specified credal network. To obtain that, we need the credal sets $\mathcal{M}(X_2)$, $\mathcal{M}(X_1|x_2)$ and $\mathcal{M}(X_1|\bar{x}_2)$.

Using (1) and (3), we obtain

$$\begin{aligned}\mathcal{M}(X_2) &= \mathrm{CH}\{[0.2,0.8],[0.5,0.5]\},\\ \mathcal{M}(X_1|x_2) &= \mathrm{CH}\{[1,0],[0.4,0.6]\},\\ \mathcal{M}(X_1|\bar{x}_2) &= \mathrm{CH}\{[0.25,0.75],[0.6,0.4]\}.\end{aligned}$$

The strong extension of these credal sets is

$$\begin{aligned}&\tilde{\mathcal{M}}_1(X_1X_2)\\ &\quad = \mathrm{CH}\{[0.2,0.2,0,0.6],[0.2,0.48,0,0.32],\\ &\qquad [0.08,0.2,0.12,0.6],[0.08,0.48,0.12,0.32],\\ &\qquad [0.5,0.125,0,0.375],[0.5,0.3,0,0.2],\\ &\qquad [0.2,0.125,0.3,0.375],[0.2,0.3,0.3,0.2]\}.\end{aligned}$$

which is less precise than the original model. ♢

It can be viewed as an advantage of compositional models that they are based on "local knowledge" even in cases when the credal network is not separately specified.

4.3 From Perfect Sequence to Credal Network

In this subsection we will present an algorithm for transforming a perfect sequence of credal sets to a credal network and we will illustrate its application on a simple example.

Having a perfect sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ ($\mathcal{M}_\ell$ being a credal set describing X_{K_ℓ}), we first order all of the variables for which at least one of the credal sets $\mathcal{M}_\ell$ is defined in such a way that first we order (in an arbitrary way) variables for which $\mathcal{M}_1$ is defined, then

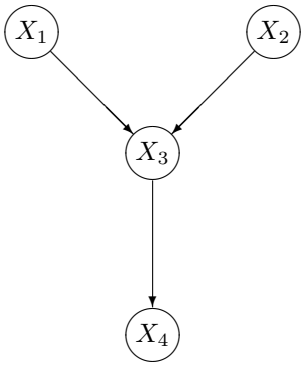


Figure 1: Graph of credal network generated from a perfect sequence

variables from $\mathcal{M}_2$ that are not contained in $\mathcal{M}_1$, etc. Finally we have

$$\{X_1, X_2, X_3, \ldots, X_n\} = \{X_i\}_{i \in K_1 \cup \ldots \cup K_m}.$$

Then we get a graph of the constructed evidential network in the following way:

(i) the nodes are all the variables $X_1, X_2, X_3, \ldots, X_n$;

(ii) there is an edge $(X_i \to X_j)$ if there exists a credal set $\mathcal{M}_\ell$ such that both $i, j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ and either $i \in K_1 \cup \ldots \cup K_{\ell-1}$ or $i < j$.

Having the structure of the credal network, i.e., graph $\mathcal{G}$, one can obtain the systems of conditional probability distributions from corresponding perfect sequences of probability distributions.

Evidently, for each j the requirement $j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ is met exactly for one $\ell \in \{1, \ldots, n\}$. It means that all the parents of node X_j must be from the respective set $\{X_i\}_{i \in K_\ell}$ and therefore the necessary conditional probability distributions $P^i(X_j|pa(X_j))$ can easily be computed from probability distribution P^i_ℓ.

Example 3 *(Continued)* From perfect sequence

$$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4,$$

we get the following ordering of variables

$$X_1, X_2, X_3, X_4$$

and the structure of the credal network as suggested in Figure 1. From six perfect sequences of probability distributions (12) one gets six systems of conditional probability distributions:

$$\begin{aligned}&P^1_1(X_1), P^1_2(X_2), P_3(X_3|X_1X_2), P^1_4(X_4|X_3),\\ &P^1_1(X_1), P^1_2(X_2), P_3(X_3|X_1X_2), P^2_4(X_4|X_3),\end{aligned}$$

$$P_1^1(X_1), P_2^2(X_2), P_3(X_3|X_1X_2), P_4^1(X_4|X_3),$$
$$P_1^1(X_1), P_2^2(X_2), P_3(X_3|X_1X_2), P_4^2(X_4|X_3),$$
$$P_1^2(X_1), P_2^2(X_2), P_3(X_3|X_1X_2), P_4^1(X_4|X_3),$$
$$P_1^2(X_1), P_2^2(X_2), P_3(X_3|X_1X_2), P_4^2(X_4|X_3),$$

where

$$P_1^1(X_1 = x_1) = 0.2, \qquad P_1^2(X_1 = x_1) = 0.5,$$
$$P_2^1(X_2 = x_2) = 0.5, \qquad P_2^2(X_2 = x_2) = 0.8,$$
$$P_3(X_3 = x_3|X_1 = x_1, X_2 = x_2) = 1,$$
$$P_3(X_3 = x_3|X_1 = x_1, X_2 = \bar{x}_2) = 0.75,$$
$$P_3(X_3 = x_3|X_1 = \bar{x}_1, X_2 = x_2) = 0.5,$$
$$P_3(X_3 = x_3|X_1 = \bar{x}_1, X_2 = \bar{x}_2) = 0.25,$$
$$P_4^1(X_4 = x_4|X_3 = x_3) = 0.8,$$
$$P_4^1(X_4 = x_4|X_3 = \bar{x}_3) = 0.4,$$
$$P_4^2(X_4 = x_4|X_3 = x_3) = 0.4,$$
$$P_4^2(X_4 = x_4|X_3 = \bar{x}_3) = 0.2.$$

The resulting model is again a credal set (11). ♢

5 Conclusions

This paper is devoted to the further development of the operator of composition for credal sets. Our main attention is paid to the relationship between so-called perfect sequences of credal sets, and those of (precise) probability distributions with the aim to find the relationship between credal compositional models and credal networks. We have proved that a perfect sequence of credal sets is a convex hull of perfect sequences of extreme points of these credal sets. We have also proved that perfect sequences of credal sets form a proper subclass of credal networks and, simultaneously, they are a proper superclass of separately specified credal networks. In other words, any separately specified credal network can be expressed in the form of credal compositional models and any perfect sequence of credal sets can be expressed as a credal network.

From the results presented in this paper it is evident that compositional models for credal sets can be seen as an alternative to credal networks. Therefore it seems desirable to further develop the composition operator within this framework. The first, and most important, task will be a definition of composition in the general case.

Acknowledgements

The support of Grant No. GAČR 13-20012S is gratefully acknowledged. The author would like to express her gratitude to anonymous referees for their inspiring comments.

References

[1] A. Antonucci and M. Zaffalon, Decision-theoretic specification of credal networks: a unified language for uncertain modeling with sets of Bayesian networks. *Int. J. Approx. Reasoning,* **49** (2008), 345–361.

[2] S. Benferhat, D. Dubois, L. Gracia and H. Prade, Directed possibilistic graphs and possibilistic logic. In: B. Bouchon-Meunier, R.R. Yager, (eds.) *Proc. of the 7th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98*, Editions E.D.K. Paris, pp. 1470–1477.

[3] B. Ben Yaghlane, Ph. Smets and K. Mellouli, Directed evidential networks with conditional belief functions. *Proceedings of ECSQARU 2003,* eds. T. D. Nielsen, N. L. Zhang, 291–305.

[4] I. Couso, S. Moral and P. Walley, Examples of independence for imprecise probabilities, *Proceedings of ISIPTA'99,* eds. G. de Cooman, F. G. Cozman, S. Moral, P. Walley, Ghent, 1999, pp. 121–130.

[5] F. G. Cozman, Sets of probability distributions, independence, and convexity, *Synthese*, **186** (2012), pp. 577–600.

[6] 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.

[7] 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.

[8] 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.

[9] R. Jiroušek and J. Vejnarová, Construction of multidimensional models by operators of composition: current state of art. *Soft Computing,* **7** (2003), pp. 328–335.

[10] R. Jiroušek, J. Vejnarová and M. Daniel, Compositional models for belief functions. *Proceedings of 5th International Symposium on Imprecise Probability: Theories and Applications ISIPTA'07*, eds. G. De Cooman, J. Vejnarová, M. Zaffalon, Praha, 2007, pp. 243–252.

[11] R. Jiroušek and J. Vejnarová, Compositional models and conditional independence in Evidence Theory, *Int. J. Approx. Reasoning,* **52** (2011), pp. 316–334.

[12] S. Moral and A. Cano, Strong conditional independence for credal sets, *Ann. of Math. and Artif. Intell.*, **35** (2002), pp. 295–321.

[13] J. Vejnarová, Composition of possibility measures on finite spaces: preliminary results. In: *Proc. of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98,* (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, 1998, pp. 25–30.

[14] J. Vejnarová, On possibilistic marginal problem, *Kybernetika* **43**, 5 (2007), pp. 657–674.

[15] J. Vejnarová, Evidential networks from a different perspective. In: *Synergies of Soft Computing and Statistics for Intelligent Data Analysis,* Soft Methods In Probability and Statistics, (2012). pp. 429–436.

[16] J. Vejnarová, Operator of composition for credal sets, *ISIPTA'13: 8th International Symposium on Imprecise Probability: Theories and Applications*, pp. 355–364.

On the Validity of Minimin and Minimax Methods for Support Vector Regression with Interval Data

Andrea Wiencierz
Department of Mathematics
University of York
andrea.wiencierz@york.ac.uk

Marco E. G. V. Cattaneo
Department of Mathematics
University of Hull
m.cattaneo@hull.ac.uk

Abstract

In the recent years, generalizations of support vector methods for analyzing interval-valued data have been suggested in both the regression and classification contexts. Standard Support Vector methods for precise data formalize these statistical problems as optimization problems that can be based on various loss functions. In the case of Support Vector Regression (SVR), on which we focus here, the function that best describes the relationship between a response and some explanatory variables is derived as the solution of the minimization problem associated with the expectation of some function of the residual, which is called the risk functional. The key idea of SVR is that even when considering an infinite-dimensional space of arbitrary regression functions, given a finite-dimensional data set, the function minimizing the risk can be represented as the finite weighted sum of kernel functions. This allows to practically determine the SVR estimate by solving a much simpler optimization problem, even in the case of nonlinear regression. In case that only interval-valued observations of the variables of interest are available, it has been suggested to minimize the minimal or maximal risk values that are compatible with the imprecise data, yielding precise SVR estimates on the basis of interval data. In this paper, we show that also in the case of an interval-valued response the optimal function can be represented as the finite weighted sum of kernel functions. Thus, the minimin and minimax SVR estimates can be obtained by minimizing the corresponding simplified expressions of the empirical lower and upper risks, respectively.

Keywords. Support Vector Regression, interval data, Representer Theorem.

1 Introduction

In this paper, we deal with the generalization of Support Vector Regression (SVR) to interval data. By SVR we denote a class of kernel-based methods for the statistical problem of regression analysis. These methods originated in the field of Machine Learning (Vapnik, 1998, 1995) and recently also gained attention in the field of Statistics (see, e.g., Hable, 2012; Christmann et al., 2009; Hofmann et al., 2008; Steinwart and Christmann, 2008). The typical goal of a regression analysis is to describe the relationship between a response variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and a number $d \in \mathbb{N}$ of explanatory variables $X \in \mathcal{X} \subseteq \mathbb{R}^d$ by a function $f : \mathcal{X} \to \mathbb{R}$. The sought-after function f is usually assumed to be a member of a particular space $\mathcal{F}$ of considered regression functions, for example, the space of all (affine) linear functions.

To identify which functions in $\mathcal{F}$ best describe the relationship between the random variables in $(X, Y) = V$, the considered regression functions are assessed by a loss function. Most common loss functions are characteristics of the distribution of (some function of) the residual R_f, which we here define by

$$R_f = |Y - f(X)|$$

for each $f \in \mathcal{F}$. In the SVR methodology, the expectation of some usually convex error function is considered as loss function, which is called risk functional. If the probability distribution P_V of the random vector V is known, the distribution of R_f can be derived from it and the best regression functions can be identified by minimizing the chosen loss function. Yet, usually the true distribution of the investigated variables is unknown, but it is assumed that P_V lies in some specific set of probability measures $\mathcal{P}_V$. Thus, the evaluation of each regression function also varies over possible distributions of V.

Given the realization of an independent sample of random variables $V_1 = (x_1, y_1), \ldots, V_n = (x_n, y_n)$, with $n \in \mathbb{N}$, where $V_i \sim P_V$ for all $i \in \{1, \ldots, n\}$, we can learn something about the distribution of the variables of interest. In SVR, the empirical distribution $\hat{P}_V$ of the observations is used as a point estimate of P_V and the (regularized) risk under this particular distribu-

tion is minimized to obtain the regression estimate. The SVR estimate is in general unique. Moreover, the so-called Representer Theorem states that the function minimizing the risk given the observations can be represented as the finite weighted sum of kernel functions. This is a key result for SVR, as it allows to practically determine the SVR estimate by solving a relatively simple optimization problem, even in the case of nonlinear regression. Further details of the SVR methodology are presented in the next section.

If the variables of interest are not observed as precise numbers but only upper and lower bounds to the values are available, the empirical distribution $\hat{P}_V$ is not revealed by the observable data. We denote the random sets describing the observables by $V_1^*, \ldots, V_n^*$ and their probability distribution by P_{V^*}. If the observed intervals are assumed to cover the unknown precise values with probability one, bounds for the empirical risk can be derived from the empirical distribution $\hat{P}_{V^*}$ of the imprecise data. How can we use this information to obtain an SVR estimate in this situation? Starting from the simplified representation of the optimal function in standard SVR, Utkin and Coolen (2011) proposed to follow a minimin or a minimax approach and to minimize either the lower or the upper (regularized) risk in order to obtain a precise regression estimate.

In this paper, we investigate the validity of their starting from the simplified representation in the generalized data situation. At first, we introduce the formal framework of the SVR methodology in detail and formally discuss Utkin and Coolen (2011)'s SVR generalization. Then, we consider the Representer Theorem in the more general data situation. We find that also in this case the optimal function can be represented as the finite weighted sum of kernel functions. Finally, after applying the discussed SVR methods to an interesting problem in the area of winemaking, a short outlook concludes the paper.

2 Methodological Framework of SVR

In this section, the formal framework of SVR with precise data is presented. In the SVR methodology, the set $\mathcal{P}_V$ is assumed to contain all probability measures on $\mathcal{V} = \mathcal{X} \times \mathcal{Y}$. In this paper, we additionally assume that $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. Furthermore, in SVR, the loss assigned to a possible regression function f and a distribution P_V is the risk $\mathcal{E}_{P_V}(f)$. Presupposing measurability, the risk functional $\mathcal{E}_{P_V}$ on $\mathcal{F}$ can be defined for each $P_V \in \mathcal{P}_V$ as

$$\mathcal{E}_{P_V} \,:\, f \mapsto \mathcal{E}_{P_V}(f) = \mathbb{E}_{P_V}(\psi(R_f)), \qquad (1)$$

where ψ is a convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ satisfying $\psi(0) = 0$ and $\mathbb{E}_{P_V}$ denotes the expectation with respect to P_V. For example, if ψ is defined by $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, the loss associated with a pair (f, P_V) is given by $\mathcal{E}_{P_V}(f) = \mathbb{E}_{P_V}(R_f^2)$. Thus, we obtain the loss function corresponding to Least Squares regression. Another famous example is the function defined by $\psi(r) = \max\{0, r - \nu\}$, for all $r \in \mathbb{R}_{\geq 0}$ and some $\nu \geq 0$, which was introduced by Vapnik (1995, Section 6.1) and represents the so-called ν-insensitive loss.

The convexity of the mapping ψ implies convexity of the risk functional $\mathcal{E}_{P_V}$, that is, the risk functional satisfies for each $\rho \in [0, 1]$

$$\mathcal{E}_{P_V}\left(\rho\, f + (1-\rho)\, f'\right) \leq \rho\, \mathcal{E}_{P_V}(f) + (1-\rho)\, \mathcal{E}_{P_V}(f'),$$

for all $f, f' \in \mathcal{F}$ (see also Steinwart and Christmann, 2008, Lemma 2.13). As explained later, this property is crucial to the existence of a unique optimal regression function.

In the SVR framework, the space $\mathcal{F}$ of considered regression functions from $\mathcal{X}$ to $\mathbb{R}$ is supposed to be a Reproducing Kernel Hilbert Space (RKHS) with associated scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$. An RKHS is uniquely associated with its reproducing kernel function. A kernel function κ is a positive semi-definite function on $\mathcal{X} \times \mathcal{X}$, that is, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i\, \alpha_j\, \kappa(x_i, x_j) \geq 0$, for all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $n \in \mathbb{N}$. Here, we only consider kernel functions that are moreover measurable and bounded. If κ is the reproducing kernel function of the RKHS $\mathcal{F}$, for each $x \in \mathcal{X}$ we have $\kappa(\cdot, x) \in \mathcal{F}$ and

$$f(x) = \langle f, \kappa(\cdot, x) \rangle_{\mathcal{F}},$$

for all $f \in \mathcal{F}$. From this property called reproducing property follows that $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x') \rangle_{\mathcal{F}}$, for all $x, x' \in \mathcal{X}$. A simple example for an RKHS and its reproducing kernel is the function space associated with the linear kernel defined by $\kappa(x, x') = \langle x, x' \rangle + 1$, for all $x, x' \in \mathcal{X}$, which is the Hilbert space of all (affine) linear functions from $\mathcal{X}$ to $\mathbb{R}$. Another common kernel function is the so-called Gaussian kernel, which is defined for all $x, x' \in \mathcal{X}$ by

$$\kappa(x, x') = \exp\left(-\tfrac{1}{\sigma^2}\, \|x - x'\|^2\right),$$

with $\sigma > 0$. The associated RKHS is a very large function space that is dense in the space of all continuous (real-valued) functions on $\mathcal{X}$. For more details on kernels and RKHSs, see, for example, Steinwart and Christmann (2008, Chapter 4).

To avoid obtaining too wiggly functions as descriptions of the relationship of interest when the regression analysis is based on a finite sample of observations,

the risk is further supplemented by an additive penalty for the complexity of the functions $f \in \mathcal{F}$. Hence, in the SVR methodology, instead of $\mathcal{E}_{P_V}$ the regularized risk functional $\mathcal{E}_{P_V,\lambda}$ is minimized, which is defined for all $f \in \mathcal{F}$ by

$$\mathcal{E}_{P_V,\lambda}(f) = \mathcal{E}_{P_V}(f) + \lambda \|f\|_{\mathcal{F}}^2,$$

where $\lambda > 0$ is a fixed parameter regulating the penalization and $\|\cdot\|_{\mathcal{F}}$ is the norm induced by the scalar product in $\mathcal{F}$. The regularization can be interpreted as minimizing $\mathcal{E}_{P_V}$ under the restriction $\|f\|_{\mathcal{F}}^2 \leq c$, for some $c \in \mathbb{R}_{\geq 0}$, but instead of choosing the bound c explicitly, we fix the value of the corresponding Lagrange multiplier λ in the constrained optimization problem.

As the functional $f \mapsto \lambda \|f\|_{\mathcal{F}}^2$ is strictly convex by general properties of norms and $\mathcal{E}_{P_V}$ is convex because of ψ, we have that $\mathcal{E}_{P_V,\lambda}$ is also a strictly convex functional on $\mathcal{F}$. Exploiting the strict convexity of $\mathcal{E}_{P_V,\lambda}$, it can be shown that an optimal function always exists and is unique, provided that some regularity conditions are fulfilled (see, e.g., Steinwart and Christmann, 2008, Lemma 5.1 and Theorem 5.2).

Given observations $(x_1, y_1), \ldots, (x_n, y_n)$ of an independent and identically distributed random sample $V_1, \ldots, V_n$, the SVR methodology consists in estimating P_V by the corresponding empirical distribution $\hat{P}_V$, before identifying the regression estimate $f_{\hat{P}_V,\lambda} \in \mathcal{F}$ by the minimization of $\mathcal{E}_{\hat{P}_V,\lambda}$, for some $\lambda > 0$. Like in the general case, there always exists a unique minimizer of the regularized risk for $\hat{P}_V$. Moreover, the so-called Representer Theorem states that this unique function $f_{\hat{P}_V,\lambda}$ can be represented as the linear combination of the corresponding functions $\kappa(\cdot, x_1), \ldots, \kappa(\cdot, x_n)$, that is, there exist weights $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f_{\hat{P}_V,\lambda}(x) = \sum_{j=1}^{n} \alpha_j \, \kappa(x, x_j), \tag{2}$$

for all $x \in \mathcal{X}$ (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5). This expression is sometimes called support vector expansion of $f_{\hat{P}_V,\lambda}$ and the optimal function $f_{\hat{P}_V,\lambda}$ is often referred to as a Support Vector Machine (SVM). This term can be explained historically, because Vapnik (1998, 1995) proposed to use functions for ψ that have the property that some of the resulting $\alpha_1, \ldots, \alpha_n$ are zero. The vectors x_j for which $\alpha_j \neq 0$ are called support vectors, whence the notion SVM. One example for such a representing function ψ is the function associated with the ν-insensitive loss mentioned before. Nevertheless, in general, SVMs are not sparse in this sense (see, e.g., Steinwart and Christmann, 2008, Section 11.1).

The result of the Representer Theorem expressed in (2) is extremely useful for the practical computation of SVR estimates as it simplifies the associated optimization problems and allows to solve them even when large RKHSs of arbitrary smooth regression functions are considered, like, for example, the RKHS associated with the Gaussian kernel. Given a data set $(x_1, y_1), \ldots, (x_n, y_n)$ with empirical distribution $\hat{P}_V$ and a fixed $\lambda > 0$, Equation (2) tells us that $f_{\hat{P}_V,\lambda}$ is an element of the set $\mathcal{F}_n \subset \mathcal{F}$, with

$$\mathcal{F}_n = \left\{ \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, x_j) \,:\, \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\}.$$

Furthermore, for all functions $f_\alpha = \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, x_j)$, with $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}} \in \mathbb{R}^n$, the squared norm is given by $\|f_\alpha\|_{\mathcal{F}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j)$. Hence, the regularized risk associated with $\hat{P}_V$ can be written for each $f_\alpha \in \mathcal{F}_n$ as

$$\begin{aligned} \mathcal{E}_{\hat{P}_V,\lambda}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \psi\big(\big|y_i - \textstyle\sum_{j=1}^{n} \alpha_j \, \kappa(x_i, x_j)\big|\big) \\ + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j). \end{aligned}$$

As $\mathcal{E}_{\hat{P}_V,\lambda}$ is convex, the SVM $f_{\hat{P}_V,\lambda}$ can be obtained by solving a convex optimization problem over $\alpha \in \mathbb{R}^n$, for which there are numerous efficient algorithms (see, e.g., Boyd and Vandenberghe, 2004). For the selection of an appropriate regularization parameter $\lambda > 0$ and of other hyper-parameters like the parameter σ of the Gaussian kernel, different strategies can be applied, for instance, cross-validation (see, e.g., Steinwart and Christmann, 2008, Section 11.3). Since we are mainly interested in the generalization of a key theoretical result about SVR to the situation with interval data, we neglect the latter issues in this paper and always consider these parameters fixed.

3 SVR with Interval Data

In this section, we investigate whether the SVR methodology can be used for regression analysis when the variables of interest cannot be observed as precise numbers but only (bounded) intervals covering the values of interest are available. Utkin and Coolen (2011) proposed a generalization of the SVR methodology to this situation. As we will see later, the suggested methods of Utkin and Coolen (2011) work well for interval-valued observations of the response variable Y, but cannot directly be extended to interval-valued observations of the variables in X. Therefore, we also consider here only the situation where instead of V the random set $V^* \in \mathcal{V}^* \subseteq 2^{\mathcal{V}}$ is observed, whose possible realizations are of the form $\{X\} \times [\underline{Y}, \overline{Y}]$, with $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\underline{Y}, \overline{Y} \in \mathcal{Y} \subset \mathbb{R}$ such that $\underline{Y} \leq \overline{Y}$.

3.1 Utkin and Coolen (2011)'s SVR Generalization

Now, we discuss the generalization of SVR proposed by Utkin and Coolen (2011) in detail. Since in the considered data situation the precise variables are not observable, it is impossible to evaluate the considered regression functions $f \in \mathcal{F}$ by $\mathcal{E}_{\hat{P}_V}(f)$, i.e., by the risk associated with the empirical distribution of the precise data. However, the probability distribution of the imprecise data P_{V^*} can be estimated on the basis of the observations.

When the probability distribution P_{V^*} of the observable data is known, as we assume that the interval $[\underline{Y}, \overline{Y}]$ covers the precise unobservable Y with probability one, we know that the unknown probability distribution of the precise data lies in the set $[P_{V^*}] \subseteq \mathcal{P}_V$ containing all distributions of the precise data, P_V, that satisfy for all measurable events $A \subseteq \mathcal{V}$ the inequalities

$$\begin{aligned} &P_V(V \in A) \geq P_{V^*}(V^* \subseteq A) \quad \text{and} \\ &P_V(V \in A) \leq P_{V^*}(V^* \cap A \neq \varnothing). \end{aligned} \tag{3}$$

By consequence, for all $f \in \mathcal{F}$, the unknown risk $\mathcal{E}_{P_V}(f)$ lies in the interval $[\underline{\mathcal{E}}_{P_{V^*}}(f), \overline{\mathcal{E}}_{P_{V^*}}(f)]$, where

$$\begin{aligned} \underline{\mathcal{E}}_{P_{V^*}}(f) &= \min_{P'_V \in [P_{V^*}]} \mathcal{E}_{P'_V}(f) \quad \text{and} \\ \overline{\mathcal{E}}_{P_{V^*}}(f) &= \max_{P'_V \in [P_{V^*}]} \mathcal{E}_{P'_V}(f). \end{aligned}$$

Hence, in the regression problem with interval-valued response, the set $[\underline{\mathcal{E}}_{P_{V^*}}(f), \overline{\mathcal{E}}_{P_{V^*}}(f)]$ of all possible risk values constitutes the loss evaluation for each $f \in \mathcal{F}$. Of course, it is in general impossible to directly determine an optimal function with respect to this imprecise criterion. The central idea of the regression methodology proposed by Utkin and Coolen (2011) is to use the minimin or the minimax rule to solve this problem, that is, to minimize either the lower risk $\underline{\mathcal{E}}_{P_{V^*}}$ or the upper risk $\overline{\mathcal{E}}_{P_{V^*}}$ in order to identify a single optimal regression function.

To derive expressions of the lower and upper risks, Utkin and Coolen (2011) describe, for each regression function $f \in \mathcal{F}$, the set of compatible probability distributions of the residual R_f given P_{V^*} by a so-called p-box and apply results from Utkin and Destercke (2009). Introduced by Ferson et al. (2003, Section 2), the notion p-box designates a convex set of probability measures for a univariate random quantity that is bounded by a lower and an upper cumulative distribution function. In the situation considered here, given P_{V^*}, also the marginal distribution of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, where

$$\begin{aligned} \underline{R}_f &= \min_{(x,y) \in V^*} |y - f(x)| \quad \text{and} \\ \overline{R}_f &= \max_{(x,y) \in V^*} |y - f(x)|, \end{aligned}$$

is known for each $f \in \mathcal{F}$. According to (3), the marginal distribution of the imprecise residual implies lower and upper bounds to the probabilities of all measurable events associated with the marginal distribution of the precise residual R_f. If we consider these lower and upper bounds for all events of the form $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, we obtain a lower and an upper cumulative distribution function that constitute a p-box. As the p-box covers all probability distributions of R_f that comply with the bounds at least for the intervals $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, some of the probability measures included in the p-box may not satisfy (3) for all measurable events, and thus, may be incompatible with the marginal distribution of the imprecise residual. However, the p-box obtained in the described way from the random set $[\underline{R}_f, \overline{R}_f]$, with $f \in \mathcal{F}$, is the tightest outer approximation by a p-box of the set of probability distributions of R_f implied by this random set (see, e.g., Destercke et al., 2008). In fact, in the present situation, for each $f \in \mathcal{F}$, the upper bound of the associated p-box corresponds to the cumulative distribution function of the lower endpoint of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, while the lower bound of the p-box corresponds to the cumulative distribution function of the upper endpoint. This can be seen by considering the corresponding bounds to the probabilities of the events $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, used to derive the p-box for all $f \in \mathcal{F}$, that is,

$$\begin{aligned} &P_V(R_f \leq r) \geq P_{V^*}([\underline{R}_f, \overline{R}_f] \subseteq (-\infty, r]) \quad \text{and} \\ &P_V(R_f \leq r) \leq P_{V^*}([\underline{R}_f, \overline{R}_f] \cap (-\infty, r] \neq \varnothing) \end{aligned}$$

It can easily be checked that the probability distributions corresponding to the bounds of the p-box comply with (3) for arbitrary measurable events, and thus, are elements of $[P_{V^*}]$. Since, according to its definition in (1), the risk functional $\mathcal{E}_{P_V}$ is the expectation of a convex function in R_f with minimum at zero, it is straightforward to conclude that $\underline{\mathcal{E}}_{P_{V^*}}$ and $\overline{\mathcal{E}}_{P_{V^*}}$ coincide with the expected errors associated with the marginal distributions of the lower and of the upper residual, that is, of $\underline{R}_f$ and of $\overline{R}_f$, respectively (see also Utkin and Destercke, 2009, Proposition 3).

Now consider that the realization of an independent sample of random sets $V_1^* = A_1, \ldots, V_n^* = A_n$ is observed, where $V_i^* \sim P_{V^*}$ for all $i \in \{1, \ldots, n\}$. Then, by analogy with standard SVR, P_{V^*} is estimated by the empirical distribution $\hat{P}_{V^*}$ of the imprecise data,

and furthermore, the complexity of the estimated functions is restricted by an additive penalty term. Hence, the optimization criteria considered in the minimin and minimax generalizations of SVR are the regularized lower and upper risk, respectively. For a fixed penalization parameter $\lambda > 0$, the regularized lower and upper risks associated with the empirical distribution $\hat{P}_{V^*}$ can, for each $f \in \mathcal{F}$, be expressed as follows:

$$\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} \min_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \|f\|_{\mathcal{F}}^2,$$
$$\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} \max_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \|f\|_{\mathcal{F}}^2, \tag{4}$$

where ψ is again the convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ representing the chosen loss.

Utkin and Coolen (2011) deduce from these expressions of the regularized empirical lower and upper risks solvable formulations of the optimization problems corresponding to both suggested strategies in the special case of linear regression for different choices of the loss function. We do not restrict the approach to this special case here and continue to consider more general RKHSs of regression functions. Moreover, Utkin and Coolen (2011) start from the support vector expansion (2) of the solution of the optimization problem corresponding to standard SVR. However, it first has to be verified that the Representer Theorem applies to or that its statements can be transferred to the setting with interval data. Only in this case, the simple expression (2) can be used for the optimal regression function in (4), providing the favorable starting point for solving the corresponding optimization problems.

3.2 The Representer Theorem for SVR with Interval-Valued Response

As mentioned in the previous subsection, the Representer Theorem implies that if an SVR analysis of a precise data set $V_1 = (x_1, y_1), \ldots, V_n = (x_n, y_n)$ with empirical distribution $\hat{P}_V$ is based on a convex representing function ψ, then, for all $\lambda > 0$, there exists a unique function minimizing $\mathcal{E}_{\hat{P}_V,\lambda}$, which can be represented as (2) (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5). In the proof of this theorem as it is presented in Steinwart and Christmann (2008, Theorem 5.5), the first steps are to show strict convexity and continuity of $\mathcal{E}_{\hat{P}_V,\lambda}$, which provide existence and uniqueness of the minimizing function $f_{\hat{P}_V,\lambda} \in \mathcal{F}$, by the corresponding arguments of the proofs of Theorem 5.2 and Lemma 5.1 of Steinwart and Christmann (2008), respectively. Then, the representation of $f_{\hat{P}_V,\lambda}$ as the kernel expansion of (2) is derived by exploiting properties of the function spaces $\mathcal{F}_n$ and $\mathcal{F}$ in addition to the existence and the uniqueness of the function $f_{\hat{P}_V,\lambda}$.

The generalized SVR methods discussed in this section differ from the standard SVR methods only in the expressions of their risks. Hence, we have to derive the crucial properties of convexity and continuity for the lower and upper risks to be able to transfer the arguments proving the simplified expression of $f_{\hat{P}_V,\lambda}$ to the situation with interval-valued response. In the following lemma, we derive for the general case that the regularized lower and upper risks have unique minimizers, before we prove Theorem 1, stating that the functions minimizing the regularized empirical lower and upper risks can be expressed as in Equation (2).

Lemma 1. *The regularized lower and upper risk functionals*

$$\underline{\mathcal{E}}_{P_{V^*},\lambda} : f \mapsto \underline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \quad \textit{and}$$
$$\overline{\mathcal{E}}_{P_{V^*},\lambda} : f \mapsto \overline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2$$

have unique minimizers $f_{P_{V^*},\lambda}^{\text{minimin}}$ and $f_{P_{V^*},\lambda}^{\text{minimax}}$ in $\mathcal{F}$, respectively.

Proof. Since κ is bounded, convergence in the norm $\|\cdot\|_{\mathcal{F}}$ implies convergence in the norm $\|\cdot\|_\infty$, because using the Cauchy–Schwarz inequality,

$$\begin{aligned}\|f\|_\infty &= \sup_{x\in\mathcal{X}} \|f(x)\| = \sup_{x\in\mathcal{X}} \|\langle f, \kappa(\cdot,x)\rangle_{\mathcal{F}}\| \\ &\leq \sup_{x\in\mathcal{X}} \|f\|_{\mathcal{F}} \sqrt{\langle \kappa(\cdot,x), \kappa(\cdot,x)\rangle_{\mathcal{F}}} \\ &= \|f\|_{\mathcal{F}} \sup_{x\in\mathcal{X}} \sqrt{\kappa(x,x)}\end{aligned}$$

for all $f \in \mathcal{F}$. Therefore, the functionals $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are continuous on $\mathcal{F}$ (with respect to the norm $\|\cdot\|_{\mathcal{F}}$), because they are the sum of the continuous functional $\lambda\|\cdot\|_{\mathcal{F}}^2$ with the lower and upper previsions of $\psi(R_f)$, respectively, and ψ is uniformly continuous on the relevant domain (since it is convex, and $\mathcal{Y}$ is bounded).

Moreover, $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are strictly convex functionals on $\mathcal{F}$, since $\lambda\|\cdot\|_{\mathcal{F}}^2$ is strictly convex, and the unregularized lower and upper risk functionals $\underline{\mathcal{E}}_{P_{V^*}}$ and $\overline{\mathcal{E}}_{P_{V^*}}$ can be shown to be convex. The proof for the upper risk functional is simple, since $\overline{\mathcal{E}}_{P_{V^*}}$ is the maximum of the convex functionals $\mathcal{E}_{P'_V}$ with $P'_V \in [P_{V^*}]$. By contrast, the proof for the lower risk functional is more involved. We start by noting that for each possible realization $A = \{x\} \times [\underline{y}, \overline{y}] \in \mathcal{V}^*$ of the random set V^*, the function

$$\underline{r}_A : z \mapsto \min_{\underline{y}\leq y\leq \overline{y}} |y - z| = \begin{cases} \underline{y} - z & \text{if } z < \underline{y}, \\ 0 & \text{if } \underline{y} \leq z \leq \overline{y}, \\ z - \overline{y} & \text{if } \overline{y} < z, \end{cases}$$

on $\mathbb{R}$ is convex, and therefore $\psi \circ \underline{r}_A$ is convex too, since ψ is convex and nondecreasing. This implies that

$$\min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V \,|\, V^*} \left(\psi(R_f) \,|\, V^* = A\right) = \left(\psi \circ \underline{r}_A\right)\left(f(x)\right)$$

is a convex functional of f, and so is

$$\begin{aligned}\underline{\mathcal{E}}_{P_{V^*}}(f) &= \min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V}\left(\psi(R_f)\right) \\ &= \mathbb{E}_{P_{V^*}}\left(\min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V \,|\, V^*}\left(\psi(R_f) \,|\, V^*\right)\right).\end{aligned}$$

So far we have proven that $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are continuous and strictly convex functionals on $\mathcal{F}$. The desired result is implied by Theorem A.6.9 of Steinwart and Christmann (2008), since the sets

$$\left\{f \in \mathcal{F} : \underline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \leq \underline{\mathcal{E}}_{P_{V^*}}(0)\right\} \quad \text{and}$$
$$\left\{f \in \mathcal{F} : \overline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \leq \overline{\mathcal{E}}_{P_{V^*}}(0)\right\}$$

are nonempty and bounded (with respect to the norm $\|\cdot\|_{\mathcal{F}}$). □

Theorem 1. *There exist* $\alpha_1^{\text{minimin}}, \ldots, \alpha_n^{\text{minimin}} \in \mathbb{R}$ *and* $\alpha_1^{\text{minimax}}, \ldots, \alpha_n^{\text{minimax}} \in \mathbb{R}$ *such that*

$$f_{\hat{P}_{V^*},\lambda}^{\text{minimin}} : x \mapsto \sum_{i=1}^{n} \alpha_i^{\text{minimin}} \kappa(x, x_i) \quad \textit{and}$$
$$f_{\hat{P}_{V^*},\lambda}^{\text{minimax}} : x \mapsto \sum_{i=1}^{n} \alpha_i^{\text{minimax}} \kappa(x, x_i)$$

are the unique minimizers of $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ *and* $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ *in* $\mathcal{F}$*, respectively.*

Proof. Let f' denote the orthogonal projection of a function $f \in \mathcal{F}$ on the subspace $\mathcal{F}_n$ spanned by the functions $\kappa(\cdot, x_i)$ with $i \in \{1, \ldots, n\}$. Then $\|f'\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}}$, and f' is of the form $\sum_{i=1}^{n} \alpha_i \kappa(\cdot, x_i)$ with $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Moreover, for each $i \in \{1, \ldots, n\}$, the orthogonality of $f' - f$ and $\kappa(\cdot, x_i)$ implies $f'(x_i) = f(x_i)$, because

$$f'(x_i) - f(x_i) = \langle f' - f, \kappa(\cdot, x_i)\rangle_{\mathcal{F}} = 0.$$

Therefore, $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f') \leq \underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f)$ and $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f') \leq \overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f)$, and the desired result is implied by Lemma 1. □

Hence, $f(x_i)$ can indeed be replaced by a support vector expansion in the expressions of $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ given in (4), and the derivation of solvable formulations of the corresponding optimization problems can be based on the thereby simplified expressions of the risks.

However, the above results cannot directly be generalized to accounting also for interval-valued observations of the explanatory variables. This is because, when V^* is of the form $[\underline{X}^{(1)}, \overline{X}^{(1)}] \times \ldots \times [\underline{X}^{(d)}, \overline{X}^{(d)}] \times [\underline{Y}, \overline{Y}]$, in general $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ is no longer convex, and moreover, Theorem 1 does not apply to $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ anymore.

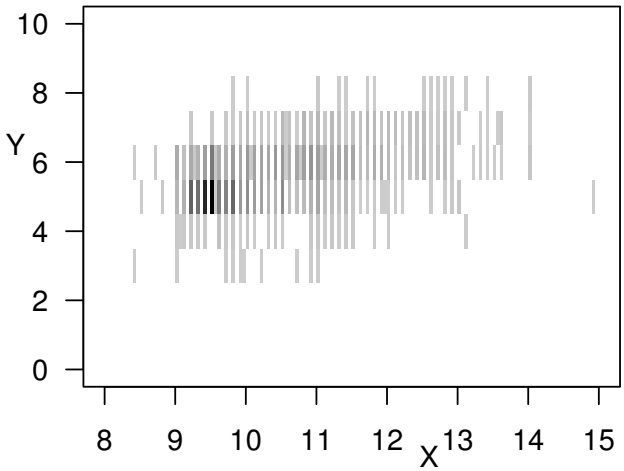


Figure 1: Histogram plot of the red wine data set with $n = 1\,599$ observations. The darker a line segment the more observations overlap this line segment.

4 SVR Analysis of Wine Quality

In this section, we analyze a data set collected to study the quality of Vinho Verde wines from Portugal. The data were obtained from wine samples that were tested by the official certification entity of the system of protected designation of origin of the Vinho Verde wines between May 2004 and February 2007. For each of the included 1 599 red and 4 898 white wines, 11 physicochemical characteristics and an evaluation of the sensory quality are available. The data set was initially analyzed by Cortez et al. (2009) and is freely available from the UC Irvine Machine Learning Repository (Lichman, 2013). Here, we focus on the subsample of red Vinho Verde wines and study the relationship between taste and alcohol content.

In the data set, the sensory quality of the wine is measured on a discrete scale ranging from 0 – *very bad* to 10 – *excellent*. These discrete quality measurements should, in fact, be considered as coarse observations of an underlying continuous variable taking values in $[0, 10]$. Therefore, instead of analyzing the discrete values as if they were precise measurements of the wine quality, we consider them to be interval data and replace the discrete values $0, 1, \ldots, 9, 10$ by the intervals $[0, 0.5], [0.5, 1.5], \ldots, [8.5, 9.5], [9.5, 10]$, respectively. The alcohol content of the wines is available as volume percent of alcohol, which we here assume to be measured with sufficient precision.

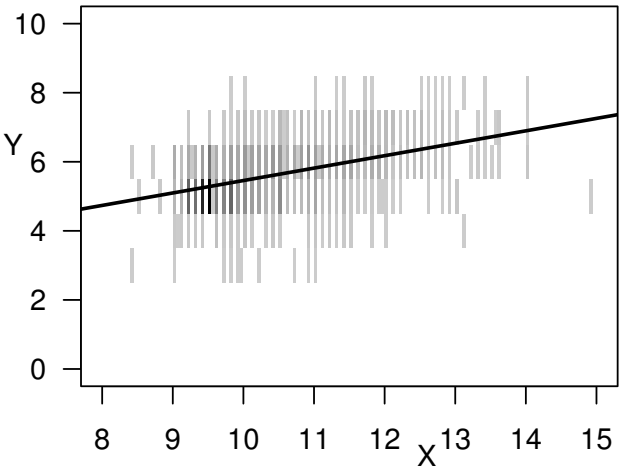


Figure 2: Minimax function of the generalized SVR analysis with linear kernel, $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, and $\lambda = 0.0001$.

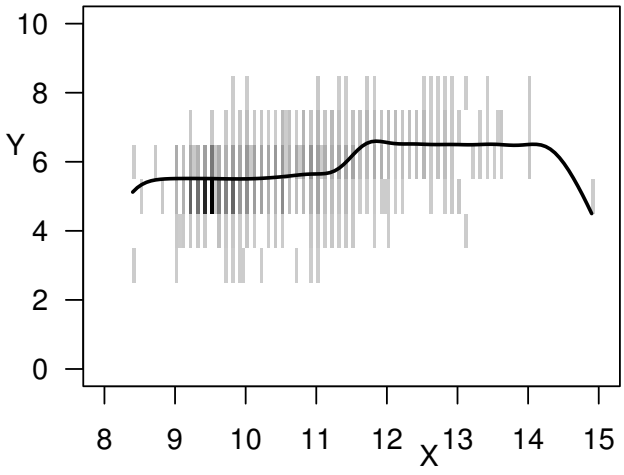


Figure 3: Minimax function of the generalized SVR analysis with Gaussian kernel, $\psi(r) = r$ for all $r \in \mathbb{R}_{\geq 0}$, and $\lambda = 0.000001$.

Hence, we analyze the relationship between the precisely observed alcohol content and the imprecisely observed sensory quality of the red Vinho Verde wine. Thus, as we consider only one explanatory variable here, the imprecise data are line segments. The analyzed data set is displayed in Figure 1, where X is the alcohol level in percent by volume and Y corresponds to the sensory quality. All graphs and computations are realized in the statistical software environment R (R Core Team, 2014), resorting amongst others to functions provided by the packages `kernlab` (Karatzoglou et al., 2004) and `quadprog` (Turlach and Weingessel, 2013).

A red wine lover would probably hypothesize that the higher the alcohol content of a red wine, the stronger and possibly better the taste of the wine. As also the data suggest a positive linear relationship, in the first instance, we choose the linear kernel function for the SVR analysis, although SVR is not limited to linear regression. Furthermore, we consider the Least Squares loss, i.e., we set $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$. This configuration of SVR corresponds to what is also known as Ridge regression. As the minimax approach appears to be more cautious, we consider the corresponding generalized SVR method of Utkin and Coolen (2011) here. Finally, for the estimation, the regularization parameter λ is set to 0.0001. The estimated regression line confirms the surmise of a positive relationship between alcohol content and sensory quality of the Vinho Verde red wines and is displayed in Figure 2.

As the assumption of a linear relationship is very strict, we alternatively consider the minimax SVR method based on the Gaussian kernel with parameter σ equal to 1. Furthermore, we consider the absolute loss here represented by ψ defined as $\psi(r) = r$ for all $r \in \mathbb{R}_{\geq 0}$ and set $\lambda = 0.000001$. The estimated regression function is depicted in Figure 3 and shows an increasing tendency in those areas of the observation space $\mathcal{V} = [8, 15] \times [0, 10]$ where most observations are. Hence, also the more general SVR analysis provides evidence for a positive relationship between alcohol content and sensory quality of red Vinho Verde wines.

5 Conclusion and Outlook

In this paper, we investigated the generalized SVR methods for regression with interval data that were initially proposed by Utkin and Coolen (2011). These methods consist in minimizing either the minimal or the maximal regularized risk compatible with the empirical distribution of the imprecise data. In this paper, we proved that the corresponding optimal functions can be represented as the weighted sum of kernel functions and thereby provide the so far lacking justification for the regression methods derived in Utkin and Coolen (2011). Hence, the minimin and minimax SVR methods constitute sensible adaptations of the SVR methodology to interval data and yield interesting results when applied to real data as in the previous section.

We here focused on the data situation where only for the response variable there are interval-valued observations, while the explanatory variables are precisely observed. Unfortunately, our findings cannot simply be generalized to account also for interval-valued observations of the explanatory variables, because then the regularized lower risk is no longer necessarily convex and the Representer Theorem cannot be transferred to the regularized upper risk anymore. This means that

for the minimin SVR method there is not necessarily a unique optimal function and that the optimal minimax function cannot be expanded as in Equation (2). This indeed limits the applicability of the minimin and minimax SVR methods to the more restrictive setting considered in this paper. Moreover, the meaning of the estimated regression functions is less clear than in the precise data case.

Furthermore, it can be argued that, in the context of the statistical analysis of imprecise data, methods yielding precise results are in general problematic, because a reasonable statistical method should reflect the imprecision of the data in its result. In addition, a responsible statistical analysis should always take the involved statistical uncertainty into account. A regression methodology for imprecise data allowing to express these two types of uncertainty at the same time constitutes the so-called Likelihood-based Imprecise Regression (LIR) methodology introduced by Cattaneo and Wiencierz (2012). In the LIR methodology, each possible regression function is evaluated by the whole set of loss values that are plausible in the light of the data and then the set of all undominated regression functions is considered as the imprecise result of the regression analysis, which can furthermore be interpreted as a confidence set. As it can be shown that, for each $f \in \mathcal{F}$, the interval $[\underline{\mathcal{E}}_{\hat{P}_{V^*}}(f), \overline{\mathcal{E}}_{\hat{P}_{V^*}}(f)]$ is the Maximum Likelihood estimate of $\mathcal{E}_{P_V}(f)$ in the situation considered in Section 3, Utkin and Coolen (2011)'s SVR methods can be further generalized by embedding them in the LIR framework.

References

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *International Journal of Approximate Reasoning* 53, 1137–1154.

Christmann, A., Van Messem, A., and Steinwart, I. (2009). On consistency and robustness properties of Support Vector Machines for heavy-tailed distributions. *Statistics and Its Interface* 2, 311–327.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 547–553.

Destercke, S., Dubois, D., and Chojnacki, E. (2008). Unifying practical uncertainty representations: I. Generalized p-boxes. *International Journal of Approximate Reasoning* 49, 649–663.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D., and Sentz, K. (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. Technical Report SAND2002-4015, Sandia National Laboratories.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis* 106, 92–117.

Hofmann, T., Schölkopf, B., and Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistics* 36, 1171–1220.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software* 11, 1–20.

Lichman, M. (2013). *UCI Machine Learning Repository*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Used R version 2.15.2.

Steinwart, I., and Christmann, A. (2008). *Support Vector Machines*. Springer.

Turlach, B., and Weingessel, A. (2013). *quadprog: Functions to solve Quadratic Programming Problems*. R package version 1.5-5.

Utkin, L., and Coolen, F. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, 371–380.

Utkin, L., and Destercke, S. (2009). Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning* 50, 778–798.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Poster Abstracts

M-Estimation with Imprecise Data

Marco E. G. V. Cattaneo
Department of Mathematics
University of Hull
m.cattaneo@hull.ac.uk

Real data often do not have the level of precision required by conventional statistical methods. In particular, a data point can be incompletely observed, in the sense that the only available observation is a set known to contain the data point. An important problem is then how to perform statistical estimation, and in particular regression, when some (or all) data points are incompletely observed. This problem has recently attracted much attention in the statistical literature in general, and at ISIPTAs in particular: see for example Cattaneo and Wiencierz (2012); Liu and Vandal (2011); Schollmeyer and Augustin (2015); Utkin and Coolen (2011).

The typical setting in these works is that instead of the precise data points $x_i \in \mathcal{X}$, only the sets $s_i \subseteq \mathcal{X}$ are observed. It is assumed that $x_i \in s_i$, but no other information about x_i is available. In particular, precisely observed data points x_i can be represented by singletons $s_i = \{x_i\}$, while missing data points x_i can be represented by observations $s_i = \mathcal{X}$. The statistical problem consists in estimating a quantity of interest $\theta \in \Theta$ on the basis of the data.

In the case of precisely observed data, most statistical estimation methods can be expressed as M-estimators (or slight generalizations thereof):

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\min_{\theta \in \Theta} \textstyle\sum_{i=1}^{n} \rho(x_i, \theta), \qquad (1)$$

where $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ describes some kind of estimation error. For example, when $\mathcal{X} = \Theta = \mathbb{R}$, the squared error $\rho(x_i, \theta) = (x_i - \theta)^2$ leads to the least squares estimation of location.

An apparently very intuitive idea for generalizing an estimator $\hat{\theta}$ to the case of incompletely observed data is to interpret

$$\{\hat{\theta}(x_1, \ldots, x_n) : x_i \in s_i\} \qquad (2)$$

as the set-valued estimate based on the observations s_i. However, for M-estimators an alternative approach is possible: replacing $\sum_{i=1}^{n} \rho(x_i, \theta)$ with

$$\{\textstyle\sum_{i=1}^{n} \rho(x_i, \theta) : x_i \in s_i\} \qquad (3)$$

(or its convex hull) in the minimization task (1). Since the quantity (3) to be minimized is set-valued, several definitions of minimum are possible and can lead to different kinds of estimators.

The present work investigates the imprecise minimization approach (3) and compares it with the set of estimates approach (2). Both approaches have interesting connections with the statistical method of estimating equations, and face some difficulties in parametric models. An important advantage of the former is the possibility, if desired, of easily obtaining a precise estimate, for example by interpreting the minimization as a minimax problem. By contrast, the interpretation of the set-valued estimates intrinsically tied to the latter approach is difficult, because they mix aspects of the different statistical concepts of point estimate and confidence region.

Keywords. M-estimator, regression, imprecise data, interval data, coarse data, missing data, estimating equations, robust statistics, set-valued estimates.

References

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *Int. J. Approx. Reasoning* 53, 1137–1154. [based on an ISIPTA '11 paper]

Liu, X., and Vandal, A. C. (2011). Bounds for self-consistent CDF estimators for univariate and multivariate censored data. In *ISIPTA '11*, 267–276.

Schollmeyer, G., and Augustin, T. (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reasoning* 56, 224–248. [based on an ISIPTA '13 paper]

Utkin, L. V., and Coolen, F. P. A. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11*, 371–380.

An Idea of Consonant Conflicts between Belief Functions

Milan Daniel
Institute of Computer Science, Academy of Sciences of the Czech Republic
milan.daniel@cs.cas.cz

Introduction General belief functions usually bear some internal conflict, which comes mainly from disjoint focal elements. Analogously there is often some conflict between two (or more) belief functions (BFs). This theoretical contribution introduces a new approach to conflicts of BFs. Conflicts between BFs are here considered independently of any combination rule and of any distance measure.

Consonant Conflicts The suggested approach is based on consonant approximations of BFs in general; two important special cases based on consonant inverse pignistic and consonant inverse plausibility transformations are discussed. Their idea is based on our previous study of conflicts of BFs [1, 2, 3].

Probabilistic approximations of belief functions were used in several previous approaches, e.g. pignistic probability in W. Liu's two-dimensional degree of conflict and in pignistic conflict [2], and normalized plausibility of singletons in plausibility conflict [1, 2].

Unfortunately, doing a probability approximation usually adds new conflicting information, which increases internal conflict of input beliefs and also resulting global conflict. There are many inverses of any probabilistic approximation, in general (a mapping back to original input BFs among them), nevertheless, there are unique consonant inverses of both pignistic and plausibility probabilistic transformations. These inverses are internally non-conflicting (they have no internal conflict). Thus the entire global conflict of these approximations is the conflict between them (there is no conflict inside them). Our present idea is use of consonant instead of probabilistic approximations.

Definitions Let *the consonant inverse contour approximation* $iC(Bel)$ of a BF Bel be the unique consonant inverse of *the normalized plausibility of singletons (normalized contour function)* corresponding to Bel.

Let *the consonant inverse pignistic approximation* $iBet(Bel)$ of a BF Bel be the unique consonant inverse of *the pignistic probability* corresponding to Bel.

Let Bel_1, Bel_2 be any belief functions on any frame Ω, $iC(Bel_i)$ and $iBet(Bel_i)$ be their consonant inverse contour and consonant inverse pignistic approximations given by consonant bbas ${}_{iC}m_i$, ${}_{iBet}m_i$. *The inverse contour conflict* is defined by the formula $iC\text{-}Conf(Bel_1, Bel_2) = \sum_{X\cap Y=\emptyset} {}_{iC}m_1(X){}_{iC}m_2(Y)$, where $X, Y \subseteq \Omega$. *The inverse pignistic conflict* is analogously defined by $iBet\text{-}Conf(Bel_1, Bel_2) = \sum_{X\cap Y=\emptyset} {}_{iBet}m_1(X){}_{iBet}m_2(Y)$, where $X, Y \subseteq \Omega$.

Properties In [4] we have proved an equivalence of the consonant conflict $iC\text{-}Conf$ with the conflict between BFs based on their con-conflicting parts [3]. For quasi Bayesian BFs (focal elements: $|X| = 1$ or $X = \Omega$) Bel_1, Bel_2 with bbas m_1, m_2 we have proved: $Conf(Bel_1, Bel_2) \leq \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y)$ for both $iC\text{-}Conf$ and $iBet\text{-}Conf$. Note that this does not hold for general BFs. For more detail, general counterexample, and other properties see [4].

Keywords. Belief functions, Dempster-Shafer theory, internal conflict of a belief function, conflict between belief functions, consonant approximation.

References

[1] M. Daniel. Conflicts within and between Belief Functions. In: E. Hüllermeier, et al. (eds.) *IPMU 2010*. LNAI 6178, 696–705, Springer, Heidelberg, 2010.

[2] M. Daniel. Belief Functions: a Revision of Plausibility Conflict and Pignistic Conflict. In: W. Liu, V. S. Subrahmanian, J. Wijsen (eds.) *SUM 2013*. LNCS (LNAI) vol. 8078, pp. 190–203. Springer, 2013.

[3] M. Daniel. Conflict between Belief Functions: a New Measure Based on their Non-Conflicting Parts. In: F. Cuzzolin (eds.) *BELIEF 2014*. LNCS (LNAI) vol. 8764, pp. 321–330. Springer, Heidelberg, 2014.

[4] M. Daniel. *An Introduction to Consonant Conflicts between Belief Functions* Technical Report ICS AS CR, Prague (In preparation).

Convergence of Continuous-Time Imprecise Markov Chains

Jasper De Bock
Ghent University, SYSTeMS Research Group
jasper.debock@ugent.be

We provide necessary and sufficient conditions for the unique convergence of a continuous-time imprecise Markov chain to a stationary distribution.

Problem Statement Consider the set of all the continuous-time non-stationary Markov chains with finite state space $\mathcal{X}$ of which the transition rate matrix Q_t is a function of time such that $Q_t \in \mathcal{Q}$, where $\mathcal{Q}$ is a closed convex set of transition rate matrices that has *separately specified rows*, meaning that

$$Q \in \mathcal{Q} \Leftrightarrow (\forall x \in \mathcal{X})\ Q(x,*) \in \mathcal{Q}_x$$

where, for all $x \in \mathcal{X}$, $\mathcal{Q}_x$ is a closed convex set of row vectors. We call such a set of Markov chains a *continuous-time imprecise Markov chain*.

Fix any $t > 0$. Then for all $f \in \mathbb{R}^{\mathcal{X}}$ and $x \in \mathcal{X}$, the expected value $E_t(f|X_0 = x)$ of f at time t, conditional on $X_0 = x$, ranges over a closed interval of which we will denote the lower bound by $\underline{T}_t(f|x)$. For all $x \in \mathcal{X}$, $\underline{T}_t(\cdot|x)$ is a *coherent lower prevision* on $\mathbb{R}^{\mathcal{X}}$. The corresponding *lower transition operator* $\underline{T}_t : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is defined by

$$\underline{T}_t f(x) \coloneqq \underline{T}_t(f|\,x) \text{ for all } x \in \mathcal{X}.$$

By a recent result of Škulj [1], $\underline{f}_t \coloneqq \underline{T}_t f$ is the solution to the differential equation

$$\frac{d}{dt}\underline{f}_t = \underline{Q}\,\underline{f}_t$$

with initial condition $\underline{f}_0 = f$, where for all $h \in \mathbb{R}^{\mathcal{X}}$:

$$\underline{Q}h(x) \coloneqq \min_{Q \in \mathcal{Q}} \sum_{y \in \mathcal{X}} Q(x,y)h(y) \text{ for all } x \in \mathcal{X}.$$

We study the limit behaviour of $\underline{T}_t$. In particular, we provide necessary and sufficient conditions for $\mathcal{Q}$ to be *Perron-Frobenius-like (PF)*, meaning that there is some $\underline{P}_\infty : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ such that, for all $x \in \mathcal{X}$:

$$\lim_{t \to +\infty} \underline{T}_t f(x) = \underline{P}_\infty f \text{ for all } f \in \mathbb{R}^{\mathcal{X}},$$

or, equivalently, for $\underline{T}_t(\cdot|\,x)$ to converge to a stationary distribution $\underline{P}_\infty$ that does not depend on x.

Results Our main result is that the following four conditions are equivalent:

1. $\mathcal{Q}$ is PF,
2. $\underline{T}_t$ is PF for some $t > 0$,
3. $\underline{T}_t$ is PF for all $t > 0$,
4. $\mathcal{Q}$ is regularly absorbing,

where (i) for any $t > 0$, we say that $\underline{T}_t$ is PF if the discrete-time imprecise Markov chain that has $\underline{T}_t$ as its lower transition operator is PF, in the sense that, for all $f \in \mathbb{R}^{\mathcal{X}}$, $\lim_{n\to\infty} \underline{T}_t^n f$ exists and is constant and (ii) '*regularly absorbing*' is a qualitative property of $\mathcal{Q}$ that is fully determined by the signs of the *upper transition rates to singletons* $\overline{Q}(x,y) \coloneqq \max_{Q\in\mathcal{Q}} Q(x,y)$ and the *lower transition rates to sets* $\underline{Q}(x,A) \coloneqq \min_{Q\in\mathcal{Q}} \sum_{y\in A} Q(x,y)$, for $x, y \in \mathcal{X}$, $x \neq y$ and $A \subset \mathcal{X} \setminus \{x\}$. See the poster for more details.

As future work, we would like to develop *coefficients of ergodicity* that characterise whether $\mathcal{Q}$ is PF and that provide—tight—bounds on the rate of convergence. So far, we have found a coefficient of ergodicity whose positivity is sufficient—but not necessary—for $\mathcal{Q}$ to be PF and which, in that case, provides a conservative bound on the rate of convergence.

Acknowledgements Many thanks to Gert de Cooman, Matthias C. M. Troffaes and Stavros Lopatatzidis for stimulating discussions on the topic of continuous-time imprecise Markov chains.

Keywords. Perron-Frobenius, continuous-time imprecise Markov chains, convergence, lower and upper transition rates, coefficients of ergodicity.

References

[1] Damjan Škulj. Efficient computation of the bounds of continuous time imprecise Markov chains. *Applied Mathematics and Computation*, 250:165–180, 2015.

Probabilistic Analysis of Sutural Lines Developed in Ammonites. An Example: Lower Jurassic Hammatoceratacea

Andrea Di Cencio
Department of Engineering and Geology,
University G.d'Annunzio, Chieti, Italy
andreadicencio@geologiaepaleontologia.eu

Serena Doria
Department of Engineering and Geology,
University G.d'Annunzio, Chieti, Italy
s.doria@dst.unich.it

Motivations Ammonites are extinct ectococled molluscs belonging to the Class Cephalopoda which lived during the Mesozoic Era. Their usefulness in Jurassic and Cretaceous paleontology and biostratigraphy study is widely proved. For this reason, they are studied by several authors worldwide in order to achieve information regarding their habitats and climate of past world. Coherent upper conditional previsions defined with respect to Hausdorff outer measures are used to make a probabilistic analysis of the paleoenvironmental causes that generated complex sutural lines. In particular, the role of hydrostatic pressure is studied.

Sutural Lines The shell of ammonites is sub-divisible in three parts: protoconch, phragmocone and body chamber. The phragmocone is divided in chambers separated by septa. The geometric projection of septum on the inner side of the shell is the sutural line. Every sutural line is characterized by alternation of several elements, named saddles and lobes, which reflect a fractal geometrical development. Sutural lines of Toarcian (lower Jurassic) ammonites are made of almost two separated groups. The first is close to mathematical model of the von Koch curve and the latter is close to mathematical model of the Cesaro curve [1]. These two models are associated to different hydrodynamic arrangements which correspond to two different life strategies. The von-Koch-sutural-line is related to good swimmer ammonites which show hydrodynamic features as oxycone section, developed keel, sinuous ribs and short body chambers [4]. The Cesaro-sutural-line is related to no good swimmer ammonites which are characterized by no hydrodynamic features as rounded sections, little keel, strong ribs, spines and very long body chambers [2].

Probabilistic Analysis In order to study the paleoenvironmental causes of the complexity of the sutural lines, we interpret ammonites as complex systems whose evolution during time is described by a finite family of contractions; the attractor of this family represents the sutural line, whose complexity is measured in terms of Hausdorff dimension. The hydrostatic pressure is represented by a random variable and we calculate the Choquet integral of this random variable given the sutural line, which is the conditioning event [3]. We consider a constant pressure and a strictly monotone pressure, corresponding to different life style. Different cases are studied according to the complexity of the sutural line.

Conclusions The results show that the Choquet integral of the hydrostatic pressure given the sutural line is a mathematical tool to describe different life styles of ammonites, which determined the complexity in the sutural lines.

Keywords. Ammonites, sutural lines, Toarcian, von Koch model, Cesaro model, hydrostatic pressure, Hausdorff dimension, Choquet integral.

References

[1] G. Damiani, Computer simulation of some ammonoid suture lines, In Pallini et al. eds. "Atti II Conv. Int. F.E.A., Pergola, 1987", 221-228, 1990.

[2] A. Di Cencio, Position of spines and tubercles in ammonites. Correlation with shape of shell and complexity of sutural line. Hypothesis of life style. In "XV Giornate di Paleontologia, 27-29 Maggio 2015", Abstract Volume, (accepted).

[3] S. Doria, Characterization of a coherent upper conditional prevision as the Choquet integral with respect to its associated Hausdorff outer measure, Annals of Operations Research, 33-48, 2012.

[4] F. Venturi and S. Rossi, Subasio, Origine e vicende di un Monte Appenninico, Porzi editoriali, 1-112, 2003.

Bayesian Updating Based on Hausdorff Outer Measures and the Role of Emotions During the Therapeutic Phase of Alliance

Serena Doria
Department of Engineering and Geology,
University G.d'Annunzio, Chieti, Italy
s.doria@dst.unich.it

Iolanda Angelucci
Psychotherapist and teacher
CTA Trainer at SSPIG-IRPIR
iolanda381@virgilio.it

Formulating diagnosis is a complex process, related to the clinician's ability to represent the patient's discomfort, to use error due to the incompleteness of the information available, to make predictions about well-being. We interpret the therapist-patient system as a complex system, whose evolution, representing the phase of alliance, is described by a finite family of contractions that, starting from certain initial conditions, evolve the system into the attractor; this set, characterized by its own complexity, measured in terms of the Hausdorff dimension, represents the state in which the therapist and patient find themselves after the phase of alliance.

Updating the Level of Knowledge and Making a Successful Diagnosis A probabilistic approach of the diagnostic process is proposed in which the subject's degree of knowledge is represented with coherent upper conditional probabilities defined by Hausdorff outer measures [1]. Using this model, the diagnosis is assumed to be positive when it produces a change, that is when the subject's level of knowledge is defined by an a posteriori Hausdorff outer measure different from the initial Hausdorff outer measure. We believe that one of the roles of the therapist in the phase of the alliance (i.e. interactive and collaborative relationship between patient and therapist, common to different psychotherapies, where both have an active role in achieving therapeutic goals) is to shorten the distance between him and the patient so that he can update the level of cognitive and emotional understanding of the problem the patient asks for his help for. The first step that the therapist must take is to realize that he is a complex system and small perturbations to the initial state, i.e. the encounter with the patient, can bring to totally unpredictable states, from which he has to assess the probability of success of the diagnosis. The phase of the therapist-patient alliance can be interpreted as the phase in which the complexity is likely to increase. In the mathematical model, the role of the therapist is represented by choosing a particular system of contractions, the similarities, that keep unchanged some geometric properties. These invariance of geometric properties aims to describe the fact that some features of the therapist are repeated at different scales, influencing the diagnostic attitude. By iterating these contractions, the patient-therapist system reaches a state represented by a self-similar set, called the attractor of the system; if the attractor has zero probability with respect to the Hausdorff measure that defines the initial level of knowledge of the patient then another measure needs to be used to represent the subject's level of knowledge conditioned to the attractor. The goal of the phase of the alliance is therefore to have the patient to confront with an unpredictable state, represented by a set having initial probability of zero value.

Conclusions According to the mathematical representation of the diagnostic process highlighted in this work, we find similarities with the idea of Matte Blanco [2] that emotion can undergo endless measurements. The attractor of the system represents the unconscious of the system therapist-patient and according to the theory developed in [2] it is characterized by symmetry and self-similarity.

Keywords. Iterated functions system, Hausdorff outer measure, coherent upper conditional probabilities, symmetry, self-similarity.

References

[1] S. Doria, Characterization of a coherent upper conditional prevision as the Choquet integral with respect to its associated Hausdorff outer measure, Annals of Operations Research, 33-48, 2012

[2] Matte Blanco, I. The Unconscious as Infinite Sets. An Essay in BiLogic. Duckworth, London 1975.

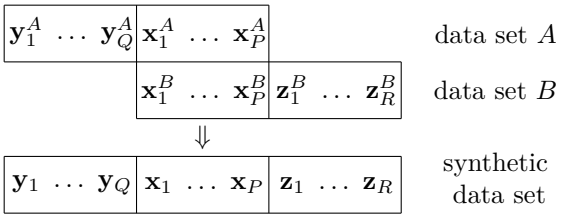
Probabilistic Graphical Models for Statistical Matching

Eva Endres
Department of Statistics, LMU Munich
eva.endres@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

In the information age a massive amount of data is available. It can be of great benefit to use this existing data for secondary analysis instead of collecting new data, which might be time-consuming and expensive. But what can be done if the required variables are not all accessible in one single data set? The solution is given by statistical matching: With the aid of statistical matching, information from different surveys can be combined.

The initial situation of statistical matching [2, e.g.] are two (or more) data sets, e.g. A and B with n_A or n_B observations, respectively, that contain information on a set of common variables $\mathbf{X}$, and specific variables $\mathbf{Y}$ and $\mathbf{Z}$ which are not jointly observed. The observation units in the different data sets are not the same.
The objective is, on the one hand, to estimate the joint probability distribution of all common and specific variables (*macro approach*) or, on the other hand, to generate one synthetic data set, that contains information on all variables of interest (*micro approach*).



The most popular statistical matching strategies are premised on the restrictive assumption of conditional independence, i.e. the independence of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$. This technical assumption makes the joint distribution of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ identifiable and, thus, estimable for $A \cup B$ ($\in \mathbb{R}^{(n_A+n_B)\times(P+Q+R)}$), where $A \cup B$ is an incomplete i.i.d. sample from $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ without joint information on $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ [2, cf.].

Here, it is proposed to perform statistical matching by graphical network models. This might be a promising alternative to existing statistical matching approaches, since it provides a natural form of representing conditional independence. In addition, the use of auxiliary information for solving the statistical matching problem remains possible.

In a first step, one Bayesian network [3, e.g.] has to be created on each of the data sets to be matched. Random variables are represented by nodes and the dependencies between them are displayed by arcs.

Afterwards, the individual networks can be linked to a single one by means of graph union or graph intersection, respectively.

The second step will be the application of credal networks [1, e.g.] in this setting. Thereby, the uncertainty of the statistical matching process can be taken into consideration by sets of compatible contingency tables. Moreover, the strict conditional independence assumption can be weakened by using independence concepts for sets of conditional probabilities.

Keywords. Statistical matching, Bayesian networks, credal networks, independence.

References

[1] A. Antonucci, C. P. de Campos, and M. Zaffalon. Probabilistic graphical models. In T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 207–229. Wiley, 2014.

[2] M. D'Orazio, M. Di Zio, and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, 2006.

[3] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Optimal Control of Linear Systems with Quadratic Cost and Imprecise Forward Irrelevant Input Noise

Alexander Erreygers and **Jasper De Bock** and **Gert de Cooman** and **Arthur Van Camp**
Ghent University, SYSTeMS Research Group
{alexander.erreygers, jasper.debock, gert.decooman, arthur.vancamp}@UGent.be

Preliminaries We consider a finite-state, discrete-time stationary linear system with a deterministic (known) initial state $X_0 = x_0$. For all $k \in \{0, \dots, n\}$, the dynamics of the system is described by

$$X_{k+1} = aX_k + bu_k(X_{0:k}) + W_k.$$

In this expression, a and b are real-valued parameters and the state X_k and noise W_k at time k are real-valued random variables. The control input u_k at time k is also real-valued and is taken to be some function of the previous states $x_{0:k}$. We call a tuple of control input functions $u_{0:n} \coloneqq (u_0, u_1, \dots, u_n)$ a *control policy*. We measure the performance of such a control policy by means of the associated linear quadratic cost

$$\eta\left[u_{0:n} \middle| x_0\right] \coloneqq \sum_{k=0}^{n} ru_k(X_{0:k})^2 + qX_{k+1}^2,$$

where r is a strictly positive real number and q is a non-negative real number.

The Precise Case If the uncertain noise terms W_k are modelled by means of a probability measure, an optimal control policy is usually required to minimise the expected value of the cost. Under some relatively weak technical assumptions, there will be a unique control policy $\hat{u}_{0:n}$ that satisfies this optimality criterion. If the noise is white—if the noise terms at different time instants are uncorrelated—then for all $k \in \{0, \dots, n\}$, this optimal control policy is given by

$$\hat{u}_k(x_{0:k}) \coloneqq -\tilde{r}_k b(m_{k+1}ax_k + h_k), \tag{1}$$

where the parameters $\tilde{r}_k$, m_{k+1} and h_k are derived from the initial conditions $m_{n+1} \coloneqq q$ and $h_{n+1} \coloneqq 0$ and, for all $k \in \{0, \dots, n\}$, the recursive expressions $\tilde{r}_k \coloneqq (r + b^2 m_{k+1})^{-1}$, $m_k \coloneqq q + a^2 \tilde{r}_k r m_{k+1}$ and

$$h_k \coloneqq m_{k+1}\mathrm{E}\left(W_k\right) + a\tilde{r}_{k+1} r h_{k+1}, \tag{2}$$

where $\mathrm{E}\left(W_k\right)$ is the expected value of W_k. In general, if the noise is not white, computing the optimal control policy $\hat{u}_{0:n}$ is intractable.

The Imprecise Case Our contribution consists in studying a generalised version of this problem, where the noise is described by an imprecise uncertainty model—a set of probability measures—and the optimal control policies are those that are E-admissible—that minimise the expected value of the cost for at least one element of this set. We show that if the model for the noise is *forward irrelevant* [1]—an imprecise notion of independence—then the corresponding set of optimal control policies is again characterised by Eq. (1). The only difference is that h_k is not given by Eq. (2), but instead takes values in some interval. If $a \geq 0$, then for all $k \in \{0, \dots, n\}$, the exact lower and upper bounds of this interval are

$$\underline{h}_k = m_{k+1}\underline{\mathrm{E}}(W_k) + a\tilde{r}_{k+1} r \underline{h}_{k+1},$$
$$\overline{h}_k = m_{k+1}\overline{\mathrm{E}}(W_k) + a\tilde{r}_{k+1} r \overline{h}_{k+1},$$

with $\overline{h}_{n+1} = \underline{h}_{n+1} \coloneqq 0$, and $\underline{\mathrm{E}}(W_k)$ and $\overline{\mathrm{E}}(W_k)$ the lower and upper expectations of W_k. If $a \leq 0$, $\underline{h}_{k+1}$ and $\overline{h}_{k+1}$ switch places. At first sight, these bounds might seem to follow trivially from Eq. (2), but this is *not* the case, because the optimisation ranges over a forward irrelevant set of probability measures, almost *none* of whose members corresponds to white noise.

In this way, for any time k and state history $x_{0:k}$, we obtain an interval of optimal control inputs. Nevertheless, in a practical control situation, a single control input has to be chosen. The most obvious or lazy choice is to apply the control input which, amongst the ones in the interval, has the lowest absolute value. In future work, we would like to investigate how this type of lazy control performs in practice.

Keywords. Linear system, quadratic cost, optimal control, imprecise noise, forward irrelevance.

References

[1] Gert de Cooman and Enrique Miranda. Forward irrelevance. *Journal of Statistical Planning and Inference*, 139(2):256 – 276, 2009.

Decision Theory Meets Linear Optimization Beyond Computation

Christoph Jansen
Department of Statistics
University of Munich
Germany
christoph.jansen@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics
University of Munich
Germany
augustin@stat.uni-muenchen.de

We consider the standard model of *finite* decision theory: An *actor* has to decide which *action* to pick from a finite set $\mathbb{A} = \{a_1, \dots, a_n\}$ of alternatives. However, the *utility* of the chosen action depends on which *state of nature* from a finite set $\Theta = \{\theta_1, \dots, \theta_m\}$ corresponds to the true description of reality. Specifically, we assume that the utility of every pair $(a, \theta) \in \mathbb{A} \times \Theta$ can be evaluated by a *known* map $u : \mathbb{A} \times \Theta \to \mathbb{R}$.

Within this framework, our goal is to determine an *optimal* action. However, any appropriate definition of optimality depends on (what we assume about) the *mechanism generating the states of nature*. Here, traditional decision theory mainly covers two *extremes*: The mechanism follows a *known* probability measure ξ on $(\Theta, \mathcal{P}(\Theta))$ or can be compared to a *game against an omniscient enemy*. Then optimality is almost unanimously defined by the *Bernoulli-criterion* (w.r.t. ξ) or the *Maximin-criterion*, respectively.

In contrast, defining optimality becomes less obvious if we consider ξ only *partially* known. Here, *imprecise probabilities* offer a powerful framework: Uncertainty is now described by the *credal set* of all the measures being compatible with our information (or by *linear partial information*, see [4]). However, criteria for optimal decision making now strongly depend on the actor's *attitude towards ambiguity*. Accordingly, many concurring decision criteria exist: Γ-*maximin,* Γ-*maximax, maximality, E-admissibility* [3, e.g.].

For determining optimal decisions w.r.t. these complex criteria *linear programming theory* (*LPT*) is well-suited: By embedding decision problems into this framework, one can draw on the whole toolbox of this well-investigated discipline. Particularly, this allows a computational treatment of complex decision making in statistical standard software (e.g. R): Proposals for linear programming based algorithms for optimizing all criteria mentioned above are given in [1] and [2].

However, the opportunities using *LPT* in decision theory are not exhausted by producing algorithms: Applying results from *LPT* provides deep theoretical insights on the connection between decision criteria as well as on the properties of optimal actions.

Firstly, we demonstrate the computational strength of *LPT* by recalling algorithms from [1] and [2] and exemplifying their implementation in R. Additionally, we introduce two algorithms for checking maximality of pure actions by solving *one single* linear program.

Secondly, we illuminate the power of *LPT* apart from algorithmic considerations: *Duality theory* from *LPT* is used to derive connections between optimal *randomized* Γ-maximin actions and *pure* Bernoulli-optimal actions w.r.t. a *least favourable* measure contained in the underlying credal set $\mathcal{M}$. We show that for every randomized $\Gamma(\mathcal{M})$-Maximin-optimal action p^*, there exists a pair $(a^*, \pi^*) \in \mathbb{A} \times \mathcal{M}$ such that $\mathbb{E}_{\pi^*}(u(a^*, \cdot))$ equals the $\Gamma(\mathcal{M})$-Maximin utility of p^*.

Keywords. decision making, imprecise probabilities, linear programming, partial information, ambiguity

References

[1] L.V. Utkin, T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In: F.G. Cozman, R. Nau, T. Seidenfeld (eds.): *ISIPTA '05*, 2005, pp. 349-358.

[2] D. Kikuti, F.G. Cozman, 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*, 2005, pp. 118-123.

[3] N. Huntley, R. Hable, M.C.M. Troffaes. Decision making. In: *Introduction to imprecise probabilities*. Ed. by T. Augustin, F.P.A. Coolen, G. de Cooman, M.C.M. Troffaes. Chichester: Wiley, 2014, pp. 190-206.

[4] E. Kofler, G. Menges. *Entscheiden bei unvollständiger Information*. Springer. Berlin (Lecture Notes in Economics and Mathematical Systems, 136), 1976.

Searching for the Most Plausible Partition: an Evidential Reasoning Approach to Clustering

Orakanya Kanjanatarakul
Heudiasyc UMR 7253
Université de Technologie de Compiègne & CNRS
Sorbonne Universités, France
Faculty of Management Sciences
Chiang Mai Rajabhat University, Thailand
okanjana@utc.fr

Thierry Denoeux
Heudiasyc UMR 7253
Université de Technologie de Compiègne & CNRS
Sorbonne Universités, France
tdenoeux@utc.fr

Clustering can be seen as the search for a "good" partition of a set of n objects described either by attributes, or by a dissimilarity matrix. Usual approaches are based either on a geometric criterion, as in the k-means algorithm, or on a finite mixture model whose parameters are estimated using, e.g., the EM algorithm. Here, we propose a different view of partitional clustering, in which dissimilarities are seen as pieces of evidence and represented as belief functions on the set of all partitions of the dataset under study. Using a technique similar to the one used in [1] for the association problem, we show that the most plausible partition can be found for small n. We then propose a heuristic algorithm that can handle large datasets.

Formalization Let $\mathcal{O}$ denote a set of n objects and let $\mathcal{R}$ be the set of equivalence relations on $\mathcal{O}$ (this set is in one-to-one correspondence with the set of partitions). We assume the existence of a true equivalence relation R_0. Dissimilarities between objects are considered as items of evidence about R_0, which can be represented by mass function m_{ij} with three focal sets: the set $\mathcal{R}_{ij}$ of equivalence relations containing objects i and j, its complement $\neg\mathcal{R}_{ij}$, and $\mathcal{R}$, and corresponding masses $m_{ij}(\mathcal{R}_{ij}) = \alpha_{ij}$, $m_{ij}(\neg\mathcal{R}_{ij}) = \beta_{ij}$ and $m_{ij}(\mathcal{R}) = 1 - \alpha_{ij} - \beta_{ij}$. After combining these $n(n-1)/2$ mass functions by Dempster's rule, we get a mass function m on $\mathcal{R}$ with contour function pl defined by the following equation,

$$\ln pl(R) = C + \sum_{i<j} R_{ij} \ln \frac{1-\beta_{ij}}{1-\alpha_{ij}}, \qquad (1)$$

where C is a constant. The most plausible partition can thus be found exactly, for small n (until, say, $n \leq 100$) using a binary linear programming solver.

Hopfield Model To make the above approach feasible for large n, we need a heuristic optimization method. We show that a local maximum of $\ln pl(R)$ defined by (1) can be found by a Hopfield neural network model [2] with n neurons, in which each neuron can be in one of c states, where c is the desired number of clusters. The weight v_{ij} of the connection between neurons i and j is the coefficient of R_{ij} in (1). Starting from random initial states, the state of each neuron i is updated at asynchronous times, by finding k such that $\sum_{j\neq i} v_{ij} s_{jk}$ is maximum, where $s_{jk} = 1$ if neuron j is in state k, and $s_{jk} = 0$ otherwise. This algorithm is shown to converge to a global network state that corresponds to a local maximum of (1).

Results and Conclusions The above clustering algorithm was applied to several datasets with different characteristics, including: large numbers of objects and/or clusters, non-metric dissimilarities, and complex cluster shapes, showing good performances as compared to existing algorithms. The definition of constants α_{ij} and β_{ij} is problem-specific and is an important step for ensuring good performances of the method. The application of this approach to semi-supervised clustering is currently under study.

Keywords. Clustering, Dempster-Shafer theory, Evidence theory, belief functions, Hopfield network.

Acknowledgement This research was supported by the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02).

References

[1] T. Denœux, N. E. Zoghby, V. Cherfaoui, and A. Jouglet. Optimal object association in the Dempster-Shafer framework. *IEEE Transactions on Cybernetics*, 44(22):2521–2531, 2014.

[2] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.

Computational Methods for Imprecise Continuous-Time Birth-Death Processes: a Preliminary Study of Flipping Times

Stavros Lopatatzidis and **Jasper De Bock** and **Gert de Cooman**
Ghent University, SYSTeMS Research Group
{Stavros.Lopatatzidis,Jasper.DeBock,Gert.deCooman}@UGent.be

We introduce the notion of flipping times for imprecise continuous-time birth-death processes, show how to obtain them, and explain how they lead to new computational methods.

The Precise Case Consider a continuous-time Markov processes where, at any time t, the stochastic matrix of the process P_t is derived from a transition rate matrix Q. When Q is bounded, P_t satisfies the Kolmogorov backward equation

$$\frac{d}{dt}P_t = QP_t. \tag{1}$$

If we let $f_t(x) \coloneqq E_t(f|X_0 = x)$, with f a real-valued function on the finite state space $\mathcal{X}$ and $x \in \mathcal{X}$ an initial state, then we can rewrite Equation (1) as follows:

$$\frac{d}{dt}f_t = Qf_t. \tag{2}$$

Combined with the boundary condition $f_0 = f$, the unique solution of Equation (2) is $f_t = e^{Qt}f$.

Instead of considering a time-invariant Q, we can also let Q_t be a function of the time t. In that case, Equation (2) can be rewritten as

$$\frac{d}{dt}f_t = Q_t f_t. \tag{3}$$

In general, Equation (3) has no analytical solution.

The Imprecise Case We focus on the case where every state in $\mathcal{X} \coloneqq \{0, \ldots, L\}$, has an interval-valued birth and/or death rate. The transition rate matrix is then a tridiagonal matrix of the form

$$\begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \mu_i & -(\mu_i+\lambda_i) & \lambda_i & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & \mu_L & -\mu_L \end{pmatrix}$$

where, for all $i \in \{0, \ldots, L-1\}$ and $j \in \{1, \ldots, L\}$, $\lambda_i \in [\underline{\lambda}, \overline{\lambda}]$ and $\mu_j \in [\underline{\mu}, \overline{\mu}]$. We use $\mathcal{Q}$ to denote the the set that consists of all these transition rate matrices.

At any time t, the only assumption we make about Q_t is that it is an element of $\mathcal{Q}$. Every such possible choice of non-stationary transition rate matrices will, by Equation (3), result in a—possibly different—solution f_t. Our goal is to calculate exact lower and upper bounds for the set of all these solutions f_t, as denoted by $\underline{f}_t$ and $\overline{f}_t$; we focus on the lower bound here. As proved by Škulj [1], $\underline{f}_t$ is the solution to

$$\frac{d}{dt}\underline{f}_t = \min_{Q \in \mathcal{Q}} Q\underline{f}_t, \tag{4}$$

with boundary condition $\underline{f}_0 = f$. If $\mathcal{Q}$ is the convex hull of a *finite* number of extreme transition rate matrices—as in our case—then since the solution to the above differential equation is continuous, we find that there must be time points $0 = t_0 < t_1 < \ldots < t_i < t_{i+1} < \ldots$ such that, for all $t \in [t_i, t_{i+1}]$, the minimum in Equation (4) is obtained by the same extreme transition rate matrix $Q_i \in \mathcal{Q}$. We call these time points t_i *flipping times*. The differential equation (4) is then piecewise linear, and the solution is therefore given by

$$\underline{f}_t = e^{Q_i(t-t_i)}e^{Q_{i-1}(t_i-t_{i-1})}\ldots e^{Q_1(t_2-t_1)}e^{Q_0(t_1)}f,$$

for $t \in [t_i, t_{i+1}]$. The difficult part is now to find the flipping times t_i and the corresponding extreme transition rate matrices Q_i. We provide computational methods that are able to do so.

Keywords. Imprecise continuous-time Markov process, birth-death process, flipping time. birth-death process, flipping time.

References

[1] Damjan Škulj. Efficient computation of the bounds of continuous time imprecise markov chains. *Applied Mathematics and Computation*, 250:165–180, 2015.

Bayesian Nonparametric Tests Based on the Imprecise Dirichlet Process

Francesca Mangili and **Alessio Benavoli** and **Giorgio Corani** and **Marco Zaffalon**
IPG IDSIA, Switzerland
{francesca, alessio, giorgio, zaffalon}@idsia.ch

Empirical results of almost all scientific research are analyzed based on frequentist null hypothesis significance tests, even though the shortcomings of such methods are well known (consider, for instance, the recent decision of a psychology journal to ban null hypothesis significance tests from their articles [4]). This work on hypothesis testing based on the Imprecise Dirichlet Process (IDP) [2] aims to change this perspective by providing Bayesian versions of nonparametric frequentist tests.

The Imprecise Dirichlet Process The Dirichlet process (DP) is a natural prior for developing nonparametric tests in a Bayesian framework. It is completely defined by its prior strength s (a scalar) and its normalized base measure α (a probability measure). To overcome the problem of eliciting its infinite dimensional parameter α in case of lack of prior information, we have developed a prior near-ignorance DP model (IDP) that consists of the set of all DPs with fixed s and α free to vary in the set of all probability measures. Beside solving the prior elicitation problem, this model reduces the computational costs and provides posterior inference which are more robust with respect to the choice of the prior.

Nonparametric Hypothesis Tests Based on the IDP model we have developed imprecise Bayesian tests that share strong similarities with a number of frequentist statistics, and thus provide a Bayesian justification of many traditional nonparametric tests: the sign test [3], the Wilcoxon signed test [1], the Mann-Whitney-Wilcoxon rank-sum test [2] (including the case for censored data [5]), the Friedman test [1] and the Kendall tau test. In this Bayesian framework, tests are formulated as decision problems where the goal is to minimize the expected loss. Such a principled way of balancing significance and power of the test is lacking in the frequentist setting. Moreover, IDP based tests automatically inform the analyst when the decision minimizing the expected loss changes depending on the DP base measure. In these prior-dependent cases the test issues an indeterminate outcome. We have empirically verified that, often, traditional tests virtually behave as random guessers in these indeterminate instances.

Conclusions By making the elicitation of the DP prior easier, computations faster and posterior inferences more reliable, the IDP model allows performing simple and efficient nonparametric hypothesis tests in a Bayesian way. These tests have several advantages: they avoid the shortcomings of the frequentist ones, formulate the hypothesis test as a decision problem, are conservative with respect to the choice of the prior and automatically inform when the decision is difficult (and thus traditional tests are not reliable). Due to all these qualities, these test can challenge the widespread use of nonparametric frequentist test in all areas of scientific research.

Keywords. Dirichlet process, nonparametric hypothesis testing, prior near-ignorance.

References

[1] A. Benavoli et al. A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In *Proc. of ICML-14*, pages 1026–1034, 2014.

[2] A. Benavoli et al. Imprecise Dirichlet process with application to the hypothesis test on the probability that x < y. *J. Stat. Theory Pract.*, 2014.

[3] A. Benavoli et al. A Bayesian nonparametric procedure for comparing algorithms. Subm. to ICML-15, 2015.

[4] T. David and M. Michael. Editorial. *Basic and Applied Social Psychology*, 37(1):1–2, 2015.

[5] F. Mangili et al. Reliable survival analysis based on the Dirichlet process. *Biom. J.* in-press.

Hyperbolic Systems with Random Set Coefficients

Jelena Nedeljković
University of Innsbruck, Austria
jelena.nedeljkovic@student.uibk.ac.at

Michael Oberguggenberger
University of Innsbruck, Austria
michael.oberguggenberger@uibk.ac.at

This contribution addresses linear hyperbolic systems with random set coefficients. We consider the problem

$$(\partial_t + \Lambda(x,t)\partial_x)u = F(x,t)u + G(x,t), \quad (x,t) \in \mathbb{R}^2,$$
$$u(x,0) = a(x), \quad x \in \mathbb{R},$$

where $u = (u_1, \ldots, u_n), G = (G_1, \ldots, G_n)$, Λ and F are $(n \times n)$-matrix functions.

The coefficient matrix Λ is real-valued and diagonal, with entries $\lambda_j, j = 1, \ldots, n$, given by any of the following: (a) a random set; (b) a random field (a stochastic process in higher dimensions); (c) a random field whose parameters are random sets. Applications: The addressed problem is a prototype model for wave propagation in random media. Coefficients describing material properties may have non-differentiable paths and their statistical parameters might be imprecise.

Method of characteristics In the deterministic case, the problem is often solved using the method of characteristics [2]. After introducing random sets as coefficients, we are still able to use this method, obtaining a set-valued solution U.

A random set is a map X which assigns to every ω from a probability space (Ω, Σ, P) a subset $X(\omega)$ of a target space $\mathbb{E}$ such that the upper inverses $X^-(B) = \{\omega \in \Omega : X(\omega) \cap B \neq \emptyset\}$ are measurable for every Borel subset B of $\mathbb{E}$. An important tool is the *fundamental measurability theorem* that states (if $\mathbb{E}$ is a Polish space) the equivalence of the defining measurability property of $X^-(B)$ for Borel, open, and closed subsets B as well as the equivalence with the existence of a *Castaing representation*. A set-valued random variable such that $X^-(B)$ is measurable for every open set B is called *Effros-measurable*. One of the goals of this contribution is to prove that the solution given by

$$U(\omega) = \{u_{l_1,\ldots,l_n} : l_j \in \lambda_j(\omega), j = 1, \ldots, n\} \qquad (1)$$

is a random set in the space of continuous functions.

Thanks to the results of [2] and continuous dependence $l_j \mapsto u_{l_1,\ldots,l_n}$, a Castaing representation can be immediately obtained, which leads to the Effros measurability; the fundamental measurability theorem completes the argument.

In the case of random field coefficients whose paths are at least Lipschitz continuous, the continuous dependence of the deterministic solution on its coefficients is enough to prove that the stochastic solution is a random field as well.

Random fields with non-Lipschitz paths If we wish to include random field whose paths are not Lipschitz continuous, we are no longer able to use the method of characteristics in a simple way.

We manage to overcome this difficulty by changing the entire setting and entering *the algebras of Colombeau generalized functions*, combining approaches described in [1, 2]. Colombeau generalized functions are defined as equivalence classes of families of smooth functions, depending on a regularization parameter ε. Measurability is understood componentwise on representatives. The Colombeau algebra is a complete metric space, but not separable. Random fields and random sets valued in the Colombeau algebra constitute a new concept.

Keywords. Random sets, random fields, hyperbolic systems, Colombeau algebra of generalized functions.

References

[1] M. Oberguggenberger, D. Rajter. Stochastic differential equations driven by generalized positive noise. *Publ. Inst. Math. Beograd*, 77(91):7–19, 2005.

[2] M. Oberguggenberger. *Multiplication of Distributions and Applications to Partial Differential Equations*. Pitman Res. Notes Math. 259, Longman, Harlow, 1992.

Partial Partial Preference Order Orders

Erik Quaeghebeur
Centrum Wiskunde & Informatica
Algorithms & Complexity Group
Erik.Quaeghebeur@cwi.nl

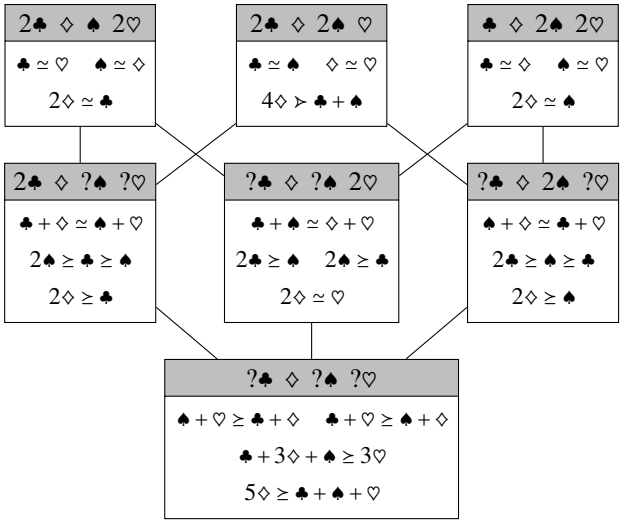
Consider the following collection of six-sided dice:

> *There are four faces, each present at least once: clubs ♣, spades ♠, diamonds ♢, and hearts ♡. A face only becomes visible after applying a drop of white wine to its side. There are at least three black faces. There are either more hearts than diamonds or an equal number of clubs and spades. A die is fair unless it has more black than white-faced sides, then each of the latter is equally more likely to land up than each of the former.*

Because of the *exclusive disjunctions*—either/or statements—in this description, the uncertainty we must model when gambling with dice from this collection cannot be handled using a single convex credal set, set of desirable gambles, preference order, or other such uncertainty model. Arguably, also non-convex credal sets are inadequate here.

I wish to discuss the following conceptual approach for dealing with this modeling issue:

- The *possibility space* is restricted to observables only (♣, ♢, ♠, and ♡) and so should not involve, e.g., the die variant. (There are three such variants; see the gray boxes in the top row of the diagram.)
- We consider the *partial order* X generated by the exclusive disjunctions. (See the gray boxes and their interconnections in what is in fact a Hasse diagram.)
- We attach an *uncertainty model* to each element of X, e.g., a partial preference order, that reflects the information common to its upset in X. (In the diagram we use $\geq$ for non-strict acceptance, $\succ$ for strict preference, and $\simeq$ for indifference [1]. Also, in the expressions, the faces denote the corresponding indicator gamble.)
- We can furthermore assign an *optimality criterion* to each element of X. Maximality and maximin variants thereof are natural candidates, E-admissibility perhaps less so, due to its use of individual probability measures, which can be replaced by exclusive disjunctions.
- With any set of decision options, we can then associate the corresponding partial order of optimal options. *Choice functions* [cf. 2] may be derived as functions thereof, for example the union of optimal options for the maximal elements of X.



Acknowledgments Research financed by the *Safe Statistics* project of the Netherlands Organisation for Scientific Research (NWO). Thanks go to Jasper De Bock, Gert de Cooman, and Arthur Van Camp for stimulating discussion.

Keywords. Exclusive disjunction, partial order, uncertainty model, credal set, set of desirable gambles, preference order, optimality criterion, choice function.

References

[1] Erik Quaeghebeur, Gert de Cooman & Filip Hermans. Accept & reject statement-based uncertainty models. *International Journal of Approximate Reasoning* 57 (Feb. 2015), 69–102. DOI: 10.1016/j.ijar.2014.12.003. arXiv: 1208.4462.

[2] Teddy Seidenfeld, Mark J. Schervish & Joseph B. Kadane. Coherent choice functions under uncertainty. *Synthese* 172 (2010), 157–176. DOI: 10.1007/s11229-009-9470-7.

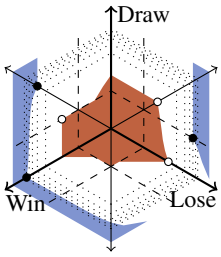
Eliciting Sets of Acceptable Gambles — The CWI World Cup Competition

Erik Quaeghebeur, **Chris Wesseling**, **Emma Beauxis-Aussalet**, **Teresa Piovesan**, and **Tom Sterkenburg**
Centrum Wiskunde & Informatica
Science Park 123, 1098 XG Amsterdam

We present an interface for eliciting sets of acceptable gambles on a three-outcome possibility space, discuss an experiment conducted for testing this interface, and present the results of this experiment.

Elicitation When uncertainties are elicited from experts, they are typically quantified as probabilities. Eliciting probabilities directly from domain experts as precise numbers is often problematic due to a lack of familiarity with probability theory and the absence of a concrete context. Moreover, even some of the most ardent 'precise' probabilists agree that imprecise-probabilistic techniques are better suited to deal with the results of an elicitation procedure (O'Hagan & Oakley 2004, Sec. 3.3). Sets of acceptable gambles form a representation for imprecise probabilities that is close to human behavior and eliciting them directly may improve the quality of the resulting uncertainty model.

Interface As a first step towards testing this hypothesis, we designed an interface for eliciting sets of acceptable gambles on three-outcome possibility spaces. We started from a two-dimensional representation of the three-dimensional space of gambles that was inspired by the flexibility afforded under the coherence axioms: We used the set of gambles with minimal value -1. This set of gambles was projected onto the plane and a logarithmic transformation was applied to obtain a representation with a sufficiently wide range of gamble values. To implement this representation, we needed to apply a discretization and had to develop a set of techniques for efficiently calculating the natural extension of an assessment in the context of a web browser, our chosen implementation environment.



Experiment We organized a betting competition for the 2014 FIFA World Cup. For each match, sets of acceptable gambles were elicited from participants; using the assessments so obtained, we computed a bet between them, i.e., a gamble was assigned to each participant. We were inspired by (Walley 1991, App. I), who ran an experiment for eliciting lower and upper probabilities concerning the outcome of matches of the 1982 FIFA World Cup.

Whereas (Walley 1991, App. I) used pairwise fair bets between the participants to score them, we designed a new algorithm for generating a single fair bet between all the participants in the betting pool. The algorithm's objective was to maximize lower expected payoff over all participants, while keeping the sum of the payoffs equal to zero.

Results Participant feedback indicated that reducing the complexity of the task and the interface would ease the elicitation procedure. The experiment's results underlined that imprecision is an essential aspect of real-life uncertainty modeling: most assessments made were imprecise. An interesting observation: the few participants who used complete, 'precise' models almost exclusively all had greater global losses than winnings.

Acknowledgments Erik Quaeghebeur was an ERCIM "Alain Bensoussan" Fellow, a program receiving funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016. Currently, his work and that of Tom Sterkenburg is part of the *Safe Statistics* project financed by the Netherlands Organisation for Scientific Research (NWO). Teresa Piovesan is partially funded by the European Project SIQS.

References

O'Hagan A. & Oakley J. E. (2004). Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety* 85.1–3, 239–248. DOI: 10.1016/j.ress.2004.03.014.

Walley P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman & Hall.

Radically Elementary IP Theory Based on Extensive Measurement

Teddy Seidenfeld, **Mark J. Schervish**, **Joseph B. Kadane**, **Rafael Stern**, and **Jessi Cisewski**
Carnegie Mellon University
{teddy,mark,kadane,rafaelst}@stat.cmu.edu, jcisewsk@andrew.cmu.edu

Extensive Structures *A Closed Archimedean Extensive Structure* [**CAES**] provides a (positive) real-valued, scalar representation of a binary relation that is additive in a concatenation operation. We summarize that theory as follows.

Let $\mathscr{D} = \{d_1, d_2, \dots\}$ be a domain of objects. Let $\oplus$ be a function from $\mathscr{D} \times \mathscr{D} \to \mathscr{D}$, understood as a concatenation operation on pair of objects. Finally, let $\succeq$ be a binary relation on $\mathscr{D} \times \mathscr{D}$. Five axioms for a **CAES** are these:

Axiom$_1$ $\succeq$ is a transitive, complete weak order, with symmetric $\approx$ and asymmetric $\succ$ parts.
Axiom$_2$ (Cancellation) $d_1 \succeq d_2$ iff $d_1 \oplus d_3 \succeq d_2 \oplus d_3$.
Axiom$_3$ (Associativity and Commutativity) $d_1 \oplus (d_2 \oplus d_3) \approx (d_2 \oplus d_1) \oplus d_3$.
Axiom$_4$ (Positivity) $d_1 \oplus d_2 \succeq d_1$.

Let $\boldsymbol{n}d = d \oplus d \oplus \cdots \oplus d$ with $\boldsymbol{n}-1$ concatenations.

Axiom$_5$ (Archimedes) If $d_2 \succ d_1$, and given d_3 and d_4, there exists $\boldsymbol{n}$ such that $[\boldsymbol{n}d_2] \oplus d_3 \succeq [\boldsymbol{n}d_1] \oplus d_4$.

Theorem$_1$ [1] Given a **CAES**, there exists a positive, real-valued function $\boldsymbol{g} : \mathscr{D} \to \Re^+$ where

- $\boldsymbol{g}(d_1) \succeq \boldsymbol{g}(d_2)$ iff $d_1 \succeq d_2$,
- $\boldsymbol{g}(d_1 \oplus d_2) = \boldsymbol{g}(d_1) + \boldsymbol{g}(d_2)$.

and $\boldsymbol{g}$ is unique up to scalars, $\boldsymbol{g}' = \alpha\boldsymbol{g}$ $(\alpha > 0)$.

We call a system that satisfies all but **Axiom$_5$** a *Radically Elementary Closed Extensive Structure* [**RECES**].

Theorem$_2$ [2] Given a **RECES**, there exists a positive, non-standard ${}^*\Re^+$ valued function ${}^*\boldsymbol{g} : \mathscr{D} \to {}^*\Re^+$ where

- ${}^*\boldsymbol{g}(d_1) \geq {}^*\boldsymbol{g}(d_2)$ iff $d_1 \succeq d_2$,
- ${}^*\boldsymbol{g}(d_1 \oplus d_2) = {}^*\boldsymbol{g}(d_1) + {}^*\boldsymbol{g}(d_2)$.

Regular Probability on a Finite Set as a CAES Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a finite partition and let $\mathscr{I}$ be a domain of *favorable investments* $\mathscr{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3, \dots\}$ where each investment scheme pays a determinate, non-negative dollar return $\mathscr{I}_i(\omega_j) = x_{ij} \geq 0$, as a function of ω. Define concatenation as $\boldsymbol{I}_1 \oplus \boldsymbol{I}_2 = \boldsymbol{I}_3$ where $x_{3j} = x_{1j} + x_{2j}$, $j = 1, \dots, n$. Let $\succeq$ be a binary preference relation between such favorable investment opportunities.

Application$_1$ With a simple modification of **Axiom$_4$** to include the constant $\boldsymbol{I}_0 = \boldsymbol{0}$, by *Theorem*$_1$, if this system is a **CAES** over $\mathscr{I}$, there exists a unique regular probability $\boldsymbol{P}$ on Ω, $\boldsymbol{P}(\omega_j) > 0$, where preference is represented by expected value: $\boldsymbol{g}(\boldsymbol{I}_i) = \sum_j \boldsymbol{P}(\omega_j) x_{ij}$. Let $\emptyset \neq E \subseteq \Omega$. Then, in the usual fashion, $\succeq_E$, called-off preference given E, suffices to define the conditional probability, $\boldsymbol{P}(\cdot|E)$.

Non-Standard Probability on Ω as a RECES

Application$_2$ Drop the Archimedean **Axiom$_5$** from *Application*$_1$ and, by *Theorem*$_2$, preference is a **RECES** that is represented through ${}^*\boldsymbol{g}$ by a non-standard probability ${}^*\boldsymbol{P}$ with non-standard expected value, and non-standard, conditional expected value.
Application$_3$ Modify **Axiom$_1$** in *Application*$_1$ so that strict preference is a strict partial order, $\succ$, as in [3, §4 in particular]. A corollary to *Theorem*$_1$ is **IP** theory, where a convex set of probabilities represents $\succ$ and $\succ_E$.
Application$_4$ Continue *Application*$_3$ by dropping **Axiom$_5$**. A corollary to *Theorem*$_2$ is non-standard ***IP** theory, where a convex set of non-standard probabilities and non-standard conditional probabilities represent strict preference and strict called-off preference.
Application$_5$ Continue *Application*$_4$. Replace modified **Axiom$_1$** with **Axioms 1a** and **1b** from [4, p. 164] in the theory of coherent choice functions.

Conjecture This modified **RECES** structure characterizes *all* ***IP** sets of non-standard probabilities on the finite set Ω.

References

[1] Krantz, Luce, Suppes, and Tversky. *Foundations of Measurement. Academic Press*, 1971.

[2] Narens. *Measurement without Archimedean Axioms. Phil. Sci. 41: 374-393*, 1974.

[3] Seidenfeld, Schervish, and Kadane. *Decisions without Ordering*. In Sieg (ed.) Acting and Reflecting. *Kluwer Academic: 143-170*, 1990.

[4] Seidenfeld, Schervish, and Kadane. Coherent Choice Functions under Uncertainty. *Synthese 172: 157-176*, 2010.

System Reliability Estimation under Prior-Data Conflict

Gero Walter
Eindhoven University of Technology
g.m.walter@tue.nl

Frank P.A. Coolen
Durham University
frank.coolen@durham.ac.uk

Simme Douwe Flapper
Eindhoven University of Technology
S.D.P.Flapper@tue.nl

In reliability engineering, data about failure events is often scarce. To arrive at meaningful estimates for the reliability of a system, it is therefore often necessary to also include expert information in the analysis, which can be dealt with straightforwardly via a Bayesian approach using an informative prior distribution.

A problem that then can arise is called prior-data conflict, see, e.g., [3]: from the viewpoint of the expert, the observed data seem surprising, i.e., the information derived from observed data is in conflict with prior assumptions. Models based on conjugate priors can be insensitive to prior-data conflict, in the sense that the spread of the posterior distribution does not increase in case of such a conflict [see 4, §A.1.2 for two examples], thus conveying a false sense of certainty by communicating that we know the reliability of a system quite precisely when in fact we do not.

As was shown in [5], models using sets of conjugate priors (generated through sets of canonical parameters) can mitigate this issue, by leading to larger sets of posteriors, and thus to more cautious inferences, in case of a prior-data conflict. [See 4, §§3.1, 3.2 for the general framework and its comparison with other models based on sets of priors.]

Building on previous work about reliability estimation for a simplified parallel system using sets of priors [6], we generalize the approach presented in [1] by considering sets of conjugate priors for expressing prior knowledge on component lifetimes. Through use of the recently developed survival signature [2], we obtain lower and upper bounds for the system reliability function. These posterior predictive bounds adequately represent our knowledge on the system reliability, giving more precise probability statements as data accumulate, and appropriately reflecting prior-data conflict by wider bounds.

As an example, we consider the problem of forecasting the reliability of a currently running new one of a kind system, where we have vague prior information on the lifetimes of the components the system is made of, where the only available data consists of observed behaviour of the system components so far, that is, the failure times of the components that have already failed, and the fact that the remaining components still function, whose failure time is thus right-censored. We present a method for taking into account surprisingly early or late component failures in the system reliability prediction, and analyse its effect on decisions about replacements of failed components.

Keywords. System reliability, survival signature, imprecise probability, generalized Bayesian Inference.

References

[1] L. J. M. Aslett, F. P. A. Coolen, and S. P. Wilson. Bayesian inference for reliability of systems and networks using the survival signature. *Risk Analysis*, 2015. doi:10.1111/risa.12228.

[2] F. P. A. Coolen and T. Coolen-Maturi. Generalizing the signature to systems with multiple types of components. In Wojciech Zamojski et al., editors, *Complex Systems and Dependability*, pages 115–130. Springer, 2012. doi:10.1007/978-3-642-30662-4_8.

[3] M. Evans and H. Moshonov. Checking for prior-data conflict. *Bayesian Analysis*, 1:893–914, 2006. URL: http://projecteuclid.org/euclid.ba/1340370946.

[4] G. Walter. *Generalized Bayesian Inference under Prior-Data Conflict*. PhD thesis, Department of Statistics, LMU Munich, 2013. URL: http://edoc.ub.uni-muenchen.de/17059/.

[5] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009. doi:10.1080/15598608.2009.10411924.

[6] G. Walter, A. Graham, and F. P. A. Coolen. Robust Bayesian estimation of system reliability for scarce and surprising data. Paper submitted to European Safety and Reliability Conference (ESREL), 2015.

Updated Network Analysis of the Imprecise Probability Community Based on ISIPTA Electronic Proceedings

Gero Walter
Eindhoven University of Technology
g.m.walter@tue.nl

Christoph Jansen
Ludwig-Maximilians-Universität München
christoph.jansen@stat.uni-muenchen.de

Thomas Augustin
Ludwig-Maximilians-Universität München
augustin@stat.uni-muenchen.de

In the last 15 years, the biennial ISIPTA symposia have established themselves as a central forum for the presentation and discussion of recent research in the field of interval or imprecise probability (IP). Revisiting our previous contribution for ISIPTA'11, where we derived and analyzed the research network in the IP community based on co-authorships of ISIPTA papers until and including ISIPTA'09 [2], we want to investigate more closely whether the population of ISIPTA contributors, or the structure of the contributor population, has changed. We thus update the research network by considering also the papers of subsequent ISIPTAs, updating our **R** package [3] accordingly.

Besides drawing the current network graph and updating the network characteristics usually studied in scientific collaboration networks [4, 5, 7] (like, e.g., the distribution of the number of collaborators, the number of papers per author, or the number of authors per paper), we want to focus on the network evolution [1]. We wish to identify trends and recent developments in network characteristics, especially with regards to the contributor population, and study the in- and outflow of authors in more detail by analyzing their position in the network. We also investigate whether trends or 'hot topics' are emerging from the symposia contributions, by analyzing the paper's keywords.

Furthermore, we consider models for scientific collaboration networks, like random graphs with preferential attachment [6, §8], to analyze the network dynamics of the ISIPTA coauthorship network.

Keywords. Network analysis, imprecise probability, scientific collaboration networks, network evolution.

References

[1] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614, 2002. doi:10.1016/S0378-4371(02)00736-7.

[2] Manuel J. A. Eugster, Gero Walter, and Thomas Augustin. A network analysis of the imprecise probability community based on ISIPTA electronic proceedings. Poster for ISIPTA'11, 2011. URL: http://www.sipta.org/isipta11/proceedings/posters/pa008poster.pdf.

[3] Manuel J. A. Eugster, Gero Walter, and Thomas Augustin. *A Network Analysis of the Imprecise Probability Community based on ISIPTA Electronic Proceedings*, 2012. Package for the statistical programming environment **R**. URL: https://github.com/mjaeugster/ISIPTA.

[4] Jerrold W. Grossman. Patterns of collaboration in mathematical research. *SIAM News*, 35(9):8–9, 2002. URL: http://www.appliedmathematician.org/pdf/news/485.pdf.

[5] Mark E. J. Newman. The structure of scientific collaboration networks. *Proceedings of The National Academy of Sciences*, 98(2):404–409, 2001. doi:10.1073/pnas.98.2.404.

[6] Remco van der Hofstad. *Random Graphs and Complex Networks*, volume I. 2014. URL: http://www.win.tue.nl/~rhofstad/NotesRGCN.html.

[7] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, 1998. doi:10.1038/30918.

Indices

Keyword Index

Author Index