

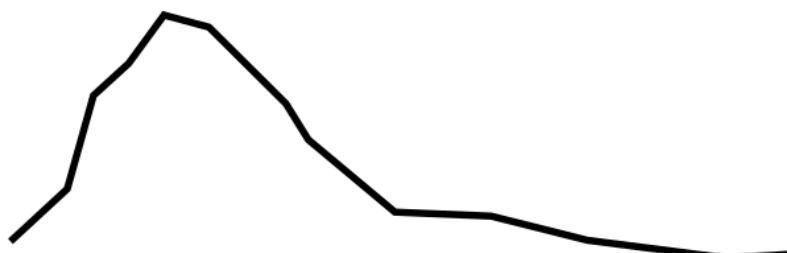
Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '17)

**Volume 62 of the Proceedings of Machine Learning
Research:**

ISIPTA '17, 10–14 July 2017

Lugano (Switzerland)

<http://isipta.idsia.ch>



ISIPTA '17



Conference Organization

Program Chairs

Alessandro Antonucci
Giorgio Corani
Inés Couso
Sébastien Destercke

Steering Committee

Alessandro Antonucci
Giorgio Corani
Inés Couso
Sébastien Destercke
Marco Zaffalon

Program Committee

Thomas Augustin
Alessio Benavoli
Mik Bickis
Seamus Bradley
Andrey Bronevich
Marco Cattaneo
Giulanella Coletti
Fabio G. Cozman
Fabio Cuzzolin
Milan Daniel
Jasper De Bock
Cassio de Campos
Gert de Cooman
Thierry Denœux
Serena Doria
Didier Dubois
Love Ekenberg

Scott Ferson
Thomas Fetz
Brian Hill
Radim Jiroušek
Jim Joyce
Alexander Karlsson
Vladik Kreinovich
Tomáš Kroupa
Francesca Mangili
Denis Mauá
Andrés Masegosa
Enrique Miranda
Ignacio Montes
Serafín Moral
Michael Oberguggenberger
Arthur Paul Pedersen
Renato Pelessoni

Erik Quaeghebeur
Fabrizio Ruggeri
Teddy Seidenfeld
Damjan Škulj
Michael Smithson
Joerg Stoye
David Sundgren
Matthias Troffaes
Lev Utkin
Barbara Vantaggi
Jiřina Vejnarová
Paolo Vicig
Gregory Wheeler
Andrea Wiencierz
Marco Zaffalon

Local Organization

Alessandro Antonucci
Giorgio Corani
Jasper De Bock
David Huber

Organizational Support

We gratefully acknowledge organizational support from Istituto Dalle Molle di Studi sull’Intelligenza Artificiale (IDSIA), Università della Svizzera Italiana (USI), and Scuola Universitaria Professionale della Svizzera Italiana (SUPSI).

CONFERENCE ORGANIZATION

Preface

The ISIPTA (*International Symposium on Imprecise Probability: Theories and Applications*) meetings are the primary forum for research on imprecise probability. 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, Compiègne and Pescara. The 2017 edition was held in Lugano (Switzerland) on July 10–14, 2017.

The proceedings of this edition are published for the first time within the *Proceedings of Machine Learning Research* (PMLR) series. This is by itself an acknowledgment of the scientific quality of the symposium.

Each submitted paper has been assigned to three program committee members. Eventually we accepted 32 papers. We would like to thank the 49 members of the program committee for their outstanding job during the reviewing process.

This edition of ISIPTA has been, for the first time, co-located with ECSQARU 2017, *the Fourteenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. ECSQARU and ISIPTA are both biennial events with a significant overlap in their scopes. The co-location should be regarded as a first attempt to promote a cross-fertilization of work from researchers of these two communities.

Besides the technical program, keynote lectures were given by five distinguished invited speakers, namely: Leila Amgoud (IRIT, France), Alessio Benavoli (IDSIA, Switzerland), Jim Berger (Duke University, USA), Didier Dubois (IRIT, France), and Eyke Hüllermeier (Paderborn University, Germany).

Lugano, July 2017

Alessandro Antonucci
Giorgio Corani
Inés Couso
Sébastien Destercke

PREFACE

The Editorial Team:

Alessandro Antonucci
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Manno-Lugano (Switzerland)
alessandro@idsia.ch

Giorgio Corani
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)
Manno-Lugano (Switzerland)
giorgio@idsia.ch

Inés Couso
University of Oviedo
Oviedo (Spain)
couso@uniovi.es

Sébastien Destercke
University of Technology of Compiègne
Compiègne (France)
sebastien.destercke@hds.utc.fr

Differences of Opinion

Dionissi Aliprantis

Federal Reserve Bank of Cleveland (US)

DIONISSI.ALIPRANTIS@CLEV.FR.B.ORG

Abstract

This paper considers the resolution of ambiguity according to the scientific ideal of direct observation when there is a practical necessity for social learning. An agent faces ambiguity when she directly observes low-quality data yielding set-identified signals. I suppose the agent's objective is to choose the single belief replicating what would occur with high-quality data yielding point-identified signals. I allow the agent to solve this missing data problem using signals observed through her network in combination with a model of social learning. In some cases the agent's belief formation reduces to DeGroot updating and beliefs in a network reach a consensus. In other cases the agent's updating can generate polarization and sustain clustered disagreement, even on a connected network where everyone observes the same data and processes that data with the same model.

Keywords: belief formation; subjective probability; social learning; partial identification; causal inference; DeGroot learning rule; bounded confidence.

1. Introduction

We all hold beliefs based on limited personal experience. This is often due to logistical, and not philosophical, limitations. The scientific ideal of “*seeing for one’s self*” is subject to time and resource constraints that make it infeasible to personally verify all claims. How do we form beliefs based on evidence beyond our personal experience?

This paper studies scenarios of partial identification in which personal experience offers no guidance for belief formation beyond a range of possibilities. Consider the example of a high school guidance counselor advising minority students on whether to attend a selective or non-selective college. What is the probability that an advisee will graduate from the selective college? The counselor would face partial identification if the high school had not tracked the experiences of recent graduates, or had sent few students to selective colleges.

When facing partial identification, the counselor could provide his students with a range of probabilities. Alternatively, the counselor could provide a single probability based on information beyond his directly-observed data. The choice of a single probability would use the counselor’s judgment to combine his own experience; his discussions with others like counselors or teachers; and the conflicting estimates in the literature ([Arcidiacono and Lovenheim, 2016](#); [Alon and Tienda, 2005](#)). This paper models the counselor’s choice of a single probability.

The general setting begins with an agent who must form beliefs about a set of propositions. The agent can use a model to translate data into signals about the truth of each proposition. Under frequentist inference she may form beliefs as the mean of her signals observed over discrete time.¹

There are many situations in which the available data might only allow the agent to partially identify a signal. An obvious scenario pertains to causal propositions when one cannot easily ob-

1. For independent and identically distributed (iid) signals, the Law of Large Numbers ensures such beliefs will converge to the mean of the signal distribution.

serve the Data Generating Process (DGP) under controlled intervention. This situation is widespread in economics, with many important counterfactual outcomes waiting to be definitively quantified.² Beliefs of an agent observing iid partially-identified signals will converge to a set when formed by averaging signals observed over discrete time ([Artstein and Vitale, 1975](#)).

When the agent has a set of possible beliefs, or faces ambiguity, prominent decision rules instruct her to choose the single belief generating an extreme utility ([Gilboa and Marinacci, 2013](#)). For example, the maxmin expected utility decision rule maximizes expected utility after choosing the belief that would be set by a malevolent nature minimizing the agent's utility for any decision ([Gilboa and Schmeidler, 1989](#)). The minimax regret decision rule maximizes expected utility after choosing the belief maximizing the agent's lost utility from not knowing the true state of the world ([Manski, 2011](#)).

This paper separates belief formation from preferences: When choosing one belief, the objective is to accurately represent the DGP. While the scientific ideal is to attain this objective based on direct observation, no single belief cannot satisfy this ideal when directly-observed data are only capable of partial identification. However, a single belief can approximate the scientific ideal if data yielding point-identification can be inferred from second-hand observations.

I specify the agent's problem as an attempt to replicate the beliefs she would have formed had she directly observed data yielding point-identified signals. The agent's problem can be viewed as a missing data problem to be solved with signals observed through her social network. I assume that communication is imperfect, so that socially-observed signals are communicated alone, without the data or model used in their construction.

I first show that if the agent uses linear opinion pooling of signals, a common method for combining forecasts and estimates, she will follow the canonical [DeGroot \(1974\)](#) learning rule under a special case of observed data. I then show that such DeGroot updating solves the agent's problem under additional assumptions on the homogeneity of data and models in the agent's network.

Two issues argue for pushing beyond the assumptions necessary for DeGroot updating to solve the agent's problem. The first is that the assumptions justifying DeGroot updating are strong. For example, individuals can be justified in using different models to interpret the same data ([Al-Najjar, 2009](#)), and the agent might observe new data over time ([Jadbabaie et al., 2012](#)).

Second, while DeGroot updating is the benchmark for non-Bayesian learning on social networks, a combination of theory and evidence motivates the desideratum of an alternative capable of generating polarization ([Golub and Sadler, 2016](#)). DeGroot learning and many of its generalizations converge to a degenerate distribution for connected networks ([Jackson, 2008; Dandekar et al., 2013](#)).³ However, an empirical analogue of a connected network - individuals exposed to sources of information contradicting their beliefs - is often observed together with persistent disagreement. Examples include scientific opinions when journals publish opposing research and public opinion when individuals are exposed to diverse news sources ([Gentzkow and Shapiro, 2011](#)).⁴ The emergence of "fake news" highlights this limitation of DeGroot updating.

-
2. In microeconomics alone it has proven difficult to ascertain outcomes under controlled interventions to neighborhood characteristics ([Ludwig et al., 2008; Aliprantis, 2017](#)), teacher characteristics ([Rothstein, 2010; Kinsler, 2012](#)), educational attainment ([Angrist and Krueger, 1991; Aliprantis, 2012](#)), minimum wages ([Card and Krueger, 1994; Neumark and Wascher, 2000](#)), unemployment benefits ([Hagedorn et al., 2013; Farber and Valletta, 2015](#)), income taxes ([Manski, 2014](#)), and right-to-carry laws ([Manski and Pepper, 2015; Durlauf et al., 2016](#)).
 3. Time to consensus, though, is not invariant across all connected network structures ([Golub and Jackson \(2012\)](#)).
 4. For example, there is persistent disagreement over propositions like [Iraq had an active WMD program](#), [President Obama was born in the US](#), [vaccines cause autism](#), and [global warming is occurring](#) despite public debate.

I show that linear opinion pooling can still solve the agent's problem after weakening the assumptions justifying DeGroot updating. In contrast to DeGroot updating, though, this solution requires a first stage in which signals are properly-transformed. I present the selection of a model that properly interprets signals as a statistical learning problem, and show that this problem is not well-posed. That is, frictions from communication generate a fundamental problem of inference, in that signals do not convey the same information as directly-observed data, and the agent cannot know whether she is properly interpreting signals without this information.

The agent might nevertheless choose a model for interpreting signals, just as methods for causal inference attempt to overcome the fundamental problem of evaluation. I study how the agent might use the model implied by a “reasonable” heuristic. The agent first interprets signals according to the model. The agent then uses the relative entropy of disagreement over all propositions to assess the credibility of applying the heuristic to each sender. The agent then combines interpreted signals, giving more weight to the interpreted signals from senders deemed most credible.

Although the updating rule tends to reach a consensus, I show that the rule is also capable of generating polarization and can sustain clustered disagreement, even on a connected network where everyone directly-observes the same data and processes that data with the same model. A key mechanism is generated by the use of relative entropy to assess the credibility of interpreted signals. If a given agent tends to agree with those in a widely-distributed cluster (unbiased but imprecise), but tends to disagree with those in a tightly-distributed cluster (biased but precise), that agent will rely more on interpreted signals from the disagreeing cluster, and this can cause her to overcompensate when they provide her with unbiased signals.

Polarization is possible because in contrast to updating in DeGroot or bounded confidence models, the agent can update her beliefs away from a signal if it comes from a sender with whom she tends to disagree. In other words, the agent's updating rule need not lead to constricting belief updating (Mueller-Frank, 2015). Two keys for generating polarization are low-quality data and perceptions about the distribution of models for interpreting directly-observed data.

The paper proceeds as follows: Section 2 sets the stage for the agent's problem, describing how she could arrive at a set of beliefs when directly observing data. Section 3 explores one way the agent might try to resolve the ambiguity she faces, using the signals she receives from individuals in her social network to form her beliefs. In the full paper I also show why finding a model of social learning to solve the agent's problem is an ill-posed problem, and describe the implications of a heuristic the agent might use to specify a model of social learning. I further investigate the implications of this heuristic in greater detail, studying belief dynamics under one specification of the updating rule for several parameterizations under various network and proposition structures. Section 4 concludes.

2. Belief Formation via Directly-Observed Data

Suppose there is a finite set of propositions $\{p^1, p^2, \dots, p^K\} = \mathcal{K}$, none of which can be written as a compound proposition using other propositions in the set.⁵ An agent must determine the truth value of the statements, $T(p^k) \in \{0, 1\}$, and agent i 's beliefs at time t are denoted by $\lambda_{it}^k = \Pr(T(p^k) = 1)$.⁶ The agent directly observes data W_{it} .

5. This greatly simplifies the analysis. See Paris and Vencovská (1990) and Wilmers (2010) for implications of propositional calculus when considering propositions formed as compound propositions.

6. A proposition is a statement that is either true ($T(p^k) = 1$) or false ($T(p^k) = 0$).

Consider a classical (frequentist) setting. With high-quality data W_{it}^* , the agent would be able to use her model φ_i^k to translate her data into an independent and identically distributed (iid) sequence of signals $\{\sigma_{it}^{k*}\}_{t=1}^T$, where

$$\sigma_{it}^{k*} = \varphi_i^k(W_{it}^*) \in [0, 1].$$

The law of large numbers ensures convergence to the mean of the signal distribution, which I will denote by μ_i^{k*} , for beliefs formed as

$$\begin{aligned} \lambda_{it+1}^{k*} &= \frac{1}{t} \sum_{n=1}^t \sigma_{in}^{k*} \\ &= \beta_t \sigma_{it}^{k*} + (1 - \beta_t) \lambda_{it}^{k*} \quad \text{where} \quad \beta_t = 1/t. \end{aligned} \quad (1)$$

Now consider a setting in which the agent's directly-observed data W_{it} only allows her to set identify the true iid signal σ_{it}^{k*} . Inspired by the literature on partial identification ([Manski, 2007](#); [Tamer, 2010](#)), suppose the agent's model and data allow her to determine a signal σ_{it}^k and its quality θ_{it}^k ,

$$(\sigma_{it}^k, \theta_{it}^k) = \varphi_i^k(W_{it}) \in [0, 1]^2,$$

where the true signal is related to the observed signal by

$$\sigma_{it}^{k*} \in [\max\{0, \sigma_{it}^k - (1 - \theta_{it}^k)\}, \min\{\sigma_{it}^k + (1 - \theta_{it}^k), 1\}] \equiv [\underline{\sigma}_{it}^{k*}, \bar{\sigma}_{it}^{k*}]. \quad (2)$$

The agent then knows from her signals of imperfect quality that the average

$$\lambda_{it+1}^{k*} = \frac{1}{t} \sum_{n=1}^t \sigma_{in}^{k*} \in \Lambda_{it+1}^{k*} = \left[\frac{1}{t} \sum_{n=1}^t \underline{\sigma}_{in}^{k*}, \frac{1}{t} \sum_{n=1}^t \bar{\sigma}_{in}^{k*} \right],$$

where the sets $[\underline{\sigma}_{it}^{k*}, \bar{\sigma}_{it}^{k*}]$ and Λ_{it+1}^{k*} are often referred to as “imprecise probabilities” ([Coolen et al., 2011](#)). The set $[\underline{\sigma}_{it}^{k*}, \bar{\sigma}_{it}^{k*}]$ is what can be learned about p^k from the directly-observed data under the most credible assumptions. While the agent can also determine σ_{it}^k , doing so requires less credible assumptions, so the agent cannot be sure that $\mathbb{E}[\sigma_{it}^k] = \mu_i^{k*}$ unless $\theta_{it}^k = 1$.

The canonical example of the proposition p^1 = “A given coin will land Heads.” helps to illustrate the difference between these settings. Suppose that high-quality data maps into signals generated by iid draws from a binomial distribution with probability 0.5 where $\sigma = 1$ if the coin lands Heads and $\sigma = 0$ if the coin lands Tails. In the case of high-quality data where $\theta_{it}^1 = 1$ for all t , $\sigma_{it}^k = \sigma_{it}^{k*}$, and so λ_{it+1}^{k*} can be calculated from (1) as the relative frequency of Heads, and will converge to 0.5 as $t \rightarrow \infty$.

In contrast, an agent with low-quality data mapping into signals represented by $\theta_{it}^1 = 0.2$ for all t will be subject to ambiguity in addition to risk.⁷ If the observed signal is Heads, then the agent can bound the true signal to be within $[0.2, 1]$. If the observed signal is Tails, then the agent bounds the true signal to be in $[0, 0.8]$. Thus as $t \rightarrow \infty$, the agent will infer that the mean of the true signals is $\mu_i^{k*} \in \Lambda_i^{k*} = [0.1, 0.9]$.⁸

7. In this context a point-valued belief $\lambda_{it}^k \in (0, 1)$ represents risk, while a set-valued belief $\lambda_{it}^k \in \Lambda_{it}^k \subseteq [0, 1]$ represents Knightian uncertainty or ambiguity.

8. Confidence intervals for the identified set Λ_i^{k*} are studied in [Imbens and Manski \(2004\)](#) and [Stoye \(2009\)](#), more generally as confidence regions in [Chernozhukov et al. \(2007\)](#) and [Romano and Shaikh \(2010\)](#), and using Bayesian methods in [Moon and Schorfheide \(2012\)](#) and [Bollinger and van Hasselt \(2008\)](#).

In addition to describing signals, throughout the analysis I will use “high-quality” (relative to the agent’s model) to describe data yielding point-identified signals ($\theta_{it}^k = 1$), and “low-quality” to describe data yielding set-identified signals ($\theta_{it}^k < 1$). For causal propositions, the difficulty of achieving identification is an obvious interpretation of signals having low quality. Examples abound of counterfactual outcomes that are difficult to quantify in microeconomics, macroeconomics, and finance because one cannot easily observe the Data Generating Process (DGP) under controlled intervention.⁹

Non-causal propositions can also have low-quality signals for reasons like survey non-response (Manski, 2015). Another interpretation of an extremely low-quality signal, $\theta_{it}^k = 0$, is that the agent does not directly observe any data for a given proposition p^k , so that $\varphi_i^k(\emptyset) = (\sigma_{it}^k, 0) \Rightarrow \sigma_{it}^{k*} \in [0, 1]$. It could also be the case that the agent’s model φ_i^k is not capable of extracting information from data. For example, an agent ignorant of genetics and molecular biology would likely have a model incapable of interpreting data on the human genome. In such cases, one could assign $\varphi_i^k(W_{it}^*) = \varphi_i^k(W_{it}) = (\sigma_{it}^k, 0) \Rightarrow \sigma_{it}^{k*} \in [0, 1]$ for any data set. For this analysis I will assume that the agent’s model produces a point-identified signal given a high-quality data set.

3. Belief Formation via Social Learning

A criticism of Bayesian decision theory is that in some circumstances, it might not be possible for the agent to express her beliefs using a distribution over the set Λ_{it}^{k*} . Bayesian decision theory is difficult to apply to these circumstances, since an imprecise probability cannot be used to make decisions according to the standard Savage axioms (Gilboa and Marinacci, 2013).

When holding beliefs represented by an imprecise probability Λ_{it}^{k*} , several approaches to decision making can be interpreted as picking one belief from the set Λ_{it}^{k*} , and then using this probability as a subjective belief with which to make decisions following the Savage axioms. The chosen probability is typically pessimistic, assuming the worst case in some sense of utility. For example, the Γ -maxmin utility decision rule maximizes expected utility after choosing the belief that would be set by a malevolent nature minimizing the agent’s utility for any decision (Gilboa and Schmeidler, 1989). Similarly, the Γ -minimax regret decision rule chooses the single belief that maximizes the loss from making decisions with the chosen belief rather than the true probability when the agent makes decisions to minimize this loss (Manski, 2011).

The subsequent model explores belief formation when the agent chooses one belief from Λ_{it}^{k*} using information from her social network.

3.1 The Agent’s Problem

Suppose the agent is a member of a network of $J + 1$ individuals from which she might gather information. The agent directly-observes the information

$$\mathcal{I}_{it} \equiv \left\{ (\lambda_{it}^1, \sigma_{it}^1, \theta_{it}^1), \dots, (\lambda_{it}^K, \sigma_{it}^K, \theta_{it}^K) \right\}.$$

To initialize the process we might let $\lambda_{i1}^k = \sigma_{i1}^k$; assume that the agent observes point identified signals from $t = -T$ until $t = 1$ and then set identified signals for $t > 1$; or else assume that the agent has just randomly reset $t = 1$ (as a random mutation in an evolutionary algorithm). The agent

9. See Footnote 2 for some examples from microeconomics.

also observes information in her social network about the truth of propositions. We denote the set of others in the agent's network as \mathcal{J} . However, the agent does not directly observe the data individuals in her network ($j \in \mathcal{J}$) directly observe. Instead, the agent observes individuals' beliefs and their interpreted data in the form of their signals. Thus, the socially-observed information available to the agent is

$$\mathcal{I}_{jt} \equiv \left\{ \{\lambda_{jt}^1, \sigma_{jt}^1\}_{j \in \mathcal{J}^1}, \dots, \{\lambda_{jt}^K, \sigma_{jt}^K\}_{j \in \mathcal{J}^K} \right\},$$

where the agent receives information about proposition p^k from individuals in $\mathcal{J}^k \subseteq \mathcal{J}$.

The agent might try Bayesian updating, or Bayesian social learning, according to Bayes' rule:

$$Pr(T(p^k) = 1 | \sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k}) = \frac{Pr(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k} | T(p^k) = 1) Pr(T(p^k) = 1)}{Pr(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k})}$$

Using beliefs λ_{it}^k as the agent's prior, this would imply updating as

$$\lambda_{it+1}^k = \frac{f(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k} | T(p^k) = 1) \lambda_{it}^k}{f(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k} | T(p^k) = 1) \lambda_{it}^k + f(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k} | T(p^k) = 0) (1 - \lambda_{it}^k)}.$$

[Acemoglu et al. \(2016\)](#) show in a related setting that strong restrictions would be required on the conditional pdfs $f(\cdot | T(p^k))$ for there to be asymptotic agreement across agents. More fundamentally, correctly specifying the likelihood function $f(\sigma_{it}^k, \{\sigma_{jt}^k\}_{j \in \mathcal{J}^k} | T(p^k))$ can require unrealistic assumptions about the information and computation available to the agent ([Acemoglu and Ozdaglar, 2011](#)).¹⁰ Weakening these assumptions is a key motivation of the literature on non-Bayesian social learning ([Molavi et al., 2015](#)).

Correctly specifying the likelihood function is the same as specifying

$$f(\varphi_{it}^k(W_{it}^*), \{\varphi_j^k(W_{jt})\}_{j \in \mathcal{J}^k} | T(p^k)),$$

which would require not only that the agent know the sampling processes for W_{it}^* and W_{jt}^* conditional on $T(p^k)$, but also the models $\{\varphi_j^k\}_{j \in \mathcal{J}^k}$. I rule out Bayesian social learning by restricting social information to beliefs and signals, assuming that the agent does not observe the additional information required to specify the likelihood function:

- (A1) Imperfect Communication: Agent i can only observe point estimates λ_{jt}^k and σ_{jt}^k . She cannot observe measures of the sender's ambiguity $\Lambda_{jt}^{k*}, \theta_{jt}^k$ or their model $\varphi_j^k \forall j, t, k$

The issue captured by A1 is that data must be transformed into information using a model, and it is difficult for individuals to communicate this process. Therefore, valuable details are lost relative to directly observing the data when information is obtained socially. Among other reasons, this assumption is positively appealing because there is a well-documented tendency for researchers and statistical agencies to focus on communicating their point estimates σ_{it}^k without communicating about their models φ_i^k or measures of uncertainty θ_{it}^k ([Manski, 2007, 2015](#)).

10. [Benoît and Dubra \(2015\)](#) and [Andreoni and Mylovanov \(2012\)](#) study polarization under private learning when agents disagree about $f(\sigma_{it}^{k*} | T(p^k))$. Alternatively, in this context their analyses could be interpreted as agents having different models for private learning φ_i^k , each proposition p^k being a conjunction of simple propositions $p^k = p^{k'} \wedge p^{k''}$, and W_{it}^* being revealed at different subperiods of t for $p^{k'}$ and $p^{k''}$.

With A1 ruling out Bayesian social learning, I assume that the agent uses signals in an effort to replicate classical inference. Given a loss function \mathcal{L} , the agent's problem is to choose functions f^k from some set \mathcal{F} to solve the problem

$$\begin{aligned} \min_{f^1, \dots, f^K \in \mathcal{F}} \quad & \sum_{k=1}^K \mathcal{L}\left(\mathbb{E}\left[\mu_i^{k*} - \lim_{t \rightarrow \infty} \lambda_{it+1}^k\right]\right) \\ \text{s.t. } & (\mathcal{I}_{it}, \mathcal{I}_{Jt}) \\ & \widehat{\sigma}_{it}^k = f^k(\mathcal{I}_{it}, \mathcal{I}_{Jt}) \quad \text{for } k = 1, \dots, K \\ & \lambda_{it+1}^k = \beta_t \widehat{\sigma}_{it}^k + (1 - \beta_t) \lambda_{it}^k \quad \text{for } k = 1, \dots, K \end{aligned} \quad (3)$$

I will refer to the agent's construction of her unobserved, high-quality signals $\widehat{\sigma}_{it}^k$ as her inferred signals. A natural restriction on \mathcal{F} is to make inferred signals a weighted average of directly- and socially-observed signals. In this case, f^k can be written as

$$\widehat{\sigma}_{it}^k = \underbrace{\theta_{it}^k}_{\substack{\text{share of signal} \\ \text{directly-observed}}} \sigma_{it}^k + \underbrace{(1 - \theta_{it}^k)}_{\substack{\text{share of signal} \\ \text{socially-observed}}} \sigma_{Jt}^k.$$

This restriction reframes the choice of f^k as the choice of σ_{Jt}^k .¹¹ Posing the inferred signals as weighted averages also gives an interpretation to θ_{it}^k as the agent's subjective judgment about the credibility of her modeling assumptions and/or a measure of the quality of her data.

3.2 Some Solutions to the Agent's Problem

When faced with problems like the agent's problem, a popular set \mathcal{F} is linear opinion pooling ([Ranjan and Gneiting, 2010](#)). It turns out that using repeated linear opinion pooling to solve the agent's problem results in DeGroot updating if data are only observed in the first period, and signals continue to be sent in later periods.

Proposition 1 (DeGroot) *If data are only observed once at $t = 1$, the agent sets $\lambda_{i1}^k = \sigma_{i1}^k$, $\theta_{i1}^k = \theta_{i1}^k$ for all $t > 1$, and subsequent signals are interchangeable with beliefs ($\sigma_{it}^k = \lambda_{it}^k$ and $\sigma_{jt}^k = \lambda_{jt}^k$ for $j \geq 2$), then linear opinion pooling where the agent constructs her inferred signals for $t \geq 2$ as*

$$\widehat{\sigma}_{it}^k = \theta_i^k \sigma_{it}^k + (1 - \theta_i^k) \sigma_{Jt}^k \quad \text{where} \quad (4)$$

$$\sigma_{Jt}^k = \sum_{j \in \mathcal{J}^k} \underbrace{w_j^k}_{\substack{\text{share of social signal} \\ \text{from individual } j}} \sigma_j^k \quad \text{with } w_j^k \geq 0 \quad \forall j \in \mathcal{J}^k, \quad \sum_{j \in \mathcal{J}^k} w_j^k = 1 \quad (5)$$

is equivalent to DeGroot updating where $\boldsymbol{\lambda}_{t+1}^k = \Omega_t^k \boldsymbol{\lambda}_t^k$ and the entries of Ω_t^k are

$$\begin{aligned} \omega_{itt}^k &= \beta_t \theta_i^k + (1 - \beta_t) \\ \omega_{ijt}^k &= \beta_t (1 - \theta_i^k) w_j^k. \end{aligned}$$

11. Assuming that $\{W_{it}\}_{t=1}^\infty$ and $\{\varphi_i^k\}_{k=1}^K$ are exogenous, both $\{\sigma_{it}^k\}_{t=1}^\infty$ and $\{\theta_{it}^k\}_{k=1, t=1}^{K, \infty}$ are given. Thus, in an abuse of notation, I will refer to f^k both as the function determining $\widehat{\sigma}_{it}^k$ and as the function determining σ_{Jt}^k .

Proof As hypothesized, set $\lambda_{i1}^k = \sigma_{i1}^k$. For $t \geq 2$, the equality of beliefs and signals, together with the updating equation in the agent's problem (3) imply that

$$\begin{aligned}\sigma_{it+1}^k &= \beta_t \hat{\sigma}_{it}^k + (1 - \beta_t) \sigma_{it}^k \\ &= \beta_t \theta_i^k \sigma_{it}^k + (1 - \beta_t) \sigma_{it}^k + \beta_t (1 - \theta_i^k) \sum_{j \in \mathcal{J}^k} w_j^k \sigma_{jt}^k.\end{aligned}$$

■

Furthermore, when the data observed in $t = 1$ generate unbiased point-estimates of signals, repeated linear opinion pooling/DeGroot updating solves the agent's problem.

Proposition 2 (Unbiased Signals) *Assume again, as we did in the case of private learning, that*

(A2) *Averaging Signals: $\beta_t = 1/t$, so that $\beta_t \hat{\sigma}_{it}^k + (1 - \beta_t) \lambda_{it}^k = \frac{1}{t} \sum_{n=1}^t \hat{\sigma}_{in}^k$*

If the observed data yield unbiased signals

(A3) *Private signals are iid with $\mathbb{E}[\sigma_{it}^{k*}] \equiv \mu_i^{k*} = \mu_i^k \equiv \mathbb{E}[\sigma_i^k]$, and*

(A4a) *Social signals are iid for each $j \in \mathcal{J}^k$ with $\mathbb{E}[\sigma_{it}^{k*}] \equiv \mu_j^{k*} = \mu_j^k \equiv \mathbb{E}[\sigma_{jt}^k] \quad \forall j \in \mathcal{J}^k$,*

then repeated linear opinion pooling/DeGroot updating following Equations 4 and 5 solves the agent's problem.

Proof Proposition 6 in Golub and Sadler (2016) states that as long as Ω^k is strongly connected and primitive, then

$$\lim_{t \rightarrow \infty} \sigma_{it+1}^k = \sum_{n=1}^{J+1} \pi_n^k \sigma_{n1}^k$$

where π_n^k is n 's left-hand eigenvector centrality in Ω^k . Since $\sum_{n=1}^{J+1} \pi_n^k = 1$ and $\mathbb{E}[\sigma_{n1}^k] = \mu_i^{k*}$ for all n , we know that

$$\mathbb{E}[\mu_i^{k*} - \lim_{t \rightarrow \infty} \lambda_{it+1}^k] = \mathbb{E}[\mu_i^{k*} - \sum_{n=1}^{J+1} \pi_n^k \sigma_{n1}^k] = \mu_i^{k*} - \mu_i^{k*} = 0.$$

■

We can imagine scenarios in which the agent observes data and signals in each period, but this additional information is potentially biased. In this case, the agent can still solve her problem if she has a model capable of accurately interpreting the social signals she receives.

Proposition 3 (Biased Social Signals) *Now suppose that the agent receives biased signals in the sense that $\mathbb{E}[\sigma_{jt}^k] \neq \mu_j^{k*}$, but that the agent has successfully engaged in statistical learning in the following sense:*

(A4b) *The agent has a model of social learning g^k that interprets social signals as $s_{jt}^k = g^k(\mathcal{I}_{it}, \mathcal{I}_{Jt})$. The s_{jt}^k are iid for each $j \in \mathcal{J}^k$ with $\mathbb{E}[\sigma_{it}^{k*}] \equiv \mu_i^{k*} = \mathbb{E}[s_{jt}^k] \quad \forall j \in \mathcal{J}^k$.*

Then linear opinion pooling where the agent constructs unobserved high-quality signals with her model as

$$\hat{\sigma}_{it}^k = \theta_{it}^k \sigma_{it}^k + (1 - \theta_{it}^k) \sigma_{Jt}^k \quad \text{where} \quad (6)$$

$$\sigma_{Jt}^k = \sum_{j \in \mathcal{J}^k} w_{jt}^k s_{jt}^k \quad \text{with } w_{jt}^k \geq 0 \quad \forall j \in \mathcal{J}^k, \quad \sum_{j \in \mathcal{J}^k} w_{jt}^k = 1 \quad (7)$$

$$s_{jt}^k = g^k(\mathcal{I}_{it}, \mathcal{I}_{Jt}) \quad (8)$$

solves the agent's problem.

Proof By A2 we know that $\lim_{t \rightarrow \infty} \lambda_{it+1}^k = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=1}^t \hat{\sigma}_{in}^k$. If the signals are iid, then since the sum of iid random variables is itself an iid random variable, by the law of large numbers we know that $\lim_{t \rightarrow \infty} \lambda_{it+1}^k = \mathbb{E}[\hat{\sigma}_{it}^k]$. After repeatedly applying the linearity of the expectations operator, A3 and A4a imply that

$$\begin{aligned} \lim_{t \rightarrow \infty} \lambda_{it+1}^k &= \mathbb{E}[\hat{\sigma}_{it}^k] = \mathbb{E}[\bar{\theta}_i^k \sigma_{it}^k + (1 - \bar{\theta}_i^k) \sigma_{Jt}^k] = \bar{\theta}_i^k \mathbb{E}[\sigma_{it}^k] + (1 - \bar{\theta}_i^k) \mathbb{E}[\sigma_{Jt}^k] \\ &= \bar{\theta}_i^k \mathbb{E}[\sigma_{it}^k] + (1 - \bar{\theta}_i^k) \mathbb{E}\left[\sum_{j \in \mathcal{J}^k} w_{jt}^k \sigma_{jt}^k\right] = \bar{\theta}_i^k \mathbb{E}[\sigma_{it}^k] + (1 - \bar{\theta}_i^k) \sum_{j \in \mathcal{J}^k} w_{jt}^k \mathbb{E}[\sigma_{jt}^k] \\ &= \bar{\theta}_i^k \mu_{it}^k + (1 - \bar{\theta}_i^k) \sum_{j \in \mathcal{J}^k} w_{jt}^k \mu_{jt}^k \\ &= \mu_i^{k*}. \end{aligned} \quad (9)$$

■

4. Conclusion

This paper presented a positive theory of belief formation. I proposed one way that an agent might choose a single subjective probability from a set of possible probabilities. When the agent faces ambiguity because her directly-observed data only allow her to partially identify a signal about the truth of a proposition, she might seek to learn from individuals in her social network. Assuming that communication is imperfect, so that individuals can only communicate a point estimate of their signals and beliefs, the agent must determine how to combine the signals she observes. I showed that when signals are unbiased, linear opinion pooling of signals generates DeGroot updating, and is able to replicate classical inference with high-quality data yielding point-identified signals.

Acknowledgment

Many of the ideas in this paper originated from discussions with Alon Bergman and Gregorio Cetano. I also thank Michalis Haliassos, Charles Manski, Jan-Peter Siedlarek, Alireza Tahbaz-Salehi, and seminar participants at the Cleveland Fed and Goethe University Frankfurt for helpful comments.

The opinions expressed are those of the author and do not represent views of the Federal Reserve Bank of Cleveland or the Board of Governors of the Federal Reserve System.

References

- D. Acemoglu and A. Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.
- D. Acemoglu, V. Chernozhukov, and M. Yildiz. Fragility of asymptotic agreement under Bayesian learning. *Theoretical Economics*, 11:187–225, 2016.
- N. I. Al-Najjar. Decision makers as statisticians: Diversity, ambiguity, and learning. *Econometrica*, 77(5):1371–1401, 2009.
- D. Aliprantis. Redshirting, compulsory schooling laws, and educational attainment. *Journal of Educational and Behavioral Statistics*, 37(2):316–338, 2012.
- D. Aliprantis. Assessing the evidence on neighborhood effects from Moving to Opportunity. *Empirical Economics*, 2017. doi:[10.1007/s00181-016-1186-1](https://doi.org/10.1007/s00181-016-1186-1).
- S. Alon and M. Tienda. Assessing the “mismatch” hypothesis: Differences in college graduation rates by institutional selectivity. *Sociology of Education*, 78(4):294–315, 2005.
- J. Andreoni and T. Mylovanov. Diverging opinions. *American Economic Journal: Microeconomics*, 4(1):209–232, 2012.
- J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.
- P. Arcidiacono and M. Lovenheim. Affirmative action and the quality-fit trade-off. *Journal of Economic Literature*, 54(1):3–51, 2016.
- Z. Artstein and R. A. Vitale. A strong law of large numbers for random compact sets. *The Annals of Probability*, 3(5):879–882, 1975.
- J.-P. Benoît and J. Dubra. A theory of rational attitude polarization. *Mimeo., London Business School*, 2015.
- C. R. Bollinger and M. van Hasselt. A Bayesian analysis of binary misclassification: Inference in partially identified models. *Mimeo., University of Kentucky*, 2008.
- D. Card and A. B. Krueger. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4):772–793, 1994.
- V. Chernozhukov, H. Hong, and E. Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007.
- F. P. Coolen, M. C. Troffaes, and T. Augustin. Imprecise probability. In M. Lovric, editor, *International Encyclopedia of Statistical Science*. Springer, 2011.
- P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013. doi:[10.1073/pnas.1217220110](https://doi.org/10.1073/pnas.1217220110).

- M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.
- S. N. Durlauf, S. Navarro, and D. A. Rivers. Model uncertainty and the effect of shall-issue right-to-carry laws on crime. *European Economic Review*, 81:32–67, 2016.
- H. S. Farber and R. G. Valletta. Do extended unemployment benefits lengthen unemployment spells?: Evidence from recent cycles in the U.S. labor market. *Journal of Human Resources*, 50(4):873–909, 2015.
- M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011. doi:[10.1093/qje/qjr044](https://doi.org/10.1093/qje/qjr044).
- I. Gilboa and M. Marinacci. Ambiguity and the Bayesian paradigm. In D. Acemoglu, M. Arellano, and E. Dekel, editors, *Advances in Economics and Econometrics: Tenth World Congress: Economic Theory*, volume I, chapter 7, pages 179–242. Cambridge University Press, 2013.
- I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.
- B. Golub and M. O. Jackson. How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338, 2012. doi:[10.1093/qje/qjs021](https://doi.org/10.1093/qje/qjs021).
- B. Golub and E. Sadler. Learning in social networks. In Y. Bramoullé, A. Galeotti, and B. Rogers, editors, *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.
- M. Hagedorn, F. Karahan, I. Manovskii, and K. Mitman. Unemployment benefits and unemployment in the great recession: The role of macro effects. Technical report, 2013.
- G. W. Imbens and C. F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.
- M. O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, 2008.
- A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210 – 225, 2012.
- J. Kinsler. Assessing Rothstein’s critique of teacher value-added models. *Quantitative Economics*, 3(2):333–362, 2012.
- J. Ludwig, J. B. Liebman, J. R. Kling, G. J. Duncan, L. F. Katz, R. C. Kessler, and L. Sanbonmatsu. What can we learn about neighborhood effects from the Moving to Opportunity experiment? *American Journal of Sociology*, 114(1):144–188, 2008.
- C. F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007.
- C. F. Manski. Choosing treatment policies under ambiguity. *Annual Review of Economics*, 3:25–49, 2011.
- C. F. Manski. Identification of income-leisure preferences and evaluation of income tax policy. *Quantitative Economics*, 5(1):145–174, 2014.

- C. F. Manski. Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. *Journal of Economic Literature*, 53(3):631–653, 2015.
- C. F. Manski and J. V. Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *NBER Working Paper 21701*, 2015.
- P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie. Foundations of non-Bayesian social learning. *Mimeo., Columbia University*, 2015.
- H. R. Moon and F. Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.
- M. Mueller-Frank. Reaching consensus in social networks. *IESE Working Paper 1116-E*, 2015.
- D. Neumark and W. Wascher. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. *The American Economic Review*, 90(5):1362–1396, 2000.
- J. B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4(3):183–223, 1990.
- R. Ranjan and T. Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society. Series B*, 72(1):71–91, 2010.
- J. P. Romano and A. M. Shaikh. Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211, 2010.
- J. Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.
- J. Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.
- E. Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010.
- G. Wilmers. The social entropy process: Axiomatising the aggregation of probabilistic beliefs. *Mimeo., Manchester Institute for Mathematical Sciences*, 2010.

Kurt Weichselberger's Contribution to Imprecise Probabilities

Thomas Augustin

AUGUSTIN@STAT.UNI-MUENCHEN.DE

*Department of Statistics, Ludwigs-Maximilians Universität München (LMU Munich)
Munich (Germany)*

Rudolf Seising

R.SEISING@LRZ.UNI-MUENCHEN.DE

*The Research Institute for the History of Technology and Science, Deutsches Museum Munich
Munich (Germany)*

Abstract

Kurt Weichselberger, one of the influential senior members of the imprecise probability community, passed away on February 7, 2016. Almost throughout his entire academic life the major focus of his research interest has been on the foundations of probability and statistics. The present article is a first attempt to trace back chronologically the development of Weichselberger's work on interval probability and his symmetric theory of logical probability based on it. We also try to work out the intellectual background of his different projects together with some close links between them.

Keywords: Weichselberger, Kurt; interval probability; imprecise probabilities; logical probability; symmetric theory; history of probability and statistics.

1. Introduction

Kurt Weichselberger, who passed away last year, has been “a man of the first hour” of the ISIPTA meetings, perceiving them as the natural place to discuss the foundations of probability. He enthusiastically participated in the first six ISIPTAs, from the 1999 Ghent symposium to the Durham meeting in 2009, contributing several papers, a tutorial in 2005 and a special session in 2009. At least from the mid sixties of the last century onwards, the foundations of statistics and probability have always been Weichselberger's great passion – although he had worked on a variety of different topics¹, and had been intensively engaged in academic self-administration and societies.

This paper is a first attempt to trace back fundamental aspects of Weichselberger's ideas as well as their links constituting his challenging research program. Our work is embedded into the *HiStaLMU* project (History of Statistics at LMU Munich). Among other activities, its members interview former leading personalities of the Department of Statistics as oral history and build up an archive around Kurt Weichselberger's office estate.² The structure of presentation in this paper is chosen mainly chronologically. After a brief biographic sketch (Section 2), we look at Weichselberger's foundational work and distinguish four main periods: the first intensive research on logical probability (see Section 3), the work on probability intervals in the context of modelling uncertain expert knowledge (Section 4), the axiomatic foundation of an interpretation independent theory of interval probability (Section 5), and eventually the aim to synthesize the previous results towards the *symmetric theory of logical probability* (Section 6). Section 7 concludes.

1. The work on applied statistics includes among others research on survey and census methodology (e.g. Weichselberger (1962)), regional price indices (Weichselberger and Wulsten, 1978), quality control (Weichselberger, 1971), and time series (Weichselberger, 1994).

2. See also the workshop in March 2016 (https://statsoz-neu.userweb.mwn.de/research/ws_historystatistik_2016/index.html).

2. A Short Biographic Sketch

In this section we quickly summarize the main stages of Weichselberger's career.³ Kurt Weichselberger was born on April 13, 1929, in Vienna. He studied mathematics there, and earned his PhD (*Dr. phil.*) in 1953 for a thesis supervised by Johann Radon ([Weichselberger, 1953](#)). Weichselberger started his academic career at the Department of Statistics in Vienna at Wilhelm Winkler's chair, worked at a social research institute in Dortmund, as well as at Johann Pfanzagl's chair in Cologne, where he received his *Habilitation* in 1962 with a thesis on controlling census results ([Weichselberger, 1962](#)). From 1963 to 1969 Weichselberger held the chair in statistics at the Technische Universität Berlin. In 1967 he was elected rector of this university and substantially contributed to the then vivid public debate about the role of education and scientists in the modern society.

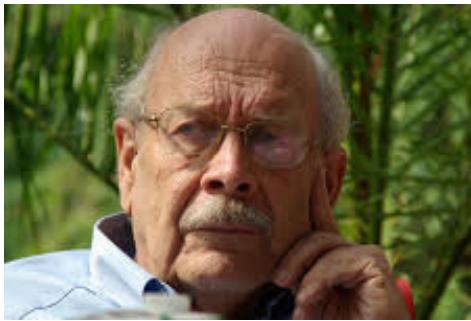


Figure 1: Kurt Weichselberger (photo kindly provided by Weichselberger's family)

From 1969 on, for almost 50 years, Weichselberger has been a member of the Ludwig-Maximilians-Universität München (LMU Munich). During this time he has played a leading role in the sustainable development of statistics as a discipline of its own – far beyond LMU. In particular, he co-founded the Department of Statistics and Philosophy of Science at LMU (see the end of the next section), and substantially contributed, also as Chairman of the Education Committee of the German Statistical Society for more than 10 years, to establishing first study programs for a major in statistics in Germany. From 1997 on, Weichselberger continued his research activities as a professor emeritus. On February 7, 2016, he passed away in his house in Grafing among close family.

3. Logical Probability I

Already at his inaugural speech ([Weichselberger, 1968](#)) as rector in Berlin, Weichselberger set out for his great scientific mission and passion: the development of a new theory of statistical inference, putting Fisher's fiducial argument back on its feet and substantially extending it. This theory has to be founded on what Weichselberger called *logical probability*: a non-subjective probability in its literal sense,⁴ evaluating, as a two-place function, the reasoning from premises to conclusions and, the other way round,⁵ finally allowing to describe the degree of support data give to statistical models. In the last section of his inaugural speech he explices:

3. For more details see in particular [Rüger \(1995\)](#) and Rüger's obituary ([Rüger, 2016](#)). Many students and academic companions until the mid nineties are assembled in the Festschrift edited by [Rinne, Rüger, and Strecker \(1995\)](#).
4. Note the etymological basis of the word probability: *prove-ability*, as well as the constituents of the corresponding German word *Wahr-schein-lich-keit*, i.e. the extent to which something seems to be true.
5. That is the reason, why Weichselberger called his theory *symmetric theory of probability*.

[... We are challenged with the task to reconceptualise the foundations of probability. The question is whether we can make progress towards a broader concept designation without losing key benefits of the previous – objectivistic – concept. For that matter we have to decide which properties of the objectivistic probability concept we consider to be essential.

[...] two essential properties of Kolmogorov's probability concept that consequently may not be waived [are the following]:

1. *The embedding into modern mathematics.*

This needs to remain ensured by determining the mathematical properties of the concept with a consistent system of axioms.

2. *The possibility of the frequency interpretation of probabilities because this presents to date the only known mindset that enables an explanation of the concept and thus guarantees that the ideas of different persons on the meaning of probability can be adapted. Taking these issues into account we are challenged with the task as follows:*

We have to develop a system of axioms that

1. *includes Kolmogorov's system of axioms as a special case;*
2. *associates probabilities not with events but with inferences from premises to conclusions;*
3. *enables the frequency interpretation of the probability concept;*
4. *enables probability propositions in both directions in cases in which the Fisher theory and the Neyman-Pearson theory yield the same results; for example in the case of a sample from a population, it associates a probability with the inference from the population to the sample as well as with the inference from the sample to the population.* ([Weichselberger, 1968](#), p. 46-47) [translation by TA & RS]

Weichselberger is already very clear about the fact that such a theory has to go beyond the restrictions precise probabilities imply, and therefore continues:

As in many cases in the history of science it is shown also here that — as a form of compensation for desired benefits — we have to abandon a “habit of thinking” (Denkgewohnheit). In the present case this is the habit of thinking that the probability is always a number. We must instead allow sets of numbers — say the interval between 0.2 and 0.3 — to act as the probability of the inference from the proposition B to the proposition A. However, we continue to demand that the set of numbers lies in the interval between 0 and 1.

This extension of the probability concept from a number to a set of numbers is encouraged as soon as we try to formalize Fisher's fiducial probability. Therefore a similar approach has already taken the American Henry Kyburg Jr. in his works in the years 1961 to 1964. However, Kyburg's concept is inconsistent at a decisive point, and, to his own statements, it does not lead to useful results in detail. His view is mainly of philosophical and not of mathematical nature.

In fact, the definition of probability as a set of numbers — normally an individual number or an interval — leads to mathematical problems. We need a system of

calculation rules for algebraic operations with such sets. This prompts us to similar considerations as the systematization of calculations with inexact or error-prone quantities: one could call it a “theory of tolerance space” (Spielraum-Theorie) because we associate tolerance spaces instead of individual numbers. I think that it is possible that this view may give rise to interesting inner mathematical questions. ([Weichselberger, 1968](#), p. 47) [translation by TA & RS]

During this period Weichselberger had accepted an offer from LMU Munich on a newly installed chair for *Special Topics in Statistics (Spezialgebiete der Statistik)*. In Munich he was strongly engaged in changing the institutional alignment of statistics within the university. In 1974, the Institute for Statistics and Philosophy of Science was founded, as a member of the new Faculty of Philosophy, Philosophy of Science and Statistics. Weichselberger has stayed in intensive intellectual contact with his colleagues from philosophy all the time. Clearly, there have been common scientific interests in particular with Wolfgang Stegmüller, who held the Chair in Philosophy of Science and did research among other topics also on subjective probability and Carnap’s concept of logical probability.⁶ In the first Munich years Weichselberger worked intensively on a book on logical probability. According to his former assistant Christina Schneider (personal communication), a manuscript of several hundred pages evolved, but, unfortunately, never got published.⁷

4. Probability Intervals, *Uncertainty in Knowledge-Based Systems*

In the mid eighties Weichselberger’s research experienced a shift, which gave his interests in imprecise probabilities new impetus, where he had been in-depth engaged in the vivid discussion about modelling uncertain expert knowledge in artificial intelligence. He has understood it as a question of life and death for statistics as a discipline whether statistics can contribute here. Weichselberger agreed with many other researchers mainly from computer science that the problem how to model uncertain knowledge produces a big challenge, where statistics, in its traditional form, reaches its limits. However, he also warned not to throw out the baby with the bath water and end up in a wild arbitrariness of conclusions, by leaving the field to ad hoc calculi. Weichselberger stood for a very clear position: there will be an important contribution of statistics and probability in this area, if, but also radically only if, the concept of probability is ready to overcome the dogma of precision.

Therefore, the book *A Methodology for Uncertainty in Knowledge-Based Systems* ([Weichselberger and Pöhlmann, 1990](#)), published by Weichselberger together with his post-doctoral researcher Sigrid Pöhlmann, aims at reconciling probability theory with the objectives of flexible modern uncertainty calculi. In their preface they argue:

First of all it must be stated that although the basic ideas prevailing in some considerations about diagnostic systems sound convincing, they violate fundamental requirements for reasonable handling of uncertainty. [...] We] shall demonstrate that negligence with respect to [...] some basic principles] may result in the inclusion of information into a diagnostic system which is equivalent to ruining it. ([Weichselberger and Pöhlmann, 1990](#), pp. 1-2)

-
6. To which extent a concrete co-operation in research has taken place between Weichselberger and Stegmüller is still an open question, which shall be studied further within the HiStaLMU project.
 7. Tragically, that manuscript is not part of Weichselberger’s office estate. Up to now it is unclear whether this manuscript still exists.

Indeed, after fundamentally criticising the Dempster-Shafer combination rule and the MYCIN certainty factors, Weichselberger and Pöhlmann develop, in the context of a prototypical special case⁸, a neat probabilistic alternative to handle different sources of information in diagnostic systems. It has consistently tried to be based on a generalisation of probability, synthesising the well-founded concept of probability with the flexibility needed for modelling uncertain knowledge:

[...] An argument against a possible application of probability theory [, understood in its traditional, precise form here,] in diagnostic systems is as follows: While probability theory affords statements, using real numbers as measures of uncertainty, the informative background of diagnostic systems is often not strong enough to justify statements of this type. [...] However, it is possible to expand the framework of probability theory in order to meet these requirements without violating its fundamental assumptions. [...] We believe that the weakness of estimates for measures of uncertainty as used in diagnostic systems represents a stimulus to enrich probability theory and the methodological apparatus derived from it, rather than an excuse for avoiding its theoretical claims. (Weichselberger and Pöhlmann, 1990, p. 2)

Technically, Weichselberger & Pöhlmann do not yet use interval probability in its full generality, but confine themselves to the case which they call *probability intervals (PRI)*.⁹ There an interval-valued probability is assigned to the singletons only, and natural extension is applied to calculate the probabilities of the other events. Moreover, speaking often of “interval *estimates* of probabilities” (italics by TA & RS), Weichselberger & Pöhlmann implicitly rely exclusively on the sensitivity analysis (epistemic) point of view of imprecise probabilities. The book was published one year before Peter Walley’s book (Walley, 1991) on general imprecise probability appeared. In Weichselberger and Pöhlmann’s book the notions of R- and F-probability (“R” for *reasonable*, corresponding to *avoiding sure loss* to use Walley’s terminology, and “F” for *feasible*, corresponding to *coherent*) were developed for the first time. Having been perceived well, mainly in the artificial intelligence community, the book was also criticized strongly as “a little too unfinished” and too example-based in a review in the Journal of the American Statistical Association (Wasserman, 1991). Convenient expressions to work with PRIs were extended in Weichselberger (2001a, Chapter 3.3 and Appendix A.5). The construction of least favourable pairs for testing hypotheses described by PRIs is considered in Martin Gümbel’s dissertation (Gümbel, 2009), supervised by Weichselberger.

5. Interval Probability: *Elementare Grundbegriffe* ...

5.1 The Book and the ISIPTA '99 Paper Including its IJAR Extension

Immediately after having finished the book with Pöhlmann, Weichselberger started to develop the theory of interval probability as a “one-place assignment”, i.e. assigning probability to events, in its generality.¹⁰ No later than 1992, a first version of a book was ready, which already contained the core of the theory. The material grew and grew in its dimensions, and Weichselberger decided to split the book project into three volumes. Finally, in 2001 the first volume, *Elementare*

8. The general case was later solved in Pöhlmann’s Habilitation thesis (Pöhlmann, 1995).

9. See de Campos, Huete, and Moral (1994) for an independent development of almost the same framework.

10. Weichselberger, however, always has been stressing the importance of the “two-place concept” (logical probability with premises and conclusions as functional arguments, see Section 3) as the ideal, calling it still “[...] without doubt the most challenging [...]” concept (Weichselberger, 2001a, p. 33) [translation by TA & RS]. Unless mentioned differently, the term ‘probability’ is used throughout this section in its one-place meaning as probability of events.

Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept (Elementary Fundamentals of a More General Calculus of Probability I: Interval Probability as a Comprehensive Concept) ([Weichselberger, 2001a](#)), appeared.¹¹ Soon this book became Weichselberger's most influential publication, together with the paper *The theory of interval probability as a unifying concept for uncertainty* ([Weichselberger, 2000](#)), which arose from his ISIPTA '99 contribution and serves as an English language reference summarizing some of the main concepts. The title of the book, an immediate allusion to Kolmogorov's *Grundbegriffe ...* ([Kolmogorov, 1933](#)), founding traditional probability theory, formulated the research program. Weichselberger develops thoroughly the theory of interval probability by generalizing the Kolmogorovian concept to interval-valued assignments. The book consists of four main chapters:¹²

The first chapter elaborates the background of the theory. It starts with embedding the theory into the historical development of the concept of probability, including other generalizations of probability. Then motives for the paradigmatic shift from traditional probability to interval probability and major objectives of the theory are discussed in-depth.

The second chapter contains the axioms of R- and F-probability. Weichselberger characterises interval-valued assignments $P(\cdot) = [L(\cdot), U(\cdot)]$ on a σ -field \mathcal{A} by their relation to the set \mathcal{M} of classical probabilities (in the sense of Kolomogorov) $p(\cdot)$ they induce. If this set is not empty, then $P(\cdot)$ is an *R*-probability, and \mathcal{M} is its *structure*. An R-probability is *F*-probability if $P(\cdot)$ and the structure uniquely correspond to each other by

$$L(A) = \inf_{p(\cdot) \in \mathcal{M}} p(A) \quad \text{and} \quad U(A) = \sup_{p(\cdot) \in \mathcal{M}} p(A), \quad \forall A \in \mathcal{A}.$$

In the light of Walley's lower envelope theorems, R-probability and F-probability technically correspond, in essence, to lower and upper probabilities avoiding sure loss and being coherent, respectively ([Walley, 1991](#), Chapters 2 and 3), where, however, Weichselberger, in the spirit of Kolmogorov, demands σ -additivity. In conformity with Walley, Weichselberger stresses that there is no need to require additional restrictive properties (like two- or total monotonicity of the lower bound), but in contrast to him, Weichselberger focuses on interval-valued assignments of *events*, instead of random variables/gambles. For Weichselberger, probability of events is the constitutive entity (of a one-place probability assignment¹³); he sees expectations/previsions as derived entities, explicitly needing an underlying metrical scale. The most important difference for Weichselberger to Walley is that his axiomatisation is, just as the Kolmogorovian approach in traditional probability theory, strictly independent of any interpretation of probability. By this, he emphasises, it provides a sound mathematical basis for expressing all different interpretations of (one-place) generalized probabilities, from subjective to frequentist, which eventually is the key to overcome the methodological antagonisms in statistical inference. Chapter 2.6 reflects on decision criteria based on probabilistic evaluations of events. There Weichselberger also argues that behaviour following Hurwicz-like

11. The book title has the rare addendum "unter Mitarbeit von (with the cooperation of) T. Augustin und (and) A. Wallner", which tributes to the special way the book was written: Augustin entered the project in 1993, Wallner in 1995, both as young PhD students and assistants. They were rather intensively engaged with the book, but rarely as co-authors (Wallner, and to an even smaller extent Augustin, contributed some shorter, clearly marked parts of the book only, listed in [Weichselberger \(2001a, p. x\)](#)), but as critical discussions partners. Weichselberger extended and developed further the theory step by step, and in several meetings per week these steps were immediately and intensively discussed.

12. See also the review by [Coolen \(2003\)](#).

13. See also Footnote 10.

criteria (e.g. [Huntley, Hable, and Troffaes \(2014\)](#), p. 193)) challenges the betting interpretation of imprecise probabilities, which he judges to rely solely on a Γ -maximin point of view.¹⁴

Chapter 3 generalizes the setting to situations where the limits of an interval probability firstly are only specified on certain subsets of the σ -field \mathcal{A} , and then natural extension is applied (*partially determinate probability*). This gives rise to a list of interesting special cases, including PRIs (see Section 4) and a kind of general p-boxes (*cumulative R/F-probability*¹⁵). Supplementing natural extension, which already appears in [Weichselberger and Pöhlmann \(1990\)](#) (*derived F-PRI*), Weichselberger also proposes a *cautious standpoint* to proceed from a given R-probability $[L(\cdot), U(\cdot)]$ that is not F-probability to a uniquely defined F-probability $[L^*(\cdot), U^*(\cdot)]$, now such that the original limits $L(\cdot)$ and $U(\cdot)$ are always respected, in the sense that $L^*(\cdot) \leq L(\cdot)$ and $U^*(\cdot) \geq U(\cdot)$.

Specific issues of interval probabilities on finite spaces are in the focus of Chapter 4, see also [Weichselberger \(1996\)](#). In Chapter 4.1 algorithms are developed to check whether assignments constitute R- and F-probability, as well as to calculate the natural extension and its counterpart from the cautious standpoint. Interestingly, linear programming is here not only utilized powerfully for calculations, but also, by duality results, as a mathematical tool for elegant proofs.

5.2 Preceding First Contributions to General Interval Probability; Strongly Related Work and Co-operations



Figure 2: Participants of the Foundations of Statistics Workshop organized by Frank Hampel in 1994: From left to right: Walley, Goldstein, Smets, Coolen, Weichselberger, Morgenstahler, Hampel, Augustin (photo kindly provided by Frank Coolen)

In this section we collect some of Weichselberger's activities when working on his book. His axioms and further core elements of his theory were presented at several workshops, including a workshop in June 1993 honouring Peter J. Huber ([Weichselberger, 1996](#)), the Second Gauss Symposium in August 1993 ([Weichselberger, 1995a](#)), and a workshop on the foundations of statistics in September 1994 in Zurich organized by Frank Hampel. By that workshop and an associated research retreat to the mountains nearby, Hampel connected researchers interested in the foundations of statistics (see also Figure 2), who only partially knew each other personally. The participants' ex-

14. See also [Coolen \(2003\)](#), p. 254).

15. Compare also [Destercke, Dubois, and Chojnacki \(2008\)](#) for a related concept.

cited discussions had had a sustainable impact on their further research. Particularly close remained over all the years the relationship of Weichselberger (and Augustin) with Frank Coolen.

In 1995 also a paper on the implications of the rich framework of interval probability on sampling appeared ([Weichselberger \(1995b\)](#), see also [Weichselberger \(2001a, Chapter 4.3\)](#)), which in our eyes by far did not receive the attention it actually deserves. Only with interval probability it becomes possible to express the distinction between different types of symmetry, called *epistemic* versus *physical symmetry* by Weichselberger. Epistemic symmetry relies merely on the lack of knowledge of asymmetry, while for physical symmetry knowledge is available actively supporting symmetry. Only the latter in its purest form justifies the use of precise, traditional probabilities. By these concepts, Weichselberger develops nothing less than a generalization of the principle of insufficient reason, replacing precise uniform probability by a continuum of uniform probabilities, adequately expressing the knowledge on the system under consideration.

Decision theoretic implications of imprecise probabilities are discussed in 1998 in a contribution ([Weichselberger and Augustin, 1998](#)) to a Festschrift honouring Weichselberger's Munich long-standing colleague Hans Schneeweiss, working out how interval probability provides an immediate description of the preferences observed in Ellsberg's seminal experiments ([Ellsberg, 1961](#)), violating the axioms of traditional subjective utility theory.

As a preparation for the third volume, which was originally devoted to statistical implications of interval probability, the Huber-Strassen theory on robust testing of hypotheses described by neighbourhood models had been intensively discussed by the members of Weichselberger's chair and Helmut Rieder, who spent in 1994 one semester at LMU. Augustin, who originally had started a dissertation about the historical roots of imprecise probability, took over the topic and developed under Weichselberger's supervision a Neyman-Pearson theory under general interval probability, where the hypotheses are described by F-probability instead of two-monotone capacities. In his thesis ([Augustin, 1998](#)) it is shown that Weichselberger's condition of continuity of F-fields ([Weichselberger, 2001a, p. 152f.](#)) is both necessary and sufficient for the structure to be uniformly dominated. Furthermore, Augustin derives results on different types of least favourable pairs (published in a generalized form for the first ISIPTA and in [Augustin \(2002a\)](#), based on it) and a representation of the optimal test by a single linear program (published later in a decision theoretic context in [Augustin \(2002b, 2004\)](#)), including a Neyman-Person lemma form obtained from duality arguments.

5.3 Further Planned Volumes, Work on Interval Probability After the Book

When the book appeared, a second volume was already in a rather advanced stage. Originally it was devoted to a closer study of types of assignments that lead to two- or totally-monotone capacities (probability intervals, cumulative probabilities, belief-functions), concepts of conditional probabilities and independence, parametric statistical models and a law of large numbers.¹⁶

In Weichselberger's ISIPTA '01 contribution ([Weichselberger, 2001b](#)), *indicator fields* are studied, i.e. interval probabilities that can be understood as basic building blocks for more complex models. In 2002, Lev Utkin visited the Weichselberger chair for almost two years, and a very close relationship with Weichselberger (and Augustin) was established that has endured over all the years. At ISIPTA '03 (cf. [Weichselberger and Augustin \(2003\)](#)), Weichselberger presented his research on conditional probability. He strongly argued in favour of the idea that there cannot be a single concept of conditional probability; several, conceptually different concepts are needed which happen to

16. Some concepts are already briefly sketched in [Weichselberger \(2000\)](#).

coincide in the case of a precise probability. In particular, he elaborates his – rather controversially perceived – *canonical concept of conditional interval probability*, derived from a canon of desirable properties, like a commutative combination of marginals and conditional probabilities.¹⁷

In autumn of that year, Weichselberger abruptly stopped his research on one-place probability and radically turned all his interest to the foundation of logical probability again.

Anton Wallner, who worked together very closely with Weichselberger at that time (see also Footnote 11), first continued his research on the one-place interval probability and prepared a dissertation under Weichselberger's supervision ([Wallner, 2002](#)). There he develops a series of characterisations of interval probabilities in general as well as of uniform interval probabilities, and studies neighbourhood models based on distorted probabilities. Furthermore he presents a rather involved proof that also under general interval probability the structure of an R-probability on a space with cardinality k has, interpreted as a polyhedron in \mathbb{R}^{k-1} , at most $k!$ vertices. Related articles, presented at ISIPTA '03 and '05, are [Wallner \(2003\)](#) and [Wallner \(2007\)](#).

6. Logical Probability II

All the development of one-place interval probability described in the previous two sections, as interesting it may be on its own, has been understood by Weichselberger mainly as a preparation for his concept of logical probability, and thus for his general inference theory. Therefore, from 2003 on Weichselberger had devoted all his energy to this topic. Supported by Wallner, Weichselberger started to (re)build a neat framework for logical probabilities, now finding a neat basis in the theory of interval probability, pushing the vision of a closed theory of inference closer to reality. In one of his last public presentations, a special session on the symmetric theory at ISIPTA '09 ([Weichselberger, 2009](#), p. 9), he characterises his major objective in simplified terms as follows:

A comprehensive methodology of probabilistic modelling and statistical reasoning, which makes possible hierarchical modelling with information gained from empirical data.

To achieve the goals of Bayesian approach — but without the pre-requisite of an assumed prior probability. ([Weichselberger, 2009](#), p. 3)

Many of the constituents already mentioned in his inaugural speech as rector in Berlin ([Weichselberger, 1968](#)), see also Section 3 above, are revisited in the light of the new foundation. The fundamental idea of logical probability as a two-place function on the reasoning from a premise to the conclusion is formalized in a system of axioms ([Weichselberger \(2009, p. 8\)](#), see [Weichselberger \(2016, Chapter 4\)](#) for more details), while the inference is developed in the context of a duality theory ([Weichselberger \(2009, p. 8\)](#), for the detailed arguments see [Weichselberger \(2016, Chapter 6\)](#)). Also the idea of a frequency interpretation of logical probability could be made rigorous (see [Weichselberger \(2009, p. 9\)](#) and [Weichselberger \(2016, Chapter 5\)](#)). Special aspects have been published in advance at the previous ISIPTAs ([Weichselberger, 2005, 2007](#)).

7. Concluding Remarks

We presented a preliminary summary of Kurt Weichselberger's contribution to the theory of imprecise probability. As already emphasized, this paper is a report on current research within the

17. Some aspects are already discussed in [Weichselberger \(2000, Section 3\)](#).

HiStaLMU project, an interdisciplinary project involving statisticians and historians of science to chronicle the history of statistics at LMU Munich. Concerning the research on Weichselberger's scientific biography, the next practical step is to build up the necessary infrastructure by establishing an archive of his office estate, and we hope that we can integrate further material from his family and friends. We also started to prepare a bibliometric network analysis on the spread and influence of Weichselberger's ideas. Far beyond the historical interest, a detailed rework of Weichselberger's unfinished opus and his scattered results will enable a deeper scientific discussion of his scientific inheritance. His results and ideas provide a big challenge, still promising a substantial impact on – nay a paradigmatic shift of – probability and statistics.

Acknowledgments

We are thankful to the three anonymous reviewers for helpful remarks and support. A few small parts of this text are based on a short obituary written by TA last year for the SIPTA list. Many thanks are due to Christina Schneider and Frank Coolen for their help with a draft of the obituary, and to Eva Endres for her comments on a draft of this paper. We thank Christa Jürgensonn, Hans Schneeweiß and Christina Schneider for their cooperativeness to be interviewed for the HiStaLMU project in 2013 and 2014, Nina Krem, Georgios Mechteridis and Theresa Parstorfer for their work on preparing and documenting the interviews, and Christiane Didden for research assistance. We also are very grateful to Wolfgang J Smolka and Claudius Stein from the Archive of LMU Munich for their help to start our archive project on Kurt Weichselberger.

References

- T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck and Ruprecht, Göttingen, 1998.
- T. Augustin. Neyman-Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105:149–173, 2002a. Based on an ISIPTA '99 paper.
- T. Augustin. Expected utility within a generalized concept of probability: a comprehensive framework for decision making under ambiguity. *Statistical Papers*, 43:5–22, 2002b.
- T. Augustin. Optimal decisions under complex uncertainty—basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. *Zeitschrift für Angewandte Mathematik und Mechanik.*, 84:678–687, 2004.
- F. Coolen. Book review “Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung, vol. I, Intervallwahrscheinlichkeit als umfassendes Konzept” by Weichselberger. *Journal of the Royal Statistical Society: Series D*, 52:253–254, 2003.
- L. de Campos, J. Huete, and S. Moral. Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.
- S. Destercke, D. Dubois, and E. Chojnacki. Unifying practical uncertainty representations – I: Generalized p-boxes. *International Journal of Approximate Reasoning*, 49:649–663, 2008.

- D. Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.
- M. GÜMBEL. *Über die effiziente Anwendung von F-PRI: ein Beitrag zur Statistik im Rahmen eines allgemeineren Wahrscheinlichkeitsbegriffs*. Pinus, Augsburg, 2009.
- N. Huntley, R. Hable, and M. Troffaes. Decision making. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 190–206. Wiley, Chichester, 2014.
- A. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. (English translation: *Foundations of Probability*, Chelsea, Providence, RI, 1950).
- S. Pöhlmann. *Kombination von unsicherem Wissen in Form von Wahrscheinlichkeitsintervallen*. Habilitationsschrift: Universität München, 1995.
- H. Rinne, B. Rüger, and H. Strecker, editors. *Grundlagen der Statistik und ihre Anwendungen (Festschrift für Kurt Weichselberger)*, Heidelberg, 1995. Physika.
- B. Rüger. Kurt Weichselberger. In H. Rinne, B. Rüger, and H. Strecker, editors, *Grundlagen der Statistik und ihre Anwendungen (Festschrift für Kurt Weichselberger)*, pages 3–14, Heidelberg, 1995. Physika.
- B. Rüger. *Nachruf auf Kurt Weichselberger*, 2016. see: <https://statsoz-neu.userweb.mwn.de/research/MemoryKurtWeichselberger/Weichselberger.pdf>.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, London, 1991.
- A. Wallner. *Beiträge zur Theorie der Intervallwahrscheinlichkeit: der Blick über Kolmogorov und Choquet hinaus*. Kovač, Hamburg, 2002.
- A. Wallner. Bi-elastic neighbourhood models. In J. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA '03*, pages 593–607, Waterloo, 2003. Carleton Scientific.
- A. Wallner. Extreme points of coherent probabilities on finite spaces. *International Journal of Approximate Reasoning*, 44:339–357, 2007. Based on an ISIPTA '05 paper.
- L. Wasserman. Book review: “A Methodology for Uncertainty in Knowledge-based Systems” by Weichselberger & Pöhlmann. *Journal of the American Statistical Association*, 86:546–547, 1991.
- K. Weichselberger. *Bernstein-Polynomapproximation in höheren Räumen*. Dissertation: Universität Wien, 1953.
- K. Weichselberger. *Kontrollen der Ergebnisse von Volkszählungen*. Habilitationsschrift: Universität zu Köln, 1962.
- K. Weichselberger. Einige Grundprobleme der Statistik und der Wahrscheinlichkeitstheorie, Rektoratsübergabe, 24.11.1967. *Technische Universität Berlin: Akademische Reden*, 47:34–50, 1968. Digitalized by Historische Kommission bei der Bayerischen Akademie der Wissenschaften see: www.historische-kommission-muenchen-editionen.de/rektoratsreden.

- K. Weichselberger. Über die statistischen Eigenschaften von Korrekturen aufgrund repräsentativer Verfahrenskontrollen. *Metrika*, 17:159–188, 1971.
- K. Weichselberger. Über eine Theorie der gleitenden Durchschnitte und verschiedene Anwendungen dieser Theorie. *Metrika*, 8:185–230, 1994.
- K. Weichselberger. Axiomatic foundations of the theory of interval probability. In V. Mammitzsch and H. Schneeweiss, editors, *Proc. 2nd Gauss Symp. B*, pages 47–64, Berlin, 1995a. de Gruyter.
- K. Weichselberger. Stichproben und Intervallwahrscheinlichkeit. *ifo studien*, 41:653–676, 1995b.
- K. Weichselberger. Interval probability on finite sample spaces. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, pages 391–409. Springer, New York, 1996.
- K. Weichselberger. The theory of interval probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000. Based on an ISIPTA '99 paper.
- K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Physica, Heidelberg, 2001a.
- K. Weichselberger. The status of F-indicator-fields within the theory of interval-probability. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA '01*, pages 362–369, Maastricht, 2001b. Shaker.
- K. Weichselberger. The logical concept of probability and statistical inference. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05*, pages 396–405, Manno, 2005. SIPTA.
- K. Weichselberger. The logical concept of probability: Foundation and interpretation. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *ISIPTA '07*, pages 455–463, Manno, 2007. SIPTA.
- K. Weichselberger. Symmetric probability theory, presentation at a *Special Session at ISIPTA '09*, Durham (UK), 2009.
- K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung II. Symmetrische Wahrscheinlichkeitstheorie*. (In preparation, unpublished manuscript.), 2016.
- K. Weichselberger and T. Augustin. Analysing Ellsberg's paradox by means of interval-probability. In R. Galata and H. Küchenhoff, editors, *Econometrics in Theory and Practice. (Festschrift for Hans Schneeweiss)*, pages 291–304. Physica, Heidelberg, 1998.
- K. Weichselberger and T. Augustin. On the symbiosis of two concepts of conditional interval probability. In J. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA '03*, pages 608–630, Waterloo, 2003. Carleton Scientific.
- K. Weichselberger and S. Pöhlmann. *A Methodology for Uncertainty in Knowledge-based Systems*. Springer, Heidelberg, 1990.
- K. Weichselberger and A.-R. Wulsten. Preisindizes für nicht-kommerzielle Forschung in der Bundesrepublik Deutschland 1968-1977. *Seminar für Spezialgebiete der Statistik, Universität München*, 1978.

SOS for Bounded Rationality

Alessio Benavoli

ALESSIO@IDSIA.CH

Alessandro Facchini

ALESSANDRO.FACCHINI@IDSIA.CH

Dario Piga

DARIO@IDSIA.CH

Marco Zaffalon

ZAFFALON@IDSIA.CH

Istituto Dalle Molle di Studi Sull'Intelligenza Artificiale (IDSIA), Lugano (Switzerland)

Abstract

In the gambling foundation of probability theory, rationality requires that a subject should always (never) find desirable all nonnegative (negative) gambles, because no matter the result of the experiment the subject never (always) decreases her money. Evaluating the nonnegativity of a gamble in infinite spaces is a difficult task. In fact, even if we restrict the gambles to be polynomials in \mathbb{R}^n , the problem of determining nonnegativity is NP-hard. The aim of this paper is to develop a *computable theory of desirable gambles*. Instead of requiring the subject to accept all nonnegative gambles, we only require her to accept gambles for which she can efficiently determine the nonnegativity (in particular SOS polynomials). We call this new criterion *bounded rationality*.

Keywords: bounded rationality; polynomial gambles; Sum-Of-Squares.

1. Introduction

The subjective foundation of probability by [de Finetti \(1937\)](#) is based on the notion of rationality (coherence or equiv. self-consistency). A subject is considered rational if she chooses her odds so that there is no bet that leads her to a sure loss (no Dutch books are possible). In this way, since odds are the inverse of probabilities, de Finetti provided a justification of Kolmogorov's axiomatisation of probability as a rationality criterion on a gambling system.¹

Later [Williams \(1975\)](#) and [Walley \(1991\)](#) shown that it is possible to justify probability in a simpler and more elegant way. This approach is nowadays known as the *theory of desirable gambles*. To understand this gambling framework, we introduce a subject, Alice, and an experiment whose result ω belongs to a possibility space Ω (e.g., the experiment may be tossing a coin or determining the future value of a derivative instrument). When Alice is uncertain about the result ω of the experiment, we can model her beliefs about this value by asking her whether she accepts to engage in certain risky transactions, called *gambles*, whose outcome depends on the actual outcome of the experiment ω . Mathematically, a gamble is a bounded real-valued function on Ω , $g : \Omega \rightarrow \mathbb{R}$, and if Alice accepts a gamble g , this means that she commits herself to receive $g(\omega)$ utiles² if the experiment is performed and if the outcome of the experiment eventually happens to be the event $\omega \in \Omega$. Since $g(\omega)$ can be negative, Alice can also lose utiles and hence the desirability of a gamble depends on Alice's beliefs about Ω . Denote by \mathcal{L} the set of all the gambles on Ω . Alice examines gambles in \mathcal{L} and comes up with the subset \mathcal{K} of the gambles that she finds desirable. How can we characterise the rationality of the assessments represented by \mathcal{K} ?

1. De Finetti actually considered only finitely additive probabilities, while σ -additivity is assumed in Kolmogorov's axiomatisation.
2. A theoretical unit of measure of utility, for indicating a supposed quantity of satisfaction derived from an economic transaction. It is expressed in some linear utility scale.

Two obvious rationality criteria are: Alice should always accept (reject) gambles such that $g \geq 0$ ($\sup g < 0$), because no matter the result of the experiment she never (always) decreases her utiles. There is a world of difference between saying and doing. For instance, let us consider an infinite space of possibilities like $\Omega = \mathbb{R}^2$ and the gamble: $g(x_1, x_2) = 4x_1^4 + 4x_1^3x_2 - 3x_1^2x_2^2 + 5x_2^4$. Should Alice accept this gamble? In practice the answer to this question does not only depend on Alice's beliefs about the value of x_1 and x_2 . We can in fact verify that the above polynomial can be rewritten as $(2x_1^2 - 2x_2^2 + x_1x_2)^2 + (x_2^2 + 2x_1x_2)^2$ and, thus, is always nonnegative. Hence, rationality implies that Alice should always accept it. However, in these cases, we must also take into account the inherent difficulty of the problem faced by Alice when she wants to determine whether a given gamble is nonnegative or not. In other words, we need to quantify the amount of computational resources needed to address rationality.

The aim of this paper is to develop a *computable theory of desirable gambles* by relaxing the two rationality criteria discussed above. In particular, instead of requiring Alice to accept all nonnegative gambles, we only require Alice to accept gambles for which she can efficiently determine the nonnegativity. We call this new criterion *bounded rationality*. The term bounded rationality was proposed by Herbert A. Simon – it is the idea that when individuals make decisions, their rationality is limited by the tractability of the decision problem, the cognitive limitations of their minds, and the time available to make the decision. Decision-makers in this view act as “satisficers”, seeking a satisfactory solution rather than an optimal one. We do not propose our model as a realistic psychological model of Alice's behaviour, but we embrace the idea that the actual rationality of an agent is determined by its computational intelligence.

In this paper, we exploit the results on SOS polynomials and theory-of-moments relaxation to make numerical inferences in our theory of bounded rationality and to show that the theory of bounded rationality can be used to approximate the theory of desirable gambles. At the same time, we provide a gambling interpretation of SOS optimization. Some preliminary applications of the theoretical ideas presented in this paper can be found in Lasserre (2009); Benavoli and Piga (2016); Piga and Benavoli (2018). It is worth mentioning that a relaxation of the rationality criteria for desirability has also been investigated in Schervish et al. (2000); Pelessoni and Vicig (2016). In the first case, the work focuses on relaxations of the “avoiding sure loss” axiom, while in the second on two different criteria (additivity and positive scaling).

2. Theory of desirable gambles

In this section, we briefly introduce the theory of desirable gambles. Let us denote by $\mathcal{L}^+ = \{g \in \mathcal{L} : g \geq 0\}$ the subset of the *nonnegative gambles* and with $\mathcal{K} \subset \mathcal{L}$ the subset of the gambles that Alice finds desirable. How can we characterise the rationality of the assessments in \mathcal{K} ?

Definition 1 *We say that \mathcal{K} is a coherent set of (almost) desirable gambles (ADG) when it satisfies the following rationality criteria:*

- A.1** *If $\inf g > 0$ then $g \in \mathcal{K}$ (Accepting Sure Gains);*
- A.2** *If $g \in \mathcal{K}$ then $\sup g \geq 0$ (Avoiding Sure Loss);*
- A.3** *If $g \in \mathcal{K}$ then $\lambda g \in \mathcal{K}$ for every $\lambda > 0$ (Positive Scaling);*
- A.4** *If $g, h \in \mathcal{K}$ then $g + h \in \mathcal{K}$ (Additivity);*
- A.5** *If $g + \delta \in \mathcal{K}$ for every $\delta > 0$ then $g \in \mathcal{K}$ (Closure).*

Note that A.1 and A.5 imply that $\mathcal{L}^+ \subseteq \mathcal{K}$ (including the zero gamble) (Walley, 1991; Miranda and Zaffalon, 2010). The criterion A.5 does not actually follow from rationality and can be omitted (Seidenfeld et al., 1990; Walley, 1991; Miranda and Zaffalon, 2010). However, it is useful to derive a connection between the theory of desirable gambles and probability theory and for this reason we consider it in this paper. This connection will be briefly discussed in Section 3.

To explain these rationality criteria, let us introduce a simple example: the toss of a fair coin $\Omega = \{\text{Head}, \text{Tail}\}$. A gamble g in this case has two components $g(\text{Head}) = g_1$ and $g(\text{Tail}) = g_2$. If Alice accepts g then she commits herself to receive/pay g_1 if the outcome is Heads and g_2 if Tails. Since a gamble is in this case an element of \mathbb{R}^2 , $g = (g_1, g_2)$, we can plot the gambles Alice accepts in a 2D coordinate system with coordinate g_1 and g_2 .

A.1 says that Alice is obviously willing to accept any gamble $g = (g_1, g_2)$ with $g_i > 0$ – Alice always accepts the first quadrant, Figure 1(a). Similarly, Alice does not accept any gamble $g = (g_1, g_2)$ with $g_i < 0$. In other words, Alice always rejects the interior of the third quadrant, Figure 1(b). This is the meaning of A.2. Then we ask Alice about $g = (-0.1, 1)$ – she loses 0.1 if Heads and wins 1 if Tails. Since Alice knows that the coin is fair, she accepts this gamble as well as all the gambles of the form νg with $\nu > 0$, because this is just a “change of currency” (this is A.3). Similarly, she accepts all the gambles $g + h$ for any $h \in \mathcal{L}^+$, since these gambles are even more favourable for her (this is basically A.4). Now, we can ask Alice about $g = (1, -0.1)$ and the argument is symmetric to the above case. We therefore obtain the following set of desirable gambles (see Figure 1(c)): $\mathcal{K}_2 = \{g \in \mathbb{R}^2 \mid 10g_1 + g_2 \geq 0 \text{ and } g_1 + 10g_2 \geq 0\}$. Finally, we can ask Alice about $g = (-1, 1)$ – she loses 1 if Heads and wins 1 if Tails. Since the coin is fair, Alice may accept or not accept this gamble. A.5 implies that she must accept it (closure). A similar conclusion can be derived for the symmetric gamble $g = (1, -1)$. Figure 1(d) is her final set of desirable gambles about the experiment concerned with the toss of a fair coin, which in a formula becomes $\mathcal{K}_3 = \{g \in \mathbb{R}^2 \mid g_1 + g_2 \geq 0\}$. Alice does not accept any other gamble. In fact, if Alice would also accept for instance $h = (-2, 0.5)$ then, since she has also accepted $g = (1.5, -1)$, i.e., $g \in \mathcal{K}_3$, she must also accept $g + h$ (because this gamble might also be favourable to her). However, $g + h = (-0.5, -0.5)$ is always negative, Alice always loses utiles in this case. In other words, by accepting $h = (-2, 0.5)$ Alice incurs a sure loss – she is irrational (A.2 is violated).

In this example, we can see that Alice’s set of desirable gambles is a closed half-space, but this does not have to be the case. For instance, if Alice does not know anything about the coin, she should only accept nonnegative gambles: $\mathcal{K} = \mathcal{L}^+$. This corresponds to a state of complete ignorance, but all intermediate cases from complete belief on the probability of the coin to complete ignorance are possible. In general, \mathcal{K} is a pointed (whose vertex is the origin) closed convex cone that includes \mathcal{L}^+ and exclude the interior of the negative orthant (this follows by A.1–A.5).

For the coin, the space of possibilities is finite and in this case Alice can check if a gamble g is nonnegative by simply examining the elements of the vector g . In this paper, we are interested in infinite spaces, in particular $\Omega = \mathbb{R}^n$, where applying the above rationality criteria is far from easy. We aim to develop a theory of *bounded rationality* for this case. Before doing that, we briefly recall the connection between ADG and probability theory.

3. Duality for ADG

Duality can be defined for general space of possibilities Ω (Walley, 1991). However, for the purpose of the present paper, we consider gambles that are bounded real-valued function on \mathbb{R}^n , i.e., $g :$

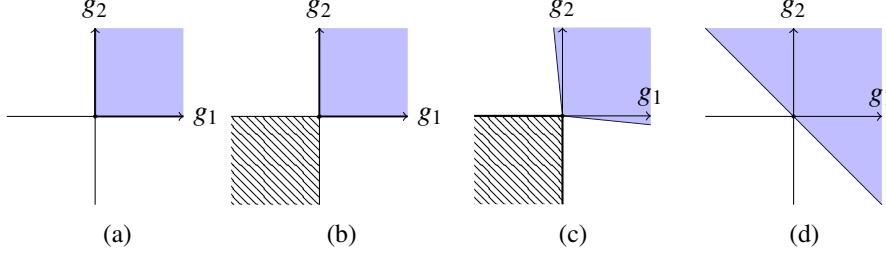


Figure 1: Alices' sets of coherent almost desirable gambles for the experiment of tossing a fair coin.

$\mathbb{R}^n \rightarrow \mathbb{R}$. Let \mathcal{A} be an algebra of subsets of \mathbb{R}^n and $\mu : \mathcal{A} \rightarrow [-\infty, \infty]$ denotes a charge: that is μ a finitely additive set function of \mathcal{A} ([Aliprantis and Border, 2007](#), Ch.11). Let $\mathcal{A}_{\mathbb{R}}$ denote the algebra generated in \mathbb{R} by the collection of all half open intervals ([Aliprantis and Border, 2007](#), Th.11.8):

Theorem 2 *Every bounded $(\mathcal{A}, \mathcal{A}_{\mathbb{R}})$ -measurable function is integrable w.r.t. any finite charge.*

For any $g \in \mathcal{L}$ and charge μ we can define $\int gd\mu$, that we can interpret as a linear functional $\langle \cdot, \mu \rangle$ on \mathcal{L} . We denote the set of all finite charges on \mathcal{A} as \mathcal{M} and the set of nonnegative charges as \mathcal{M}^+ . We can then define the dual of the coherent set of desirable gambles \mathcal{K} as:

$\{\mu \in \mathcal{M} : \int gd\mu \geq 0, \forall g \in \mathcal{K}\}$, and it can be proven that the above set is equivalent to

$$\mathcal{K}^* = \left\{ \mu \in \mathcal{M}^+ : \int gd\mu \geq 0, \forall g \in \mathcal{K} \right\}. \quad (1)$$

This follow by observing that: (i) $g = I_{\{x\}}$ (with I_x being the indicator function on $x \in \mathbb{R}^n$), is a nonnegative gamble and, therefore, is always in \mathcal{K} ; (ii) if μ is negative for some value of $x \in \mathbb{R}^n$, i.e., $x = \tilde{x}$, then $\int I_{\tilde{x}} d\mu$ is negative too and, thus, μ cannot be in \mathcal{K}^* . Hence, we can only focus on $\mu \in \mathcal{M}^+$. If we also impose the further requirement to $\langle \cdot, \mu \rangle$ to preserve constant gambles, in the sense that $\int cd\mu = c$, we obtain

$$\mathcal{P} = \left\{ \mu \in \mathcal{M}^+ : \int gd\mu \geq 0, \int d\mu = 1, \forall g \in \mathcal{K} \right\}. \quad (2)$$

We have imposed that $\int d\mu = 1$, i.e., μ is a probability charge. Hence, it can be observed that the dual of an ADG \mathcal{K} is a convex set of probability charges. The other direction of this result can be obtained by applying Hahn-Banach Theorem.

4. Finite assessments

The goal of this and next sections is to define a practical notion of desirability. To this end, we first assume that the set of gambles that Alice finds to be desirable is finitely generated. By this, we mean that there is a finite set of gambles $G = \{g_1, \dots, g_{|G|}\}$ such that $\mathcal{K} = \text{posi}(G \cup \mathcal{L}^+)$, where the posi of a set $A \subset \mathcal{L}$ is defined as $\text{posi}(A) := \left\{ \sum_{j=1}^{|G|} \lambda_j g_j : g_j \in A, \lambda_j \geq 0 \right\}$, and where by $|G|$ we denote the cardinality of G . By using this definition, it is clear that whenever \mathcal{K} is finitely generated, it includes all nonnegative gambles and satisfies A.3, A.4 and A.5. Once Alice has defined G and so

\mathcal{K} via posi , ADG assumes that she is able to perform the following operations: to check that \mathcal{K} avoids sure loss (A.2 is also satisfied); to determine the implication of desirability. It is easy to show that all above operations in ADG imply the assessment of the nonnegativity of a gamble.

Proposition 3 *Given a finite set $G \subset \mathcal{L}$ of desirable gambles, the set $\text{posi}(G \cup \mathcal{L}^+)$ includes the gamble f if and only if there exist $\lambda_j \geq 0$ for $j = 1, \dots, |G|$ such that*

$$f - \sum_{j=1}^{|G|} \lambda_j g_j \geq 0. \quad (3)$$

There are two subcases of (3) that are particularly interesting. The first is when $f = h - \lambda_0$ for some $\lambda_0 \in \mathbb{R}$ that allows us to define the concept of lower prevision [Walley \(1991\)](#); [Miranda \(2008\)](#).

Definition 4 *Assume that $\mathcal{K} = \text{posi}(G \cup \mathcal{L}^+)$ is an ADG, then the solution of the following problem*

$$\sup_{\lambda_0 \in \mathbb{R}, \lambda_j \geq 0} \lambda_0, \quad \text{s.t. } h - \lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j \geq 0, \quad (4)$$

is called the lower prevision of h and denoted as $\underline{P}[h]$.

From a behavioural point of view, we can reinterpret this by saying that Alice is willing to buy gamble h at price λ_0 , since she is giving away λ_0 utiles while gaining h . The lower prevision is the supremum buying price for h . We can equivalently define the upper prevision of h as $\bar{P}[h] = -\underline{P}[-h]$. From Section 3, it can be easily shown that $\underline{P}[h]$ is the lower expectation of h computed w.r.t. the probability charges in \mathcal{P} . As a matter of fact, the dual of (4) is the moment problem: $\inf_{\mu \in \mathcal{P}} \int h d\mu$. The second subcase allows us to formulate sure loss as nonnegativity of a gamble ([Walley et al., Alg.2](#)). Let us consider $\mathcal{K} = \text{posi}(G \cup \mathcal{L}^+)$ and the following problem:

$$\sup_{0 \leq \lambda_0 \leq 1, \lambda_j \geq 0} \lambda_0, \quad \text{s.t. } -\lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j \geq 0. \quad (5)$$

\mathcal{K} incurs a sure loss iff the above problem has solution $\lambda_0^* = 1$ and avoids sure loss iff $\lambda_0^* = 0$.

4.1 Complexity of inferences

When Ω is finite (e.g., coin toss), then a gamble g can also be seen as a vector in $\mathbb{R}^{|\Omega|}$, where ($|\Omega| = 2$ for the coin). Then (3) can be expressed as a linear programming problem, thus its complexity is polynomial: *Alice can check her coherence in polynomial time*. In case $\Omega = \mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$, solving (3) means to check the existence of real parameters $\lambda_j \geq 0$ ($j = 1, \dots, |G|$) such that the function

$$F := f - \sum_{j=1}^{|G|} \lambda_j g_j \quad (6)$$

is non-negative in \mathbb{R}^n . In order to study the problem from a computational viewpoint, and avoid undecidability results, it is clear that we must impose further restrictions on the class of functions F . At the same time we would like to keep the problem general enough, in order not to lose expressiveness of the model. A good compromise can be achieved by considering the case of

multivariate polynomials. The decidability of $F \geq 0$ for multivariate polynomials can be proven by means of the Tarski–Seidenberg quantifier elimination theory [Tarski \(1951\)](#); [Seidenberg \(1954\)](#).

Let $d \in \mathbb{N}$. By $\mathbb{R}_{2d}[x_1]$ we denote the set of all polynomials up to degree $2d$ in the indeterminate variable $x_1 \in \mathbb{R}$ with real-valued coefficients. With the usual definitions of addition and scalar multiplication, $\mathbb{R}_{2d}[x_1]$ becomes a vector space over the field \mathbb{R} of real numbers. We can introduce a basis for $\mathbb{R}_{2d}[x_1]$ that we denote as $v_{2d}(x_1)$ where $v_j(x_1) = [1, x_1, x_1^2, \dots, x_1^j]^\top$. We denote the dimension of $v_j(x_1)$ as $s_1(j)$ for $j = 0, 1, 2, \dots$, e.g., $s_1(2d) = 2d+1$. Any polynomial in $\mathbb{R}_{2d}[x_1]$ can then be written as $p(x_1) = b^\top v_{2d}(x_1)$ being $b \in \mathbb{R}^{s_1(2d)}$ the vector of coefficients. We may also be interested in some subsets of $\mathbb{R}_{2d}[x_1]$ that are: (1) the subset of nonnegative polynomials that we will be denoted as $\mathbb{R}_{2d}^+[x_1]$; (2) the subset of polynomials

$$\Sigma_{2d}[x_1] = \left\{ p(x_1) \in \mathbb{R}_{2d}[x_1] \mid p(x_1) = v_d^\top(x_1) Q v_d(x_1) \text{ with } Q \in \mathbb{R}_s^{s_1(d) \times s_1(d)}, Q \geq 0 \right\}, \quad (7)$$

where $\mathbb{R}_s^{s_1(d) \times s_1(d)}$ is the space of $s_1(d) \times s_1(d)$ real-symmetric matrices. The polynomial $\Sigma_{2d}[x_1]$ are also called SOS polynomials, because any polynomial in $\mathbb{R}_{2d}[x_1]$ that is a sum of squares of polynomials belongs to $\Sigma_{2d}[x_1]$ and viceversa ([Lasserre, 2009](#), Prop.2.1).

We can extend the previous framework to multivariate polynomials $\mathbb{R}_{2d}[x_1, \dots, x_n]$, by noticing that any polynomial in $\mathbb{R}_{2d}[x_1, \dots, x_n]$ can be written as $p(x_1, \dots, x_n) = b^\top v_{2d}(x_1, \dots, x_n)$ with

$$v_{2d}(x_1, \dots, x_n) = [1, x_1, \dots, x_n, x_1^2, x_1 x_2, \dots, x_{n-1} x_n, x_n^2, \dots, x_1^{2d}, \dots, x_n^{2d}]^\top, \quad (8)$$

$b \in \mathbb{R}^{s_n(2d)}$ with $s_n(j) = \binom{n+j}{j}$ for $j = 0, 1, 2, \dots$. Similarly to the univariate case we can define the nonnegative polynomials $\mathbb{R}_{2d}^+[x_1, \dots, x_n]$ and the SOS polynomials $\Sigma_{2d}[x_1, \dots, x_n]$. In the multivariate case, it is in general not true that every nonnegative polynomial is SOS or, in other words, in general $\Sigma_{2d}[x_1, \dots, x_n] \subset \mathbb{R}_{2d}^+[x_1, \dots, x_n]$. For instance $g(x_1, x_2) = x_1^2 x_2^2 (x_1^2 + x_2^2 - 1) + 1$ is a nonnegative polynomial that does not have a SOS representation ([Lasserre, 2009](#), Sec.2.4). [Hilbert \(1888\)](#) showed the following.

Proposition 5 $\mathbb{R}_{2d}^+[x_1, \dots, x_n] = \Sigma_{2d}[x_1, \dots, x_n]$ holds iff either $n = 1$ or $d = 1$ or $(n, d) = (2, 2)$.

The problem of testing global nonnegativity of a polynomial function is in general *NP-hard*. If Alice wants to avoid the complexity associated with this problem, an alternative option is to consider a subset of polynomials for which a nonnegativity test is not *NP-hard*. The problem of testing if a given *polynomial* is SOS has polynomial complexity (we only need to check if the matrix of coefficients Q in (7) is positive-semidefinite).

5. Bounded rationality

In the bounded rationality theory we are going to represent we will work with $\mathcal{Q} = \mathbb{R}^n$ and make two important assumptions. We assume that \mathcal{L} is the set of multivariate polynomials of n variables and of degree less than or equal to $2d$, with $d \in \mathbb{N}$. We denote \mathcal{L} as \mathcal{L}_{2d} and the nonnegative polynomials as \mathcal{L}_{2d}^+ . Note that \mathcal{L}_{2d} is a vector space and A.1–A.5 are well-defined in \mathcal{L}_{2d} . This restriction is useful to define the computational complexity of our bounded rationality theory as a function of n and d . We now define our bounded rationality criteria, and point out the two assumptions.

Definition 6 We say that $C \subset \mathcal{L}_{2d}$ is a **bounded-rationality** coherent set of almost desirable gambles (BADG) when it satisfies A.2–A.5 and:

bA.1 If $g \in \Sigma_{2d}$ then $g \in C$ (bounded accepting sure gain);

where $\Sigma_{2d} \subset \mathcal{L}_{2d}^+$ is the set of SOS of degree less than or equal to $2d$.

We have seen that A.1 and A.5 imply that a coherent set of gambles must include all nonnegative gambles (and, therefore, \mathcal{L}_{2d}^+ that is the set of all nonnegative polynomials). Here, we restrict A.1 imposing bounded-rationality that implies that the set must only include SOS polynomials up to degree $2d$. In BADG theory, we ask Alice only to accept SOS polynomials, i.e., gambles for which she can efficiently determine the nonnegativity. Note that in Walley's terminology (Walley, 1991, Sec. 3.7.8, Appendix F) the set C is coherent relative to the vector subspace of quadratic forms $v_{2d}(x_1, \dots, x_n)^T Q v_{2d}(x_1, \dots, x_n)$ defined by the symmetric real matrices Q (SOS are the nonnegative gambles in this subspace, i.e., $Q \geq 0$).

In the multivariate case, we have seen that there are nonnegative polynomials that do not have a SOS representation. These polynomials should be in principle desirable for Alice in the ADG framework, but in BADG we do not enforce Alice to accept them. For this reason, BADG is a theory of bounded rationality. Note that Alice may not be able to prove that her set of desirable gambles satisfies A.2. In fact, as it has been shown in (5) this requires to check the nonnegativity of a gamble. Note however that, the requirement A.2 is weaker than A.1. In fact, while A.1 requires Alice to accept all nonnegative gambles, A.2 only requires Alice to carefully choose the gambles in G so that a sure loss is not possible. We will return on A.2 later in the section.

A BADG set C that satisfies A.2 but not A.1 can (theoretically) be turned to an ADG in \mathcal{L}_{2d} by considering its extension $\text{posi}(C \cup \mathcal{L}_{2d}^+)$ and also to an ADG in \mathcal{L} by considering its extension $\text{posi}(C \cup \mathcal{L}^+)$ (note in fact that it holds $\Sigma_{2d} \subseteq \mathcal{L}_{2d}^+ \subset \mathcal{L}^+$). This is important because, as it will be shown in the next sections, it will allow us to use BADG as a computable approximation of ADG.

In BADG theory, Proposition 3 is reformulated as follows.

Theorem 7 Given a finite set $G \subset \mathcal{L}_{2d}$ of desirable gambles, the set $\text{posi}(G \cup \Sigma_{2d})$ includes the gamble f if and only if there exist $\lambda_j \geq 0$ for $j = 1, \dots, |G|$ such that

$$f - \sum_{j=1}^{|G|} \lambda_j g_j \in \Sigma_{2d}. \quad (9)$$

Also in this case we can consider the gamble $f = h - \lambda_0$ for some $\lambda_0 \in \mathbb{R}$ and define the concept of lower prevision.

Definition 8 Let $G \subset \mathcal{L}_{2d}$ be a finite set, and let $C = \text{posi}(G \cup \Sigma_{2d})$. Assume that C is BADG, then the solution of the following problem

$$\sup_{\lambda_0 \in \mathbb{R}, \lambda_j \geq 0} \lambda_0, \quad s.t. \quad h - \lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j \in \Sigma_{2d}, \quad (10)$$

is called the lower prevision of h and denoted as $\underline{P}^*[h]$.

We can similarly use (10) to prove that $C = \text{posi}(G \cup \Sigma_{2d})$ incurs a sure loss by solving the problem

$$\sup_{0 \leq \lambda_0 \leq 1, \lambda_j \geq 0} \lambda_0, \quad s.t. \quad -\lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j \in \Sigma_{2d}. \quad (11)$$

We have that if $\lambda_0^* = 1$ then C incurs a sure loss. A similar reasoning holds for any $0 \leq \lambda_0^* < 1$ since, as it will be shown in Section 5.2, λ_0^* is always smaller or equal than the solution obtained in (5). This means that we cannot use (11) to prove that C avoids a sure loss. An alternative way to guarantee that $C = \text{posi}(G \cup \mathcal{L}_{2d})$ avoids sure loss, is to relax (5) as

$$\lambda_0^{**} = \sup_{0 \leq \lambda_0 \leq 1, \lambda_j \geq 0} \lambda_0, \quad s.t. \quad -\lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j(x_k) \geq 0, \quad k = 1, \dots, M, \quad (12)$$

thus by enforcing that the constraint $-\lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j \geq 0$ only holds in M (randomly generated) points $x_k \in \mathbb{R}^n$. Indeed, if $\lambda_0^{**} < 1$, then the solution of problem (5) cannot be 1, thus C avoids sure loss. We will discuss this case with an example in Section 7.

5.1 Duality for BADG

We can also define the dual of a BADG. In this case, the gambles g are polynomials and the non-negative gambles that Alice accepts are SOS. Polynomials on \mathbb{R}^n are not bounded functions and, therefore, we cannot use Theorem 2.³ However, the rationality criteria A.1–A.5 do not explicitly need boundedness, but boundedness is essential to show the duality between ADG and closed convex set of probability charges, as shown in Section 3. However, since we are dealing with a vector space, we can consider its dual space \mathcal{L}_{2d}^\bullet , defined as the set of all linear maps $L : \mathcal{L}_{2d} \rightarrow \mathbb{R}$ (linear functionals). The dual of $C \subset \mathcal{L}_{2d}$ is defined as

$$C^\bullet = \{L \in \mathcal{L}_{2d}^\bullet : L(g) \geq 0, \forall g \in C\}. \quad (13)$$

Since \mathcal{L}_{2d} has a basis, i.e., the monomials, if we introduce the scalars

$$y_{\alpha_1 \alpha_2 \dots \alpha_n} := L(x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}) \in \mathbb{R}, \quad (14)$$

and we further assume that $y_0 = L(1) = 1$ (the linear functionals preserve constants), then we can rewrite $L(g)$ for any polynomial g as a function of the vector of variables $y \in \mathbb{R}^{s_n(2d)}$, whose components are the real variables $y_{\alpha_1 \alpha_2 \dots \alpha_n}$ defined above. This means that \mathcal{L}_{2d}^\bullet is isomorphic to $\mathbb{R}^{s_n(2d)}$. We can then rewrite the dual in a simpler form. Before doing that we define the matrix $M_{n,d}(y) := L(v_d(x_1, \dots, x_n)v_d(x_1, \dots, x_n)^\top)$, where the linear operator is applied component-wise. For instance, in the case $n = 1$ and $d = 2$, we have that

$$M_{1,2}(y) = L(v_2(x_1)v_2(x_1)^\top) = L\left(\begin{bmatrix} 1 & x_1 & x_1^2 \\ x_1 & x_1^2 & x_1^3 \\ x_1^2 & x_1^3 & x_1^4 \end{bmatrix}\right) = \begin{bmatrix} y_0 & y_1 & y_2 \\ y_1 & y_2 & y_3 \\ y_2 & y_3 & y_4 \end{bmatrix}.$$

We have then the following result (see for instance Lasserre, 2009).

Theorem 9 *Let C be a BADG. Then its dual is*

$$C^\bullet = \{y \in \mathbb{R}^{s_n(2d)} : L(g) \geq 0, M_{n,d}(y) \geq 0, \forall g \in C\}. \quad (15)$$

where $L(g)$ is completely determined by y via the definition (14).

3. For an extension of the theory of desirable gambles to unbounded gambles see Troffaes and De Cooman (2003)

Proof We have seen that any SOS in Σ_{2d} can be written as $v_d(x_1, \dots, x_n)^\top Q v_d(x_1, \dots, x_n)$ (see eq. (7)). By exploiting matrix algebra we have $v_d(x_1, \dots, x_n)^\top Q v_d(x_1, \dots, x_n)$ is equal to $Tr(Q v_d(x_1, \dots, x_n) v_d(x_1, \dots, x_n)^\top)$ with $Q \geq 0$. Now observe that because of linearity of L and trace

$$L(Tr(Q v_d(x_1, \dots, x_n) v_d(x_1, \dots, x_n)^\top)) = Tr(QL(v_d(x_1, \dots, x_n) v_d(x_1, \dots, x_n)^\top)) = Tr(QM_{n,d}(y))$$

where $M_{n,d}(y) = L(v_d(x_1, \dots, x_n) v_d(x_1, \dots, x_n)^\top)$. From $L(g) \geq 0$ in (13) for any $g \in \Sigma_{2d}$ we have $Tr(QM_{n,d}(y)) \geq 0$. This means that $Tr(QM_{n,d}(y)) \geq 0 \ \forall Q \geq 0$. This implies that $M_{n,d}(y) \geq 0$ (it can be proven by using the eigenvalue-eigenvector decomposition of $M_{n,d}(y)$). ■

The other direction follows by Hahn-Banach Theorem. Note that when $C = \Sigma_{2d}$, its dual is

$$C^* = \{y \in \mathbb{R}^{s_n(2d)} : M_{n,d}(y) \geq 0\}, \quad (16)$$

which corresponds to a state of ignorance: Alice only accepts nonnegative gambles.

In Section 3, by considering the space of all bounded gambles, we have showed that the dual of an ADG is a closed convex set of probability charges. In (15) there is no reference to probability. However, if the integral $\int x_1^{\alpha_1} x_2^{\alpha_2}, \dots, x_n^{\alpha_n} d\mu$ is well-defined, we can interpret $y_{\alpha_1 \alpha_2 \dots \alpha_n}$ as the expectation of $x_1^{\alpha_1} x_2^{\alpha_2}, \dots, x_n^{\alpha_n}$ w.r.t. the charge μ . Note that, $y_0 = L(1) = 1$ implies that $\int 1 d\mu = 1$ under this interpretation (normalization). Therefore, we can interpret $M_{n,d}(y)$ as a truncated moment matrix. However, since C does not include all nonnegative gambles, we cannot conclude that the charges are non-negative or, in other words, that μ is a probability charge. The constraint $M_{n,d}(y) \geq 0$ is not strong enough to guarantee non-negativity of μ (it is only a necessary condition). Negative probabilities are a manifestation of incoherence, that is they are a manifestation of the assumption of bounded rationality. Finally, the dual of the lower prevision problem (10) is then given by the convex SDP problem: $\inf_{y \in \mathbb{R}^{s_n(2d)}} L(h), \quad s.t. \quad L(g) \geq 0, \quad L(1) = 1, \quad M_{n,d}(y) \geq 0$.

Example 1 Consider the case $n = 1, d = 1$. The matrix $M_{1,2}(y)$ is in this case

$$M_{1,2}(y) = L\left(\begin{bmatrix} 1 & x_1 \\ x_1 & x_1^2 \end{bmatrix}\right) = \begin{bmatrix} 1 & y_1 \\ y_1 & y_2 \end{bmatrix}.$$

Assume that $G = \{g_1, g_2\} = \{x_1 - 0.5, -x_1 + 0.5\}$ and so $L(g_1) = L(x_1 - 0.5) = y_1 - 0.5$ and $L(g_2) = L(-x_1 + 0.5) = -y_1 + 0.5$. Hence, we have that

$$C^* = \{[y_1, y_2]^\top \in \mathbb{R}^2 : y_1 - 0.5 \geq 0, -y_1 + 0.5 \geq 0, M_{1,2}([y_1, y_2]^\top) \geq 0\}. \quad (17)$$

The first two constraints imply that $y_1 = 0.5$ and so we are left with the only constraint $\det(M_{1,2}([y_1, y_2]^\top)) = y_2 - 0.25 \geq 0$. Assume that we aim at computing $\underline{P}^*[-x_1(1-x_1)]$. The solution of (10) is $\underline{P}^*[-x_1(1-x_1)] = -0.25$ and it is attained for instance by the charge $0.352\delta_{0.367} + 0.786\delta_{0.521} - 0.138\delta_{0.281}$ (that is not a probability), here δ_a denoted an atomic charge (Dirac's delta) centred on a .

5.2 BADG as an approximating theory for ADG

We are going to show that we can use BADG as a computable approximating theory for ADG. So let us consider the BADG set $C = \text{posi}(G \cup \Sigma_{2d})$ and the corresponding ADG set $\mathcal{K} = \text{posi}(G \cup \mathcal{L}^+)$ (same G). We have the following result.

Theorem 10 Assume that \mathcal{K} avoids sure loss and let $f \in \mathcal{L}_{2d}$, then BADG is a conservative approximation of ADG theory in the sense that $\underline{P}^*(f) \leq \underline{P}(f)$.

Proof Let λ_0^* be the supremum value of λ_0 such that $h - \lambda_0 - \sum_j^{|G|} \lambda_j g_j \in \Sigma_{2d}$ and λ_0^{**} the value such that $h - \lambda_0 - \sum_j^{|G|} \lambda_j g_j \geq 0$. Since the constraint $h - \lambda_0 - \sum_j^{|G|} \lambda_j g_j \in \Sigma_{2d}$ is more demanding than $h - \lambda_0 - \sum_j^{|G|} \lambda_j g_j \geq 0$, it follows that $\lambda_0^* \leq \lambda_0^{**}$. ■

The fact that $\underline{P}[f]$ is equal to the minimum of f when G is empty, i.e., Alice is in a state of full ignorance, explains why SOS polynomials are used in optimization, i.e., $\underline{P}^*[f]$ provides a lower bound for the minimum of f (Lasserre, 2009).

6. Updating

We assume that Alice considers an event “indicated” by a certain finite set of polynomial constraints $A = \{h_1(x) \geq 0, \dots, h_{|A|}(x) \geq 0\}$: that means that Alice knows that x belongs to the set $A = \{x \in \mathbb{R}^n : h_1(x) \geq 0, \dots, h_{|A|}(x) \geq 0\}$. In ADG we will use this information to update (conditioning) her set of desirable gambles based on A (Walley, 1991; Couso and Moral, 2011) : $\mathcal{K}_{|A|} = \{g \in \mathcal{L} : gI_A \in \mathcal{K}\}$, where I_A is the indicator function on A . How do we do that in the BADG framework? In BADG we cannot completely use this information because Σ_{2d} does not include indicator functions. However, we can still exploit the information in A in a weaker way. In fact, if we know that $h_i(x) \geq 0$ in A , then we also know that $\sigma_1(x)h_1(x) + \dots + \sigma_{|A|}(x)h_{|A|}(x) \geq 0 \forall x \in A$ and for every $\sigma_i \in \Sigma_{2d}$ with degree equal to $d - \lceil n_{h_i}/2 \rceil$, where n_{h_i} is the degree of $h_i(x)$ (so that the degree is less than $2d$).

Definition 11 Let G be a finite subset of \mathcal{L}_{2d} , and $C = \text{posi}(G \cup \Sigma_{2d})$ be a set of BADG. Assume A is a finite set of polynomial constraints. Then, the set $C_{|A|}$ that includes all the gambles $f \in \mathcal{L}_{2d}$ such that there exist $\lambda_i \geq 0$, with $i = 1, \dots, |G|$, and $\sigma_0, \sigma_1, \dots, \sigma_{|A|}, \sigma_{|A|+1} \in \Sigma_{2d}$:

$$f - \sum_{i=1}^{|G|} \lambda_i g_i = \sigma_0 + \sum_{i=1}^{|A|} \sigma_i h_i \quad \text{and} \quad - \sum_{i=1}^{|G|} \lambda_i g_i = \sigma_{|A|+1} \quad (18)$$

is called the **updated set of desirable gambles based on A** .

In the state of full ignorance, since G is empty, there is only one constraint $f = \sigma_0 + \sum_{i=1}^{|A|} \sigma_i h_i$.

Theorem 12 Let G be a finite subset of \mathcal{L}_{2d} , and A be a finite set of polynomial constraints. Assume that $\mathcal{K} = \text{posi}(G \cup \mathcal{L}^+)$ avoids sure loss and let $f \in \mathcal{L}_{2d}$. Then we have that $\underline{P}_{C_{|A|}}(f) \leq \underline{P}_{\mathcal{K}_{|A|}}(f)$ where $C = \text{posi}(G \cup \Sigma_{2d})$.

Proof From the definition of conditioning for ADG we aim to find the supremum λ_0 such that $(f - \lambda_0)I_A - \sum_{j=1}^{|G|} \lambda_j g_j(x) \geq 0 \quad \forall x \in \mathbb{R}^n$. It can be rewritten as the two constraints on the left and relaxed to the constraints on the right:

$$\begin{aligned} - \sum_{j=1}^{|G|} \lambda_j g_j(x) &\geq 0 \quad \forall x \notin A, & - \sum_{j=1}^{|G|} \lambda_j g_j(x) &= \sigma_{|A|+1} \quad \forall x \in \mathbb{R}^n, \\ f - \lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j(x) &\geq 0 \quad \forall x \in A, & f - \lambda_0 - \sum_{j=1}^{|G|} \lambda_j g_j(x) &= \sigma_0 + \sum_{i=1}^{|A|} \sigma_i h_i \quad \forall x \in \mathbb{R}^n. \end{aligned}$$

■

Corollary 13 *The dual of $C_{|A}$ is*

$$C_{|A}^{\bullet} = \left\{ y \in \mathbb{R}^{s_n(d)} : L(g) \geq 0, M_{n,d}(y) \geq 0, M_{n,d-\lceil n_h/2 \rceil}(hy) \geq 0 \forall g \in C_{|A} \right\}$$

where $M_{n,d-\lceil n_h/2 \rceil}(hy) = L(h(x)v_{d-\lceil n_h/2 \rceil}(x)v_{d-\lceil n_h/2 \rceil}(x)^{\top})$ is called localizing matrix (Lasserre, 2009).

7. Numerical example

Consider the case $n = 2, d = 3$ and assume that Alice finds these gambles to be desirables

$$G = \{g_1, \dots, g_7\} = \{-x_1^4 x_2^2 - x_1^2 x_2^4 + x_1^2 x_2^2 - 1, x_1, 1 - x_1, x_2, 1 - x_2, 10 - x_1^2, 10 - x_2^2\}$$

Alice first checks if her set of desirable gambles satisfies A.2 by solving (11). The solution is $\lambda_0^* = 0.0062$ and, therefore, since $\lambda_0^* \approx 0$ Alice may think that G does not incur in sure loss. To numerically verify this statement, she can increase the degree d . For $d = 4$, Alice gets $\lambda_0^* = 0.0774$ that is greater than previous solution and for $d = 5$ $\lambda_0^* = 1$. Therefore, this shows that G actually incurs a sure loss. In this case, since $\arg \max_{i>0} \lambda_i^* = 1$, the polynomial that contributes more to the sure loss is g_1 .

Alice can verify if g_1 is negative by computing the BADG lower prevision of $-g_1$ for an empty G (this gives the minimum of $-g_1$ in ADG). The solution of (10) is $\underline{P}^*[-g_1] = -5.056$ for $d = 3$. For $d = 4$ we obtain $\underline{P}^*[-g_1] = 0.596$, for $d \geq 7$ $\underline{P}^*[-g_1] = 0.963$. Therefore, g_1 is strictly negative. It can be verified that 0.963 is the minimum of $-g_1$ and, therefore, $\underline{P}^*[-g_1] = \underline{P}[-g_1]$. So we have generated a family of BADG approximations (relaxations of coherence) that converge to ADG. Why can BADG obtain a lower “lower prevision” than ADG? In ADG $\underline{P}[-g_1]$ is attained by an atomic charge on the values of x_1, x_2 corresponding to the minimum of $-g_1$. Conversely, in BADG, since we allow mixtures of atomic charges with possibly negative weights, then we have more freedom in the minimization.

Alice can then remove g_1 from G and check if the following set satisfies A.2:
 $G \setminus g_1 = \{g_2, \dots, g_7\} = \{x_1, 1 - x_1, x_2, 1 - x_2, 10 - x_1^2, 10 - x_2^2\}$. To prove that, Alice can solve the linear programming problem (12) that gives the solution $\lambda_0^* \approx 10^{-17}$ and shows that G avoids sure loss.

Let $f = x_1^4 + 4x_1^3 + 5.375x_1^2 + 2.75x_1 + 0.41016$ and assume Alice aims to solve (10), i.e., to compute the BADG lower prevision of f . The result is $\underline{P}^*[f] = 0.41016$ for $d \geq 3$. Now let us assume $h(x_1) = 0.0025 - (x_1 + 0.425)^2$ and compute the conditional lower prevision. The solution is $\underline{P}^*[f|A] = -0.0625$ that gives the conditional lower prevision for BADG. This is also the minimum of f in $h(x) > 0$ and coincides with the conditional lower prevision for ADG $\underline{P}[f|A]$.

8. Conclusions

In this paper we have presented a computable theory of desirable gambles by imposing bounded rationality. To achieve that we have exploited recent results from Sum-Of-Square (SOS) polynomials optimization. As future work, we plan to further develop this theory by introducing other probabilistic operations such as marginalisation and structural judgements such as epistemic independence.

References

C. Aliprantis and K. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer, 2007.

- A. Benavoli and D. Piga. A probabilistic interpretation of set-membership filtering: Application to polynomial systems through polytopic bounding. *Automatica*, 70:158 – 172, 2016.
- I. Couso and S. Moral. Sets of desirable gambles: Conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52:1034–1055, 2011.
- B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l’Institut Henri Poincaré*, 7:1–68, 1937.
- D. Hilbert. Über die darstellung definiter formen als summe von formenquadrate. *Mathematische Annalen*, 32(3):342–350, 1888.
- J. B. Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2009.
- E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008. doi:[10.1016/j.ijar.2007.12.001](https://doi.org/10.1016/j.ijar.2007.12.001).
- E. Miranda and M. Zaffalon. Notes on desirability and conditional lower previsions. *Annals of Mathematics and Artificial Intelligence*, 60(3–4):251–309, 2010.
- R. Pelessoni and P. Vicig. 2-coherent and 2-convex conditional lower previsions. *International Journal of Approximate Reasoning*, 77:66–86, 2016.
- D. Piga and A. Benavoli. A unified framework for deterministic and probabilistic D -stability analysis of uncertain polynomial matrices. *IEEE Transactions on Automatic Control*, In press, 2018.
- M. J. Schervish, T. Seidenfeld, and J. B. Kadane. How sets of coherent probabilities may serve as models for degrees of incoherence. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(03):347–355, 2000.
- A. Seidenberg. A new decision method for elementary algebra. *Annals of Mathematics*, pages 365–374, 1954.
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Decisions without ordering. In W. Sieg, editor, *Acting and reflecting*, volume 211 of *Synthese Library*, pages 143–170. Kluwer, Dordrecht, 1990.
- A. Tarski. A decision method for elementary algebra and geometry. 1951.
- M. Troffaes and G. De Cooman. Extension of coherent lower previsions to unbounded random variables. *Intelligent systems for information processing: from representation to applications*, pages 277–288, 2003.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.
- P. Walley, R. Pelessoni, and P. Vicig. Direct algorithms for checking coherence and making inferences from conditional probability assessments. *Journal of Statistical Planning and Inference*, pages 119–151.
- P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975.

A Polarity Theory for Sets of Desirable Gambles

Alessio Benavoli

ALESSIO@IDSIA.CH

Alessandro Facchini

ALESSANDRO.FACCHINI@IDSIA.CH

Marco Zaffalon

ZAFFALON@IDSIA.CH

Istituto Dalle Molle di Studi Sull’Intelligenza Artificiale (IDSIA), Lugano (Switzerland)

José Vicente-Pérez

JOSE.VICENTE@UA.ES

Departamento de Fundamentos del Análisis Económico, Universidad de Alicante (Spain)

Abstract

Coherent sets of almost desirable gambles and credal sets are known to be equivalent models. That is, there exists a bijection between the two collections of sets preserving the usual operations, e.g. conditioning. Such a correspondence is based on the polarity theory for closed convex cones. Learning from this simple observation, in this paper we introduce a new (lexicographic) polarity theory for general convex cones and then we apply it in order to establish an analogous correspondence between coherent sets of desirable gambles and convex sets of lexicographic probabilities.

Keywords: desirability; credal sets; lexicographic probabilities; separation theorem; polarity.

1. Introduction

De Finetti (1937) established a foundation of probability theory based on the notion of “coherence” (self-consistency). The idea was that a subject is considered rational if she chooses her odds so that there is no bet that leads her to a sure loss (no Dutch books are possible). In this way, since numerically odds are the inverse of probabilities, de Finetti’s approach provides a justification of Kolmogorov’s axioms of probability as a rationality criterion on a gambling system.

Later, building on de Finetti’s betting setup, Williams (1975) and then Walley (1991) have shown that it is possible to justify probability in a way that is even simpler, more general and elegant. The basic idea is that an agent’s knowledge about the outcome of an experiment to be performed (e.g. tossing a coin) is provided by her set of *desirable* gambles, that is the set of gambles she is ready to accept. A gamble is modelled as a real-valued function g on the set Ω of outcomes of the experiment. Hence by accepting a gamble g , an agent commits herself to receive $g(\omega)$ *utils* in case the experiment is performed and the outcome of the experiment eventually happens to be the event $\omega \in \Omega$. Among all the sets of desirable gambles, we are able to find those satisfying some properties, and called *coherent* sets of desirable gambles, as they represent rational choices. Mathematically, those properties boil down to ask for a coherent set of desirable gambles to be a convex cone without the origin that contains all positive gambles, and thus avoids the negative ones (avoids partial loss). In spite of its simplicity, the *theory of desirable gambles* encompasses not only the Bayesian theory of probability but also other important mathematical models like upper and lower previsions or (credal) sets of probabilities.

An important variant of the traditional theory of probability is the probabilistic model of lexicographic probabilities (Blume et al., 1991), that is a sequence of standard probability measures. Developed to deal with the problem of conditioning on events of measure 0, it shares several features not only with models such as conditional probabilities or non-standard probabilities, but also with the theory of desirable gambles (see, e.g., Seidenfeld et al., 1990; Seidenfeld, 2000; Coz-

man, 2015; Van Camp et al., 2017). In particular Cozman (2015) notices that (conditional) sets of desirable gambles expressed via preference relations can be represented by sets of (conditional) lexicographic probabilities. This fact leads us to wonder whether, analogously to the case of sets of almost desirable gambles and sets of probabilities, a stronger, more fundamental correspondence exists between sets of desirable gambles and sets of lexicographic probabilities.

The goal of the present paper is to show that this is the case. That is, we verify that (conditional) sets of lexicographic probabilities and (conditional) sets of desirable gambles are isomorphic structures. In doing so, we provide a duality transformation (via orthogonal matrices) that allows us to go from a coherent set of desirable gambles to an equivalent set of lexicographic probabilities and vice versa. This transformation is an important contribution to uncertainty modelling because having access to dual models of uncertainty enables greater freedom of expression. In particular, we believe that the possibility of transferring through duality constructions from one theory to the other can be used to better understand issues related to lexicographic probabilities, such as defining independence.

2. Preliminaries

We start by introducing the necessary notation and basic definitions to be used later. Assume that the set of outcomes of an experiment is finite, say $\Omega = \{\omega_1, \dots, \omega_n\}$, and that there is an unknown true value in Ω . A gamble g on Ω is a mapping $g : \Omega \rightarrow \mathbb{R}$, and so $g(\omega)$ represents the reward the gambler would obtain if ω is the true unknown value. As the cardinality of Ω is n (a natural number), every gamble g on Ω can be thought as a point in the Euclidean space \mathbb{R}^n , and hence write $g = (g_1, \dots, g_n)$ with $g_i \in \mathbb{R}$ for every $i \in N := \{1, \dots, n\}$. In line with the tradition within the imprecise probability community, the set of all gambles defined on Ω is denoted by $\mathcal{L}(\Omega)$, although at times we simply write \mathbb{R}^n .

The elements of \mathbb{R}^n will be considered column vectors and the symbol \top will mean transpose. We denote by 0_n (-1_n , respectively) the vector whose components are all equal to 0 (−1, respectively). The vectors e^1, \dots, e^n stand for the canonical basis of \mathbb{R}^n , that is, e^i is the vector of zeros with a one in the i -th position, for all $i \in N$. Given $g, f \in \mathbb{R}^n$, the standard inner product of g and f is $\langle g, f \rangle := g^\top f$ and the Euclidean norm of g is $\|g\| := \sqrt{\langle g, g \rangle}$. For any subset $C \subset \mathbb{R}^n$, we denote by $\text{posi}(C)$ the set of all positive linear combinations of gambles in C , that is, $\text{posi}(C) := \{\sum_{j=1}^m \lambda_j g^j : g^j \in C, \lambda_j > 0, m \in \mathbb{N}\}$. We say that g is *less than or equal to* f (in short, $g \leq f$) whenever $g_i \leq f_i$ for all $i \in N$, and we will write $g < f$ whenever $g \leq f$ and $g \neq f$. The set of non-negative gambles is $\mathbb{R}_+^n := \{g \in \mathbb{R}^n : g \geq 0_n\}$. Furthermore, g is said to be *lexicographically less than* f (in short, $g <_L f$) if $g \neq f$ and $g_k < f_k$ for $k := \min\{i \in N : g_i \neq f_i\}$. We also write $g \leq_L f$ if either $g <_L f$ or $g = f$.

The following properties for a subset $\mathcal{K} \subset \mathbb{R}^n$ will be needed below.

- A1.** If $g \in \mathcal{K}$ and $f \in \mathcal{K}$, then $g + f \in \mathcal{K}$ (addition).
- A2.** If $g \in \mathcal{K}$ and $\lambda > 0$, then $\lambda g \in \mathcal{K}$ (positive homogeneity).
- A3.** If $g > 0_n$, then $g \in \mathcal{K}$ (accepting partial gain).
- A4.** $0_n \notin \mathcal{K}$ (avoiding status quo).
- A5.** If $g < 0_n$, then $g \notin \mathcal{K}$ (avoiding partial loss).
- A6.** $-1_n \notin \mathcal{K}$ (avoiding sure loss).
- A7.** If $g + f \in \mathcal{K}$ for all $f > 0_n$, then $g \in \mathcal{K}$ (closure).

A8. $0_n \in \mathcal{K}$ (accepting status quo).

Definition 1 A subset $\mathcal{K} \subset \mathbb{R}^n$ is said to be a coherent set of

- desirable gambles if it satisfies properties A1, A2, A3, A4;
- almost desirable gambles if it satisfies properties A1, A2, A3, A6, A7.

Thus, it easily follows that a coherent set of desirable gambles also satisfies properties A5 and A6, and a coherent set of almost desirable gambles also satisfies property A8. By definition, one has that the elements of \mathbb{D}_n , the family of all coherent sets of desirable gambles on Ω , are convex cones in \mathbb{R}^n omitting their apex (the origin), whereas the elements of \mathbb{A}_n , the family of all coherent sets of almost desirable gambles on Ω , are closed convex cones (containing the origin) in \mathbb{R}^n . However, not every convex cone omitting its apex (closed convex cone, respectively) belongs to \mathbb{D}_n (\mathbb{A}_n , respectively).

A crucial tool for duality within the framework of Convex Analysis is the polarity operator. Given a convex cone $K \subset \mathbb{R}^n$, the (*positive*) polar of K is defined to be

$$K^\circ := \{v \in \mathbb{R}^n : \langle v, g \rangle \geq 0 \text{ for all } g \in K\}.$$

Note that K° is a closed convex cone (containing the origin). Furthermore, one has $K^{\circ\circ} = \text{cl } K$ (see Rockafellar, 1970), and for closed convex cones $K_1, K_2 \subset \mathbb{R}^n$, one has $K_1 \subset K_2$ if and only if $K_2^\circ \subset K_1^\circ$.

Let $m \in \mathbb{N}$ with $m \leq n$. The symbol $\mathbb{M}_{m,n}$ denotes the space of real matrices with m rows and n columns, whereas $\mathbb{O}_{m,n}$ denotes the subset of matrices in $\mathbb{M}_{m,n}$ with orthonormal rows, that is, those matrices A satisfying $AA^\top = I$ (where I is the identity matrix of appropriate order). For $A \in \mathbb{M}_{m,n}$ we denote by a_{ij} the element of A in row i and column j , the i -th row of A is denoted by $a_{i\cdot}$, whereas its j -th column is denoted by $a_{\cdot j}$. Given $A \in \mathbb{M}_{n,n}$, we write $A \geq_L (>_L) 0_n$ (in the sense of Martínez-Legaz, 1984) if each column of A satisfies $a_{\cdot j} \geq_L (>_L) 0_n$ for all $j \in N$.

A probability mass function over Ω is any vector belonging to the set

$$\mathbb{P}_n := \left\{ p \in \mathbb{R}^n : 0 \leq p_i \leq 1, \sum_{i \in N} p_i = 1 \right\}.$$

Any closed convex subset of \mathbb{P}_n is called a *credal set*. We shall denote by \mathbb{C}_n the family of all credal sets within \mathbb{P}_n . A *lexicographic probability* over Ω is a sequence $\{p^j\}_{j=1}^m$ with $p^j \in \mathbb{P}_n$. We usually identify lexicographic probabilities over Ω with *stochastic matrices*, that is,

$$\mathbb{S}_{m,n} := \{P \in \mathbb{M}_{m,n} : p_{i\cdot} \in \mathbb{P}_n \text{ for all } i = 1, \dots, m\}.$$

We shall denote by $\mathbb{T}_{m,n}$ the subset of $\mathbb{S}_{m,n}$ containing all the full-rank stochastic matrices.

3. Almost desirability and probability

Walley (1991) showed that there is a one-to-one correspondence between coherent sets of almost desirable gambles and credal sets, say $C : \mathbb{A}_n \rightarrow \mathbb{C}_n$. Moreover, it is often claimed that this correspondence actually shows that the theory of almost desirable gambles and the theory of credal sets are equivalent. In this section, we first recall the bijection C which is based on the polarity theory for closed convex cones (Rockafellar, 1970). Second, by using the point of view of model theory (see e.g. Hodges, 1997), we explain how one has to understand the claim that the theory of almost desirable gambles and the theory of credal sets are equivalent. Finally, we prove the claim.

3.1 Polarity for almost desirability

The underlying tool for getting the aforementioned bijection is the *classical separation theorem for closed convex sets*: if $\mathcal{K} \subset \mathbb{R}^n$ is a nonempty closed convex cone, then for every $\bar{g} \notin \mathcal{K}$ there exists $v \in \mathbb{R}^n$ (non-null) such that $\langle v, g \rangle \geq 0 > \langle v, \bar{g} \rangle$ for all $g \in \mathcal{K}$. Thus, every closed convex cone $\mathcal{K} \subset \mathbb{R}^n$ can be written as $\mathcal{K} = \{g \in \mathbb{R}^n : \langle v^t, g \rangle \geq 0, t \in T\}$ for certain $v^t \in \mathbb{R}^n$ and T an arbitrary index set. In such a case, a well-known result in Convex Analysis (see Rockafellar, 1970) states that \mathcal{K}° coincides with the closure of the conic convex hull of the $\{v^t, t \in T\}$. In particular, if $\mathcal{K} = \{g \in \mathbb{R}^n : \langle v, g \rangle \geq 0\}$ with $v \in \mathbb{R}^n$, then $\mathcal{K}^\circ = \mathbb{R}_+v = \{\lambda v : \lambda \geq 0\}$. Concerning the geometry of coherent sets of almost desirable gambles, any set $\mathcal{K} \in \mathbb{A}_n$ is characterised as a closed convex cone containing the set \mathbb{R}_+^n (or equivalently, containing all indicator gambles). Thus, as a particular case, since any $\mathcal{K} \in \mathbb{A}_n$ is a closed convex cone containing $\{e^1, \dots, e^n\}$, the following proposition holds.

Proposition 2 *Let $\mathcal{K} \in \mathbb{A}_n$ and $\bar{g} \notin \mathcal{K}$. Then, there exists $v \in \mathbb{R}^n$ with $v > 0_n$ and $\|v\| = 1$ such that $\langle v, g \rangle \geq 0_n > \langle v, \bar{g} \rangle$ for all $g \in \mathcal{K}$.*

Corollary 3 *For every $\mathcal{K} \in \mathbb{A}_n$, there exist an index set T and vectors $v^t > 0_n$ with $\|v^t\| = 1$ for all $t \in T$ such that $\mathcal{K} = \{g \in \mathbb{R}^n : \langle v^t, g \rangle \geq 0, t \in T\}$.*

Recall that a set $\mathcal{K} \in \mathbb{A}_n$ is said to be *maximal* if there is no other element $\mathcal{K}' \in \mathbb{A}_n$ such that $\mathcal{K} \subsetneq \mathcal{K}'$. Thus, we have that the maximal elements in \mathbb{A}_n are the closed halfspaces containing the origin in the boundary and determined by vectors with non-negative components and norm 1. Hence, if we denote by $\text{Max}(\mathbb{A}_n)$ the set of all maximal elements in \mathbb{A}_n , given $\mathcal{K} \in \mathbb{A}_n$ one has

$$\mathcal{K} \in \text{Max}(\mathbb{A}_n) \iff \exists v > 0_n, \|v\| = 1 \text{ (unique) such that } \mathcal{K} = \{g \in \mathbb{R}^n : \langle v, g \rangle \geq 0\}. \quad (1)$$

This means that there is a one-to-one correspondence between maximal coherent sets of almost desirables gambles and non-negative vectors with norm 1. Since a bijection between the set of non-negative vectors with norm 1 and \mathbb{P}_n exists, then there is a one-to-one correspondence between maximal coherent sets of almost desirables gambles and probability mass functions over Ω . Furthermore, as a consequence of Proposition 2, for any $\mathcal{K} \in \mathbb{A}_n$ one can write

$$\mathcal{K} = \bigcap \{\mathcal{K}' \in \text{Max}(\mathbb{A}_n) : \mathcal{K} \subset \mathcal{K}'\}.$$

The above equality and the one in (1) imply a reformulation of Proposition 2: if $\mathcal{K} \in \mathbb{A}_n$ and $g \notin \mathcal{K}$, then there exists $\mathcal{K}' \in \text{Max}(\mathbb{A}_n)$ such that $\mathcal{K} \subset \mathcal{K}'$ and $g \notin \mathcal{K}'$.

Next we define the function $\mathbf{C} : \mathbb{A}_n \rightarrow \mathbb{C}_n$ which maps coherent sets of almost desirable gambles into credal sets and it is the key for the equivalence of both theories. For a coherent set of almost desirable gambles $\mathcal{K} \in \mathbb{A}_n$, we associate the credal set

$$\mathbf{C}(\mathcal{K}) := \mathcal{K}^\circ \cap \mathbb{P}_n. \quad (2)$$

Observe that if $\mathcal{K} \in \text{Max}(\mathbb{A}_n)$ is determined by v as in (1), then $\mathbf{C}(\mathcal{K}) = (\sum_{i \in N} v_i)^{-1}v$.

Theorem 4 *The mapping $\mathbf{C} : \mathbb{A}_n \rightarrow \mathbb{C}_n$ defined in (2) is a bijection whose inverse is given by $\mathbf{C}^{-1}(\mathcal{P}) := \mathcal{P}^\circ$ for every credal set $\mathcal{P} \in \mathbb{C}_n$.*

Proof First, it is easy to see that, for any $\mathcal{K} \in \mathbb{A}_n$, the set $\mathbf{C}(\mathcal{K})$ is a credal set. Since $\mathbb{R}_+^n \subset \mathcal{K}$, one has $\mathcal{K}^\circ \subset (\mathbb{R}_+^n)^\circ = \mathbb{R}_+^n$. Moreover, \mathcal{K}° does not reduce to 0_n (this fact just happens whenever $\mathcal{K} = \mathbb{R}^n$, which does not belong to \mathbb{A}_n indeed) and so, \mathcal{K}° contains non-null non-negative vectors, and particularly, at least one vector with the sum of its components equal to 1 (up to normalisation). Thus, the set $\mathcal{K}^\circ \cap \mathbb{P}_n \subset \mathbb{P}_n$ is nonempty. Moreover, since both \mathcal{K}° and \mathbb{P}_n are closed convex sets and closedness and convexity are preserved under intersection, then $\mathbf{C}(\mathcal{K}) \in \mathbb{C}_n$.

We have shown that the mapping \mathbf{C} is well-defined, associating a credal set to each coherent set of almost desirable gambles. Next, we verify that \mathbf{C} is a bijection, that is, for any credal set $\mathcal{P} \in \mathbb{C}_n$, there exists a unique $\mathcal{K} \in \mathbb{A}_n$ such that $\mathbf{C}(\mathcal{K}) = \mathcal{P}$.

Given a credal set $\mathcal{P} \in \mathbb{C}_n$, it follows that $\mathbb{R}_+ \mathcal{P}$ is a closed convex cone contained in \mathbb{R}_+^n . Thus, by taking polars one has $\mathbb{R}_+^n = (\mathbb{R}_+^n)^\circ \subset (\mathbb{R}_+ \mathcal{P})^\circ = \mathcal{P}^\circ$ and so, $\mathbf{C}^{-1}(\mathcal{P}) \in \mathbb{A}_n$ as \mathcal{P}° is a closed convex cone containing \mathbb{R}_+^n . Indeed, $\mathbf{C}^{-1}(\mathcal{P}) \in \mathbb{A}_n$ is the unique coherent set of almost desirable gambles satisfying $\mathbf{C}(\mathbf{C}^{-1}(\mathcal{P})) = \mathcal{P}$. Furthermore, for any $\mathcal{K} \in \mathbb{A}_n$ one has $\mathbf{C}^{-1}(\mathbf{C}(\mathcal{K})) = \mathcal{K}$. ■

3.2 Theories as structures, and equivalence as isomorphism

The fact that \mathbf{C} establishes a bijection between coherent sets of almost desirable gambles and credal sets is clearly not enough for claiming that the two theories are equivalent. We also need to verify that such a mapping preserves all considered operations (like conditioning and marginalisation) and relations (like independence). In other words, we have to verify that it is an isomorphism, once the two theories, from the point of view of model theory, are formulated as structures on the same signature. To illustrate this point, let us assume that we are only interested in conditioning. From a model-theoretic point of view, this means that we are considering a signature consisting of only a unary functional symbol. The next steps are thence the following: (i) we have to state how the considered operation is defined over coherent sets of almost desirable gambles and over credal sets (in model-theoretic terms, we have to specify how the elements of the signature – in this case its unique element – must be interpreted in both cases), and then (ii) we have to show that the map \mathbf{C} preserves the considered operation (in model-theoretic terms, we have to verify that the map is a homomorphism).

Here below we thence recall the definition of this operation within the theory of almost desirable gambles as given in [De Cooman and Quaeghebeur \(2012\)](#), a slightly different but completely equivalent version as the one in [Walley \(1991\)](#). To this aim, given a subset $\Pi \subsetneq \Omega$ of cardinality $m < n$, we shall denote by Π^c the set of outcomes which are not in Π , that is, $\Pi^c := \Omega \setminus \Pi$. For a gamble $g \in \mathbb{R}^m$ we define the gamble $(g|_{\Pi^c}) \in \mathbb{R}^n$ as $(g|_{\Pi^c})(\omega) := g(\omega)$ if $\omega \in \Pi$ and $(g|_{\Pi^c})(\omega) := 0$ if $\omega \in \Pi^c$.

Definition 5 Let $\mathcal{K} \subset \mathbb{R}^n$. The conditioned set of \mathcal{K} with respect to Π is the set

$$(\mathcal{K}|_\Pi) := \{g \in \mathbb{R}^m : (g|_{\Pi^c}) \in \mathcal{K}\}.$$

Notice that conditioning does not necessarily preserve coherent sets of almost desirable gambles (see [Miranda and Zaffalon \(2010, Section 4\)](#) for a thorough discussion on this point). As an example, consider the sets $\Omega = \{1, 2\}$, $\Pi = \{2\}$ and $\mathcal{K} = \{g \in \mathbb{R}^2 : g_1 \geq 0\}$. Whereas $\mathcal{K} \in \mathbb{A}_2$, it holds that $(\mathcal{K}|_\Pi) = \mathbb{R} \notin \mathbb{A}_1$.

For a probability mass function p over Ω , let $p(\cdot|\Pi)$ denote the usual conditioning of p with respect to $\Pi \subset \Omega$. Hence, if $\mathcal{P} \subset \mathbb{P}_n$ is a credal set over Ω , the conditioning of \mathcal{P} on Π is the projection on Π of all $p(\cdot|\Pi) \in \mathbb{P}_n$, with $p \in \mathcal{P}$; that is $(\mathcal{P})_{|\Pi} := \{p \in \mathbb{P}_m : \exists q \in \mathcal{P} \text{ such that } (p|_{\Pi^c}) = q(\cdot|\Pi)\}$. Notice that this definition is completely equivalent as the usual definition of conditioning for credal sets as given in [Cousu and Moral \(2011\)](#).

We can thence formulate the missing property for the mapping \mathbf{C} to be called an isomorphism, and thus to be claimed to show the equivalence between the two theories (when the considered operation is conditioning only).

Theorem 6 *Let $\mathcal{K} \in \mathbb{A}_n$ and $\Pi \subset \Omega$. The following statements hold:*

- (i) $(\mathcal{K})_{|\Pi} \in \mathbb{A}_m$ if and only if $(\mathbf{C}(\mathcal{K}))_{|\Pi} \in \mathbb{C}_m$.
- (ii) If $(\mathcal{K})_{|\Pi} \in \mathbb{A}_m$, then $\mathbf{C}(\mathcal{K})_{|\Pi} = (\mathbf{C}(\mathcal{K}))_{|\Pi}$.

Proof It is enough to prove both claims for $\mathcal{K} \in \text{Max}(\mathbb{A}_n)$. Let $\{p\} = \mathbf{C}(\mathcal{K}) \in \mathbb{C}_n$. With i_Π we should denote the indicator gamble on Π . Since $\langle p, i_\Pi f \rangle = \langle i_\Pi p, f \rangle$ and Theorem 4, the following holds:

$$(\mathcal{K})_{|\Pi} = \{g \in \mathbb{R}^m : \langle i_\Pi p, f \rangle \geq 0, \text{ for } f \in \mathbb{R}^n \text{ such that } i_\Pi f = g|_{\Pi^c}\}. \quad (3)$$

Hence, for both points we conclude by applying Theorem 4 to Equation 3. ■

4. Desirability and lexicographic probabilities

As discussed by [Cozman \(2015\)](#), coherent sets of desirable gambles and lexicographic probabilities seem to share several properties. We wonder whether these two models are somehow equivalent, that is, if there is a one-to-one correspondence $\mathbf{G} : \mathbb{D}_n \rightarrow \mathbb{G}_n$ between coherent sets of desirable gambles and certain sets (to be defined later) of lexicographical probabilities, similar to the one existing for credal sets and coherent sets of almost desirable gambles described in Section 3.

4.1 Polarity for desirability

As done in Section 3, the following (lexicographic) separation theorem for convex sets will be now the key result for getting the aforementioned equivalence.

Theorem 7 (Martínez-Legaz (1983)) *Let $G \subset \mathbb{R}^n$ be a nonempty convex set and $\bar{g} \notin G$. Then, there exists $A \in \mathbb{M}_{n,n}$ and $b \in \mathbb{R}^n$ such that $Ag >_L b \geq_L A\bar{g}$ for all $g \in G$.*

The matrix A in the above theorem can be assumed to be full-rank, or even orthonormal. Consequently, every convex set $G \subset \mathbb{R}^n$ can be written as $G = \{g \in \mathbb{R}^n : A^t g >_L b^t, t \in T\}$ for certain $A^t \in \mathbb{M}_{n,n}$, $b^t \in \mathbb{R}^n$ and T an arbitrary index set. In particular, if $\mathcal{K} \subset \mathbb{R}^n$ is a convex cone omitting its apex, one can take $b = 0_n$ in Theorem 7 and write $\mathcal{K} = \{g \in \mathbb{R}^n : A^t g >_L 0_n, t \in T\}$ for certain $A^t \in \mathbb{M}_{n,n}$ (even in $\mathbb{O}_{n,n}$) and T an arbitrary index set.

At this point, we recall that in \mathbb{R}^n there exist maximal convex cones excluding their vertices which are called *semispaces (at the origin)* (see [Hammer, 1955](#)). Thus, a convex set $\mathcal{K} \subset \mathbb{R}^n$ is a semispace if and only if $0_n \notin \mathcal{K}$ and for all $g \in \mathbb{R}^n \setminus \{0_n\}$, exactly one of g and $-g$ belongs to

\mathcal{K} . Furthermore, according to [Singer \(1984, Lemma 1.1\)](#), $\mathcal{K} \subset \mathbb{R}^n$ is a semispace if and only if there exists $A \in \mathbb{O}_{n,n}$ (unique, as follows from [Martínez-Legaz and Singer \(1988, p. 139\)](#)) such that $\mathcal{K} = \{g \in \mathbb{R}^n : Ag >_L 0_n\}$. Thus, every convex cone omitting its apex can be written as an intersection of semispaces.

Concerning the geometry of coherent sets of desirable gambles, any set $\mathcal{K} \in \mathbb{D}_n$ is characterised as a convex cone omitting its apex and containing the set $Q := \mathbb{R}_+^n \setminus \{0_n\}$. Thus, as a consequence of the above statement, since any $\mathcal{K} \in \mathbb{D}_n$ is a convex cone containing $\{e^1, \dots, e^n\}$, the following proposition follows.

Proposition 8 *Let $\mathcal{K} \in \mathbb{D}_n$ and $\bar{g} \notin \mathcal{K}$. Then, there exists $A \in \mathbb{O}_{n,n}$ with $A >_L 0_n$ such that $Ag >_L 0_n \geq_L A\bar{g}$ for all $g \in \mathcal{K}$.*

Corollary 9 *For every $\mathcal{K} \in \mathbb{D}_n$, there exist an index set T and matrices $A^t \in \mathbb{O}_{n,n}$ with $A^t >_L 0_n$ for all $t \in T$ such that $\mathcal{K} = \{g \in \mathbb{R}^n : A^t g >_L 0_n, t \in T\}$.*

Next we characterise the matrices which are lexicographically greater than 0_n . We understand that a matrix is unitary if it has ones in the main diagonal.

Lemma 10 *Given $A \in \mathbb{M}_{n,n}$, the following statements are equivalent:*

- (i) $A >_L 0_n$.
- (ii) $Ag >_L 0_n$ for all $g > 0_n$.
- (iii) $A = LP$ for some unitary lower-triangular matrix L and some $P \in \mathbb{M}_{n,n}$ such that $p_{.j} > 0_n$ for all $j \in N$.

Proof (i) \Leftrightarrow (ii). If $Ag >_L 0_n$ for all $g > 0_n$, then in particular we have $a_{.j} = Ae^j >_L 0_n$ for all $j \in N$ since $e^j > 0_n$, and that is the definition of $A >_L 0_n$. Conversely, assume that $A >_L 0_n$ and so, $Ae^j >_L 0_n$ for all $j \in N$. Since any $g = (g_1, \dots, g_n) > 0_n$ can be written as $g = \sum_{i \in N} g_i e^i$ with $g_i \geq 0$ for all $i \in N$ and there is at least one index j such that g_j is strictly positive, then $Ag = \sum_{i \in N} g_i Ae^i >_L 0_n$.

(i) \Leftrightarrow (iii). Observe that $A >_L 0_n$ if and only if $A \geq_L 0_n$ and $a_{.j} \neq 0_n$ for each $j \in N$. According to [Martínez-Legaz \(1984, Proposition 2\)](#), $A \geq_L 0_n$ if and only if $A = LP$ for some unitary lower-triangular matrix $L \in \mathbb{M}_{n,n}$ and some $P \in \mathbb{M}_{n,n}$ such that $p_{ij} \geq 0$ for all $i, j \in N$. Since $a_{.j} = L(p_{.j})$ and L is a regular lower-triangular matrix, then $a_{.j} = 0_n$ if and only if $p_{.j} = 0_n$. Thus, the conclusion follows. \blacksquare

We say that a coherent set of desirable gambles $\mathcal{K} \in \mathbb{D}_n$ is *maximal* if there is no other element $\mathcal{K}' \in \mathbb{D}_n$ such that $\mathcal{K} \subset \mathcal{K}'$. Thus, we have that the maximal elements in \mathbb{D}_n are the semispaces (at the origin) given by matrices $A \in \mathbb{O}_{n,n}$ satisfying $A >_L 0_n$. Hence, if we denote by $\text{Max}(\mathbb{D}_n)$ the set of all maximal elements in \mathbb{D}_n , given $\mathcal{K} \in \mathbb{D}_n$ one has

$$\mathcal{K} \in \text{Max}(\mathbb{D}_n) \iff \exists A \in \mathbb{O}_{n,n}, A >_L 0_n \text{ (unique) such that } \mathcal{K} = \{g \in \mathbb{R}^n : Ag >_L 0_n\}. \quad (4)$$

This means that there is a one-to-one correspondence between maximal coherent sets of desirables gambles and orthonormal matrices whose columns are lexicographically positive. Furthermore, as a consequence of Proposition 8, for any $\mathcal{K} \in \mathbb{D}_n$ one can write

$$\mathcal{K} = \bigcap \{\mathcal{K}' \in \text{Max}(\mathbb{D}_n) : \mathcal{K} \subset \mathcal{K}'\}, \quad (5)$$

recovering thus the characterisation given in [Couso and Moral \(2011, Theorem 21\)](#). The above equality and the one in (4) imply a reformulation of Proposition 8: if $\mathcal{K} \in \mathbb{D}_n$ and $g \notin \mathcal{K}$, then there exists $\mathcal{K}' \in \text{Max}(\mathbb{D}_n)$ such that $\mathcal{K} \subset \mathcal{K}'$ and $g \notin \mathcal{K}'$.

The following notions will be useful in the sequel.

Definition 11 We say that $\mathcal{A} \subset \mathbb{M}_{n,n}$ is *L-convex* if $\mathcal{A} = \{A \in \mathbb{M}_{n,n} : Ag^t >_L b^t, t \in T\}$ for certain vectors $g^t, b^t \in \mathbb{R}^n$ for all $t \in T$. In other words, $\mathcal{A} \subset \mathbb{M}_{n,n}$ is *L-convex* if and only if for every $\bar{A} \notin \mathcal{A}$ there exist $g, b \in \mathbb{R}^n$ such that $Ag >_L b \geq_L \bar{A}g$ for all $A \in \mathcal{A}$.

Analogously, we say that $\mathcal{A} \subset \mathbb{M}_{n,n}$ is an *L-convex cone* (omitting its apex) if $\mathcal{A} = \{A \in \mathbb{M}_{n,n} : Ag^t >_L 0_n, t \in T\}$ for certain $g^t \in \mathbb{R}^n$ for all $t \in T$. For any $\mathcal{A} \subset \mathbb{M}_{n,n}$, we define the set $\text{Lposi}(\mathcal{A}) := \{B \in \mathbb{M}_{n,n} : Bg >_L 0_n \text{ for any } g \in \mathbb{R}^n \text{ satisfying } Ag >_L 0_n \text{ for all } A \in \mathcal{A}\}$. Thus, $B \notin \text{Lposi}(\mathcal{A})$ if and only if there is $g \in \mathbb{R}^n$ such that $Ag >_L 0_n \geq_L Bg$ for all $A \in \mathcal{A}$.

Next we define a new polarity operator which is suitable for general convex cones in \mathbb{R}^n .

Definition 12 For a set $\mathcal{K} \subset \mathbb{R}^n$, we define $\mathcal{K}^\blacklozenge := \{A \in \mathbb{M}_{n,n} : Ag >_L 0_n \text{ for all } g \in \mathcal{K}\}$. Furthermore, for a set $\mathcal{A} \subset \mathbb{M}_{n,n}$ we also define $\mathcal{A}^\lozenge := \{g \in \mathbb{R}^n : Ag >_L 0_n \text{ for all } A \in \mathcal{A}\}$.

The following facts can be derived from these definitions:

1. \mathcal{A}^\lozenge is a convex cone omitting its apex in \mathbb{R}^n . Moreover, $\mathcal{A} = (\mathcal{A}^\lozenge)^\blacklozenge$ if and only if \mathcal{A} is an *L-convex cone* omitting its apex in $\mathbb{M}_{n,n}$.
2. $\mathcal{K}^\blacklozenge$ is an *L-convex cone* omitting its apex in $\mathbb{M}_{n,n}$. Moreover, $\mathcal{K} = (\mathcal{K}^\blacklozenge)^\lozenge$ if and only if \mathcal{K} is a convex cone omitting its apex in \mathbb{R}^n . In particular, this equality holds whenever $\mathcal{K} \in \mathbb{D}_n$.
3. For any $\mathcal{K}, \mathcal{H} \subset \mathbb{R}^n$, if $\mathcal{K} \subset \mathcal{H}$ then $\mathcal{H}^\blacklozenge \subset \mathcal{K}^\blacklozenge$. Analogously, for any $\mathcal{A}, \mathcal{B} \subset \mathbb{M}_{n,n}$, if $\mathcal{A} \subset \mathcal{B}$ then $\mathcal{B}^\lozenge \subset \mathcal{A}^\lozenge$.
4. $\mathcal{K}^\blacklozenge = \{A \in \mathbb{M}_{n,n} : \mathcal{K} \subset A^\lozenge\}$ and $\mathcal{A}^\lozenge = \{g \in \mathbb{R}^n : \mathcal{A} \subset g^\blacklozenge\}$.

Proposition 13 The following statements hold:

- (i) If $\mathcal{A} = \{A \in \mathbb{M}_{n,n} : Ag^t >_L 0, t \in T\}$, then $\mathcal{A}^\lozenge = \text{posi}\{g^t, t \in T\}$.
- (ii) If $\mathcal{K} = \{g \in \mathbb{R}^n : A^t g >_L 0, t \in T\}$, then $\mathcal{K}^\blacklozenge = \text{Lposi}\{A^t, t \in T\}$.

Proof (i) Clearly, $g^t \in \mathcal{A}^\lozenge$ for all $t \in T$. Since \mathcal{A}^\lozenge is a convex cone omitting its apex, then $\text{posi}\{g^t, t \in T\} \subset \mathcal{A}^\lozenge$. To prove the converse statement, assume that there is $\bar{g} \in \mathcal{A}^\lozenge$ such that $\bar{g} \notin \text{posi}\{g^t, t \in T\}$. By the separation theorem, there exists $A \in \mathbb{M}_{n,n}$ such that $Ag >_L 0_n \geq_L A\bar{g}$ for all $g \in \text{posi}\{g^t, t \in T\}$. In particular, $Ag^t >_L 0_n$ for all $t \in T$, which implies that $A \in \mathcal{A}$. Thus, as $\bar{g} \in \mathcal{A}^\lozenge$, one has $A\bar{g} >_L 0_n$, which entails a contradiction. The proof of (ii) follows the same reasoning as for (i). \blacksquare

Remark 14 As a consequence of the above result, if we consider the sets $\mathcal{H} := \{g \in \mathbb{R}^n : g > 0_n\}$ and $\mathcal{B} := \{A \in \mathbb{M}_{n,n} : A >_L 0_n\}$, then one has $\mathcal{H}^\blacklozenge = \mathcal{B}$ and $\mathcal{B}^\lozenge = \mathcal{H}$.

As this point, we establish an important correspondence between orthonormal matrices with lexicographically positive columns and equivalence classes of full-rank stochastic matrices. Next result guarantees the existence of a full-rank stochastic matrix determining the same semispace as a given orthonormal matrix $A >_L 0_n$, and the proof provides a method for obtaining such a matrix.

Proposition 15 *Let $A \in \mathbb{O}_{n,n}$ be such that $A >_L 0_n$. Then, there exists a full-rank stochastic matrix $P \in \mathbb{T}_{n,n}$ such that $P^\diamond = A^\diamond$.*

Proof In virtue of Lemma 10, one can write $A = LQ$ with L a unitary lower-triangular matrix and Q such that $q_{\cdot j} > 0_n$ for all $j \in N$. Thus, one has $a_{1\cdot} = q_{1\cdot}$ and $a_{i\cdot} = \sum_{j=1}^{i-1} l_{ij} q_{j\cdot} + q_{i\cdot}$ for $i \in N \setminus \{1\}$. Since A is orthonormal, then it follows that $q_{i\cdot} > 0_n$ for all $i \in N$, that is, Q does not have null rows, and clearly Q is full-rank as A is. By normalising each row so as that each row becomes a probability mass function, that is, by dividing each row by its sum, one gets the existence of a $P \in \mathbb{T}_{n,n}$. Finally, we observe that $A^\diamond = Q^\diamond = P^\diamond$. ■

The following proposition studies the way of getting an orthonormal matrix being lexicographically greater than 0_n from a full-rank stochastic one.

Proposition 16 *Let $P \in \mathbb{T}_{n,n}$ be a full-rank stochastic matrix. Then, there exists $A \in \mathbb{O}_{n,n}$ with $A >_L 0_n$ such that $A^\diamond = P^\diamond$.*

Proof We shall denote by $\text{GS}(P)$ the orthogonal matrix obtained from the full-rank stochastic matrix $P \in \mathbb{T}_{n,n}$ by applying the Gram–Schmidt orthogonalisation procedure according to the row order. Let $A \in \mathbb{O}_{n,n}$ be the orthonormal matrix obtained from $\text{GS}(P)$ by normalising each row. Since P have neither null rows nor null columns, it follows that $\text{GS}(P) >_L 0_n$ and so, $A >_L 0_n$. Finally, the Gram–Schmidt procedure guarantees that $A^\diamond = P^\diamond$. ■

The next example illustrates that the matrix whose existence has been guaranteed in the Proposition 15 is not necessarily unique.

Example 1 *Let us consider the maximal coherent set of desirable gambles $\mathcal{K} = \{g \in \mathbb{R}^3 : Ag >_L 0_3\}$, where $A = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \\ 1 & 0 & 0 \end{bmatrix}$. Since $A >_L 0_3$, following Lemma 10 A can be written as*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \tau & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & (-1-\tau)/\sqrt{2} & (1-\tau)/\sqrt{2} \\ 1 & 0 & 0 \end{bmatrix}$$

for any $\tau \leq -1$, $l_{31}, l_{32} \in \mathbb{R}$. According to Proposition 15, by normalising each row of the second matrix in the right-hand side of the equality above, we get that every matrix

$$P(\tau) = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & (\tau+1)/2\tau & (\tau-1)/2\tau \\ 1 & 0 & 0 \end{bmatrix},$$

with $\tau \leq -1$, is a full-rank stochastic matrix which determines \mathcal{K} . Finally, it can be checked that $\text{GS}(P(\tau)) = A$ holds for any $\tau \leq -1$ (after normalisation).

The above results suggest the definition of the \diamond -equivalence class of a given matrix $A \in \mathbb{M}_{n,n}$ as the set of matrices having the same polar that A , that is, $[A]_{\diamond} := \{P \in \mathbb{M}_{n,n} : P^{\diamond} = A^{\diamond}\}$. According to this definition, we have that there is a one-to-one correspondence between maximal coherent sets of desirable gambles and \diamond -equivalence classes of stochastic matrices of full rank.

Definition 17 *We say that a nonempty subset of $\mathbb{M}_{n,n}$ is an L-credal set if it is the intersection with $\mathbb{T}_{n,n}$ of some L-convex cone in $\mathbb{M}_{n,n}$. We shall denote by \mathbb{G}_n the family of all L-credal sets.*

We are now in position to define the function $\mathbf{G} : \mathbb{D}_n \rightarrow \mathbb{G}_n$ which maps coherent sets of desirable gambles into L-credal sets and it is the key for the equivalence of both theories. For a coherent set of desirable gambles $\mathcal{K} \in \mathbb{D}_n$, we associate the L-credal set

$$\mathbf{G}(\mathcal{K}) := \mathcal{K}^{\diamond} \cap \mathbb{T}_{n,n}. \quad (6)$$

We aim at showing that \mathbf{G} is a bijection.

Theorem 18 *The mapping $\mathbf{G} : \mathbb{D}_n \rightarrow \mathbb{G}_n$ defined in (6) is a bijection whose inverse is given by $\mathbf{G}^{-1}(\mathcal{P}) := \mathcal{P}^{\diamond}$, for every $\mathcal{P} \in \mathbb{G}_n$.*

Proof From the definition of the \diamond -polarity operator, $\mathbf{G}(\mathcal{K})$ is an L-credal set, for any $\mathcal{K} \in \mathbb{D}_n$. As $\mathcal{H} \subset \mathcal{K}$, then $\mathcal{K}^{\diamond} \subset \mathcal{H}^{\diamond} = \mathcal{B}$ (see Remark 14). One also has $\mathcal{K}^{\diamond} = \{A \in \mathbb{M}_{n,n} : \mathcal{K} \subset A^{\diamond}\}$. Since \mathcal{K} is determined by orthonormal matrices, then \mathcal{K}^{\diamond} contains orthonormal matrices with lexicographically positive columns and, as a consequence of Proposition 15, \mathcal{K}^{\diamond} also contains full-rank stochastic matrices, which shows that $\mathbf{G}(\mathcal{K})$ is nonempty. Now, if $\mathcal{P} \in \mathbb{G}_n$, one has that $\mathbf{G}^{-1}(\mathcal{P}) = \mathcal{P}^{\diamond}$ is a convex cone omitting its apex. On the other hand, as $\mathcal{P} \subset \mathbb{T}_{n,n} \subset \mathcal{B}$, then $Q = \mathcal{B}^{\diamond} \subset \mathcal{P}^{\diamond}$ and so, $\mathbf{G}^{-1}(\mathcal{P}) \in \mathbb{D}_n$.

To see that \mathbf{G} is one-to-one, we just need to show $\mathbf{G}(\mathbf{G}^{-1}(\mathcal{P})) = \mathcal{P}$ for any $\mathcal{P} \in \mathbb{G}_n$ and also $\mathbf{G}^{-1}(\mathbf{G}(\mathcal{K})) = \mathcal{K}$ for $\mathcal{K} \in \mathbb{D}_n$. First, $\mathbf{G}(\mathbf{G}^{-1}(\mathcal{P})) = \mathbf{G}(\mathcal{P}^{\diamond}) = \mathcal{P}^{\diamond\diamond} \cap \mathbb{T}_{n,n} = \text{Lposi}(\mathcal{P}) \cap \mathbb{T}_{n,n} = \mathcal{P}$. On the other hand, $\mathbf{G}^{-1}(\mathbf{G}(\mathcal{K})) = \mathbf{G}^{-1}(\mathcal{K}^{\diamond} \cap \mathbb{T}_{n,n}) = (\mathcal{K}^{\diamond} \cap \mathbb{T}_{n,n})^{\diamond} = \mathcal{K}^{\diamond\diamond} = \mathcal{K}$ as \mathcal{K} is a convex cone omitting its apex. ■

4.2 Closing the circle, or preserving conditioning

As for almost desirability, one wants to verify that \mathbf{G} is not only a bijection but also an isomorphism. To make sense of this claim, we thus have first to specify which operations and relations we decide to consider (in model-theoretic terms, the signature), and how they are defined over sets of gambles and over sets of stochastic matrices (in model-theoretic terms, the interpretation). Finally, we have to verify that the map \mathbf{G} preserves the considered operations and relations. As before, here we are only interested in conditioning.

Without loss of generality we assume that $\Pi \subsetneq \Omega$ has cardinality m . In the case of stochastic matrices, conditioning has to be defined by slightly modifying the approach by Blume et al. (1991). This is because we want to be sure that the result of the operation is a square stochastic matrix. With this aim in mind, we first define the following reduction rule for matrices:

- (R) Given $A \in \mathbb{M}_{n,m}$, for every $i \in N$, discard the i -th row a_i . whenever it is a linear combination of a_1, \dots, a_{i-1} . (and thus in particular when it is equal to 0_m).

Let $P' \in \mathbb{M}_{n,m}$ be the matrix obtained by projecting on Π the conditioning $p(\cdot|\Pi)$, or taking 0_m when it is undefined, for each row p of $P \in \mathbb{T}_{n,n}$. Define $P|_{\Pi}$ as the matrix obtained from P' by applying rule (R). By an immediate application of properties of minors and cofactors, we get that $P|_{\Pi} \in \mathbb{T}_{m,m}$. Moreover $(P|_{\Pi})|_{\Delta} = (P|_{\Delta})$, for $\Delta \subset \Pi$. Hence, the following operation is always defined.

Definition 19 Let $\mathcal{P} \subset \mathbb{T}_{n,n}$, with $n > 1$. Its conditioning on Π is the set $(\mathcal{P}|_{\Pi}) := \{(P|_{\Pi}) \mid P \in \mathcal{P}\} \subset \mathbb{T}_{m,m}$.

From Definition 5, it is immediate to verify that $(\mathcal{K}|_{\Pi}) \in \mathbb{D}_m$ whenever $\mathcal{K} \in \mathbb{D}_n$, and that \mathbb{D}_n is closed under conditioning. Moreover, $(\mathcal{K}|_{\Pi}) \in \text{Max}(\mathbb{D}_m)$ whenever $\mathcal{K} \in \text{Max}(\mathbb{D}_n)$. To conclude, we verify that polarity preserves conditioning.

Theorem 20 Let $\mathcal{K} \in \mathbb{D}_n$, then $(\mathbf{G}(\mathcal{K})|_{\Pi}) = \mathbf{G}(\mathcal{K}|_{\Pi}) \in \mathbb{G}_m$.

Proof It is enough to prove the claim for maximal consistent sets of desirable gambles. Hence, let $\mathcal{K} \in \text{Max}(\mathbb{D}_n)$. We first define a conditioning operation on orthogonal matrices. Let $A \in \mathbb{O}_{n,n}$. Its conditioning on Π is the matrix $A|_{\Pi}$ obtained by the following procedure: (i) erase all k -th column from A , with $k \in \{m+1, \dots, n\}$; (ii) apply rule (R) to the matrix obtained after the previous point; (iii) assume the matrix you obtained after the previous point is B . By linear algebra, $B \in \mathbb{U}_{m,m}$. Hence, $A|_{\Pi} := \text{GS}(B) \in \mathbb{O}_{m,m}$. Note that the operation also preserves the property of being lexicographic positive for columns. Thus, let $A \in \mathbb{O}_{n,n}$, $A >_L 0_n$, such that $\mathcal{K} = A^{\diamond}$. Both $(\mathcal{K}|_{\Pi}), (A^{\diamond}|_{\Pi}) \in \text{Max}(\mathbb{D}_m)$. This means that, in order to show that $(\mathcal{K}|_{\Pi}) = (A^{\diamond}|_{\Pi})$, it is enough to verify one of the two inclusions. So, let $f \in (\mathcal{K}|_{\Pi})$. By definition $f|_{\Pi^c} \in \mathcal{K}$, and thus $A(f|_{\Pi^c}) >_L 0_n$. But this means that $Bf >_L 0_n$, since $f|_{\Pi^c}$ agrees on Π with f , and is 0 elsewhere. Thence $\text{GS}(B)f >_L 0_n$, meaning that $f \in A^{\diamond}|_{\Pi}$. Now, because of the properties of the procedures given by Propositions 15 and 16, it holds that $P \in [A]_{\diamond}$ if and only if $P|_{\Pi} \in [A|_{\Pi}]_{\diamond}$, for $P \in \mathbb{T}_{n,n}$. Finally, we can apply Theorem 18 and conclude that $(\mathbf{G}(\mathcal{K})|_{\Pi}) = \mathbf{G}(\mathcal{K}|_{\Pi})$. ■

5. Conclusions

In this paper we have shown that (conditional) sets of lexicographic probabilities and (conditional) sets of desirable gambles are isomorphic structures. In doing so, we have provided a duality transformation (via orthogonal and stochastic matrices) that allows us to go from a coherent set of desirable gambles to an equivalent (convex) set of lexicographic probabilities and vice versa. As future work we plan to complete this analysis by including other operations, such as marginalisation (this should be straightforward), and structural judgements such as independence. It would be also of great interest to study what are the geometric properties of lexicographic convex sets of stochastic matrices, and what happens for gambles on infinite sample spaces.

Acknowledgments

The authors are grateful to the referees for their constructive comments and helpful suggestions which have contributed to the final preparation of the paper. J. Vicente-Pérez was partially supported by MINECO of Spain and ERDF of EU, Grants MTM2014-59179-C2-1-P and ECO2016-77200-P.

References

- L. Blume, A. Brandenburger, and E. Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59(1):61–79, 1991.
- I. Couso and S. Moral. Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011.
- F. G. Cozman. Some remarks on sets of lexicographic probabilities and sets of desirable gambles. In *9th ISIPTA*, Pescara, Italy, 2015.
- G. De Cooman and E. Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53(3):363–395, 2012.
- B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l’Institut Henri Poincaré*, 7:1–68, 1937.
- P. C. Hammer. Maximal convex sets. *Duke Math. J.*, 22:103–106, 1955.
- W. Hodges. *A shorter model theory*. Cambridge University Press, 1997.
- J. Martínez-Legaz. Exact quasiconvex conjugation. *Zeitschrift für Operations-Research*, 27(1):257–266, 1983.
- J. E. Martínez-Legaz. Lexicographical order, inequality systems and optimization. In *Proceedings of the 11th IFIP Conference on System Modelling and Optimization*, volume 59 of *Lecture Notes in Control and Inform. Sci.*, pages 203–212. Springer, 1984.
- J. E. Martínez-Legaz and I. Singer. The structure of hemispaces in \mathbb{R}^n . *Linear Algebra and its Applications*, 110:117–179, 1988.
- E. Miranda and M. Zaffalon. Notes on desirability and conditional lower previsions. *Annals of Mathematics and Artificial Intelligence*, 60(3-4):251–309, 2010.
- R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- T. Seidenfeld. Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. 2000.
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Decisions without ordering. In *Acting and reflecting*, pages 143–170. Springer, 1990.
- I. Singer. Generalized convexity, functional hulls and applications to conjugate duality in optimization. In *Selected topics in operations research and mathematical economics*, volume 226 of *Lecture Notes in Econ. and Math. Systems*, pages 49–79. Springer, 1984.
- A. Van Camp, E. Miranda, and G. De Cooman. Lexicographic choice functions without archimedeanicity. In *Soft Methods for Data Science*, pages 479–486. Springer, 2017.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. 1991.
- P. M. Williams. Notes on conditional previsions. Technical report, University of Sussex, 1975.

Modeling Markov Decision Processes with Imprecise Probabilities Using Probabilistic Logic Programming

Thiago P. Bueno

Denis D. Mauá

Leliane N. de Barros

Instituto de Matemática e Estatística - Universidade de São Paulo

Rua do Matão, 1010

São Paulo, S (Brazil)

TBUENO@IME.USP.BR

DDM@IME.USP.BR

LELIANE@IME.USP.BR

Fabio G. Cozman

FGCOZMAN@USP.BR

Escola Politécnica - Universidade de São Paulo

Av. Prof. Mello Moraes, 2231

São Paulo, P (Brazil)

Abstract

We study languages that specify Markov Decision Processes with Imprecise Probabilities (MDPIPs) by mixing probabilities and logic programming. We propose a novel language that can capture MDPIPs and Markov Decision Processes with Set-valued Transitions (MDPSTs); we then obtain the complexity of one-step inference for the resulting MDPIPs and MDPSTs. We also present results of independent interest on the complexity of inference with probabilistic logic programs containing interval-valued probabilistic assessments. Finally, we also discuss policy generation techniques.

Keywords: Markov decision processes; MDP; MDPIP; MDPST; imprecise probabilities; non-determinism; probabilistic logic programming; credal semantics.

1. Introduction

To be able to plan, one must be able to represent the relation between actions and their consequences on the world. Operator-based languages such as STRIPS or PDDL ([Fikes and Nilsson, 1971](#); [McDermott et al., 1998](#)) have been devised so as to encode *deterministic* sequential decision problems, with a specific solution in mind (heuristic search). Action languages such as \mathcal{A} or \mathcal{C} ([Giunchiglia and Lifschitz, 1998](#)), as well as programming languages such as GOLOG ([Levesque et al., 1997](#)), add more expressiveness, but also focus primarily on deterministic problems. Other languages focus on decision under uncertainty; for instance, PPDDL ([Younes and Littman, 2004](#)), RDDL ([Sanner, 2010](#)), DT-GOLOG ([Boutilier et al., 2000](#)). In particular, languages based on probabilistic logic programming ([Kersting and De Raedt, 2003](#); [Nitti et al., 2015](#); [Srivastava et al., 2014](#); [Bueno et al., 2016](#)) allow for probabilities, while \mathcal{C}^+ ([Giunchiglia et al., 2004](#)) and \mathcal{K} ([Eiter et al., 2004](#)) allow for nondeterminism. There are languages that even allow both probabilities and nondeterminism ([Halpern and Tuttle, 1993](#); [Eiter and Lukasiewicz, 2003](#); [Trevizan et al., 2008](#); [Iocchi et al., 2009](#)).

In this paper, we study the properties of planning domain description languages that have enough power so as to encode Markov Decision Processes with Imprecise Probabilities (MDPIPs) ([White III and Eldeib, 1994](#); [Delgado et al., 2009, 2011](#)). We propose a novel language based on probabilistic logic programming, enhanced with decision theoretic constructs such as actions, state fluents and

utilities. We consider interval-valued probabilities attached to independent facts, and we adopt a semantics given by Lukasiewicz (2007) within the context of probabilistic description logics. The semantics assigns probability measures over answer sets (Gelfond and Lifschitz, 1988). As has been recently noted by Cozman and Mauá (2016), this semantics induces an infinitely-monotone Choquet capacity on interpretations of atoms. We show that our language can be used to specify Markov Decision Processes with Set-valued Transitions (MDPSTs) when all probabilities are point-valued. This class of MDPIPs encompass a wide spectrum of planning tasks ranging from the classical, deterministic case to the probabilistic setting in which actions have stochastic and/or uncertain effects (Trevizan et al., 2007, 2008). We derive the complexity of one-step inference with the resulting languages; we also present results of independent interest on the complexity of inference with probabilistic logic programs containing interval-valued probabilistic assessments. We also discuss how to generate optimal policies from a specification in our language, in this paper focusing on MDPSTs.

The paper is organized as follows. We offer some background knowledge on MDPIPs and MDPSTs, and on probabilistic logic programming, in Section 2. We then present our language in Section 3. We discuss the complexity of one-step inference in Section 4, and describe policy generation algorithms in Section 5. Finally, Section 6 concludes the paper.

2. Background

In this section we review the main concepts behind Markov Decision Processes and some of their variants. We also summarize the main ideas in probabilistic logic programming.

2.1 MDPs, MDPIPs and MDPSTs

Markov Decision Processes (MDP) represent a class of sequential decision-making problems in a stochastic environment (Puterman, 2014). Intuitively, a planning agent has to deliberate over his/her model of the world to choose an optimal action in each decision stage in order to maximize his/her accumulated reward (or minimize the accumulated cost) given the immediate and long-term uncertain effects of available actions.

Formally, an MDP consists of (i) a finite set of *states* \mathcal{S} ; (ii) a finite set of applicable *actions* $\mathcal{A}(s)$ for each state s ; (iii) a Markovian *transition model* $\mathcal{T}(s, a, s') = \mathbb{P}(s'|s, a)$ specifying the probability that after executing action a in state s the next state is s' ; (iv) a *reward model* $\mathcal{R}(s, a, s')$ specifying the reward (or cost) of executing action a in state s and transitioning to state s' ; and (v) a set of decision stages $D = 1, \dots, H$. The solution of an MDP with infinite horizon (i.e., $H \rightarrow \infty$) is a stationary, deterministic *optimal policy* $\pi^* : \mathcal{S} \rightarrow \mathcal{A}(s)$ that prescribes an optimal action a in state s in order to maximize the expected cumulative reward of state s defined by the *optimal value function* $V^* : \mathcal{S} \rightarrow \mathbb{R}$ given by:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \left\{ \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) (\mathcal{R}(s, a, s') + \gamma V^*(s')) \right\}, \quad (1)$$

where $\gamma \in [0, 1[$ is the discount factor necessary for convergence.

There are situations in which it is not easy (or even possible) to define a precise probability measure for a given transition. In this case, it is necessary to consider a more general version of an MDP known as **Markov Decision Processes with Imprecise Probabilities** (MDPIP) (White III and Eldeib, 1994; Satia and Lave Jr, 1973). In this model, the probability parameters are imprecise

and therefore the transition model cannot be specified by a single conditional distribution, but it must be defined by *sets of probabilities* for each state transition. These sets are commonly referred to as *transition credal sets* $\mathcal{K}(\cdot|s, a)$ (Delgado et al., 2009). All other components of the MDP are unchanged (i.e., finite state and action space, reward function).

There are several objective criteria for solving an MDPIP with infinite horizon. In this paper, we only consider the Γ -maximin criterion (Delgado et al., 2009) which selects a robust policy that yields the supremum of the lower expected reward. The optimal value function of state s is:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \left\{ \min_{\mathbb{P}(\cdot|s, a) \in \mathcal{K}(\cdot|s, a)} \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) (\mathcal{R}(s, a, s') + \gamma V^*(s')) \right\}. \quad (2)$$

Finally, another interesting variant model of MDP is the **Markov Decision Process with Set-valued Transition** (MDPST). This model is a particular instance of an MDPIP aimed at representing the transition model of an MDP with (separate) components for probabilistic and non-deterministic action effects (Trevizan et al., 2007, 2008). In an MDPST, the transition model is defined by the probability mass function $m(k|s, a)$ and the non-deterministic function $F(s, a) \subseteq 2^{\mathcal{S}}$, such that $k \in F(s, a)$. Its semantics is that after applying action a to state s the probability that the next state s' is in the *reachable set* $k \in F(s, a)$ is given by $m(k|s, a)$. These components together induce the imprecise probabilities over next states constrained by the following set of inequalities:

$$0 \leq m(\{s'\}|s, a) \leq \mathbb{P}(s'|s, a) \leq \sum_{k \in F(s, a) \text{ s.t. } s' \in k} m(k|s, a) \leq 1, \quad (3)$$

$$0 \leq \sum_{s' \in \mathcal{D}(k, s, a)} \mathbb{P}(s'|s, a) \leq m(k|s, a) \leq \sum_{s' \in k} \mathbb{P}(s'|s, a) \leq 1, \quad (4)$$

where $\mathcal{D}(k, s, a) = k - \bigcup_{k' \in F(s, a), k' \neq k} k'$.

Inequalities 3 and 4 define a transition credal set $\mathcal{K}(\cdot|s, a)$ as demonstrated by Trevizan et al. (2007) therefore proving that an MDPST is indeed an MDPIP. Though, the contrary does not necessarily holds since the class of MDPIPs is much more general than that of MDPSTs.

The solution of an MDPST under the minimax criteria is an optimal policy with respect to the optimal value function, given by:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \left\{ \sum_{k \in F(s, a)} m(k|s, a) \min_{s' \in k} (\mathcal{R}(s, a, s') + \gamma V^*(s')) \right\}. \quad (5)$$

Throughout the paper we assume a factored representation of the state in which a state s is given by a set of state fluents $\{x_1, \dots, x_n\}$ which are state properties whose truth value changes with the actions; the factored transition function is $\mathbb{P}(s'|s, a) = \prod_{i=1}^n \mathbb{P}(x'_i|x_1, \dots, x_n, a)$ and the reward function is also factored. This representation implies a dynamic Bayesian network in which next-state fluents are independent given the current-state fluents and action.

2.2 Probabilistic Logic Programming and the Credal Semantics

Probabilistic Logic Programming (PLP) extends Logic Programming (LP) by assigning *probability measures* to logical facts. It is typically assumed a fixed vocabulary of constants and relations. An atom is a predicate $r(t_1, \dots, t_n)$ representing a n-arity relation over terms t_1, \dots, t_n where a term is

either a logical variable or a constant from the vocabulary. We use lowercase to denote constants and uppercase to denote variables. A ground atom is an atom with no variables as one of its terms.

A **probabilistic logic program** is a pair $L_p = \langle \mathbf{BK}, \mathbf{PF} \rangle$ consisting of a set of logical rules **BK** called *background knowledge* and a set of probabilistic facts **PF**. A logical rule is of the form $h :- b_1, \dots, b_m, \text{not } b_{m+1}, \dots, \text{not } b_n$, where atom h is called the *head* and the atoms $b_i, i = 1, \dots, n$ are called the *body*. The reserved symbol *not* is to be interpreted as *negation as failure*, i.e., *not* b_i is *true* in the absence of information that justifies b_i being *true*. A probabilistic fact denoted by $\alpha :: f$ is an atom f annotated with probability $\alpha \in [0, 1]$. All probabilistic facts are probabilistically independent and cannot be unified with any rule's head atom.

A *total choice* denoted by θ is a complete truth assignment to the probabilistic facts of L_p . Each total choice θ induces a logical program denoted by L^θ containing the background knowledge of L_p and only the facts with a *true* value in θ . This semantics defines each probabilistic fact $\alpha_i :: f_i$ as a boolean random variable f_i distributed accordingly to the Bernoulli distribution with mean α_i . Since the probabilistic facts $\alpha_i :: f_i$ are independent, the probability of the induced logic program L^θ is given by:

$$\mathbb{P}(L^\theta | L_p) = \prod_{f_i \in \theta} \alpha_i \prod_{f_i \notin \theta} (1 - \alpha_i). \quad (6)$$

The semantics of a probabilistic logic program L_p is given by the set of all probability models of L_p , accordingly to its **credal semantics** (Lukasiewicz, 2007; Cozman and Mauá, 2016). A *probability model* for a program L_p is a probability measure \mathbb{P} over logical interpretations of its atoms such that (i) every interpretation I with $\mathbb{P}(I) > 0$ is a stable model of the induced program L^θ for the total choice θ that is consistent with I on the set **PF**; and (ii) the probability of the induced program L^θ is given by Equation 6. If the probabilistic logic program L_p is acyclic or stratified (Lloyd, 2012) then the *credal set* for program L_p consists of a single probability model related to its unique stable model.

An interpretation I over the set of atoms of a logic program L is a **stable model** if and only if I is the minimal model of the reduct program L^I . The reduct program L^I is the set of positive rules $\{H(r) :- B^+(r) \mid r \in L \text{ and } B^-(r) \cap I = \emptyset\}$ where $H(r)$ is the head of rule r ; $B^+(r)$ and $B^-(r)$ are the sets of positive and negative atoms in the body of rule r . A typical logic program with more than one stable model is the non-stratified program $L = \{p :- \text{not } q. q :- \text{not } p.\}$ which has two stable models, namely the set of models $\{p\}, \{q\}$.

Given a probabilistic logic program L_p whose credal semantics is given by the credal set \mathcal{K}_{L_p} , the inference tasks of computing the *lower conditional probability* of query **Q** given evidence **E** denoted by $\underline{\mathbb{P}}(\mathbf{Q}|\mathbf{E})$ and respectively the *upper conditional probability* denoted by $\bar{\mathbb{P}}(\mathbf{Q}|\mathbf{E})$ are given by:

$$\underline{\mathbb{P}}(\mathbf{Q}|\mathbf{E}) = \inf_{\mathbb{P} \in \mathcal{K}_{L_p}} \mathbb{P}(\mathbf{Q}|\mathbf{E}) \quad (7)$$

$$\bar{\mathbb{P}}(\mathbf{Q}|\mathbf{E}) = \sup_{\mathbb{P} \in \mathcal{K}_{L_p}} \mathbb{P}(\mathbf{Q}|\mathbf{E}) \quad (8)$$

where **Q** and **E** are consistent sets of literals and it is assumed that $\mathbb{P}(\mathbf{E}) > 0$.

3. A Language to Specify MDPIPs and MDPSTs

One can specify an MDP through a probabilistic logic program, by annotating atoms with special meanings so as to distinguish actions, state fluents and rewards. This has been, for instance, the

approach taken by LOMDP (Kersting and De Raedt, 2003), DTBLOG (Srivastava et al., 2014), and DDC (Nitti et al., 2015). In a previous work, we devised the MDP-PROBLOG specification language for sequential decision problems based on the PROBLOG language (Bueno et al., 2016). Here we extend the language so as incorporate incomplete and imprecise assessments.

An **MDP-PROBLOG program** consists of three parts: a program L_{MDP}^{SPACE} declaring state fluents and actions, a program $L_{MDP}^{TRANSITION}$ encoding a transition model, and a program L_{MDP}^{REWARD} encoding the reward model.

The *dependency graph* of an MDP-PROBLOG program is the signed directed graph over the ground atoms of the program; there is a positive (resp., negative) arc $B \rightarrow A$ if there is a rule with B in the body and A in the head, and B is non-negated (resp., negated). In our previous work, we showed that MDP-PROBLOG programs with acyclic dependency graphs represents a factored MDP, whose transition model for each action is a dynamic Bayesian network: each ground atom is a variable; probabilistic facts are root nodes associated with corresponding probabilities and non-probabilistic facts are internal nodes associated with deterministic functions. We also showed that MDP-PROBLOG with *positive cycles* in its dependency graph still represent factored MDP (note that dynamic Bayesian networks do not allow cycles). We did not define the semantics of programs with cycles; we close this gap here.

We use the following running example to illustrate concepts:

Example 1 *In the Viral Marketing (VM) domain, we are given an information about individuals and their trust relationships, and we are interested in selecting individuals to market a certain product. The goal is to maximize the long-term profit by increasing the likelihood of sales while decreasing the cost of marketing. We assume that a person might buy the product after being marketed or because she trusts someone who already bought it. Also, if a person has not been the target of a marketing action in the current step, but she has been marketed in the past, then the delayed effect of past marketing actions should be accounted for.*

The program L_{MDP}^{SPACE} consists of (invariant) facts and two types of rules: *state fluent declarations* and *action fluent declarations*. State fluent declarations are of the form $\text{state_fluent}(A) :- B_1, \dots, B_n.$, where A is an atom representing a state fluent and B_1, \dots, B_n are literals mentioning action fluents (actions that may or may not occur) or non state fluents (state properties whose truth value does not change, i.e. invariants). Action fluent declarations are of the form $\text{action_fluent}(A) :- B_1, \dots, B_n.$, where A is an atom representing an action and B_1, \dots, B_n are as before. The state fluents are distinguished between *current state* and *next state*. Current-state fluents take an extra argument 0 to indicate the current stage, while next-state fluents take an extra argument 1 to indicate the next stage.

Consider our running example. We declare individuals by a set of (invariant) ground facts $\text{person}(p_i)$, a state and action fluents by:

```

state_fluent(market(P)) :- person(P).
state_fluent(buys(P)) :- person(P).
action_fluent(market(P)) :- person(P).
    
```

Given persons p_1 and p_2 , we have 4 state fluents: $\text{marketed}(p_1)$, $\text{marketed}(p_2)$, $\text{buys}(p_1)$ and $\text{buys}(p_2)$. Thus, the program above defines 2^4 states. For example, we have a state where $\text{marketed}(p_1)$

is true, and all of $\text{marketed}(p_2)$, $\text{buys}(p_1)$ and $\text{buys}(p_2)$ are false. Similarly, we have 2 actions fluents: $\text{market}(p_1)$ and $\text{market}(p_2)$. Thus, the program defines 2^2 actions. For example, we have an action where $\text{market}(p_1)$ is true and $\text{market}(p_2)$ is true¹.

The program $L_{\text{MDP}}^{\text{TRANSITION}}$ contains a set of rules such that no action fluents nor current-state fluents unify with head atoms.

The transition model of our running example is given by the program:

```

0.5 :: decay(Person).
marketed(Person, 1) :- market(Person).
marketed(Person, 1) :- not market(Person), marketed(Person, 0), decay(Person).

0.2 :: buy_from_marketing(Person).
0.3 :: buy_from_trust(Person).
buys(Person, 1) :- marketed(Person, 1), buy_from_marketing(Person).
buys(Person, 1) :- trusts(Person, Person2), buys(Person2, 1), buy_from_trust(Person).

```

According to this program, an individual is under the effect of a marketing action if she has either been targeted in the current stage, or, with probability 0.5, if she was under the effect in a previous stage. There is also the idea that a person buys the product with a certain probability if she has been the target of marketing, and with a different probability if some of her trustees was the target of marketing.

The transition program induces a transition credal set $\mathcal{K}(s'|s, a)$, where s is an interpretation of current-state fluents, a is an interpretation of action fluents and s' is an interpretation of next-state fluents. Each conditional distribution in the transition credal set specifies a transition model $\mathcal{T}(s, a, s')$ assigned with probability $\mathbb{P}(s'|s, a)$ given by the credal semantics of the program.

The program $L_{\text{MDP}}^{\text{REWARD}}$ contains a set of rules of the form $\text{utility}(A, c) :- B_1, \dots, B_n$, where A is state or action fluent, c is a value denoting reward/cost, and each B_i is a literal.

In our running example, every product bought contributes with a reward of 5, and every marketing action costs -1:

```

utility(buys(Person, 1), 5).
utility(market(Person), -1).

```

Finally, the program $L_{\text{MDP}}^{\text{REWARD}}$ specifies an additive reward model $\mathcal{R}(s, a, s')$ over current states (interpretation of current-state fluents), actions (interpretation of action fluents) and next states (interpretations of next-state fluents). A rule $\text{utility}(A, c) :- B_1, \dots, B_n$ contributes with (additive) reward c if and only if A, B_1, \dots, B_n are all true in the interpretation.

Since we adopted the credal semantics for the transition program, the transition credal set is the dominating credal set of an infinitely monotone Choquet capacity (Cozman and Mauá, 2016); that is, each transition is governed by a probabilistic transition into a reachable set that consists of the stable models. To get some intuition on this result, consider that for each fixed total choice, we obtain a logic program that may have more than one stable model (if it has no stable model, the whole probabilistic logic program has no semantics). And recall that over the total choices we have

1. Note that the semantics of the probabilistic logic programming allows concurrent actions just like in RDDL (Sanner, 2010).

a product measure. Hence we have a multi-valued mapping from one sample space endowed with a probability measure (the space of total choices) into another space (the space of stable models); this implies that over the latter space we have an infinitely monotone Choquet capacity (Augustin et al., 2014). Thus we have the following surprising (and pleasant) consequence:

Theorem 1 *An MDP-PROBLOG program specifies a factored MDPST.*

Although an MDPST is a particular case of MDPIP, so far we have assumed that every probability value is known with absolute precision. This is obviously unrealistic in practice. The natural solution then is to allow a fact to be associated with a probability interval. We denote these extended probabilistic facts by $[\alpha, \beta] :: p$, where p is an atom and parameters α and β are probability bounds such as $0 \leq \alpha \leq \beta \leq 1$. In the case of $\alpha = \beta$, we have a standard probabilistic fact. The semantics of a probabilistic logic program with interval-valued facts is the credal set that consists of all probability distributions that satisfy the constraints (that is, whose marginal probabilities for facts lies within given intervals).

For example, in the viral marketing domain, we might be uncertain about the probabilities that an individual will buy a product given different scenarios:

$$\begin{aligned} [0.1, 0.3] &:: \text{buy_from_marketing}(Person). \\ [0.2, 0.4] &:: \text{buy_from_trust}(Person). \end{aligned}$$

Now suppose we have an MDP-PROBLOG program, possibly with interval-valued probabilistic facts and negative cycles². Suppose also the current state S_0 is given, and possibly an additional set of grounded atoms E on the current time step; finally suppose we have a set of grounded atoms Q of next state, and we wish to compute $\bar{\mathbb{P}}(Q|E, S_0)$. By using arguments that apply to inference in credal networks, we have that the value of $\bar{\mathbb{P}}(Q|E, S_0)$ is attained at a selection of extreme points of the probability intervals, together with a selection of reachable set for all resulting probabilities (Augustin et al., 2014). That is, to compute an upper probability, we must go through all extreme points of probability intervals, and all possible extreme points of the induced infinitely monotone Choquet capacities. The same result obtains for the computation of lower probabilities. We will use these results in Section 5.

4. The Complexity of One-Step Inference

In this section we will need a number of concepts from complexity theory; most of them are standard: we use *languages*, *decision problems*, *many-one reductions*, and *complexity classes* such as P and NP (Papadimitriou, 2003). The complexity class PP consists of those languages \mathcal{L} such that: there is a polynomial time nondeterministic Turing machine M such that $\ell \in \mathcal{L}$ if and only if more than half of the computations of M on input ℓ end up accepting). We consider oracle machines and complexity classes such as Σ_i^P , recursively defined as $\Sigma_i^P = \text{NP}^{\Sigma_{i-1}^P}$ with $\Sigma_0^P = P$. We also use classes from Wagner's *polynomial counting hierarchy*: that is, the smallest set of classes containing P and, recursively, for any class C in the polynomial counting hierarchy, the classes PP^C , NP^C , and coNP^C (Torán, 1991; Wagner, 1986).

2. One could suppose that an MDP-PROBLOG program with interval-valued probabilities defines a BMDP (Givan et al., 1997), however in our language the imprecision is over state fluents while in BMDPs the imprecision is over states.

We are interested here in the complexity of *one-step inference*; that is, if we have the state at time t , then what is the computational cost of computing the probability of $\{X_{t+1} = x\}$? We start by analyzing a problem of independent interest: the complexity of inferences in probabilistic logic programs with interval-valued probabilistic facts (Section 4.1) and then we look at the complexity of one-step inference (Section 4.2).

4.1 Credal logic programs with interval-valued probabilistic facts

Suppose we have a credal logic program, possibly disjunctive and non-stratified, but not necessarily aimed at modeling planning scenarios. That is, we just have a disjunctive logic program associated with a number of interval-valued probabilistic facts. The only restriction we impose is that there is a bound on predicate arity. Suppose that additionally we have, as *input*, a set \mathbf{Q} of truth assignments to grounded atoms, and another set \mathbf{E} of truth assignments to grounded atoms; additionally we have a rational number γ in $[0, 1]$. We refer to (\mathbf{Q}, \mathbf{E}) as the *query*, and to \mathbf{E} as the *evidence*. As *output* we have the decision as to whether $\bar{\mathbb{P}}(\mathbf{Q}|\mathbf{E}) > \gamma$ where the probabilities are computed with respect to the input credal logic program. Consider the strings describing a credal logic program, a query, and a rational, and denote by \mathcal{C} the language consisting of all such strings that satisfy $\bar{\mathbb{P}}(\mathbf{Q}|\mathbf{E}) > \gamma$. Note that if we restrict our programs to be non-disjunctive and acyclic, then they specify credal networks (Cozman, 2005), and therefore deciding \mathcal{C} is at least a NP^{PP} -hard problem. It is remarkable that we can also decide \mathcal{C} in NP^{PP} ; that is:

Theorem 2 *Deciding whether a string is in \mathcal{C} is a NP^{PP} -complete problem.*

Proof Hardness follows, as already noted, from the fact that inference with credal networks is NP^{PP} -complete (De Campos and Cozman, 2005). Membership is a consequence of the following construction. First, guess the extreme point of each interval-valued probability assessment (this requires a nondeterministic Turing machine, but given that predicate arity is bounded, there is a polynomial number of guesses to be made). Then call, as an oracle, a counting Turing machine that guesses the truth assignment for all grounded probabilistic facts; by counting the number of such assignments that leads to satisfaction of \mathbf{Q} and \mathbf{E} , we can decide whether the base nondeterministic choice satisfies or not the inequality of interest. The problem is that, for each selected truth assignment for ground probabilistic facts, we must decide whether it is possible to satisfy the query; for a disjunctive logic program this can be made using a Σ_3^P oracle. That is, our problem can be solved in $\text{NP}^{\text{PP}^{\Sigma_3^P}}$. However, due to a remarkable result by Toda and Watanabe (Toda and Watanabe, 1992), we have that $\text{P}^{\text{PP}^{\Sigma_k^P}} = \text{P}^{\text{PP}}$; consequently, our decision problem is in NP^{PP} and the proof is finished. ■

4.2 One-step transitions

Now consider the specification of a planning problem using a credal logic program as described in Section 3. That is, we have a logic program with added interval-valued probabilistic facts. Denote by \mathcal{PC} the language that consists of strings encoding a credal logic program with a bound on predicate arity, a query, and a rational as in Section 4.1, but now the credal logic program is the description of a planning scenario as in Section 3, and with the following additional restrictions. The query must now refer only to grounded atoms at the next time step (not at current time step), and a string is

in the language if and only if $\bar{P}(\mathbf{Q}|\mathbf{E}, \mathbf{S}_0) > \gamma$ where \mathbf{S}_0 is the current state. That is, we focus on one-step, from current to next state, and we wish to compute an inference about the time step.

Using the result in the previous section we immediately have:

Theorem 3 *Deciding whether a string is in \mathcal{PC} is a NP^{PP} -complete problem.*

Proof Note that when we fix \mathbf{S}_0 , we obtain a decision problem for a credal probabilistic program. Then Theorem 2 implies the result. ■

Now suppose we restrict ourselves to point-valued probabilistic assessments; that is, every probabilistic facts is of the form $\alpha :: A..$. As discussed in Section 3, such assessments allow us to define MDPSTs when programs can be disjunctive/non-stratified. Now denote by \mathcal{PM} the language defined exactly as \mathcal{PC} , with the difference that every probabilistic assessment is point-valued. It is known that the complexity of inference in non-disjunctive probabilistic logic programs that can be non-stratified is $\text{PP}^{\Sigma_2^P}$ -complete, while the complexity of inference in disjunctive probabilistic logic programs is $\text{PP}^{\Sigma_3^P}$ -complete (Cozman and Mauá, 2017), submitted. Hence we obtain, as a direct consequence:

Theorem 4 *Deciding whether a string is in \mathcal{PM} is a $\text{NP}^{\Sigma_3^P}$ -complete problem.*

5. Dynamic Programming for MDP-PROBLOG programs

In this section, we discuss how dynamic programming can be applied to solve sequential decision problems specified by MDP-PROBLOG programs. To emphasize: we allow programs with (negative and positive) cycles in the dependency graph and interval-valued probabilistic facts.

For simplicity, we consider grounded programs. So consider a (ground) MDP-PROBLOG program, a current state s (i.e., an interpretation of state fluents) and action $a \in \mathcal{A}(s)$ (i.e., an interpretation of actions). Due to Theorem 1, given evidence s, a , the transition model induces a set of probability mass functions $m(k|s, a)$ over sets of stable models $k \in F(s, a)$. One can show that the robust (i.e., maximin) policy is given by the argument of the following modified Bellman equation:

$$V^*(s) = \max_{a \in \mathcal{A}(s)} \left\{ \min_{m(\cdot|s,a) \in \mathcal{K}(\cdot|s,a)} \sum_{k \in F(s,a)} m(k|s, a) \min_{s' \in k} (\mathcal{R}(s, a, s') + \gamma V^*(s')) \right\} \quad (9)$$

The outer (i.e., leftmost) minimization can be solved by considering all extremes of the interval-valued probabilities; after each choice is made, the resulting program specifies an MDPST whose transition is governed by the stable models of the transition program: this is the inner (rightmost) minimization in the equation above.

When all reachable sets are singletons (i.e., $\forall k \in F(s, a), |k| = 1$) there is no need to perform the inner minimization over the states in k and then we have the traditional case of MDPIPs given by Equation 2. On the other hand, if all interval-valued probabilistic facts degenerate to point-valued standard probabilistic facts the outer minimization over the probabilistic models of the credal set $\mathcal{K}(\cdot|s, a)$ is not need and then we have the Equation 5 for precise MDPSTs. Finally, when both assumptions hold we are back to the classical MDP case of Equation 1.

The traditional dynamic programming scheme for solving the set of equations defining the state value function is the **Value Iteration** algorithm (Puterman, 2014). Essentially, it assigns an initial

value to all states and iteratively updates all state values until the convergence by using Equation 9 as an update rule known as Bellman backup. A number of optimizations exist for avoiding redundant calculations and restricting the computation for only the most promising states regarding the optimal policy. Nevertheless, virtually all of these techniques has to deal one or more backup calculations.

6. Conclusion

In this paper, we addressed the problem of modeling MDPIPs and MDPSTs using probabilistic logic programming. Our contributions are:

- an extension of the MDPPROBLOG language that aimed at representing imprecise probabilities and non-determinism;
- novel results about the complexity of one-step inference in credal logic programs with interval-valued probabilistic facts (and on the complexity of probabilistic logic programs with interval-valued probabilistic facts); and
- a scheme for generating optimal policy for MDPIPs and MDPSTs encoded by probabilistic logic programming.

For the future, we plan to implement and test algorithms for policy generation. In order to do so, it would be valuable to maximize expected values with respect to the credal sets encoding transitions. Given that heuristics are important in state-of-art algorithms for MDPs, we believe that similar heuristics must be developed for MDPIPs and MDPSTs. In particular, it should be important to import techniques from logical reasoning into the realm of probabilistic logic programming.

Acknowledgments

This work was partially supported by CNPq (grants 870666/1998-3, 308433/2014-9) and FAPESP (grants 2015/01587-0, 2016/01055-1, 2016/22900-1).

References

- T. Augustin, F. P. Coolen, G. de Cooman, and M. C. Troffaes. *Introduction to imprecise probabilities*. 2014.
- C. Boutilier, R. Reiter, M. Soutchanski, and S. Thrun. Decision-Theoretic, High-Level Agent Programming in the Situation Calculus. In *AAAI/IAAI*, pages 355–362, 2000.
- T. Bueno, D. Mauá, L. N. de Barros, and F. Cozman. Markov Decision Processes Specified by Probabilistic Logic Programming: Representation and Solution. In *BRACIS*, pages 337–342, 2016.
- F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.
- F. G. Cozman and D. D. Mauá. The structure and complexity of credal semantics. In *Workshop on Probabilistic Logic Programming*, pages 3–14, 2016.

- F. G. Cozman and D. Mauá. The Complexity of Inferences and Explanations in Probabilistic Logic Programming. *Proceedings ISIPTA, Lugano, Switzerland*, 2017.
- C. P. De Campos and F. G. Cozman. The inferential complexity of bayesian and credal networks. In *IJCAI*, volume 5, pages 1313–1318, 2005.
- K. V. Delgado, L. N. de Barros, F. G. Cozman, and R. Shirota. Representing and solving factored Markov decision processes with imprecise probabilities. *ISIPTA*, pages 169–178, 2009.
- K. V. Delgado, S. Sanner, and L. N. De Barros. Efficient solutions to factored MDPs with imprecise transition probabilities. *Artificial Intelligence*, 175(9-10):1498–1527, 2011.
- T. Eiter and T. Lukasiewicz. Probabilistic Reasoning About Actions in Nonmonotonic Causal Theories. In *UAI*, pages 192–199, 2003.
- T. Eiter, W. Faber, N. Leone, G. Pfeifer, and A. Polleres. A Logic Programming Approach to Knowledge-state Planning: Semantics and Complexity. *ACM Trans. Comput. Logic*, 5(2):206–263, 2004.
- R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3-4):189–208, 1971.
- M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *ICLP/SLP*, volume 88, pages 1070–1080, 1988.
- E. Giunchiglia and V. Lifschitz. An Action Language based on Causal Explanation: Preliminary Report. In *AAAI/IAAI*, pages 623–630, 1998.
- E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153:49–104, 2004.
- R. Givan, S. Leach, and T. Dean. Bounded parameter Markov decision processes. In *European Conference on Planning*, pages 234–246, 1997.
- J. Y. Halpern and M. R. Tuttle. Knowledge, Probability, and Adversaries. *J. ACM*, 40(4):917–960, 1993.
- L. Iocchi, T. Lukasiewicz, D. Nardi, and R. Rosati. Reasoning About Actions with Sensing Under Qualitative and Probabilistic Uncertainty. *ACM Trans. Comput. Logic*, 10(1):5:1–5:41, 2009.
- K. Kersting and L. De Raedt. Logical Markov decision programs. In *Proceedings of the IJCAI'03 Workshop on Learning Statistical Models of Relational Data*, pages 63–70, 2003.
- H. Levesque, R. Reiter, Y. Lesperance, F. Lin, and R. Scherl. GOLOG: A logic programming language for dynamic domains. *Journal of Logic Programming*, 31(1-3):59–83, 1997.
- J. W. Lloyd. *Foundations of logic programming*. 2012.
- T. Lukasiewicz. Probabilistic description logic programs. *International Journal of Approximate Reasoning*, 45(2):288–307, 2007.

- D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins. PDDL—the Planning Domain Definition Language. 1998.
- D. Nitti, V. Belle, and L. De Raedt. Planning in discrete and continuous Markov decision processes by probabilistic programming. In *ML and KD in Databases*, pages 327–342. 2015.
- C. H. Papadimitriou. *Computational complexity*. 2003.
- M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. 2014.
- S. Sanner. Relational dynamic influence diagram language (RDDL): Language description. *Unpublished ms. Australian National University*, page 32, 2010.
- J. K. Satia and R. E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.
- S. Srivastava, S. J. Russell, P. Ruan, and X. Cheng. First-Order Open-Universe POMDPs. In *UAI*, pages 742–751, 2014.
- S. Toda and O. Watanabe. Polynomial-time 1-turing reductions from #PH to #P. *Theoretical Computer Science*, 100(1):205–221, 1992.
- J. Torán. Complexity classes defined by counting quantifiers. *Journal of the ACM (JACM)*, 38(3):752–773, 1991.
- F. W. Trevizan, F. G. Cozman, and L. N. de Barros. Planning under Risk and Knightian Uncertainty. In *IJCAI*, pages 2023–2028, 2007.
- F. W. Trevizan, F. G. Cozman, and L. N. De Barros. Mixed probabilistic and nondeterministic factored planning through Markov decision processes with set-valued transitions. In *Workshop on A Reality Check for Planning and Scheduling Under Uncertainty at ICAPS*, 2008.
- K. W. Wagner. The complexity of combinatorial problems with succinct input representation. *Acta informatica*, 23(3):325–356, 1986.
- C. C. White III and H. K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.
- H. L. Younes and M. L. Littman. PPDDL1. 0: An extension to PDDL for expressing planning domains with probabilistic effects. *Techn. Rep. CMU-CS-04-162*, 2004.

Empirical Interpretation of Imprecise Probabilities

Marco E. G. V. Cattaneo

M.CATTANEO@HULL.AC.UK

University of Hull

Kingston upon Hull (United Kingdom)

Abstract

This paper investigates the possibility of a frequentist interpretation of imprecise probabilities, by generalizing the approach of Bernoulli's *Ars Conjectandi*. That is, by studying, in the case of games of chance, under which assumptions imprecise probabilities can be satisfactorily estimated from data. In fact, estimability on the basis of finite amounts of data is a necessary condition for imprecise probabilities in order to have a clear empirical meaning. Unfortunately, imprecise probabilities can be estimated arbitrarily well from data only in very limited settings.

Keywords: imprecise probabilities; frequentist interpretation; empirical meaning; bag of marbles; strong estimability; consistent estimators; empirical recognizability.

1. Introduction

Imprecise probabilities mostly have a subjective, epistemic interpretation (Walley, 1991; Troffaes and de Cooman, 2014), while in this paper we will study the possibility of a frequentist, empirical interpretation for them. As regards precise probabilities, empirical interpretations are dominant in science and statistics. They are usually related to Bernoulli's law of large numbers, which connects the probabilities of events with the relative frequencies of the events' occurrence in sequences of independent repetitions of experiments.

This connection can be used asymptotically, by defining probabilities as limits of relative frequencies (Venn, 1866; von Mises, 1928, 1957; Reichenbach, 1935, 1949), but the empirical meaning of such probabilities for finite samples is then problematic. In order to have probabilities with a direct empirical meaning, the connection in Bernoulli's law of large numbers can be used in a finite-sample way, by defining probabilities as approximately equal to relative frequencies in large, but finite samples. The difficulty of this approach comes from the fact that the exact meaning of "approximately equal" is probabilistic, and therefore this definition of probability is circular.

A possible answer to this circularity consists in accepting it and interpreting probability as an abstract concept, whose meaning comes from the possibility of statistically falsifying probabilistic statements (Popper, 1935, 1959). An alternative, but related answer to the above circularity is the original approach of Bernoulli (1713, 2006): define probability only for games of chance (where the definition is unproblematic) and extend it to other fields by analogy. This analogy is empirically meaningful because Bernoulli's law of large numbers provides a way of estimating probabilities arbitrarily well (and thus also a way of statistically falsifying probabilistic statements).

In this paper, we will see if Bernoulli's approach can be extended to imprecise probabilities. That is, practically we will focus on games of chance: for example drawing colored marbles at random from a bag. In this situation, the precise probability of a certain color corresponds to the proportion of marbles of this color in the bag, and if we draw several marbles (with replacement) from the bag, we obtain probabilistic independence automatically from the noninteraction of the

drawings. How should we interpret an imprecise probability in this setting? We will see that several different interpretations may be reasonable.

The spectacular achievement of Bernoulli was to prove, through his law of large numbers, that precise probabilities are estimable from finite amounts of data, and therefore have an empirical meaning. Analogously, a frequentist, empirical interpretation of imprecise probabilities is possible only if these are estimable from finite amounts of data. The core of the present paper consists of mathematical results about the estimability of imprecise probabilities, depending on their exact interpretation in the case of games of chance. These results are given in Section 3 (due to space limitations, proofs are omitted, and will appear only in an extended version of the paper), while the next section provides a quick overview of frequentist interpretations, and the last section concludes the paper and points to an open problem.

2. Interpretations of Imprecise Probabilities

The interpretations of (precise) probabilities can be roughly grouped in two main classes, often called subjective and frequentist (see for example [Gillies, 2000](#)). With a subjective (or epistemic, Bayesian, personalistic, ...) interpretation, probabilistic statements are about the degrees of belief or knowledge of an individual. By contrast, with a frequentist (or empirical, objective, scientific, ...) interpretation, probabilistic statements are about the material world. For this reason, frequentist interpretations of probabilities are the dominant interpretations in science and in statistics.

In particular, according to the subjective interpretation of [de Finetti \(1931, 1974–1975\)](#), a probability is an individual's fair price for a bet. This interpretation can quite naturally be extended to an interpretation of lower and upper probabilities as an individual's maximum buying price and minimum selling price for a bet ([Williams, 1975, 2007; Walley, 1991; Troffaes and de Cooman, 2014](#)). In fact, it can certainly be argued that with this subjective interpretation, imprecise probabilities are more natural than precise ones. However, the topic of the present paper is frequentist interpretations for imprecise probabilities, which, contrary to what happens for precise probabilities, are far less common than subjective ones.

Since usual imprecise probability measures correspond mathematically to sets of precise ones, they appear often in classical statistics, which is based on frequentist interpretations of probabilities. In particular, imprecise probabilities can be used to describe what has been learnt so far from data (see for example [Cattaneo and Wiencierz, 2012; Antonucci et al., 2012](#)), but in this case their interpretation is in reality epistemic, although more properly intersubjective than subjective. However, a truly frequentist interpretation is indeed obtained in classical statistics when imprecise probabilities do not describe what has been learnt, but what can potentially be learnt from infinite amounts of incomplete data (see for instance [Manski, 2003; Dempster, 1967](#)). Anyway, this frequentist interpretation of imprecise probabilities is limited to particular situations involving incomplete data, while we are looking for a generally valid interpretation.

In general, frequentist interpretations of precise probabilities are related to laws of large numbers implying that the relative frequency of an event's occurrence in a sequence of independent repetitions of an experiment converges to the probability of the event. Although laws of large numbers have been generalized to the case of imprecise probabilities ([Walley and Fine, 1982; Cozman and Chrisman, 1997; Marinacci, 1999; de Cooman and Miranda, 2008; Peng, 2010; Chen and Wu, 2011](#)), the generalization of frequentist interpretations is not straightforward.

If we would simply interpret the probability of an event as the limit of the relative frequency of its occurrence in an infinite sequence of independent repetitions of an experiment ([Venn, 1866](#); [von Mises, 1928, 1957](#); [Reichenbach, 1935, 1949](#)), then we could interpret lower and upper probabilities as limits inferior and superior of such a sequence, respectively. That is, the imprecise probability interpretation would extend the precise one to the case of nonconvergent sequences of relative frequencies, and also with this frequentist interpretation (besides the above subjective one) it could be argued that imprecise probabilities are more natural than precise ones. However, this interpretation is problematic for imprecise as well as precise probabilities, since no finite part of a sequence of relative frequencies has any connection at all with the limit of the sequence, and thus strictly speaking the interpretation has no empirical meaning.

In order to have an empirical meaning, a frequentist interpretation must make probabilistic statements falsifiable on the basis of finite amounts of data. Of course, probabilistic statements are in general not strictly falsifiable, but they can be methodologically falsifiable in the sense of [Popper \(1935, 1959\)](#) if they can be rejected through some reasonable statistical test with arbitrarily low significance level (see also [Gillies, 1995, 2000](#)). Such a test for the probability of an event could be based on Bernoulli's law of large numbers, which is a probabilistic statement connecting the probability of the event to the relative frequency of the event's occurrence in a finite sequence of independent repetitions. That is, we could consider frequentist probability as an abstract concept deriving its meaning from the theory surrounding it, which makes probabilistic statements (methodologically) falsifiable.

However, in the present paper we will follow a related, but more direct approach to frequentist probability, corresponding to the original interpretation of Bernoulli's law of large numbers in the *Ars Conjectandi* ([Bernoulli, 1713, 2006](#)). This book represents the starting point of modern probability theory, and interestingly also the (temporary) end point of imprecise probability ([Shafer, 1978](#)). Citing [Sylla \(2014\)](#): “before Bernoulli's work, there existed a mathematics of games of chance but that mathematics did not involve probability—not the Latin word *probabilis*, not relative frequencies and not degrees of certainty.”

Bernoulli's law of large numbers is a theorem in the mathematics of games of chance. That is, a theorem about probabilities interpreted as ratios between the numbers of favorable and possible outcomes. Bernoulli extended the concept of probability to other fields by analogy with games of chance, an idea already present in the *Logique de Port-Royal* ([Arnauld and Nicole, 1662, 1996](#)). According to this approach, the probability of an event is interpreted through an analogy with a game of chance: for example as corresponding to the probability of drawing a black marble at random from a bag containing white and black marbles. Bernoulli's law of large numbers implies that it is possible to learn with arbitrarily high precision the probability of an event from the relative frequency of its occurrence in sufficiently many independent repetitions of an experiment.

3. Empirical Meaning of Imprecise Probabilities

The approach to frequentist probability of the *Ars Conjectandi* consists of two parts: the interpretation of probabilities by analogy with games of chance, and their estimability on the basis of finite amounts of data. In this section we will study how far this approach can be generalized to the case of imprecise probabilities. For the sake of simplicity, we will focus on a sequence of Bernoulli trials, whose outcomes are described by the binary random variables $X_1, X_2, \dots \in \{0, 1\}$ (that is,

we consider only the interpretation and estimability of imprecise probabilities of single events, not of whole imprecise probability measures on arbitrary sample spaces).

Each Bernoulli trial corresponds for instance to drawing a black or white marble (described by $X_i = 1$ or $X_i = 0$, respectively) at random from a bag containing only white and black marbles with a proportion $p_i \in [0, 1]$ of black ones (strictly speaking, all probabilities should be rational numbers, but for the sake of simplicity we will ignore this technical detail, since rational numbers are dense in the reals). The sequence of Bernoulli trials corresponds thus to drawings from a sequence of bags with possibly different proportions of black marbles. The noninteraction of the drawings corresponds to an assumption of independence of the random variables X_i , in the usual sense of (precise) probability theory (see also [Chen and Wu, 2011](#); [De Bock and de Cooman, 2012](#)). We have a precise probability model when $p_i = p$ does not depend on i , and an imprecise one when $p_i \in [\underline{p}, \bar{p}]$ is not completely determined.

For example, in Section 2 we have considered two ways in which imprecise probability measures often appear in classical statistics. The first one, related to an intersubjective epistemic interpretation, is as descriptions of what has been learnt so far from data: this would be the case for instance if $[\underline{p}, \bar{p}]$ was obtained as a confidence interval for the precise probability p . The second one, related to a truly frequentist but limited interpretation, is as descriptions of what can potentially be learnt from infinite amounts of incomplete data: this would be the case for instance if $X_i = 1$ and $X_i = 0$ were observed with probabilities \underline{p} and $1 - \bar{p}$, respectively, while with probability $\bar{p} - \underline{p}$ we would have a missing observation (independently of i). In this case, without making any assumptions about the noninformativity of the missing data, $[\underline{p}, \bar{p}]$ is the identification region of the precise probability p : that is, values of p in this interval cannot be discriminated on the basis of any amount of (incomplete) data.

In the general case without missing data, there are several possible interpretations of an imprecise probability $[\underline{p}, \bar{p}]$ with $0 \leq \underline{p} \leq \bar{p} \leq 1$ (where $\underline{p} = \bar{p}$ corresponds to the case of a degenerate interval representing a precise probability). In particular, [Walley and Fine \(1982\)](#) distinguish between an *ontological indeterminacy interpretation*, where

$$p_i \in [\underline{p}, \bar{p}] \quad (1)$$

is the only assumption about the sequence p_i , and an *epistemological indeterminacy interpretation*, where

$$p_i = p \in [\underline{p}, \bar{p}] \quad (2)$$

does not depend on i . The latter can also be seen as the special case in which the sequence of drawings (with replacement) is from the same bag, which contains a not completely determined proportion of black marbles. The interpretations (1) and (2) appear also in the theory of Markov chains with imprecise probabilities (which can be seen as generalizations of sequences of Bernoulli trials): for example in [Hartfiel \(1998\)](#) and [Kozine and Utkin \(2002\)](#), respectively. Moreover, the ontological indeterminacy interpretation (1) plays a prominent role in the theory of probabilistic graphical models with imprecise probabilities (which can be seen as further generalizations of Markov chains): see for instance [Cozman \(2005\)](#).

From the point of view of the estimability of the imprecise probability $[\underline{p}, \bar{p}]$, both interpretations (1) and (2) are problematic, because in general the sequence p_i does not determine the interval $[\underline{p}, \bar{p}]$. That is, with these interpretations the imprecise probability is only partially identified and therefore cannot in general be estimated with arbitrarily high precision. In order to make the imprecise probability identifiable, we can interpret it as allowing only the sequences $p_i \in [\underline{p}, \bar{p}]$ that determine in

a certain sense the interval $[\underline{p}, \bar{p}]$. However, we would most likely betray the intuitive meaning of imprecise probabilities if we would exclude any starting sequence $p_1, \dots, p_n \in [\underline{p}, \bar{p}]$. Similarly, assigning some kind of degree of plausibility to the starting sequences $p_1, \dots, p_n \in [\underline{p}, \bar{p}]$ would also lead to a new model, different from the one of imprecise probabilities (such as the chaotic probability model of [Fierens et al., 2009](#)).

On the basis of these considerations, we obtain an *identifiable ontological indeterminacy interpretation*, where

$$p_i \in [\underline{\alpha}(p_1, p_2, \dots), \bar{\alpha}(p_1, p_2, \dots)] = [\underline{p}, \bar{p}] \quad (3)$$

is a condition on the sequence p_i , determined by two functions $\underline{\alpha}, \bar{\alpha} : [0, 1]^{\mathbb{N}} \rightarrow [0, 1]$ that do not depend on any finite number of their arguments (that is, each function would assign the same value to sequences differing only at a finite number of positions). These functions are considered to be fixed, but we do not need to further specify them in order to obtain the results of the present paper (that is, these results are valid for any particular choice of the above functions $\underline{\alpha}, \bar{\alpha}$). An example of such pairs of functions is the limits inferior and superior of the sequence p_i , implying that the whole width of the interval

$$[\underline{p}, \bar{p}] = \left[\liminf_{i \rightarrow \infty} p_i, \limsup_{i \rightarrow \infty} p_i \right] \quad (4)$$

is used by the sequence p_i , and infinitely many times (in the sense that the sequence gets infinitely many times arbitrarily close to both endpoints of the interval). A related example is the limits inferior and superior of the Cesàro means of the sequence p_i (that is, the limits inferior and superior of the averages of the starting sequences p_1, \dots, p_n), implying that the whole width of the interval

$$[\underline{p}, \bar{p}] = \left[\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n p_i, \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n p_i \right] \quad (5)$$

is used by the sequence p_i , not only infinitely many times, but also not too rarely (in order to bring not only the sequence, but also its Cesàro means infinitely many times arbitrarily close to both endpoints of the interval).

However, it is intuitively clear that imprecise probabilities are not estimable in full generality, because for instance any finite amount of data from Bernoulli trials would always be perfectly compatible with the vacuous imprecise probability $[0, 1]$, independently of the considered interpretation (1), (2), or (3). This difficulty in discriminating between the vacuous and other imprecise probabilities is related to the more general difficulty in comparing imprecise probability models with different degrees of imprecision (see also [Seidenfeld et al., 2011](#); [Cattaneo, 2013](#)). Anyway, imprecise probabilities are estimable under additional assumptions about the possible intervals $[\underline{p}, \bar{p}]$. Let \mathcal{I} be the set of all imprecise probabilities that are considered possible in a given situation: that is, let \mathcal{I} be a nonempty set of intervals of the form $[\underline{p}, \bar{p}]$ with $0 \leq \underline{p} \leq \bar{p} \leq 1$.

Bernoulli's law of large numbers implies the *uniformly consistent estimability* of the precise probability $p_i = p \in [0, 1]$ on the basis of the outcomes X_1, X_2, \dots of the Bernoulli trials. An estimator $\underline{\pi}_n : \{0, 1\}^n \rightarrow [0, 1]$ (or more precisely, a sequence of estimators $\underline{\pi}_n$) of \underline{p} is said to be uniformly consistent when for all $\varepsilon > 0$ and all $\delta > 0$ there is an N such that

$$P(|\underline{\pi}_n(X_1, \dots, X_n) - \underline{p}| > \varepsilon) \leq \delta \quad (6)$$

for all $n \geq N$, all $[\underline{p}, \bar{p}] \in \mathcal{I}$, and all (precise) probability measures P corresponding to the sequences p_i compatible with the imprecise probability $[\underline{p}, \bar{p}]$ according to the considered interpretation (1), (2), or (3) (since we are in the setting of games of chance, the interpretation of P is

unproblematic: probabilities are ratios between the numbers of favorable and possible outcomes). The definition of a uniformly consistent estimator $\underline{\pi}_n$ of \underline{p} is analogue, and $[\underline{p}, \bar{p}] \in \mathcal{I}$ is said to be uniformly consistently estimable when there are uniformly consistent estimators $\underline{\pi}_n$ and $\bar{\pi}_n$ of \underline{p} and \bar{p} , respectively.

The uniform consistency of an estimator $\underline{\pi}_n$ of \underline{p} is particularly important, because it implies that $[\underline{\pi}_N(X_1, \dots, X_N) - \varepsilon, \underline{\pi}_N(X_1, \dots, X_N) + \varepsilon]$ is a confidence interval for \underline{p} with coverage probability at least $1 - \delta$ (an analogous result is implied by the uniform consistency of an estimator $\bar{\pi}_n$ of \bar{p}). That is, uniformly consistent estimators provide us with arbitrarily short confidence intervals of arbitrarily high confidence level, when we have a sufficiently large amount of data. In this sense, uniformly consistent estimability endows imprecise probabilities $[\underline{p}, \bar{p}] \in \mathcal{I}$ with a clear empirical meaning. However, the next theorem states that this is the case only when all nondegenerate intervals in \mathcal{I} are isolated in \mathcal{I} . An interval $[\underline{p}, \bar{p}] \in \mathcal{I}$ is said to be nondegenerate when $\underline{p} < \bar{p}$, and it is said to be isolated in \mathcal{I} when there is a $\gamma > 0$ such that $[\underline{p} - \gamma, \bar{p} + \gamma]$ does not intersect any other interval in \mathcal{I} (degenerate or nondegenerate). If all nondegenerate intervals in \mathcal{I} are isolated in \mathcal{I} , then all intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint, while the converse is not true. For example, if \mathcal{I} consists of the nondegenerate interval $[0, \frac{1}{2}]$ and all the degenerate intervals $[\underline{p}, \bar{p}]$ with $\frac{1}{2} < \underline{p} \leq 1$, then all elements of \mathcal{I} are pairwise disjoint, but $[0, \frac{1}{2}]$ is not isolated in \mathcal{I} .

Theorem 1 *The following four statements are equivalent:*

- (i) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is uniformly consistently estimable under the ontological indeterminacy interpretation (1),
- (ii) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is uniformly consistently estimable under the epistemological indeterminacy interpretation (2),
- (iii) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is uniformly consistently estimable under the identifiable ontological indeterminacy interpretation (3),
- (iv) all nondegenerate intervals in \mathcal{I} are isolated in \mathcal{I} .

Theorem 1 implies in particular Bernoulli's law of large numbers, which corresponds to the case where \mathcal{I} is the set of all degenerate intervals $[\underline{p}, \bar{p}]$ with $\underline{p} = \bar{p} \in [0, 1]$. More precisely, Bernoulli (1713, 2006) proved the result only in the case where \mathcal{I} is the set of the $m + 1$ degenerate intervals $[\underline{p}, \bar{p}]$ such that $\underline{p} \in [0, 1]$ is a rational number with (arbitrarily large) denominator m . For this case, he also provided an explicit way of calculating a value for the quantity N appearing in the definition of uniform consistency (6), thus obtaining a clear empirical meaning for precise probabilities through what we now call confidence intervals. Anyway, Theorem 1 shows that this is possible for imprecise probabilities only in very limited settings, independently of their exact interpretation.

In order to endow imprecise probabilities with a clear empirical meaning in more general settings, we can moderate our requirements for their estimability. In particular, Walley and Fine (1982) introduced the concept of *strong estimability*, which weakens uniformly consistent estimability (6) by allowing N to depend on the interval $[\underline{p}, \bar{p}]$, besides on ε and δ . When we weaken strong estimability further by allowing N to depend also on the probability measure P , we get the concept of *consistent estimability*. That is, strong estimability lies between consistent estimability and uniformly consistent estimability, and must not be confused with strongly consistent estimability (which corresponds to consistent estimability when convergence in probability is replaced by almost sure convergence). Anyway, strong estimability can also be interpreted as the generalization of consistent

estimability to imprecise probabilities: in fact, strong estimability and consistent estimability are equivalent when all intervals in \mathcal{I} are degenerate (that is, in the case of precise probabilities).

Theorem 2 *The following four statements are equivalent:*

- (i) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is strongly estimable under the ontological indeterminacy interpretation (1),
- (ii) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is strongly estimable under the epistemological indeterminacy interpretation (2),
- (iii) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is strongly estimable under the identifiable ontological indeterminacy interpretation (3),
- (iv) all intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint.

Contrary to uniformly consistent estimability, strong estimability does not guarantee the existence of arbitrarily short confidence intervals of arbitrarily high confidence level for \underline{p} and \bar{p} , but is nonetheless important because it is required in order for imprecise probabilities to be empirically recognizable, in the following sense. Given an imprecise probability $[\underline{p}, \bar{p}] \in \mathcal{I}$ and a desired level of precision for the estimators, we can choose n such that if the data X_1, \dots, X_n are generated according to $[\underline{p}, \bar{p}]$ (that is, according to any sequence p_i compatible with it), then $[\underline{p}, \bar{p}]$ can be estimated to the desired level of precision on the basis of X_1, \dots, X_n (in other words, an imprecise probability can be recognized arbitrarily well on the basis of finite amounts of data generated according to it). However, Theorem 2 shows that imprecise probabilities are empirically recognizable only in very limited settings, independently of their exact interpretation. In fact, requiring only strong estimability instead of uniformly consistent estimability as in Theorem 1 weakened only slightly the necessary and sufficient condition on \mathcal{I} . As a side result, the following corollary of Theorem 2 completes a basic result of [Walley and Fine \(1982\)](#) about the strong estimability of imprecise probability measures on finite sample spaces.

Corollary 3 *The necessary condition in Theorem 5.1 of [Walley and Fine \(1982\)](#) is sufficient as well, also in the case of infinitely many imprecise probability measures.*

Although consistent estimability (with respect to precise probability measures) is too weak to endow imprecise probabilities with a clear empirical meaning (in the sense that it does not guarantee their empirical recognizability), for completeness we can look at the consequences of requiring only this level of estimability. The next theorem shows that there is no difference between consistent estimability and strong estimability of imprecise probabilities, when only the interpretations (1) and (2) are considered. However, there is a difference when the identifiable ontological indeterminacy interpretation (3) is considered, as we will see in a moment.

Theorem 4 *The following three statements are equivalent:*

- (i) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is consistently estimable under the ontological indeterminacy interpretation (1),
- (ii) $[\underline{p}, \bar{p}] \in \mathcal{I}$ is consistently estimable under the epistemological indeterminacy interpretation (2),
- (iii) all intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint.

Theorems 1, 2, and 4 give necessary and sufficient conditions for the estimability of imprecise probabilities, but there is a difference between knowing that something is estimable and knowing how to estimate it. The next theorem closes this gap by explicitly giving examples of estimators with the required properties.

Theorem 5 *The following estimators of \underline{p} and \bar{p} satisfy all the properties considered in Theorems 1, 2, and 4, when the corresponding necessary and sufficient conditions on \mathcal{I} are fulfilled:*

$$\underline{\pi}_n(x_1, \dots, x_n) = \inf \left\{ \underline{p} : [\underline{p}, \bar{p}] \in \mathcal{I}, \bar{p} + c_n > \frac{1}{n} \sum_{i=1}^n x_i \right\}, \quad (7)$$

$$\bar{\pi}_n(x_1, \dots, x_n) = \sup \left\{ \bar{p} : [\underline{p}, \bar{p}] \in \mathcal{I}, \underline{p} - c_n < \frac{1}{n} \sum_{i=1}^n x_i \right\}, \quad (8)$$

for all $x_1, \dots, x_n \in \{0, 1\}$, where c_n is any sequence of real numbers such that $\lim_{n \rightarrow \infty} c_n = 0$ and $\lim_{n \rightarrow \infty} \sqrt{n} c_n = +\infty$, while $\inf \emptyset$ and $\sup \emptyset$ can be defined arbitrarily.

The estimators (7) and (8) exploit the fact that the relative frequency $\frac{1}{n} \sum_{i=1}^n X_i$ of the occurrence of the event $X_i = 1$ will lie in $[\underline{p} - c_n, \bar{p} + c_n]$ with arbitrarily high probability when n is sufficiently large, independently of the considered interpretation (1), (2), or (3). Theorems 4 and 5 imply that when all intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint, the estimators (7) and (8) are also consistent under the identifiable ontological indeterminacy interpretation (3), since this property is weaker than the consistency under the ontological indeterminacy interpretation (1). However, the next theorem implies that the pairwise disjointness of the intervals in \mathcal{I} is not a necessary condition for the consistent estimability of $[\underline{p}, \bar{p}] \in \mathcal{I}$ under the identifiable ontological indeterminacy interpretation (3), because it is sufficient that all nondeterministic intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint. An interval $[\underline{p}, \bar{p}] \in \mathcal{I}$ is said to be nondeterministic if it is not one of the two degenerate intervals $[0, 0]$ and $[1, 1]$.

Theorem 6 *A sufficient condition for $[\underline{p}, \bar{p}] \in \mathcal{I}$ to be consistently estimable under the identifiable ontological indeterminacy interpretation (3) is that all nondeterministic intervals in \mathcal{I} (degenerate or nondegenerate) are pairwise disjoint, while a necessary condition is that \mathcal{I} does not contain at the same time the interval $[0, 1]$ and another nondeterministic interval (degenerate or nondegenerate).*

The following estimators of \underline{p} and \bar{p} are consistent under the identifiable ontological indeterminacy interpretation (3), when the above sufficient condition on \mathcal{I} is fulfilled:

$$\underline{\pi}'_n(x_1, \dots, x_n) = \begin{cases} 1 & \text{if } x_1 = \dots = x_n = 1, \\ \underline{\pi}_n(x_1, \dots, x_n) & \text{otherwise,} \end{cases} \quad (9)$$

$$\bar{\pi}'_n(x_1, \dots, x_n) = \begin{cases} 0 & \text{if } x_1 = \dots = x_n = 0, \\ \bar{\pi}_n(x_1, \dots, x_n) & \text{otherwise,} \end{cases} \quad (10)$$

for all $x_1, \dots, x_n \in \{0, 1\}$, where $\underline{\pi}_n$ and $\bar{\pi}_n$ are the estimators (7) and (8), respectively.

Theorems 1, 2, and 4 characterize three different levels of estimability of imprecise probabilities according to three different ways of interpreting them. Only one of the nine possible characterizations is missing: the one of consistent estimability under the identifiable ontological indeterminacy

interpretation (3), because the necessary and sufficient conditions in Theorem 6 are different. In fact, this characterization seems to be much more difficult than the other eight, also because the exact meaning of the interpretation (3) depends on the functions $\underline{\alpha}, \bar{\alpha}$ considered. In particular, for the limits inferior and superior of the sequence p_i (4), it seems plausible that the sufficient condition of Theorem 6 is also necessary, but the proof does not seem to be straightforward.

In general, the results of the present section show that an empirical interpretation of imprecise probabilities is possible only in very limited settings, because imprecise probabilities cannot be estimated satisfactorily on the basis of finite amounts of data. This is hardly surprising when considering that imprecise probabilities are not identifiable in general under the interpretations (1) and (2), and only asymptotically identifiable under the interpretation (3). For these reasons, it can be interesting to study the estimability of the actual, finite-sample imprecise probabilities: that is, the estimability of $\min\{p_1, \dots, p_n\}$ and $\max\{p_1, \dots, p_n\}$ on the basis of the outcomes X_1, \dots, X_n of the corresponding Bernoulli trials.

The concepts of uniformly consistent estimability, strong estimability, and consistent estimability of the finite-sample imprecise probabilities $[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ can be obtained by replacing \underline{p} with $\min\{p_1, \dots, p_n\}$ in (6), and \bar{p} with $\max\{p_1, \dots, p_n\}$ in the analogue expression for $\bar{\pi}_n$ (the resulting concepts generalize the usual ones, since $\min\{p_1, \dots, p_n\}$ and $\max\{p_1, \dots, p_n\}$ are not necessarily constant). The next theorem implies that also the finite-sample imprecise probabilities have a very limited empirical meaning, since they can be estimated satisfactorily only when they are known to be precise.

Theorem 7 *The following six statements are equivalent:*

- (i) *$[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ is uniformly consistently estimable under the ontological indeterminacy interpretation (1) of $[\underline{p}, \bar{p}] \in \mathcal{I}$,*
- (ii) *$[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ is uniformly consistently estimable under the identifiable ontological indeterminacy interpretation (3) of $[\underline{p}, \bar{p}] \in \mathcal{I}$,*
- (iii) *$[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ is strongly estimable under the ontological indeterminacy interpretation (1) of $[\underline{p}, \bar{p}] \in \mathcal{I}$,*
- (iv) *$[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ is strongly estimable under the identifiable ontological indeterminacy interpretation (3) of $[\underline{p}, \bar{p}] \in \mathcal{I}$,*
- (v) *$[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ is consistently estimable under the ontological indeterminacy interpretation (1) of $[\underline{p}, \bar{p}] \in \mathcal{I}$,*
- (vi) *all intervals in \mathcal{I} are degenerate.*

The estimability of $[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ under the epistemological indeterminacy interpretation (2) is uninteresting, since it corresponds to the estimability of precise probabilities, which is implied by Bernoulli's law of large numbers. Of the other six possible characterizations, the only one missing is again the one of consistent estimability under the identifiable ontological indeterminacy interpretation (3): in fact, it can be shown that under this interpretation, the consistent estimabilities of $[\min\{p_1, \dots, p_n\}, \max\{p_1, \dots, p_n\}]$ and $[\underline{p}, \bar{p}]$ are equivalent, and so we are back to the difficulties of Theorem 6.

4. Conclusion

We have seen that in particular situations involving incomplete data, imprecise probabilities can have a clear empirical meaning as identification regions of frequentist, precise probabilities. Unfortunately, such situations are exceptional, and imprecise probabilities do not have a generally valid, clear empirical meaning, in the sense discussed in this paper.

Imprecise probabilities can be interpreted in several ways in terms of precise probabilities, as done for example in the imprecise versions of the theories of Markov chains and probabilistic graphical models. However, all these interpretations have a very limited empirical meaning, since imprecise probabilities are strongly estimable (that is, empirically recognizable) only in situations in which they are known to belong to a given set of pairwise disjoint imprecise probabilities. These results get even worse when we consider the actual, finite-sample imprecise probabilities, instead of the virtual, asymptotic ones. Anyway, examples of estimators have been given explicitly in this paper for the cases in which imprecise probabilities are satisfactorily estimable.

A mathematically interesting open problem is the question for a necessary and sufficient condition on a set of possibly degenerate probability intervals $[\underline{p}, \bar{p}]$, in order for them to be consistently estimable on the basis of any sequence of independent Bernoulli trials with precise probabilities of success $p_i \in [\underline{p}, \bar{p}]$ such that the sequence p_i has \underline{p} and \bar{p} as limits inferior and superior, respectively.

Acknowledgments

The author wishes to thank the anonymous referees for their valuable comments and suggestions.

References

- A. Antonucci, M. Cattaneo, and G. Corani. Likelihood-based robust classification with Bayesian networks. In S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, and R. R. Yager, editors, *Advances in Computational Intelligence*, volume 3, pages 491–500. Springer, 2012.
- A. Arnauld and P. Nicole. *La logique ou l'art de penser*. Savreux, 1662.
- A. Arnauld and P. Nicole. *Logic or the Art of Thinking*. Cambridge University Press, 1996.
- J. Bernoulli. *Ars Conjectandi*. Thurneysen Brothers, 1713.
- J. Bernoulli. *The Art of Conjecturing*. Johns Hopkins University Press, 2006.
- M. Cattaneo. On the robustness of imprecise probability methods. In F. G. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA '13*, pages 33–41. SIPTA, 2013.
- M. Cattaneo and A. Wiencierz. Likelihood-based Imprecise Regression. *International Journal of Approximate Reasoning*, 53:1137–1154, 2012.
- Z. Chen and P. Wu. Strong laws of large numbers for Bernoulli experiments under ambiguity. In S. Li, X. Wang, Y. Okazaki, J. Kawabe, T. Murofushi, and L. Guan, editors, *Nonlinear Mathematics for Uncertainty and its Applications*, pages 19–30. Springer, 2011.

- F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39:167–184, 2005.
- F. G. Cozman and L. Chrisman. Learning convex sets of probability from data. Technical Report CMU-RI-TR 97-25, Robotics Institute, Carnegie Mellon University, 1997.
- J. De Bock and G. de Cooman. Imprecise Bernoulli processes. In S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, and R. R. Yager, editors, *Advances in Computational Intelligence*, volume 3, pages 400–409. Springer, 2012.
- G. de Cooman and E. Miranda. Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138:2409–2432, 2008.
- B. de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329, 1931.
- B. de Finetti. *Theory of Probability. A Critical Introductory Treatment*, 2 volumes. Wiley, 1974–1975.
- A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38:325–339, 1967.
- P. I. Fierens, L. C. Rêgo, and T. L. Fine. A frequentist understanding of sets of measures. *Journal of Statistical Planning and Inference*, 139:1879–1892, 2009.
- D. Gillies. Popper’s contribution to the philosophy of probability. In A. O’Hear, editor, *Karl Popper: Philosophy and Problems*, pages 103–120. Cambridge University Press, 1995.
- D. Gillies. *Philosophical Theories of Probability*. Routledge, 2000.
- D. J. Hartfiel. *Markov Set-Chains*. Springer, 1998.
- I. O. Kozine and L. V. Utkin. Interval-valued finite Markov chains. *Reliable Computing*, 8:97–113, 2002.
- C. F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.
- M. Marinacci. Limit laws for non-additive probabilities and their frequentist interpretation. *Journal of Economic Theory*, 84:145–195, 1999.
- S. Peng. Nonlinear expectations and stochastic calculus under uncertainty—with robust central limit theorem and G -Brownian motion. arXiv:1002.4546 [math.PR], 2010.
- K. Popper. *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Springer, 1935.
- K. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.
- H. Reichenbach. *Wahrscheinlichkeitslehre. Eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Luitingh-Sijthoff, 1935.

- H. Reichenbach. *The Theory of Probability. An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, 1949.
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Forecasting with imprecise probabilities. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA '11*, pages 317–326. SIPTA, 2011.
- G. Shafer. Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19:309–370, 1978.
- E. D. Sylla. Tercentenary of *Ars Conjectandi* (1713): Jacob Bernoulli and the founding of mathematical probability. *International Statistical Review*, 82:27–45, 2014.
- M. C. M. Troffaes and G. de Cooman. *Lower Previsions*. Wiley, 2014.
- J. Venn. *The Logic of Chance. An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to Its Application to Moral and Social Science*. Macmillan, 1866.
- R. von Mises. *Wahrscheinlichkeit Statistik und Wahrheit*. Springer, 1928.
- R. von Mises. *Probability, Statistics and Truth*. Allen & Unwin, 2nd edition, 1957.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.
- P. Walley and T. L. Fine. Towards a frequentist theory of upper and lower probability. *The Annals of Statistics*, 10:741–761, 1982.
- P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematics and Physical Sciences, University of Sussex, 1975.
- P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007.

Bayesian Inference under Ambiguity: Conditional Prior Belief Functions

Giulianella Coletti

GIULIANELLA.COLETTI@UNIPG.IT

Dip. Matematica e Informatica, Università degli Studi di Perugia (Italy)

Davide Petturiti

DAVIDE.PETTURITI@UNIPG.IT

Dip. Economia, Università degli Studi di Perugia (Italy)

Barbara Vantaggi

BARBARA.VANTAGGI@SBAI.UNIROMA1.IT

Dip. S.B.A.I., “La Sapienza” Università di Roma (Italy)

Abstract

Bayesian inference under imprecise prior information is studied: the starting point is a precise strategy σ and a full B-conditional prior belief function Bel_B , conveying ambiguity in probabilistic prior information. In finite spaces, we give a closed form expression for the lower envelope \underline{P} of the class of full conditional probabilities dominating $\{Bel_B, \sigma\}$ and, in particular, for the related “posterior probabilities”. The assessment $\{Bel_B, \sigma\}$ is a coherent lower conditional probability in the sense of Williams and the characterized lower envelope \underline{P} coincides with its natural extension.

Keywords: conditional belief function; Bayesian conditioning rule; inference; ambiguity.

1. Introduction

Bayesian inference is known to naturally fit into de Finetti’s theory of coherent (finitely additive) conditional probabilities, where a coherent assessment can be always extended, generally not in a unique way, to any superset of conditional events (de Finetti, 1975; Williams, 1975).

In some application domains (e.g., decision theory, economics, game theory and forensic analysis, to cite some) the prior knowledge could be only partially specified or, even worse, it could refer to a different space of hypotheses. In these circumstances, instead of considering a single prior distribution, one is forced to take into account a set of priors (see, e.g., (Dempster, 1967; DeRoberts and Hartigan, 1981; Gilboa and Schmeidler, 1989)).

For instance, suppose that a pension system, based on the social security contributions Λ , is modified by a legal reform so that the new pension scheme takes into account the contribution’s years Θ . In order to use the previous information, we need to extract a new prior, starting from the prior distribution P of Λ by taking into account the logical relations between Λ and Θ . There could be possibly infinite probability distributions of (Λ, Θ) compatible with P , determining a lower envelope for the distribution of Θ . In particular, if the initial prior information P is a full conditional probability (Dubins, 1975), then for Θ we obtain a *full B-conditional belief function* (Coletti et al., 2016b), i.e., a conditional totally monotone uncertainty measure. Now, considering the profession X and a statistical model connecting X and Θ , the goal could be to draw inferences on Θ belonging to a set of values A (e.g., social pension) under particular values of X (e.g., a person is a clerk).

Motivated by the previous discussion, the main aim of this paper is to prove a generalized version of Bayes’ theorem, working with an ambiguous conditional prior information in the form of a full B-conditional belief function Bel_B and a precise statistical model λ , the latter uniquely determining a strategy σ (Dubins, 1975). A prior in the form of a full B-conditional belief function

is not so uncommon. For instance, in Example 2, starting from an automatic system \mathbf{S} which evolves according to a Markov chain, we show how to generate a full B-conditional prior belief function on the algebra spanned by the states of another unobservable automatic system \mathbf{T} , taking into account the logical constraints among the states of \mathbf{S} and those of \mathbf{T} .

Focusing on finite spaces, we provide a characterization of the lower envelope \underline{P} of the class of full conditional probabilities dominating $\{Bel_B, \sigma\}$. The assessment $\{Bel_B, \sigma\}$ is a Williams-coherent lower conditional probability and \underline{P} turns out to be its natural extension (Williams, 1975).

Our results are connected with those proved in (Walley, 1981, 1991; Wasserman, 1990a,b; Wasserman and Kadane, 1990): we generalize them in a finite context, since no assumption of positivity for the (lower or upper) probability of the conditioning events is required.

2. Preliminaries

Let \mathcal{A} be a Boolean algebra of *events* E 's, and denote with $(\cdot)^c$, \vee and \wedge the usual Boolean operations of negation, disjunction and conjunction, respectively, and with \subseteq the partial order of implication. The *sure event* Ω and the *impossible event* \emptyset coincide, respectively, with the top and bottom elements of \mathcal{A} . If \mathcal{A} is finite, we denote with $\mathcal{C}_{\mathcal{A}}$ the subset of its *atoms* which form the finer partition of Ω contained in \mathcal{A} . Denote $\mathcal{A}^0 = \mathcal{A} \setminus \{\emptyset\}$, \mathbb{N} is the set of natural numbers, I stands for an arbitrary index set and $\langle \{E_i\}_{i \in I} \rangle$ indicates the Boolean algebra generated by the set of events $\{E_i\}_{i \in I}$.

A *conditional event* $E|H$ is an ordered pair of events (E, H) with $H \neq \emptyset$. In particular, any event E can be identified with the conditional event $E|\Omega$. An arbitrary set of conditional events $\mathcal{G} = \{E_i|H_i\}_{i \in I}$ can always be embedded into a minimal set $\mathcal{A} \times \mathcal{A}^0$, where $\mathcal{A} = \langle \{E_i, H_i\}_{i \in I} \rangle$.

Recall the definition of *coherent conditional probability* essentially due to (de Finetti, 1975; Holzer, 1984; Regazzini, 1985; Williams, 1975).

Definition 1 Let $\mathcal{G} = \{E_i|H_i\}_{i \in I}$ be a set of conditional events. A function $P : \mathcal{G} \rightarrow [0, 1]$ is a **coherent conditional probability** if and only if, for every $n \in \mathbb{N}$, every $E_{i_1}|H_{i_1}, \dots, E_{i_n}|H_{i_n} \in \mathcal{G}$ and every real numbers s_1, \dots, s_n , denoting $\mathcal{B} = \langle \{E_{i_j}, H_{i_j}\}_{j=1, \dots, n} \rangle$ with set of atoms $\mathcal{C}_{\mathcal{B}} = \{C_1, \dots, C_m\}$, the random gain defined on $\mathcal{C}_{\mathcal{B}}$ as $G = \sum_{j=1}^n s_j (\mathbf{1}_{E_{i_j}} - P(E_{i_j}|H_{i_j})) \mathbf{1}_{H_{i_j}}$ satisfies

$$\min_{C_r \subseteq H_0^0} G(C_r) \leq 0 \leq \max_{C_r \subseteq H_0^0} G(C_r),$$

where $H_0^0 = \bigvee_{j=1}^n H_{i_j}$ and, for every $E \in \mathcal{B}$, $\mathbf{1}_E$ is its indicator defined on $\mathcal{C}_{\mathcal{B}}$ as $\mathbf{1}_E(C_r) = 1$ if $C_r \subseteq E$ and 0 otherwise.

In particular, if $\mathcal{G} = \mathcal{A} \times \mathcal{A}^0$ where \mathcal{A} is a Boolean algebra, then $P(\cdot|\cdot)$ is a coherent conditional probability if and only if it satisfies the following conditions:

- (C1) $P(E|H) = P(E \wedge H|H)$, for every $E \in \mathcal{A}$ and $H \in \mathcal{A}^0$;
- (C2) $P(\cdot|H)$ is a finitely additive probability on \mathcal{A} , for every $H \in \mathcal{A}^0$;
- (C3) $P(E \wedge F|H) = P(E|H) \cdot P(F|E \wedge H)$, for every $H, E \wedge H \in \mathcal{A}^0$ and $E, F \in \mathcal{A}$.

In this case $P(\cdot|\cdot)$ is simply said a *full conditional probability on* \mathcal{A} according to (Dubins, 1975).

If $\mathcal{G} = \{E_i|H_i\}_{i \in I}$ is an arbitrary set, the coherence condition is equivalent to the existence of a full conditional probability on the $\mathcal{A} = \langle \{E_i, H_i\}_{i \in I} \rangle$ extending the given assessment. This is

a consequence of the conditional version of the *fundamental theorem for probabilities* (de Finetti, 1975; Regazzini, 1985; Williams, 1975): every coherent conditional probability P on an arbitrary \mathcal{G} can be extended, generally not in a unique way, to every superset of conditional events \mathcal{G}' .

If \mathcal{A} is a finite Boolean algebra, every full conditional probability $P(\cdot|\cdot)$ on \mathcal{A} is in bijection with a unique linearly ordered class $\{P_0, \dots, P_k\}$ of probability measures on \mathcal{A} whose supports form a partition of Ω (Krauss, 1968). The class $\{P_0, \dots, P_k\}$ is called *complete agreeing class* and represents the full conditional probability $P(\cdot|\cdot)$ in the sense that, for every $F|K \in \mathcal{A} \times \mathcal{A}^0$, there is a minimum index $\alpha \in \{0, \dots, k\}$ such that $P_\alpha(K) > 0$ and $P(F|K) = \frac{P_\alpha(F \wedge K)}{P_\alpha(K)}$.

If \mathcal{G} is an arbitrary set, we can have more complete agreeing classes, each of them obtained by solving a suitable sequence of linear systems (Coletti and Scozzafava, 2002).

The set $\mathcal{P} = \{\tilde{P}(\cdot|\cdot)\}$ of all coherent extensions of P to a superset \mathcal{G}' is a compact subset of the space $[0, 1]^{\mathcal{G}'}$ endowed with the product topology and the projection set on each element of \mathcal{G}' is a (possibly degenerate) closed interval. The pointwise envelopes

$$\underline{P} = \min \mathcal{P} \quad \text{and} \quad \overline{P} = \max \mathcal{P},$$

are known as *coherent lower* and *upper conditional probabilities* (Coletti and Scozzafava, 2002), where coherence here is intended in the sense of (Williams, 1975) (namely, *Williams-coherence*). The envelopes \underline{P} and \overline{P} satisfy the *duality* property, i.e., $\overline{P}(E|H) = 1 - \underline{P}(E^c|H)$, for every $E|H, E^c|H \in \mathcal{G}'$, so in the following we mainly deal with \underline{P} .

In general, Williams-coherent lower conditional probabilities can be introduced without starting from a precise coherent conditional probability (Williams, 1975):

Definition 2 A function $\underline{P}(\cdot|\cdot)$ on a set of conditional events $\mathcal{G} = \{E_i|H_i\}_{i \in I}$ is a **Williams-coherent lower conditional probability** if there is a class of $\mathcal{P} = \{\tilde{P}(\cdot|\cdot)\}$ of coherent conditional probabilities on \mathcal{G} such that $\underline{P} = \inf \mathcal{P}$.

The extendibility of every coherent conditional probability implies the extendibility, generally not in a unique way, of every Williams-coherent lower conditional probability: the pointwise minimal of such extension is referred to as *natural extension* (Williams, 1975). For checking that an assessment is Williams-coherent in a finite setting and to find its natural extension see (Capotorti et al., 2003; Coletti and Scozzafava, 2002).

3. Full B-conditional belief and plausibility functions

A *belief function* Bel (Dempster, 1967; Shafer, 1976) on a finite Boolean algebra \mathcal{A} with set of atoms $\mathcal{C}_\mathcal{A}$ is a function such that $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and satisfying the *n-monotonicity* property for every $n \geq 2$, i.e., for every $E_1, \dots, E_n \in \mathcal{A}$,

$$Bel\left(\bigvee_{i=1}^n E_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1, \dots, n\}} (-1)^{|I|+1} Bel\left(\bigwedge_{i \in I} E_i\right).$$

The associated dual function Pl , defined as $Pl(E) = 1 - Bel(E^c)$, for every $E \in \mathcal{A}$, is said *plausibility function*. Both Bel and Pl are particular (*normalized*) *capacities* (Choquet, 1953), i.e., they are monotone with respect to the \subseteq relation. A belief function Bel (and so its dual Pl) on a

finite Boolean algebra is completely characterized by its Möbius inversion $m : \mathcal{A} \rightarrow [0, 1]$, also known as *basic probability assignment* (Shafer, 1976), defined, for every $E \in \mathcal{A}$, as

$$m(E) = \sum_{B \subseteq E} (-1)^{|\mathcal{C}_{E \setminus B}|} Bel(B),$$

where $\mathcal{C}_{E \setminus B} = \{D_r \in \mathcal{C}_\mathcal{A} : D_r \subseteq E \wedge B^c\}$. In particular, m satisfies $m(\emptyset) = 0$ and $\sum_{E \in \mathcal{A}} m(E) = 1$, and, for every $E \in \mathcal{A}$, it holds

$$Bel(E) = \sum_{B \subseteq E} m(E) \quad \text{and} \quad Pl(E) = \sum_{B \wedge E \neq \emptyset} m(E).$$

Denote with \mathbf{F}_{Bel} the set of *focal elements* of Bel , where an event A in \mathcal{A} is a focal element of Bel whenever $m(A) > 0$.

Given a finite partition $\mathcal{L} = \{H_1, \dots, H_n\}$ of Ω , a capacity φ on $\mathcal{A}_\mathcal{L} = \langle \mathcal{L} \rangle$ and a function $X : \mathcal{L} \rightarrow \mathbb{R}$, the *Choquet integral* of X with respect to φ is defined as

$$\oint X(H_i) \varphi(dH_i) = \sum_{i=1}^n [X(H_{\rho(i)}) - X(H_{\rho(i-1)})] \varphi(H_{\rho(i)} \vee \dots \vee H_{\rho(n)}),$$

where ρ is a permutation of $\{1, \dots, n\}$ such that $X(H_{\rho(1)}) \leq \dots \leq X(H_{\rho(n)})$ and $X(H_{\rho(0)}) := 0$. We write dH_i since we are integrating a function defined on the partition $\mathcal{L} = \{H_1, \dots, H_n\}$ with respect to a capacity defined on $\mathcal{A}_\mathcal{L} = \langle \mathcal{L} \rangle$.

We recall the definitions of C-class and full B-conditional belief and plausibility functions given in (Coletti et al., 2016b).

Definition 3 Let \mathcal{A} be a finite Boolean algebra. A linearly ordered class $\{Bel_0, \dots, Bel_k\}$ of belief functions on \mathcal{A} with sets of focal elements $\{\mathbf{F}_{Bel_0}, \dots, \mathbf{F}_{Bel_k}\}$ is said a **covering class**, or **C-class** for short, if it satisfies the following covering condition

$$\bigvee_{E \in \bigcup_{\alpha=0}^k \mathbf{F}_{Bel_\alpha}} E = \Omega. \quad (1)$$

By duality, a linearly ordered class $\{Pl_0, \dots, Pl_k\}$ of plausibility functions on \mathcal{A} is said **C-class** if the corresponding class of dual belief functions $\{Bel_0, \dots, Bel_k\}$ is. By means of a C-class of belief functions we define the concept of full B-conditional belief function, where B stands for Bayesian.

Definition 4 Let \mathcal{A} be a finite Boolean algebra. A function $Bel_B : \mathcal{A} \times \mathcal{A}^0 \rightarrow [0, 1]$ is a **full B-conditional belief function** on \mathcal{A} if there exists a C-class $\{Bel_0, \dots, Bel_k\}$ of belief functions on \mathcal{A} whose dual plausibility functions are $\{Pl_0, \dots, Pl_k\}$, such that, for every $E|H \in \mathcal{A} \times \mathcal{A}^0$, if $E \wedge H = H$ then $Bel_B(E|H) = 1$, while if $E \wedge H \neq H$

$$Bel_B(E|H) = \frac{Bel_{\alpha_{E,H}}(E \wedge H)}{Bel_{\alpha_{E,H}}(E \wedge H) + Pl_{\alpha_{E,H}}(E^c \wedge H)}, \quad (2)$$

where $\alpha_{E,H} = \min\{\alpha \in \{0, \dots, k\} : Bel_\alpha(E \wedge H) + Pl_\alpha(E^c \wedge H) > 0\}$.

The previous definition introduces a full B-conditional belief function through a *generalized Bayesian conditioning rule* corresponding to the one originally given in (Walley, 1981) for 2-monotone capacities. The Bayesian conditioning rule has been discussed for belief functions in (Dempster, 1967; Dubois and Denœux, 2012; Fagin and Halpern, 1991; Jaffray, 1992).

The difference with the previous ones is that the rule given in Definition 4 covers also the case in which $Bel(E \wedge H|\Omega) + Pl(E^c \wedge H|\Omega) = Bel_0(E \wedge H) + Pl_0(E^c \wedge H) = 0$, since it relies not on a single belief function but on a linearly ordered class of belief functions.

For $H \in \mathcal{A}^0$, the dual conditional function Pl_B of a full B-conditional belief function Bel_B on \mathcal{A} is defined, for every $E \in \mathcal{A}$, as

$$Pl_B(E|H) = 1 - Bel_B(E^c|H),$$

and is called *full B-conditional plausibility function*. By duality we immediately have $Pl_B(E|H) = 0$ when $E \wedge H = \emptyset$, while if $E \wedge H \neq \emptyset$ it holds

$$Pl_B(E|H) = 1 - Bel_B(E^c|H) = \frac{Pl_{\alpha_{E^c,H}}(E \wedge H)}{Pl_{\alpha_{E^c,H}}(E \wedge H) + Bel_{\alpha_{E^c,H}}(E^c \wedge H)}. \quad (3)$$

Notice that a full conditional probability P on \mathcal{A} is both a full B-conditional belief function and a full B-conditional plausibility function.

In this paper conditional belief functions are always intended in the sense of Definition 4: notice that full conditional probabilities reveal to be both full B-conditional belief and full B-conditional plausibility functions.

In (Coletti et al., 2016b) it is proved that the conditional measures Bel_B and Pl_B determine the non-empty compact set

$$\mathcal{P}_B = \{\tilde{\pi} : \tilde{\pi} \text{ is a full conditional probability on } \mathcal{A}, Bel_B \leq \tilde{\pi} \leq Pl_B\}, \quad (4)$$

for which it holds $Bel_B = \min \mathcal{P}_B$ and $Pl_B = \max \mathcal{P}_B$. In the same paper we prove that, for every full B-conditional belief function Bel_B on \mathcal{A} it is always possible to find a different finite Boolean algebra \mathcal{B} and a full conditional probability P on \mathcal{B} such that \mathcal{P}_B can be recovered as the set of coherent extensions of P to $\mathcal{A} \times \mathcal{A}^0$ and, thus, Bel_B and Pl_B as the envelopes of \mathcal{P}_B .

In (Coletti et al., 2016a) it is also shown that if all the belief functions in a C-class reduce to necessity measures then the corresponding full B-conditional belief function is a *full B-conditional necessity measure* and its dual is a *full B-conditional possibility measure*. In particular, interpreting the conditional measures Bel_B and Pl_B as envelopes, a necessary and sufficient condition (involving the finite Boolean algebras \mathcal{B} and \mathcal{A} and the full conditional probability P) is given for Bel_B and Pl_B to be full B-conditional necessity and possibility measures.

Since a full B-conditional belief function Bel_B determines the non-empty compact set \mathcal{P}_B of full conditional probabilities dominating it, its use in a Bayesian inferential procedure implies an ambiguous specification of a full conditional probability.

4. Bayesian inference with full B-conditional prior belief functions

Let $\mathcal{L} = \{H_1, \dots, H_n\}$ and $\mathcal{E} = \{E_1, \dots, E_m\}$ be two finite partitions of Ω and consider the Boolean algebras $\mathcal{A}_{\mathcal{L}} = \langle \mathcal{L} \rangle$, $\mathcal{A}_{\mathcal{E}} = \langle \mathcal{E} \rangle$, $\mathcal{A} = \langle \mathcal{A}_{\mathcal{L}} \cup \mathcal{A}_{\mathcal{E}} \rangle$. The partitions \mathcal{L} and \mathcal{E} play the roles of the sets of mutually exclusive and exhaustive “hypotheses” and “evidences”, respectively.

In the standard Bayesian setting, a *statistical model* (see, e.g., (Torgersen, 1991)) is given on $\mathcal{A}_{\mathcal{E}} \times \mathcal{L}$, where the latter is a function $\lambda : \mathcal{A}_{\mathcal{E}} \times \mathcal{L} \rightarrow [0, 1]$ such that, for every $H_i \in \mathcal{L}$:

- (L1) $\lambda(B|H_i) = 0$ if $B \wedge H_i = \emptyset$ and $\lambda(B|H_i) = 1$ if $B \wedge H_i = H_i$, for every $B \in \mathcal{A}_{\mathcal{E}}$;
- (L2) $\lambda(\cdot|H_i)$ is a probability on $\mathcal{A}_{\mathcal{E}}$.

Proposition 1 in (Petturiti and Vantaggi, 2017) implies that any statistical model λ on $\mathcal{A}_{\mathcal{E}} \times \mathcal{L}$ uniquely extends to a *strategy* on $\mathcal{A} \times \mathcal{L}$ (see, e.g., (Dubins, 1975)) which is a function $\sigma : \mathcal{A} \times \mathcal{L} \rightarrow [0, 1]$ such that, for every $H_i \in \mathcal{L}$:

- (S1) $\sigma(H_i|H_i) = 1$;
- (S2) $\sigma(\cdot|H_i)$ is a probability on \mathcal{A} .

By Theorem 5 in (Dubins, 1975), for every full conditional prior probability π on $\mathcal{A}_{\mathcal{L}}$, the assessment $\{\pi, \sigma\}$ (and, in particular, $\{\pi, \lambda\}$) is a coherent conditional probability, therefore it can be extended, generally not in a unique way, to a full conditional probability on \mathcal{A} . This implies that, given a full B-conditional prior belief function Bel_B on \mathcal{A} , the assessment $\{Bel_B, \sigma\}$ (and, in particular, $\{Bel_B, \lambda\}$) is a Williams-coherent lower conditional probability.

Remark 5 *The assessment $\{Bel_B, \sigma\}$ determines a Williams-coherent lower conditional probability \underline{P} on the set of conditional events $\mathcal{G} = (\mathcal{A}_{\mathcal{L}} \times \mathcal{A}_{\mathcal{L}}^0) \cup (\mathcal{A} \times \mathcal{L})$ such that $\underline{P}|_{\mathcal{A}_{\mathcal{L}} \times \mathcal{A}_{\mathcal{L}}^0} = Bel_B$ and $\underline{P}_{\mathcal{A} \times \mathcal{L}} = \sigma$ so, with a little abuse of terminology, $\{Bel_B, \sigma\}$ is directly referred to be a Williams-coherent lower conditional probability.*

Remark 6 *Recall that, in view of the finite setting adopted in this paper, the notion of conditioning for lower probabilities due to Williams coincides with that due to (Walley, 1991) since in this case the conglomerability condition is automatically satisfied.*

Let Bel_B be a full B-conditional belief function on $\mathcal{A}_{\mathcal{L}}$ and σ a strategy on $\mathcal{A} \times \mathcal{L}$ and denote with \mathcal{P}_B the set of full conditional probabilities on $\mathcal{A}_{\mathcal{L}}$ dominating Bel_B . Consider

$$\mathcal{P} = \{\tilde{P} : \tilde{P} \text{ is a full conditional probability on } \mathcal{A} \text{ extending } \{\tilde{\pi}, \sigma\}, \tilde{\pi} \in \mathcal{P}_B\},$$

which is a non-empty compact subset of $[0, 1]^{\mathcal{A} \times \mathcal{A}^0}$ endowed with the product topology, whose envelopes are $\underline{P} = \min \mathcal{P}$ and $\overline{P} = \max \mathcal{P}$. The lower envelope $\underline{P}(\cdot|\cdot)$ turns out to be the natural extension of the Williams-coherent lower conditional probability $\{Bel_B, \sigma\}$.

The following theorem provides a characterization of the lower envelope $\underline{P}(\cdot|\cdot)$, relying on the functions defined, for every $F, K \in \mathcal{A}$ and $A \in \mathcal{A}_{\mathcal{L}}^0$ such that $K \subseteq A$, as

$$L(F, K; A) = \min_{\tilde{\pi} \in \mathcal{P}_B} \left\{ \sum_{i=1}^n \sigma(FK|H_i) \tilde{\pi}(H_i|A) : \sum_{i=1}^n \sigma(F^c K|H_i) \tilde{\pi}(H_i|A) = \overline{P}(F^c K|A) \right\},$$

$$U(F, K; A) = \max_{\tilde{\pi} \in \mathcal{P}_B} \left\{ \sum_{i=1}^n \sigma(FK|H_i) \tilde{\pi}(H_i|A) : \sum_{i=1}^n \sigma(F^c K|H_i) \tilde{\pi}(H_i|A) = \underline{P}(F^c K|A) \right\},$$

where we write FK and $F^c K$ in place of $F \wedge K$ and $F^c \wedge K$ to save space.

Theorem 7 *The lower envelope $\underline{P}(\cdot|\cdot)$ is such that, for every $F|K \in \mathcal{A} \times \mathcal{A}^0$, if $F \wedge K = K$, then $\underline{P}(F|K) = 1$, otherwise:*

(i) *if $K \in \mathcal{A}_{\mathcal{L}}^0$, then*

$$\underline{P}(F|K) = \oint \sigma(F|H_i) Bel_B(dH_i|K);$$

(ii) *if $K \in \mathcal{A}^0 \setminus \mathcal{A}_{\mathcal{L}}^0$, then if there exists $A \in \mathcal{A}_{\mathcal{L}}^0$ such that $K \subseteq A$ and $\underline{P}(K|A) > 0$ we have that*

$$\underline{P}(F|K) = \min \left\{ \frac{\underline{P}(F \wedge K|A)}{\underline{P}(F \wedge K|A) + U(F^c, K; A)}, \frac{L(F, K; A)}{L(F, K; A) + \bar{P}(F^c \wedge K|A)} \right\},$$

otherwise $\underline{P}(F|K) = 0$.

Proof The statement is trivial if $F \wedge K = K$ since in this case $\tilde{P}(F|K) = 1$ for every $\tilde{P} \in \mathcal{P}$, thus suppose $F \wedge K \neq K$. Condition (i) follows since, if $K \in \mathcal{A}_{\mathcal{L}}^0$ then

$$\begin{aligned} \underline{P}(F|K) &= \min_{\tilde{P} \in \mathcal{P}} \tilde{P}(F|K) = \min_{\tilde{\pi} \in \mathcal{P}_B} \sum_{i=1}^n \sigma(F|H_i) \tilde{\pi}(H_i|K) \\ &= \min_{\tilde{\pi} \in \mathcal{C}_{Bel_B(\cdot|K)}} \sum_{i=1}^n \sigma(F|H_i) \tilde{\pi}(H_i|K) = \oint \sigma(F|H_i) Bel_B(dH_i|K), \end{aligned}$$

where $\mathcal{C}_{Bel_B(\cdot|K)} = \{\tilde{\pi}(\cdot|K)\}$ is the core of probability measures on $\mathcal{A}_{\mathcal{L}}$ induced by $Bel_B(\cdot|K)$ and the last equality follows by Proposition 3 in (Schmeidler, 1986). To prove condition (ii) let us consider $K \in \mathcal{A}^0 \setminus \mathcal{A}_{\mathcal{L}}^0$. If there exists $A \in \mathcal{A}_{\mathcal{L}}^0$ such that $K \subseteq A$ and $\underline{P}(K|A) > 0$ we have that $\tilde{P}(K|A) > 0$ for every $\tilde{P} \in \mathcal{P}$, thus $\underline{P}(F|K) = \min_{\tilde{P} \in \mathcal{P}} \frac{\tilde{P}(F \wedge K|A)}{\tilde{P}(F \wedge K|A) + \tilde{P}(F^c \wedge K|A)}$. So, the conclusion

follows since $\frac{x}{x+y}$ is increasing in x and decreasing in y , and, for every $\tilde{P} \in \mathcal{P}$, $\tilde{P}(\cdot|A)$ is the convex combination of probabilities $P_1(\cdot|A)$ and $P_2(\cdot|A)$, $P_1, P_2 \in \mathcal{P}$, attaining the lower and the upper envelopes, respectively, on $F \wedge K$ (or on $F^c \wedge K$) and $\tilde{P}(F|K) \geq \min\{P_1(F|K), P_2(F|K)\}$. The remaining case, realizing when for all $A \in \mathcal{A}_{\mathcal{L}}^0$ with $K \subseteq A$ it holds $\underline{P}(K|A) = 0$, is proven in analogy to the proof of Lemma 3 in (Petturiti and Vantaggi, 2017). ■

Restricting to a finite setting, the previous theorem generalizes some results proved in (Coletti et al., 2014) in which an ambiguous unconditional prior is considered, either in the form of a belief function or a 2-monotone capacity.

A simplification of condition (ii) of Theorem 7 is obtained when the functions on \mathcal{L} , defined as $X(\cdot) = \sigma(F \wedge H|\cdot)$ and $(1 - Y(\cdot)) = (1 - \sigma(F^c \wedge H|\cdot))$, are *comonotonic* (see, e.g., (Denneberg, 1994)), i.e., for every $H_h, H_k \in \mathcal{L}$, $[X(H_h) - X(H_k)] \cdot [(1 - Y(H_h)) - (1 - Y(H_k))] \geq 0$, as shown by the following Proposition 8. In particular, this happens for all conditional events in $\mathcal{A}_{\mathcal{L}} \times \mathcal{A}_{\mathcal{E}}^0$ related to “posterior probabilities”, obtaining, for a finite setting, a generalization of results in (Wasserman, 1990a).

Proposition 8 *For every $F|K \in \mathcal{A} \times \mathcal{A}^0$ such that $F \wedge K \neq K$, $K \in \mathcal{A}^0 \setminus \mathcal{A}_{\mathcal{L}}^0$ and there exists $A \in \mathcal{A}_{\mathcal{L}}^0$ such that $K \subseteq A$ and $\underline{P}(K|A) > 0$, if $X(\cdot) = \sigma(F \wedge H|\cdot)$ and $(1 - Y(\cdot)) = (1 - \sigma(F^c \wedge H|\cdot))$ are comonotonic then*

$$\underline{P}(F|K) = \frac{\underline{P}(F \wedge K|A)}{\underline{P}(F \wedge K|A) + \bar{P}(F^c \wedge K|A)}.$$

Proof For every $A \in \mathcal{A}_{\mathcal{L}}^0$, $Bel_B(\cdot|A)$ is a totally monotone capacity on $\mathcal{A}_{\mathcal{L}}$ inducing a core $\mathcal{C}_{Bel_B(\cdot|A)} = \{\tilde{\pi}(\cdot|A)\}$ of probability measures on $\mathcal{A}_{\mathcal{L}}$, moreover, the functions $X(\cdot)$ and $(1 - Y(\cdot))$ are comonotonic.

By Proposition 6.26 in (Troffaes and de Cooman, 2014) there exists $\tilde{\pi}(\cdot|A) \in \mathcal{C}_{Bel_B(\cdot|A)}$ such that $\sum_{i=1}^n X(H_i)\tilde{\pi}(H_i|A) = \oint X(H_i)Bel_B(dH_i|A)$ and $\sum_{i=1}^n (1 - Y(H_i))\tilde{\pi}(H_i|A) = \oint (1 - Y(H_i))Bel_B(dH_i|A)$.

Since $\sum_{i=1}^n (1 - Y(H_i))\tilde{\pi}(H_i|A) = 1 - \sum_{i=1}^n Y(H_i)\tilde{\pi}(H_i|A)$ and $\oint (1 - Y(H_i))Bel_B(dH_i|A) = 1 - \oint Y(H_i)Pl_B(dH_i|A)$, it follows $\oint Y(H_i)Pl_B(dH_i|A) = \sum_{i=1}^n Y(H_i)\tilde{\pi}(H_i|A)$ and this implies $\underline{P}(F \wedge K|A) = \oint X(H_i)Bel_B(dH_i|A) = \sum_{i=1}^n X(H_i)\tilde{\pi}(H_i|A) = L(F, K; A)$ and $\overline{P}(F^c \wedge K|A) = \oint Y(H_i)Pl_B(dH_i|A) = \sum_{i=1}^n Y(H_i)\tilde{\pi}(H_i|A) = U(F^c, K; A)$. ■

The following example shows that, though $Bel_B(\cdot|K)$ is a belief function on $\mathcal{A}_{\mathcal{L}}$, for every $K \in \mathcal{A}_{\mathcal{L}}^0$, and $\sigma(\cdot|H_i)$ is a probability measure on \mathcal{A} , for every $H_i \in \mathcal{L}$, the function $\underline{P}(\cdot|K)$ can fail 2-monotonicity, for some $K \in \mathcal{A}^0$.

Example 1 Let $\mathcal{L} = \{H_1, H_2, H_3\}$ and $\mathcal{E} = \{E_1, E_2, E_3, E_4\}$ be logically independent partitions of Ω , and take $\mathcal{A}_{\mathcal{L}} = \langle \mathcal{L} \rangle$, $\mathcal{A}_{\mathcal{E}} = \langle \mathcal{E} \rangle$ and $\mathcal{A} = \langle \mathcal{A}_{\mathcal{L}} \cup \mathcal{A}_{\mathcal{E}} \rangle$. Let Bel_B be the full B -conditional belief function on $\mathcal{A}_{\mathcal{L}}$ determined by the C -class of belief functions $\{Bel_0, Bel_1\}$ displayed below

$\mathcal{A}_{\mathcal{L}}$	\emptyset	H_1	H_2	H_3	$H_1 \vee H_2$	$H_1 \vee H_3$	$H_2 \vee H_3$	Ω
Bel_0	0	$\frac{1}{2}$	0	0	1	$\frac{1}{2}$	0	1
Bel_1	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1

where $\mathbf{F}_{Bel_0} = \{H_1, H_1 \vee H_2\}$ and $\mathbf{F}_{Bel_1} = \{H_1, H_2 \vee H_3\}$, thus condition (I) of Definition 3 is satisfied.

Let λ be the statistical model on $\mathcal{A}_{\mathcal{E}} \times \mathcal{L}$ such that

$$\lambda(E_j|H_1) = \lambda(E_j|H_3) = \frac{1}{6}, \text{ for } j = 1, 2, 3, \text{ and } \lambda(E_4|H_1) = \lambda(E_4|H_3) = \frac{1}{2},$$

$$\lambda(E_1|H_2) = \lambda(E_3|H_2) = \frac{1}{2}, \text{ and } \lambda(E_2|H_2) = \lambda(E_4|H_2) = 0.$$

which uniquely extends to a strategy σ on $\mathcal{A} \times \mathcal{L}$ by Proposition 1 in (Petturiti and Vantaggi, 2017). Let $K = H_2 \vee H_3$, $A = E_1 \vee E_2$ and $B = E_2 \vee E_3$. Simple computations show that $Bel_B(\cdot|K)$ is a belief function vacuous at K , so, we have

$$\begin{aligned} \underline{P}(A \vee B|K) &= \oint \sigma(A \vee B|H_i)Bel_B(dH_i|K) = \inf_{H_i \subseteq K} \sigma(A \vee B|K) = \frac{1}{2}, \\ \underline{P}(A|K) &= \oint \sigma(A|H_i)Bel_B(dH_i|K) = \inf_{H_i \subseteq K} \sigma(A|K) = \frac{1}{3}, \\ \underline{P}(B|K) &= \oint \sigma(B|H_i)Bel_B(dH_i|K) = \inf_{H_i \subseteq K} \sigma(B|K) = \frac{1}{3}, \\ \underline{P}(A \wedge B|K) &= \oint \sigma(A \wedge B|H_i)Bel_B(dH_i|K) = \inf_{H_i \subseteq K} \sigma(A \wedge B|K) = 0. \end{aligned}$$

Since $\underline{P}(A \vee B|K) < \underline{P}(A|K) + \underline{P}(B|K) - \underline{P}(A \wedge B|K)$, $\underline{P}(\cdot|K)$ is not 2-monotone.

Proposition 8 is a generalization of the ϵ -contamination model presented in Example 2.3 in (Huber, 1981), where the author provides a characterization of the lower envelope \underline{P} on $\mathcal{A}_{\mathcal{L}} \times \mathcal{E}$, starting from a statistical model λ and an unconditional prior belief function Bel obtained as the ϵ -contamination of a reference prior probability. In such case, in (Huber, 1981) it is stated that $\underline{P}(\cdot|E_j)$ is a 2-monotone capacity on $\mathcal{A}_{\mathcal{L}}$, for every $E_j \in \mathcal{E}$, nevertheless, as shown in our Example 2 the envelope $\underline{P}(\cdot|K)$ can fail 2-monotonicity on the whole \mathcal{A} , for some $K \in \mathcal{A}^0$.

The following example shows that a full B-conditional prior belief function can be obtained starting from a full conditional prior probability defined on a different algebra.

Example 2 We consider an automatic system \mathbf{S} that can assume three possible states s_1 , s_2 and s_3 . Let $\mathcal{S} = \{S_1, S_2, S_3\}$ be the partition of Ω , where $S_i = \text{"S is in state } s_i\text{"}$, for $i = 1, 2, 3$, and denote $\mathcal{A}_{\mathcal{S}} = \langle \mathcal{S} \rangle$. The evolution of \mathbf{S} is determined by the Markov chain whose transition matrix and graph are reported in Figure 1.

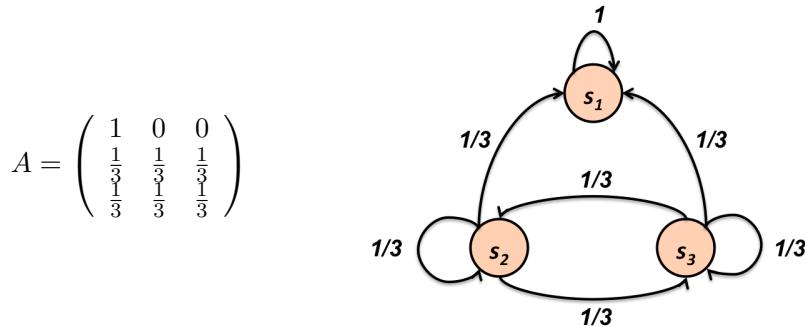


Figure 1: Transition matrix and graph of the Markov chain related to \mathbf{S}

Suppose that the initial state of \mathbf{S} is selected at random and that we observe the system evolve indefinitely in time, so, we can take the limit probabilistic behaviour as our prior information on \mathbf{S} . The starting probability distribution on the states of \mathbf{S} is $\pi^{(0)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, while after $n > 0$ steps we have $\pi^{(n)} = \pi^{(n-1)} A = \left(1 - \left(\frac{2}{3}\right)^{n+1}, \frac{1}{3} \left(\frac{2}{3}\right)^n, \frac{1}{3} \left(\frac{2}{3}\right)^n\right)$.

It is easily seen that the probability distribution $\pi^{(n)}$ is positive for every $n \geq 0$, so, it uniquely extends to a full conditional probability (still denoted with $\pi^{(n)}$) on $\mathcal{A}_{\mathcal{S}}$ setting, for every $A|B \in \mathcal{A}_{\mathcal{S}} \times \mathcal{A}_{\mathcal{S}}^0$, $\pi^{(n)}(A|B) = \frac{\pi^{(n)}(A \wedge B)}{\pi^{(n)}(B)}$. Thus, we have a sequence $\{\pi^{(n)} : n = 0, 1, 2, \dots\}$ of full conditional probabilities on $\mathcal{A}_{\mathcal{S}}$ converging pointwise to the full conditional probability $\pi^{(\infty)}$ on $\mathcal{A}_{\mathcal{S}}$ defined below

$\mathcal{A}_{\mathcal{S}}$	\emptyset	S_1	S_2	S_3	$S_1 \vee S_2$	$S_1 \vee S_3$	$S_2 \vee S_3$	Ω
$\pi^{(\infty)}(\cdot S_1)$	0	1	0	0	1	1	0	1
$\pi^{(\infty)}(\cdot S_2)$	0	0	1	0	1	0	1	1
$\pi^{(\infty)}(\cdot S_3)$	0	0	0	1	0	1	1	1
$\pi^{(\infty)}(\cdot S_1 \vee S_2)$	0	1	0	0	1	1	0	1
$\pi^{(\infty)}(\cdot S_1 \vee S_3)$	0	1	0	0	1	1	0	1
$\pi^{(\infty)}(\cdot S_2 \vee S_3)$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1
$\pi^{(\infty)}(\cdot \Omega)$	0	1	0	0	1	1	0	1

The full conditional probability $\pi^{(\infty)}$ is determined by the complete agreeing class $\{P_0, P_1\}$ of probability measures on \mathcal{A}_S such that $P_0(\cdot) = \pi^{(\infty)}(\cdot|\Omega)$ and $P_1(\cdot) = \pi^{(\infty)}(\cdot|S_2 \vee S_3)$.

Consider now a second automatic system \mathbf{T} that is not directly observable: the only information we have is that \mathbf{T} can assume three possible states t_1, t_2 and t_3 , and that if \mathbf{S} is in state s_i then \mathbf{T} is not in state t_i , for $i = 1, 2, 3$. Let $\mathcal{T} = \{T_1, T_2, T_3\}$ be the partition of Ω , where $T_i = \{\mathbf{T} \text{ is in state } t_i\}$, for $i = 1, 2, 3$, and denote $\mathcal{A}_{\mathcal{T}} = \langle \mathcal{T} \rangle$ with $T_i \wedge S_i = \emptyset$, for $i = 1, 2, 3$.

As proven in ([Coletti et al., 2016b](#)) setting, for every $B \in \mathcal{A}_{\mathcal{T}}$,

$$(B)_* = \bigvee \{S_i \in \mathcal{S} : S_i \subseteq B\}, \quad Bel_0(B) = P_0((B)_*) \quad \text{and} \quad Bel_1(B) = P_1((B)_*),$$

we obtain a C-class of belief functions $\{Bel_0, Bel_1\}$ on $\mathcal{A}_{\mathcal{T}}$ which, in turn, determines the full B-conditional belief function on $\mathcal{A}_{\mathcal{T}}$ reported below

$\mathcal{A}_{\mathcal{T}}$	\emptyset	T_1	T_2	T_3	$T_1 \vee T_2$	$T_1 \vee T_3$	$T_2 \vee T_3$	Ω
$Bel_B(\cdot T_1)$	0	1	0	0	1	1	0	1
$Bel_B(\cdot T_2)$	0	0	1	0	1	0	1	1
$Bel_B(\cdot T_3)$	0	0	0	1	0	1	1	1
$Bel_B(\cdot T_1 \vee T_2)$	0	0	0	0	1	0	0	1
$Bel_B(\cdot T_1 \vee T_3)$	0	0	0	0	0	1	0	1
$Bel_B(\cdot T_2 \vee T_3)$	0	0	0	0	0	0	1	1
$Bel_B(\cdot \Omega)$	0	0	0	0	0	0	1	1

Suppose that the state of the unobservable system \mathbf{T} can be verified through a detector \mathbf{D} that returns one of three possible values d_1, d_2 and d_3 , with d_i corresponding to the state t_i , for $i = 1, 2, 3$, with a reliability of 90% and equal chances on failures. Let $\mathcal{D} = \{D_1, D_2, D_3\}$ be the partition of Ω , where $D_i = \{\mathbf{D} \text{ returns } d_i\}$, for $i = 1, 2, 3$, and denote $\mathcal{A}_{\mathcal{D}} = \langle \mathcal{D} \rangle$. Let $\mathcal{A} = \langle \mathcal{A}_{\mathcal{T}} \cup \mathcal{A}_{\mathcal{D}} \rangle$ and consider the statistical model on $\mathcal{A}_{\mathcal{D}} \times \mathcal{T}$ singled out by

$$\lambda(D_i|T_i) = 90\%, \quad \lambda(D_j|T_i) = \lambda(D_k|T_i) = 5\%, \quad \text{for different } i, j, k \in \{1, 2, 3\},$$

that uniquely extends to a strategy σ on $\mathcal{A} \times \mathcal{T}$ by Proposition 1 in ([Petturiti and Vantaggi, 2017](#)).

The full B-conditional belief function Bel_B encodes all our prior information on \mathbf{T} and can be used together with σ to draw Bayesian inferences. At this aim, suppose that the detector \mathbf{D} shows the value d_j , for $j = 1, 2, 3$, then the lower posterior distribution on the states of \mathbf{T} can be easily determined using Proposition 8. For instance, since $\underline{P}(D_j|\Omega) > 0$, $\underline{P}(T_1 \wedge D_j|\Omega) = 0$ and $\overline{P}(T_1^c \wedge D_j|\Omega) > 0$, for $j = 1, 2, 3$, we get

$$\underline{P}(T_1|D_j) = \frac{\underline{P}(T_1 \wedge D_j|\Omega)}{\underline{P}(T_1 \wedge D_j|\Omega) + \overline{P}(T_1^c \wedge D_j|\Omega)} = 0,$$

and, analogously, we can compute $\underline{P}(T_1^c|D_j) = 1$, so, $\underline{P}(T_1|D_j) = \overline{P}(T_1|D_j) = 0$, i.e., the observation of the detector \mathbf{D} does not change our degree of belief on T_1 since it is $Bel_B(T_1|\Omega) = Pl_B(T_1|\Omega) = 0$.

5. Conclusions

We show that, as long as we consider a precise strategy σ , the introduction of ambiguity in the prior information through a full B-conditional belief function Bel_B has straightforward treatment:

a characterization for the envelopes of the class of full conditional probabilities dominating the assessment $\{Bel_B, \sigma\}$ is provided. The entire procedure lives inside Williams framework and the characterized lower envelope reveals to be the natural extension of $\{Bel_B, \sigma\}$. Our aim for future research is to introduce ambiguity also in the strategy by considering an imprecise strategy β such that $\beta(\cdot|H_i)$ is a belief function, for every $H_i \in \mathcal{L}$, possibly removing the finiteness assumption. This would lead to a theory to compare with that of (Walley, 1991).

Acknowledgments

This work was partially supported by INdAM-GNAMPA through the Project 2016 U2016/000391.

References

- A. Capotorti, L. Galli, and B. Vantaggi. Locally strong coherence and inference with lower–upper probabilities. *Soft Computing*, 7(5):280–287, 2003.
- G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953.
- G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, volume 15 of *Trends in Logic*. Kluwer Academic Publisher, Dordrecht/Boston/London, 2002.
- G. Coletti, D. Petturiti, and B. Vantaggi. Bayesian inference: the role of coherence to deal with a prior belief function. *Statistical Methods and Applications*, 23(4):519–545, 2014.
- G. Coletti, D. Petturiti, and B. Vantaggi. When upper conditional probabilities are conditional possibility measures. *Fuzzy Sets and Systems*, 304:45–64, 2016a.
- G. Coletti, D. Petturiti, and B. Vantaggi. Conditional belief functions as lower envelopes of conditional probabilities in a finite setting. *Information Sciences*, 339:64–84, 2016b.
- B. de Finetti. *Theory of Probability 1-2*. Wiley, 1975.
- A. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.
- D. Denneberg. *Non-Additive Measure and Integral*, volume 27 of *Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers, Dordrecht/Boston/London, 1994.
- L. DeRoberts and J. Hartigan. Bayesian inference using intervals of measures. *Annals of Statistics*, 9(2):235–464, 1981.
- L. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3(1):89–99, 1975.
- D. Dubois and T. Denœux. *Conditioning in Dempster-Shafer Theory: Prediction vs. Revision*, pages 385–392. Springer Berlin Heidelberg, 2012.
- R. Fagin and J. Halpern. *Uncertainty in Artificial Intelligence*, chapter A New Approach to Updating Beliefs, pages 347–374. Elsevier Science Publishers, 1991.

- I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153, 1989.
- S. Holzer. Sulla nozione di coerenza per le probabilità subordinate. In *Rendiconti dell'Istituto di Matematica dell'Università di Trieste*, volume 16, pages 46–62. 1984.
- P. Huber. *Robust Statistics*. Wiley, 1981.
- J.-Y. Jaffray. Bayesian updating and belief functions. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(5):1144–1152, 1992.
- P. Krauss. Representation of conditional probability measures on boolean algebras. *Acta Mathematica Academiae Scientiarum Hungarica*, 19(3-4):229–241, 1968.
- D. Petturiti and B. Vantaggi. Envelopes of conditional probabilities extending a strategy and a prior probability. *International Journal of Approximate Reasoning*, 81:160–182, 2017.
- E. Regazzini. Finitely additive conditional probabilities. *Rendiconti del Seminario Matematico e Fisico di Milano*, 55(1):69–89, 1985.
- D. Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97:255–261, 1986.
- G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.
- E. Torgersen. *Comparison of Statistical Experiments*, volume 36 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1991.
- M. Troffaes and G. de Cooman. *Lower Previsions*. Wiley Series in Probability and Statistics. Wiley, 2014.
- P. Walley. Coherent lower (and upper) probabilities. Technical report, Department of Statistics, University of Warwick, 1981.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
- L. Wasserman. Prior envelopes based on belief functions. *Annals of Statistics*, 18(1):454–464, 1990a.
- L. Wasserman. Belief functions and statistical inference. *Canadian Journal of Statistics*, 18(3): 183–196, 1990b.
- L. Wasserman and J. Kadane. Bayes’ theorem for choquet capacities. *Annals of Statistics*, 18(3): 1328–1339, 1990.
- P. Williams. Note on conditional previsions. Unpublished report of School of Mathematical and Physical Science, University of Sussex (Published in *International Journal of Approximate Reasoning*, 44:366–383, 2007), 1975.

Weak Dutch Books versus Strict Consistency with Lower Previsions

Chiara Corsato

CCORSATO@UNITS.IT

Renato Pelessoni

RENATO.PELESSONI@ECON.UNITS.IT

Paolo Vicig

PAOLO.VICIG@ECON.UNITS.IT

*University of Trieste
Trieste (Italy)*

Abstract

Several consistency notions for lower previsions (coherence, convexity, others) require that the suprema of certain gambles, having the meaning of gains, are non-negative. The limit situation that a gain supremum is zero is termed Weak Dutch Book (WDB). In the literature, the special case of WDBs with precise probabilities has mostly been analysed, and strict coherence has been proposed as a radical alternative. In this paper the focus is on WDBs and generalised strict coherence, termed strict consistency, with imprecise previsions. We discuss properties of lower previsions incurring WDBs and conditions for strict consistency, showing in both cases how they are differentiated by the degree of consistency of the given uncertainty assessment.

Keywords: Weak Dutch Books; (Williams') coherence; convex previsions; strict consistency.

1. Introduction

In the coherence approach to the theory of Imprecise Probabilities, consistency of an uncertainty measure is formalised requiring that the supremum of a certain gamble (a bounded random number, called *gain*) is non-negative. This is a common feature to several consistency notions, like *coherence* for lower previsions (Walley, 1991), Williams' coherence (*W-coherence*, Williams, 1975), convexity (Pelessoni and Vicig, 2005b), coherence for precise previsions or *dF-coherence* (de Finetti, 1974), and others. These concepts allow for a behavioural interpretation: the gain has the meaning of an agent's overall gain from a finite number of bets (rules for selecting the admissible gains distinguish the various consistency concepts).

Within this context, the limiting situation that the supremum of some gain G is precisely zero is termed *Weak Dutch Book* (WDB). In fact, under the behavioural interpretation, an agent whose gain is G would at best gain nothing, but otherwise lose, from the corresponding overall bet.

The literature on WDBs is not extended, and mostly focused on WDBs for *dF*-coherent probabilities. Contributions go back to the fifties of the last century (Kemeny, 1955; Shimony, 1955), when de Finetti's theory was getting widespread. In an attempt to avoid WDBs, the notion of *strict coherence* was introduced, although it became soon clear that it is subject to important constraints.

Properties of an uncertainty assessment incurring a WDB received a lesser attention, and the whole issue was rarely considered outside *dF*-coherence. The agent's beliefs of incurring a real loss were investigated in Crisma (2006) for *dF*-coherent probabilities, and in Vicig (2010) for (unconditional) coherent lower/upper previsions.

After introducing some preliminary material in Section 2, in this paper we focus precisely on the properties of WDB assessments, and on how they are differentiated under different consistency assumptions. We especially discuss *W*-coherence, convexity and *dF*-coherence. Section 3 is concerned with a 'local precision' property. This means that if the lower prevision \underline{P} satisfies in general

a certain consistency requirement, then \underline{P} complies with stronger properties, that make it closer to a precise prevision, on the set $\mathcal{D}_{\underline{G}}$ of (possibly conditional) gambles appearing in the expression of a WDB gain for \underline{P} . We prove one such relationship for the case of conditional convex lower previsions (Proposition 10). This implies that, if \underline{P} is an unconditional convex prevision, \underline{P} is the translation of a dF -coherent prevision on $\mathcal{D}_{\underline{G}}$, while if \underline{P} is a coherent lower prevision on its domain, it is precisely a dF -coherent prevision on $\mathcal{D}_{\underline{G}}$. A result for W -coherence is also supplied. Section 4 discusses the agent's beliefs about incurring a real *Dutch Book* with a WDB gain. Again, these are differentiated by consistency of the assessment, ranging from near-certainty of avoiding any losses bounded away from zero with dF -coherence to no such reassuring beliefs for convexity, with W -coherence somewhat intermediate. We also discuss interdependencies between events of positive probability and maxima for WDB gains. In Section 5 strict consistency, a generalisation of strict coherence, is explored. After recalling a result in Corsato et al. (2017) for W -coherence, the perspective is that of analysing various strict consistency conditions, which are equivalent with dF -coherence, but not necessarily so in an imprecise framework (Proposition 21). Section 6 concludes the paper. Results not proven here may be found in the extended paper Corsato et al. (2017).

2. Preliminaries

Denote with \mathcal{D} an arbitrary non-empty set of possibly conditional gambles. In the sequel, \mathcal{D} will be the domain of a (precise or imprecise) conditional or unconditional prevision.

In the *conditional* case, the generic element (conditional gamble) of \mathcal{D} is $X|B$, where X is a gamble and B is a non-impossible event. In the *unconditional* case, we shall simply term X the generic element (gamble) of \mathcal{D} .

The simplest non-trivial gamble is the *indicator* I_E of an event E . We shall not distinguish explicitly I_E and E , using the same symbol E for both. Thus we may speak of a set of events (of conditional events) \mathcal{D} , when for any $X \in \mathcal{D}$ (for any $X|B \in \mathcal{D}$), X is an (indicator of) event.

We recall the definition of dF -coherence for a precise prevision. In the sequel $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$.

Definition 1 Given $P : \mathcal{D} \rightarrow \mathbb{R}$, P is a (conditional) dF -coherent prevision on \mathcal{D} if, $\forall n \in \mathbb{N}^+$, $\forall X_1|B_1, \dots, X_n|B_n \in \mathcal{D}$, $\forall s_1, \dots, s_n \in \mathbb{R}$, defining

$$G = \sum_{i=1}^n s_i B_i (X_i - P(X_i|B_i)), \quad B = \bigvee_{i=1}^n B_i, \quad (1)$$

it holds that $\sup(G|B) \geq 0$.

If \mathcal{D} is made of unconditional gambles only, then (1) simplifies to

$$G = \sum_{i=1}^n s_i (X_i - P(X_i)) \quad (B = \Omega), \quad (2)$$

and consequently the coherence condition reduces to $\sup G \geq 0$.

The condition of dF -coherence allows a betting (or behavioural) interpretation, where $g_i = s_i(X_i - P(X_i))$ in (2) is an *elementary gain* with *stake* s_i . It represents the agent's gain from buying (if $s_i > 0$) or selling (if $s_i < 0$) $s_i X_i$ for $s_i P(X_i)$. Thus the condition $\sup G \geq 0$ requires that no finite combination of elementary gains produces an overall uniformly negative gain to the agent.

The other consistency concepts we recall here have a similar betting interpretation. Actually, they can be derived from dF -coherence simply by introducing constraints on the stakes s_i . Their definitions and a few basic properties are laid down below (for more on this topic see e.g. Pelessoni and Vicig (2005b, 2009); Troffaes and de Cooman (2014); Walley (1991); Williams (1975)). Prior to this, let us recall some properties of dF -coherent revisions to be employed later on.

Proposition 2 *If P is a dF -coherent revision on \mathcal{D} , then there exists a dF -coherent extension of P on any $\mathcal{D}' \supseteq \mathcal{D}$. Moreover, the following properties hold whenever their terms are defined:*

- (a) $P(aX + bY|B) = aP(X|B) + bP(Y|B)$, $\forall a, b \in \mathbb{R}$ (linearity).
- (b) $P(AX|B) = P(A|B)P(X|A \wedge B)$, $A \wedge B \neq \emptyset$ (product rule).

Definition 3 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be given. \underline{P} is a W -coherent lower revision on \mathcal{D} if, $\forall n \in \mathbb{N}$, $\forall X_0|B_0, X_1|B_1, \dots, X_n|B_n \in \mathcal{D}$, $\forall s_i \geq 0$, with $i = 0, 1, \dots, n$, defining*

$$\underline{G} = \sum_{i=1}^n s_i B_i (X_i - \underline{P}(X_i|B_i)) - s_0 B_0 (X_0 - \underline{P}(X_0|B_0)), \quad B = \bigvee_{i=0}^n B_i,$$

it holds that $\sup(\underline{G}|B) \geq 0$.

W -coherence was introduced in Williams (1975); the structure-free form in Definition 3 was employed in Pelessoni and Vicig (2009). In the unconditional case, it is equivalent to Walley's coherence (Walley, 1991, Section 2.5.4 (a)), while it includes (strictly) Walley's definition of coherence in (Walley, 1991, Section 7.1.4 (b)) in the conditional environment.

Proposition 4 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a W -coherent lower revision on \mathcal{D} . Then \underline{P} has a least-committal W -coherent extension \underline{E} on any $\mathcal{D}' \supseteq \mathcal{D}$, termed natural extension: $\underline{E} = \underline{P}$ on \mathcal{D} , and whatever is \underline{P}^* , W -coherent extension of \underline{P} on \mathcal{D}' , $\underline{E} \leq \underline{P}^*$ on \mathcal{D}' . Moreover, for $X|B, Y|B \in \mathcal{D}$,*

- (a) *If $X|B \leq Y|B$, then $\underline{P}(X|B) \leq \underline{P}(Y|B)$ (monotonicity).*
- (b) *$\underline{P}(X|B) \in [\inf(X|B), \sup(X|B)]$ (internality).¹*

Proposition 5 (Envelope theorem) *Given $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$, \underline{P} is a W -coherent lower revision on \mathcal{D} if and only if there exists a non-empty set \mathcal{P} of dF -coherent revisions on \mathcal{D} such that, $\forall X|B \in \mathcal{D}$, it holds that $\underline{P}(X|B) = \min\{P(X|B) : P \in \mathcal{P}\}$.*

Definition 6 *Given $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$,*

- (a) *\underline{P} is a convex lower revision on \mathcal{D} if, $\forall n \in \mathbb{N}^+$, $\forall X_0|B_0, X_1|B_1, \dots, X_n|B_n \in \mathcal{D}$, $\forall s_i \geq 0$, with $i = 1, \dots, n$, and $\sum_{i=1}^n s_i = 1$ (convexity condition), defining*

$$\underline{G}_c = \sum_{i=1}^n s_i B_i (X_i - \underline{P}(X_i|B_i)) - B_0 (X_0 - \underline{P}(X_0|B_0)), \quad B = \bigvee_{i=0}^n B_i,$$

it holds that $\sup(\underline{G}_c|B) \geq 0$.

1. Being also W -coherent, a dF -coherent revision satisfies properties (a), (b) too. Property (a) (monotonicity) also holds for a convex lower revision (Definition 6).

(b) \underline{P} is centered convex on \mathcal{D} if it is convex on \mathcal{D} and $\forall X|A \in \mathcal{D}, \emptyset|A \in \mathcal{D}$ and $\underline{P}(\emptyset|A) = 0$.

Proposition 7 (Envelope theorems with convex previsions) *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be given. Then*

- (a) (*Implicit Envelope Theorem, Pelessoni and Vicig (2005a)*) \underline{P} is a convex lower prevision on \mathcal{D} if and only if there exists a non-empty set \mathcal{P} of dF-coherent previsions such that $\forall X_0|B_0 \in \mathcal{D}, \exists P_{X_0|B_0} \in \mathcal{P} : \forall X|B \in \mathcal{D}$

$$\begin{aligned} P_{X_0|B_0}(B|B \vee B_0)(P_{X_0|B_0}(X|B) - \underline{P}(X|B)) &\geq \\ P_{X_0|B_0}(B_0|B \vee B_0)(P_{X_0|B_0}(X_0|B_0) - \underline{P}(X_0|B_0)). \end{aligned} \quad (3)$$

- (b) (*Envelope Theorem*) With unconditional lower previsions, \underline{P} is convex on \mathcal{D} if and only if there exist a non-empty set \mathcal{P} of dF-coherent previsions on \mathcal{D} and $\alpha : \mathcal{P} \rightarrow \mathbb{R}$ such that, $\forall X \in \mathcal{D}$, it holds that $\underline{P}(X) = \min\{P(X) + \alpha(P) : P \in \mathcal{P}\}$.

Moreover, \underline{P} is centered if and only if $\min\{\alpha(P) : P \in \mathcal{P}\} = 0$.

Next to lower previsions, upper previsions could be assessed. Customarily, one refers to just one type of previsions by the *conjugacy* relation: $\bar{P}(-X|B) = -\underline{P}(X|B)$. Using conjugacy, the consistency notions for lower previsions and their properties can be expressed for upper previsions.

The various gains we recalled $(G, \underline{G}, \underline{G}_c)$ are gambles themselves, being functions of a finite number of gambles in \mathcal{D} (and, in the conditional case, of indicators of their conditioning events). We mention next some other concepts regarding gains for later use.

Definition 8 Let \underline{G} be the gain in Definition 3.

Then $\mathcal{D}_{\underline{G}} = \{X_0|B_0, X_1|B_1, \dots, X_n|B_n\} \subseteq \mathcal{D}$ is the set of conditional gambles in \underline{G} .

The coarsest partition $\underline{G}|B$ is defined on B and is termed $\mathbb{P}_{\underline{G}}|B$. In other words, the atoms $\omega|B$ of $\mathbb{P}_{\underline{G}}|B$ correspond to the distinct jointly possible values of X_0, X_1, \dots, X_n that imply $B = \bigvee_{i=0}^n B_i$. We say that \underline{G} is a WDB gain if $\sup(\underline{G}|B) = 0$.

Analogous definitions apply to the other gains we considered (for instance, $\mathcal{D}_{\underline{G}_c}$ with \underline{G}_c).

Given a partition \mathbb{P} , the powerset of \mathbb{P} is called $\mathcal{A}(\mathbb{P})$. With a conditional gamble $X|B$, if X is defined on \mathbb{P} and $B \in \mathcal{A}(\mathbb{P}) \setminus \{\emptyset\}$, then $X|B$ is defined on $\mathbb{P}|B = \{\omega|B : \omega \in \mathbb{P}, \omega \Rightarrow B\}$.

3. Local Precision Properties of Weak Dutch Books

It is not difficult to obtain WDB gains, see the following simple example.

Example 1 Let $E \in \mathcal{D}$, with $\emptyset \neq E \neq \Omega$. Let $\underline{P}_1, \underline{P}_2 : \mathcal{D} \rightarrow \mathbb{R}$ be such that $\underline{P}_1(E) = 0, \underline{P}_2(E) = 1$. Then $\underline{P}_1, \underline{P}_2$ are coherent lower probabilities on $\{E\}$. Consider the gains $\underline{G}_1 = -s(E - \underline{P}_1(E)) = -sE$ and $\underline{G}_2 = s(E - \underline{P}_2(E)) = s(E - 1)$, with $s > 0$. Then $\max \underline{G}_1 = \underline{G}_1(\neg E) = 0 = \underline{G}_2(E) = \max \underline{G}_2$, that is $\underline{G}_1, \underline{G}_2$ are WDB gains associated with $\underline{P}_1, \underline{P}_2$, respectively.

In this section, we show that the existence of a WDB imposes some constraints both on convex and on coherent imprecise previsions, as for the gambles involved in the corresponding WDB gain.

Let us start with a convex lower prevision $\underline{P}(\cdot|\cdot)$. Its properties on those $\mathcal{D}_{\underline{G}_c}$ derived from WDB gains are investigated in Proposition 10.

Lemma 9 Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a conditional convex lower prevision, \underline{G}_c, B be as in Definition 6. Then, any dF-coherent prevision $P_{X_0|B_0}$ satisfying (3) is such that

$$\begin{aligned} P_{X_0|B_0}(\underline{G}_c|B) &= \\ \sum_{i=1}^n s_i P_{X_0|B_0}(B_i \vee B_0|B) [P_{X_0|B_0}(B_i|B_i \vee B_0)(P_{X_0|B_0}(X_i|B_i) - \underline{P}(X_i|B_i)) &\quad (4) \\ - P_{X_0|B_0}(B_0|B_i \vee B_0)(P_{X_0|B_0}(X_0|B_0) - \underline{P}(X_0|B_0))] &\geq 0. \end{aligned}$$

Proof Let $P_{X_0|B_0}$ satisfy (3). By Proposition 2 (a), recalling also that $\sum_{i=1}^n s_i = 1$, any dF-coherent extension of $P_{X_0|B_0}$ (still termed $P_{X_0|B_0}$) on a large enough set is such that

$$P_{X_0|B_0}(\underline{G}_c|B) = \sum_{i=1}^n s_i [P_{X_0|B_0}(B_i(X_i - \underline{P}(X_i|B_i))|B) - P_{X_0|B_0}(B_0(X_0 - \underline{P}(X_0|B_0))|B)]. \quad (5)$$

In general, we have that, for $i = 0, 1, \dots, n$,

$$\begin{aligned} P_{X_0|B_0}(B_i(X_i - \underline{P}(X_i|B_i))|B) &= P_{X_0|B_0}(B_i X_i|B - \underline{P}(X_i|B_i) B_i|B) \\ &= P_{X_0|B_0}(B_i X_i|B) - \underline{P}(X_i|B_i) P_{X_0|B_0}(B_i|B) \\ &= P_{X_0|B_0}(X_i|B_i \wedge B) P_{X_0|B_0}(B_i|B) - \underline{P}(X_i|B_i) P_{X_0|B_0}(B_i|B) \\ &= (P_{X_0|B_0}(X_i|B_i) - \underline{P}(X_i|B_i)) P_{X_0|B_0}(B_i|B), \end{aligned}$$

using Proposition 2 (a) at the second equality, Proposition 2 (b) at the third, and $B_i \wedge B = B_i$ at the fourth. Using Proposition 2 (b) again, we get also, for any $i = 1, \dots, n$,

$$\begin{aligned} P_{X_0|B_0}(B_i|B) &= P_{X_0|B_0}(B_i \vee B_0|B) P_{X_0|B_0}(B_i|B_i \vee B_0), \\ P_{X_0|B_0}(B_0|B) &= P_{X_0|B_0}(B_i \vee B_0|B) P_{X_0|B_0}(B_0|B_i \vee B_0). \end{aligned} \quad (6)$$

From (5), these derivations and (3), we obtain (4). ■

Proposition 10 Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a conditional convex lower prevision, \underline{G}_c, B be as in Definition 6 and such that \underline{G}_c is a WDB gain. Then, there exist a dF-coherent prevision $P_{X_0|B_0}$ satisfying (3) and $\alpha_{X_0|B_0} \in \mathbb{R}$ such that, for $i = 0$ and for any i such that $s_i > 0$ ($i = 1, \dots, n$), exactly one of the following holds

- (a) $P_{X_0|B_0}(B_i|B) = 0$;
- (b) $\underline{P}(X_i|B_i) = P_{X_0|B_0}(X_i|B_i) + \frac{\alpha_{X_0|B_0}}{P_{X_0|B_0}(B_i|B)}$.

Proof Take $X_0|B_0$ in \underline{G}_c . Let $P_{X_0|B_0}$ be the dF-coherent prevision in Proposition 7 (a) and define

$$\alpha_{X_0|B_0} = -P_{X_0|B_0}(B_0|B)(P_{X_0|B_0}(X_0|B_0) - \underline{P}(X_0|B_0)).$$

Since now $\sup(\underline{G}_c|B) = 0$, using Proposition 4 (b) and Footnote 1 at the first inequality, Lemma 9 at the second, we get $0 = \sup(\underline{G}_c|B) \geq P_{X_0|B_0}(\underline{G}_c|B) \geq 0$, i.e. $P_{X_0|B_0}(\underline{G}_c|B) = 0$.

Let now $s_i > 0$ ($i = 1, \dots, n$) such that $P_{X_0|B_0}(B_i|B) > 0$. Since $P_{X_0|B_0}(\underline{G}_c|B) = 0$, by Lemma 9 (the expression in square brackets in (4) is non-negative by (3)), recalling (6) at the second equality, we get

$$\begin{aligned} 0 &= P_{X_0|B_0}(B_i \vee B_0|B)[P_{X_0|B_0}(B_i|B_i \vee B_0)(P_{X_0|B_0}(X_i|B_i) - \underline{P}(X_i|B_i)) \\ &\quad - P_{X_0|B_0}(B_0|B_i \vee B_0)(P_{X_0|B_0}(X_0|B_0) - \underline{P}(X_0|B_0))] \\ &= P_{X_0|B_0}(B_i|B)(P_{X_0|B_0}(X_i|B_i) - \underline{P}(X_i|B_i)) + \alpha_{X_0|B_0}, \end{aligned}$$

hence (b). The case $i = 0$ follows immediately from the definition of $\alpha_{X_0|B_0}$. ■

Thus, for a convex \underline{P} , a WDB implies the existence of a dF -coherent prevision $P_{X_0|B_0}$ such that $\underline{P}(X_i|B_i)$ differs from $P_{X_0|B_0}(X_i|B_i)$ by a term $\frac{\alpha_{X_0|B_0}}{P_{X_0|B_0}(B_i|B)}$, for any $X_i|B_i \in D_{\underline{G}} \setminus \{X_0|B_0\}$ such that $P_{X_0|B_0}(B_i|B) \neq 0$. This latter constraint becomes irrelevant when \underline{P} is unconditional, since then $B_i = \Omega$, for $i = 0, 1, \dots, n$. Therefore $B = \Omega$ as well as $B_i|B = \Omega$ ($i = 0, 1, \dots, n$), hence $P_{X_0|B_0}(B_i|B) = P_{X_0|B_0}(\Omega) = 1$. Proposition 10 specialises then to:

Proposition 11 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be an unconditional convex lower prevision, and \underline{G}_c as in Definition 6 with $B_i = \Omega$, for $i = 0, 1, \dots, n$, be a WDB gain. Then there exist a dF -coherent prevision P_{X_0} on $\mathcal{D} \cup \{\underline{G}_c\}$ and $\alpha_{X_0} \in \mathbb{R}$ such that $\underline{P} = P_{X_0} + \alpha_{X_0}$ on $\mathcal{D}_{\underline{G}_c}^+ = \{X_0\} \cup \{X_i : s_i > 0, \text{ for } i = 1, \dots, n\}$.*

Hence, convexity of an unconditional lower prevision \underline{P} on \mathcal{D} implies that \underline{P} has a special structure on $\mathcal{D}_{\underline{G}_c}^+$, with WDBs: for each $X_i \in \mathcal{D}_{\underline{G}_c}^+$, \underline{P} differs from a dF -coherent prevision P by the same constant α_P . Perhaps surprisingly, if \underline{P} is centered convex, the preceding result does not imply that $\alpha_P = 0$ in all cases, but only if $\emptyset \in \mathcal{D}_{\underline{G}_c}^+$.

When \underline{P} is a conditional W -coherent prevision, Proposition 5 can be applied in the place of Proposition 7 (a). We get the following proposition (Corsato et al., 2017).

Proposition 12 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a W -coherent lower prevision, \underline{G}, B be as in Definition 3 and such that \underline{G} is a WDB gain. Suppose $B_1|B, \dots, B_n|B \in \mathcal{D}$. Define*

$$\mathcal{D}_{\underline{G}}^+ = \{X_0|B_0\} \cup \{X_i|B_i \in \mathcal{D}_{\underline{G}} : s_i \underline{P}(B_i|B) > 0, \text{ for } i = 1, \dots, n\}.$$

Then \underline{P} is dF -coherent on $\mathcal{D}_{\underline{G}}^+$.

The condition $B_1|B, \dots, B_n|B \in \mathcal{D}$ in Proposition 12 is not overly restrictive. If it is not met, we may consider a W -coherent extension \underline{P}' of \underline{P} on $\mathcal{D}' = \mathcal{D} \cup \{B_i|B : i = 1, \dots, n\}$ and apply Proposition 12 to \underline{P}' on \mathcal{D}' . However, the set on which \underline{P}' is dF -coherent depends then on the specific extension. It is minimal when the natural extension of \underline{P} is selected.

The result is subject to a second, more significant restriction. In fact, assuming $\underline{P}(B_i|B)$ positive is a sufficient but not necessary condition for dF -coherence of \underline{P} , i.e. \underline{P} may be dF -coherent on a set larger than $\mathcal{D}_{\underline{G}}^+$.

The important special case that \underline{P} is *unconditional*, i.e. $B_i = \Omega$, for $i = 0, 1, \dots, n$, hence $B = \Omega$, reinforces the result in Proposition 12. If all the stakes s_i , for $i = 1, \dots, n$, are non-zero, since necessarily $\underline{P}(\Omega) = 1$, we get $\mathcal{D}_{\underline{G}} = \mathcal{D}_{\underline{G}}^+$ and Proposition 12 specialises to the following statement.

Proposition 13 Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be an unconditional coherent lower prevision. Let \underline{G} be given by $\underline{G} = \sum_{i=1}^n s_i(X_i - \underline{P}(X_i)) - s_0(X_0 - \underline{P}(X_0))$, with $s_0 \geq 0$, $s_i > 0$, for $i = 1, \dots, n$, $\{X_0, X_1, \dots, X_n\} = \mathcal{D}_{\underline{G}} \subseteq \mathcal{D}$, and assume that \underline{G} is a WDB gain. Then \underline{P} is dF -coherent on $\mathcal{D}_{\underline{G}}$.

Thus a WDB implies that a coherent lower prevision is a dF -coherent prevision on $\mathcal{D}_{\underline{G}}$.

4. Further Features of Weak Dutch Books

Which are the agent's beliefs, with a WDB assessment, about suffering from a real *Dutch Book* (meaning a loss, if the gain has a maximum of zero, or a uniformly negative loss, if it does not achieve its supremum of zero)? In the simplest case, i.e. of a dF -coherent probability P , it was shown in (Crisma, 2006, Section 9.5.4) that $P(G < 0) = 0$. The generalisation to (unconditional) dF -coherent *previsions* has been investigated in Vicig (2010):

Proposition 14 Given a dF -coherent prevision P on \mathcal{D} , let the WDB gain G be as in (2). Then (any dF -coherent extension of) P is uniquely determined on certain events concerning G , and precisely:

- (a) $P(G \leq -\varepsilon) = 0, \forall \varepsilon > 0$;
- (b) if in addition X_1, \dots, X_n are all simple, we also have that $P(G < 0) = 0$.²

Thus, the results with precise previsions are rather reassuring. Take for instance case (b): although the agent cannot get any positive reward, whatever happens, she/he does not even believe that the bet will end up with a loss. However, the judgement on the potential vulnerability to Dutch books of a WDB assessment depends crucially on the kind of imprecise prevision being assessed.

In fact, the following result holds with W -coherent lower/upper previsions:

Proposition 15 Given a W -coherent lower prevision \underline{P} on \mathcal{D} , let \underline{G}, B be as in Definition 3 such that \underline{G} is a WDB gain. Then, for any W -coherent extension of \underline{P} (still termed \underline{P})

- (a) $\underline{P}(\underline{G}|B \leq -\varepsilon) = 0, \forall \varepsilon > 0$;
- (b) if $\mathcal{D}_{\underline{G}}$ is made of simple conditional gambles, $\underline{P}(\underline{G}|B < 0) = 0$.

Proposition 15, which includes also W -coherence for unconditional lower/upper previsions as a special instance, is formally analogous to Proposition 14. Yet, it replaces precise with lower previsions, as for the Dutch book evaluations. The upper probability of, say, $(\underline{G}|B < 0)$ in case (b) may well be even 1, as shown in Corsato et al. (2017). One may wonder whether it is at least *always* possible to put $\overline{P}(\underline{G}|B < 0) = 0$ or more generally (for an arbitrary $\underline{G}|B$) $\overline{P}(\underline{G}|B \leq -\varepsilon) = 0$, if wished. The answer is negative even in an unconditional environment:

Proposition 16 Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be an unconditional W -coherent lower prevision, and \underline{G} defined in Proposition 13, with $s_i > 0$ ($i = 1, \dots, n$) be a WDB gain. Then

- (a) if $\mathcal{D}_{\underline{G}} = \mathcal{D}$, it is coherent to put $\overline{P}(\underline{G} \leq -\varepsilon) = 0, \forall \varepsilon > 0$;
- (b) otherwise this choice may be incoherent.

2. The dF -coherent extension of P is mentioned explicitly because (the indicators of) the events $(G \leq -\varepsilon)$ and $(G < 0)$ need not belong to \mathcal{D} . Similar specifications will be omitted hereafter.

We may conclude that the (conditional) p -box of a WDB gain $\underline{G}|B$ for a W -coherent \underline{P} has a special structure, as for its lower distribution function $\underline{F}(x) = \underline{P}(\underline{G}|B \leq x)$, $x \in \mathbb{R}$. \underline{F} is a single-step function, identically equal to 0 for any $x < 0$, to 1 for any $x \geq 0$. On the contrary, the upper distribution function $\overline{F}(x) = \overline{P}(\underline{G}|B \leq x)$, $x \in \mathbb{R}$, is essentially unconstrained and need not coincide with $\underline{F}(x)$ if $(\underline{G}|B \leq x)$ is a non-trivial event.

Weaker notions than (W)-coherence may allow for even weaker implications about the non-occurrence of Dutch Books. In particular, in the case of centered convexity, the agent may have no strong belief that a Dutch Book associated with the gain \underline{G} will be avoided. In fact, examples may be built to show that not even $\underline{P}(\underline{G} \leq -\varepsilon)$ may be forced to be zero.

Summing up, when an uncertainty assessment incurs a WDB the agent's evaluation about avoiding a real Dutch Book depends on the degree of precision of the consistency notion the assessment satisfies. The self-protection offered by dF -coherence is maximal, whilst convexity does not ensure that $\underline{P}(\underline{G} \leq -\varepsilon)$ may be consistently set to zero.

Another facet of WDBs is concerned with conditions for the gain supremum of zero to be attained. Of course it is, if the gain involves only simple gambles, and in particular events. To explore this issue in more general situations the next result proves to be useful.

Proposition 17 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be an unconditional W -coherent lower prevision and \underline{G} as in Definition 3, with $B_i = \Omega$, for $i = 0, 1, \dots, n$. Let also \underline{G} be a WDB gain. Then, for any event $E \in \mathcal{D}$ with $\underline{P}(E) > 0$, it holds that $\sup(\underline{G}|E) = 0$.*

Now suppose that \mathcal{D} includes some atom $\omega \in \mathbb{P}_{\underline{G}}$, the coarsest partition \underline{G} is defined on. If $\underline{P}(\omega) > 0$, Proposition 17 implies (with $E = \omega$) that

$$\sup(\underline{G}|\omega) = \underline{G}(\omega) = 0,$$

hence \underline{G} achieves its supremum (at least) at ω . More generally, it holds that

Corollary 18 *Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$, \underline{G} be as in Proposition 17. Let $\mathbb{P} \subseteq \mathcal{D}$ be either $\mathbb{P}_{\underline{G}}$ or a partition finer than $\mathbb{P}_{\underline{G}}$, $e \in \mathbb{P}, \omega \in \mathbb{P}_{\underline{G}}$ be such that $e \Rightarrow \omega$ and $\underline{P}(e) > 0$. Then $\underline{G}(e) = \underline{G}(\omega) = 0$.*

Corollary 18 implies also that if $\sup \underline{G}$ is not achieved, then necessarily $\underline{P}(\omega) = 0$, $\forall \omega \in \mathbb{P}_{\underline{G}}$. Yet, there may be some $E \in \mathcal{D}$ such that $\underline{P}(E) > 0$, hence implying $\sup(\underline{G}|E) = 0$ by Proposition 17, but $E \notin \mathbb{P}_{\underline{G}}$. Letting $\mathcal{P} = \{\omega \in \mathbb{P}_{\underline{G}} : \underline{P}(\omega) > 0\}$ and $\mathcal{N} = \{\omega \in \mathbb{P}_{\underline{G}} : \underline{G}(\omega) = 0\}$, it is $\mathcal{P} \subseteq \mathcal{N}$, by Corollary 18. Thus the cardinality of \mathcal{P} is a lower bound to that of the set of atoms where \underline{G} achieves the value of zero. However, it is interesting to note that other causes may be influential too. In the next example the number of such atoms depends on the choice of the stakes.

Example 2 *Let $\mathcal{D} = \{E_1, E_0, \neg E_0, \neg E_1 \wedge E_0\}$, with $E_1 \wedge \neg E_0 = \emptyset$, $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be the vacuous lower probability and $\mathcal{D}_{\underline{G}} = \{E_1, E_0\}$. It may be checked that $\max \underline{G} = \max(s_1(E_1 - 0) - s_0(E_0 - 0)) = 0$ if $s_0 \geq s_1 > 0$. Here $\mathbb{P}_{\underline{G}} = \{E_1, \neg E_0, \neg E_1 \wedge E_0\}$, $\mathcal{P} = \emptyset$, while there are one or two atoms of $\mathbb{P}_{\underline{G}}$ where \underline{G} attains its maximum of zero, according to whether, respectively, $s_1 < s_0$ or $s_1 = s_0$. In fact $\underline{G}(E_1) = s_1 - s_0 \leq 0$ iff $s_1 \leq s_0$, in particular $\underline{G}(E_1) = 0$ iff $s_1 = s_0$, $\underline{G}(\neg E_0) = 0$, $\underline{G}(\neg E_1 \wedge E_0) = -s_0 < 0$.*

Testing Weak Dutch Books. If it is not known whether, given a coherent \underline{P} , $\sup \underline{G} = 0$ or not, we can try to rule out the possibility of a WDB by checking the sign of \underline{G} at some $\omega \in \mathbb{P}_{\underline{G}}$ such that $\underline{P}(\omega) > 0$ (if any). In fact:

- if $\underline{G}(\omega) > 0$, then obviously $\sup \underline{G} > 0$;
- if $\underline{G}(\omega) < 0$, then $\sup \underline{G} > 0$ by Corollary 18.

This method is very simple, but allows no conclusion when $\underline{G}(\omega) = 0$.³ In fact, it is clearly possible that $\underline{G}(\omega) = 0$ and $\sup \underline{G} = 0$, but even when $\underline{G}(\omega) = 0$ for all $\omega \in \mathbb{P}_{\underline{G}}$ such that $\underline{P}(\omega) > 0$, $\sup \underline{G}$ may be strictly positive.

5. Strict Consistency

As soon as the behavioural interpretation of de Finetti's theory of subjective probabilities became more widespread, the issue of whether WDBs could possibly be avoided was investigated. Thus, as early as the mid-fifties of the last century Kemeny (1955) and Shimony (1955) proposed the most immediate solution: rule out WDBs by redefining coherence. They referred to (precise) probabilities only, replacing the condition $\sup G \geq 0$ with $\sup G > 0$, for any admissible $G \neq 0$, which is what is called *strict coherence* today. As well-known, strict coherence has non-negligible drawbacks, like that of being confined to a denumerable environment. Alternative ways of tackling WDBs have also been developed. We discuss in Corsato et al. (2017) that due to Wagner (2007) and based on the interpretation of buying/selling prices in betting schemes, going back to Walley (1991), and the one resorting to desirability concepts (see e.g. Quaeghebeur (2014)). Alternatively, Pedersen (2014) introduces a qualitative version of strict coherence for comparing (not necessarily bounded, unconditional) gambles.

However, little has been said about the role and properties of (some extended version of) strict coherence with imprecise rather than precise probabilities. It is relatively simple to extend the strict coherence approach (Corsato et al., 2017):

Definition 19 Let $\mu : \mathcal{D} \rightarrow \mathbb{R}$ be a measure, whose consistency requires that $\sup(G|B) \geq 0$ for any conditional gain $G|B$ admissible according to certain rules. Then μ is strictly consistent if, for each such $G|B$, either $G|B = 0$ or $\sup(G|B) > 0$.

As for the issue of characterising strict consistency, the case of conditional coherence was hinted in Williams (1975) and is tackled for W -coherent previsions in the next proposition.

Proposition 20 Let $\underline{P} : \mathcal{D} \rightarrow \mathbb{R}$ be a W -coherent lower prevision. Then,

(a) If \underline{P} is strictly W -coherent on \mathcal{D} ,

$$\underline{P}(A|B) > 0, \text{ for all events } A|B \in \mathcal{D}, A|B \neq \emptyset|B. \quad (7)$$

(b) If \underline{P} is not strictly W -coherent on \mathcal{D} and, for any WDB gain $\underline{G}|B \neq 0$ as in Definition 8, $\exists \varepsilon > 0 : (\underline{G}|B \leq -\varepsilon) \in \mathcal{D}$ is non-impossible, then $\exists A|B \in \mathcal{D}, A|B \neq \emptyset|B : \underline{P}(A|B) = 0$.

Clearly, Proposition 20 concerns the special case of (unconditional) coherent lower previsions too. For these previsions, (7) requires *strict positivity* (*sP*) of any non-impossible event in \mathcal{D} . It can be seen that this again limits the effectiveness of strict coherence to denumerable settings. Turning to another question, in today's language the necessary and sufficient condition for strict dF -coherence

3. Alternatively, linear programming could potentially be employed in some special cases.

of a dF -coherent probability in Kemeny (1955) asks instead for *strict normalisation* (sN), i.e., that $P(E) < 1$, for any $E \neq \Omega$. In the realm of dF -coherent probabilities, this is obviously equivalent to (sP) . However, as often happens, the equivalence in a precise framework conceals a more complex situation in the field of imprecision. To see this, consider the following conditions for an uncertainty measure μ on a domain \mathcal{D} of events:

- $$\begin{aligned} (sM) \quad & \forall E, F \in \mathcal{D}, \text{ if } E \Rightarrow F, E \neq F \text{ then } \mu(E) < \mu(F) && (\text{strict Monotonicity}); \\ (sP) \quad & \forall E \in \mathcal{D}, \text{ if } E \neq \emptyset \text{ then } \mu(E) > 0 && (\text{strict Positivity}); \\ (sN) \quad & \forall E \in \mathcal{D}, \text{ if } E \neq \Omega \text{ then } \mu(E) < 1 && (\text{strict Normalisation}). \end{aligned}$$

Then, it holds that:

Proposition 21 *Let \mathcal{A} be an algebra of events (i.e., $\forall E \in \mathcal{A}, \neg E \in \mathcal{A}, \forall E, F \in \mathcal{A}, E \wedge F \in \mathcal{A}$), and let $\underline{P} : \mathcal{A} \rightarrow \mathbb{R}$ be a lower probability.*

- $$\begin{aligned} (a) \quad & \text{If } \underline{P} \text{ is W-coherent, then} && (sM) \Leftrightarrow (sP) \Rightarrow (sN), \\ & \text{while } (sN) \text{ implies neither of } (sM), (sP). \\ (b) \quad & \text{If } \underline{P} \text{ is centered convex, then} && (sM) \Rightarrow (sP) \Rightarrow (sN), \\ & \text{while } (sP) \text{ does not imply } (sM), \text{ nor does } (sN) \text{ imply } (sP). \end{aligned}$$

Proof Recall that $\emptyset, \Omega \in \mathcal{A}$ and that for any $E \in \mathcal{A}, \emptyset \Rightarrow E \Rightarrow \Omega$.

Proof of (a). $(sM) \Leftrightarrow (sP)$. Let (sM) hold, and let $E \in \mathcal{A}, E \neq \emptyset$. We have $\underline{P}(E) > \underline{P}(\emptyset) = 0$. Assume now (sP) is satisfied. Let $E, F \in \mathcal{A}$ such that $E \Rightarrow F$ and $E \neq F$. We have $\neg E \wedge F \in \mathcal{A} \setminus \{\emptyset\}$. Since then $\underline{P}(\neg E \wedge F) > 0$ by (sP) , by superlinearity we get $\underline{P}(E) < \underline{P}(E) + \underline{P}(\neg E \wedge F) \leq \underline{P}(F)$.

$(sP) \Rightarrow (sN)$. By the previous step, it is equivalent to $(sM) \Rightarrow (sN)$, which holds: taking $E \neq \Omega$, by (sM) $\underline{P}(E) < \underline{P}(\Omega) = 1$.

$(sN) \not\Rightarrow (sP)$. Let us consider a non-trivial event E , $\mathcal{A} = \{\emptyset, E, \neg E, \Omega\}$ and $\underline{P} : \mathcal{A} \rightarrow \mathbb{R}$ given by $\underline{P}(\emptyset) = \underline{P}(E) = 0$, $\underline{P}(\neg E) = \varepsilon$, for some $\varepsilon \in]0, 1[$, $\underline{P}(\Omega) = 1$. Then \underline{P} is a coherent lower probability on \mathcal{A} satisfying (sN) but not (sP) (nor its equivalent condition (sM)).

Proof of (b). $(sM) \Rightarrow (sP)$: same as in the proof of (a).

$(sP) \not\Rightarrow (sM)$. Can be shown by means of a (counter)example. For this, let $\emptyset \neq E \Rightarrow F \neq \Omega$, $E \neq F$. Thus $\mathbb{P} = \{E, \neg E \wedge F, \neg F\}$ is a partition. Let $\mathcal{A} = \mathcal{A}(\mathbb{P})$, and $\underline{P} = \min\{P_1, P_2 + 0.2\}$, where P_1, P_2 are defined in Table 1 (only the relevant events in \mathcal{A} are displayed).

	\emptyset	E	$\neg E \wedge F$	$\neg F$	F
P_1	0	0.1	0.2	0.7	0.3
P_2	0	0	0	1	0
$P_2 + 0.2$	0.2	0.2	0.2	1.2	0.2
\underline{P}	0	0.1	0.2	0.7	0.2

Table 1: Data for the (counter)example.

- \underline{P} is centered convex on \mathcal{A} (by Proposition 7), but not coherent: $\underline{P}(E) + \underline{P}(\neg E \wedge F) > \underline{P}(F)$, thus \underline{P} does not comply with superadditivity.

- \underline{P} satisfies (sP) (on the events in \mathbb{P} and hence, by monotonicity of convex lower previsions, on all events in \mathcal{A}).
- \underline{P} does not satisfy (sM) : $\neg E \wedge F \Rightarrow F$ (and $\neg E \wedge F \neq F$), while $\underline{P}(\neg E \wedge F) = \underline{P}(F) = 0.2$.

$(sP) \Rightarrow (sN)$. Let $\underline{P}(E) > 0$, $\forall E \neq \emptyset$. Recall that a centered convex lower prevision avoids sure loss (Pelessoni and Vicig, 2005b), and as such satisfies the inequality $\underline{P}(X) + \underline{P}(\mu - X) \leq \mu$, $\forall \mu \in \mathbb{R}$ (Walley, 1991, Section 2.4.7 (c)). Putting $X = E$, $\mu = 1$, the inequality boils down to $\underline{P}(E) + \underline{P}(\neg E) \leq 1$. This implies $\underline{P}(E) < 1$ if $E \neq \Omega$ since then $\underline{P}(\neg E) > 0$ by assumption.

$(sN) \not\Rightarrow (sP)$. Indeed the implication is not valid under the stronger assumption that \underline{P} is coherent, as proven in (a). ■

Comments. When $\mu = P$, a dF -coherent probability, in (sM) , (sP) , (sN) and P is defined on an algebra \mathcal{A} , then $(sM) \Leftrightarrow (sP) \Leftrightarrow (sN)$, by normalisation and linearity of P .

We may summarise the situation in the next figure (only valid implications are displayed).

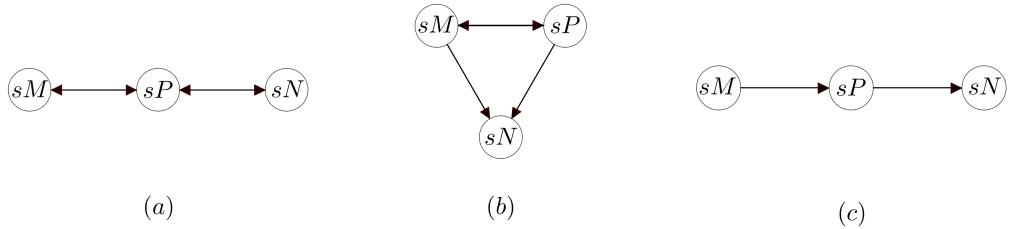


Figure 1: Comparison among the conditions (sM) , (sP) , (sN) for either a dF -coherent probability in case (a), or a lower probability which is coherent (b) or centered convex (c).

It is clear from Proposition 21 why Proposition 20 refers to strict positivity instead of strict normalisation as in Kemeny (1955): even in an unconditional frame, strict positivity is tighter. To put it differently, requiring (sN) does not prevent \underline{P} from incurring a WDB: it suffices that there is a possible E with $\underline{P}(E) = 0$ to which the WDB gain $\underline{G} = -s_0 E \leq 0$ is associated. Instead, (sM) could replace (sP) on algebras, while on more general domains (sP) is easier to work with. However, (sM) remains the only relevant condition for strict convexity.

Interestingly, these relationships may change with upper probabilities. Thus, when \overline{P} is a coherent upper probability on an algebra \mathcal{A} , (sN) and (sP) exchange their roles. Using the conjugacy relation $\underline{P}(E) = 1 - \overline{P}(\neg E)$, we deduce that

$$(sM) \Leftrightarrow (sN) \Rightarrow (sP), \quad (sP) \not\Rightarrow (sM), \quad (sP) \not\Rightarrow (sN).$$

6. Conclusions

In this paper the properties of assessments incurring WDBs have been explored by their degree of consistency. The results point out a certain differentiation and a number of perhaps surprising features of such assessments. By contrast, the more known special case of dF -coherent precise

probabilities often flattens these differences. The situation is similar for strict consistency, the generalisation of strict coherence that avoids WDBs, even though its domain of application remains restricted even with W -coherence. Thus, in general WDBs are something to coexist with.

Acknowledgements

R. Pelessoni and P. Vicig acknowledge partial support by the FRA2015 grant ‘Mathematical Models for Handling Risk and Uncertainty’.

References

- C. Corsato, R. Pelessoni, and P. Vicig. Weak Dutch Books with imprecise previsions. *International Journal of Approximate Reasoning*, 88:72–90, 2017.
- L. Crisma. *Introduzione alla teoria delle probabilità coerenti*. EUT, 2006.
- B. de Finetti. *Theory of Probability*. Wiley, 1974.
- J. G. Kemeny. Fair bets and inductive probabilities. *J. Symb. Logic*, 20:263–273, 1955.
- A. P. Pedersen. Comparative Expectations. *Stud. Logica*, 102:811–848, 2014.
- R. Pelessoni and P. Vicig. Envelope theorems and dilation with convex conditional previsions. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *Proc. ISIPTA ’05*, pages 266–275. SIPTA, 2005a.
- R. Pelessoni and P. Vicig. Uncertainty modelling and conditioning with convex imprecise previsions. *Int. J. Approx. Reason.*, 39:297–319, 2005b.
- R. Pelessoni and P. Vicig. Williams coherence and beyond. *Int. J. Approx. Reason.*, 50(4):612–626, 2009.
- E. Quaeghebeur. Desirability. In T. Augustin, F. P. A. Coolen, G. d. Cooman, and M. C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 1–27. J. Wiley & Sons, 2014.
- A. Shimony. Coherence and the axioms of confirmation. *J. Symb. Logic*, 20:1–28, 1955.
- M. C. M. Troffaes and G. de Cooman. *Lower Previsions*. Wiley, 2014.
- P. Vicig. A gambler’s gain prospects with coherent imprecise previsions. In E. Hüllermeier, R. Kruse, and F. Hoffmann, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, pages 50–59. Springer, 2010.
- C. G. Wagner. The Smith-Walley Interpretation of Subjective Probability: An Appreciation. *Stud. Logica*, 86:343–350, 2007.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
- P. M. Williams. *Notes on conditional previsions. Research Report*. School of Math. and Phys. Science, University of Sussex, 1975.

Reconciling Bayesian and Frequentist Tests: the Imprecise Counterpart

Inés Couso

Statistics and O.R.

Universidad de Oviedo

Gijón (Spain)

COUSO@UNIOVI.ES

Antonio Álvarez-Caballero

Metrology and Models

Universidad de Oviedo

Gijón (Spain)

ANALCA3@GMAIL.COM

Luciano Sánchez

Computer Sciences and A.I.

Universidad de Oviedo

Gijón (Spain)

LUCIANO@UNIOVI.ES

Abstract

Imprecise Dirichlet Process-based tests (IDP-tests, for short) have been recently introduced in the literature. They overcome the problem of deciding how to select a single prior in Bayesian hypothesis testing, in the absence of prior information. They make use of a “near-ignorance” model, that behaves a priori as a vacuous model for some basic inferences, but it provides non-vacuous posterior inferences. We perform empirical studies regarding the behavior of IDP-tests for the particular case of Wilcoxon rank sum test. We show that the upper and lower posterior probabilities can be expressed as tail probabilities based on the value of the U statistic. We construct an imprecise frequentist-based test that reproduces the same decision rule as the IDP test. It considers a neighbourhood around the U -statistic value. If all the values in the neighbourhood belong to the rejection zone (resp. to the acceptance region), the null hypothesis is rejected (resp. accepted). Otherwise, the judgement is suspended. This construction puts a step forward in the reconciliation between frequentist and Bayesian hypothesis testing.

Keywords: Wilcoxon rank sum test; imprecise tests; one-sided test; frequentist test; Bayesian test; IDP test; interval p-values.

1. Introduction

The problem of reconciling Bayesian and frequentist techniques has been extensively treated in the literature and seems to be still open. In the frequentist setting, the level of significance of the outcome against the null hypothesis is determined in terms of the p-value. Notwithstanding the “probability that the null hypothesis is true” has no meaning in this framework, but it has been argued that some practitioners attach such a meaning to the p-value (see [Casella and Berger \(1987\)](#) for further discussion). Alternatively, under the Bayesian approach, evidence takes the form of the posterior probability about the null hypothesis, based on the combination of prior evidence and the evidence provided by the dataset. The relation between the p-value and the posterior probability of the null hypothesis has been examined by different authors (see [Berger and Selke \(1987\)](#); [Casella and Berger \(1987\)](#); [DeGroot \(1973\)](#); [Pratt \(1965\)](#); [Shafer \(1982\)](#); [Jeffreys \(1939\)](#) among many oth-

ers). For two-sided tests, it has been noticed by several of them that the p-value tends to be smaller than the posterior probability of the null hypothesis (see Berger and Selke (1987); Lindley (1957)) for some collections of priors, while for the one-sided testing problem situations can be found where they are approximately equal (see Pratt (1965); Casella and Berger (1987)). In particular, Casella and Berger (1987) prove that for some classes of reasonable and impartial priors, and under some additional requirements about the distribution of X , the p-value coincides with the infimum of the posterior probability of the null hypothesis. With respect to the large discrepancies between the infimum for the posterior probability and the p-value in the two-sided problem observed in Berger and Selke (1987), Casella and Berger (1987) question the impartiality of the priors considered by the authors. The problem of selecting an appropriate prior (specially in those cases where no initial information is available) has been a subject of study of many authors. One solution to this problem has been proposed initially by Ferguson (1973) and afterwards by Rubin (1981) under the name of Bayesian Bootstrap (BB). Notwithstanding, the BB model cannot be regarded as non-informative, since it assigns zero probability to any set that does not include the observations. In order to overcome this issue, Benavoli et al. (2015) introduced a new kind of test, by means of replacing a single prior by a collection of priors based on the imprecise Dirichlet process (IDP). The combination of this near-to-ignorance prior information with our evidence obtained from the sample leads to a pair of dual upper and lower posterior probabilities. The IDP-based test has the advantage of not deciding when this decision is somehow prior-dependent. In other words, when the action that minimizes the risk (expected loss) is not the same for all the prior probabilities, the IDP suspends its judgment. The authors have exemplified their proposal with an IDP-based version of the well known Wilcoxon rank sum test, also called the Mann-Withney-Wilcoxon test, or simply, the MWW test (Mann and Whitney (1947); M.P. Fay (2010)).

Consider two random variables X and Y whose cdf's satisfy $F_X(x) = F_Y(x + \Delta)$, $\forall x \in \mathbb{R}$. The null hypothesis of the traditional MWW test is that $P(X \leq Y) \leq 0.5$ against the alternative hypothesis $P(X \leq Y) > 0.5$. When the distribution of $X - Y$ is continuous, we can interpret a significant Mann-Whitney-Wilcoxon test as showing that the median of the difference is negative (Couso et al. (2015)). The IDP-based procedure will assign a pair of upper and lower probabilities to the null hypothesis, $\bar{P}(H_0|(\vec{x}, \vec{y}))$ and $\underline{P}(H_0|(\vec{x}, \vec{y}))$, that encompass the collection of posterior probabilities associated to the selected collection of priors. The authors propose the following decision rule, for some threshold $\gamma \in (0, 1)$:

- If both the upper and the lower posterior probabilities are on one side of the threshold γ , we will either reject (left side) or accept (right side) the null hypothesis.
- Alternatively, if they satisfy the inequalities $\underline{P}(H_0|(\vec{x}, \vec{y})) < \gamma \leq \bar{P}(H_0|(\vec{x}, \vec{y}))$, then we are in an indeterminate decision, i.e, we suspend our judgement.

After presenting their new proposal, the authors have performed some empirical comparisons with respect to the Bayesian bootstrap-based test as well as with the traditional frequentist MWW test, under different conditions for the shift parameter Δ . They suggest that when the IDP based test is indeterminate, both the frequentist and the Boostrap Bayesian test behave as “random guessers”. What they check in fact is that, for some values of Δ , the proportion of rejections under those situations is nearly 50%, which coincides with the proportion of rejections of a randomized test derived from the IDP test (the one called the 50/50 test by the authors) that returns the same response as the IDP test when it is determinate, and a random answer otherwise.

Our paper deepens the study about the relations between this new imprecise test and its precedents. We will first empirically check that the p-value of the traditional MWW test coincides with the posterior probability of the null hypothesis for the BB test. Afterwards, we will show that there is a one-to-one correspondence between the upper (resp. the lower) posterior probability of the IDP test and the p-value of the MWW test. Thus the outcome of the latter is univocally determined by the upper (equivalently, by the lower) posterior probability of the former. In fact, upper and lower posterior probabilities derived from the IDP-based approach can be calculated in terms of the cdf of $U + \epsilon$ and $U - \epsilon$, for some $\epsilon > 0$, where U represents the MWW statistic. On the basis of this relation, we construct an imprecise frequentist-based test whose performance mimics the one of the IDP-based test. These findings help us to better understand the behaviour of the new IDP-based test, and put a step forward in the reconciliation between the frequentist and Bayesian approaches in this imprecise setting. In particular, this kind of imprecision over the set of priors seems to produce similar effects on the decision mechanism as an imprecision of data around the observations.

2. Preliminaries

The Mann-Withney U test (also called Wilcoxon rank sum or Mann-Whitney-Wilcoxon test) is used to check whether or not it is equally likely that a randomly selected value from one population will be less than or greater than a randomly selected value from a second one, assuming that both selections are independent from each other.

Consider two independent samples containing n_1 and n_2 elements respectively from each population. The U statistic is calculated as the sum of the ranks of the elements contained in the first sample, with the minimum value $n_1(n_1 + 1)/2$ subtracted. In other words, it counts the number of items (x_i, y_j) such that x_i is less than or equal to y_j ,

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{[X_i, \infty)}(Y_j).$$

Under the assumption $P(X \leq Y) = 0.5$, the expectation and the variance of U are respectively:

$$\mu_0 = \frac{n_1 n_2}{2} \quad \sigma_0^2 = \frac{n_1 n_2 (n_1 + n_2)}{12}.$$

Let us consider the one-sided test of $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$, where $\theta = P(X \leq Y)$. The rejection region of the Mann-Whitney U test of size α is defined in terms of U as follows:

$$R_\alpha = \left\{ (\vec{x}, \vec{y}) : \frac{U(\vec{x}, \vec{y}) - \mu_0}{\sigma_0} > z_\alpha \right\},$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the quantile $1 - \alpha$ of the distribution $N(0, 1)$. Alternatively, it can be defined as:

$$R_\alpha = \{(\vec{x}, \vec{y}) : p(\vec{x}, \vec{y}) < \alpha\},$$

where

$$p(\vec{x}, \vec{y}) = 1 - \Phi \left(\frac{U(\vec{x}, \vec{y}) - \mu_0}{\sigma_0} \right)$$

denotes the p-value of the sample, i.e.,

$$p(\vec{x}, \vec{y}) = \inf\{\alpha \in (0, 1) : (\vec{x}, \vec{y}) \in R_\alpha\}.$$

Under the Bayesian approach, the problem of hypothesis testing is seen as a decision problem where the possible actions are $a = 0$ (accept H_0) and $a = 1$ (reject H_0). We start from a prior distribution over the set parametric space $\Theta = [0, 1]$, determined by a density function $\pi : \Theta \rightarrow \mathbb{R}^+$. A loss function $\ell : \Theta \times \{0, 1\} \rightarrow \mathbb{R}$ links the action to the unknown value of the parameter $\theta = P(X \leq Y)$: when the true state of nature is $\theta \in \Theta$ and we take the action $a \in \{0, 1\}$ we incur in a loss $\ell(\theta, a)$ determined as follows:

	$a = 0$	$a = 1$
$\theta \leq 0.5$	0	K_0
$\theta \geq 0.5$	K_1	0

The decision rule d that minimizes the posterior expected loss is the one defined as follows:

- $d(\vec{x}, \vec{y}) = 1$ if $P(H_0 | (\vec{x}, \vec{y})) < \frac{K_0}{K_0 + K_1}$
- $d(\vec{x}, \vec{y}) = 0$ otherwise,

where $P(H_0 | (\vec{x}, \vec{y}))$ denotes the posterior probability of the null hypothesis calculated as follows:

$$P(H_0 | (\vec{x}, \vec{y})) = \int_{-\infty}^{0.5} L(\vec{x}, \vec{y}; \theta) \pi(\theta) d\theta,$$

and $L(\vec{x}, \vec{y}; \theta)$ represents the likelihood function.

The Dirichlet process was proposed by [Ferguson \(1973\)](#) as a second-order probability (in our context, a probability on the space of joint probability distributions for (X, Y)). Since every joint distribution determines a specific value for $\theta = P(X \leq Y)$, a Dirichlet process determines a (prior) probability distribution over the parametric space, Θ . But how do we choose this prior in case of lack of information? [Rubin \(1981\)](#) addressed this problem by means of selecting the so-called Bayesian bootstrap. It is the Bayesian analogue to the Efron's bootstrap [Efron \(1979\)](#). Instead of simulating the sampling distribution of a statistic estimating a parameter, it simulates the posterior distribution of the parameter. This choice nevertheless seems controversial (see [Rubin \(1981\)](#) and [Benavoli et al. \(2015\)](#) for detailed discussions), since it cannot be seen as a representation of a lack of knowledge. In fact, the Bayesian bootstrap assigns probability one to the collection of observations (see [Rubin \(1981\)](#)). To overcome this issue, [Benavoli et al. \(2015\)](#) proposed to use the imprecise Dirichlet process (IDP). It is considered as a prior near-ignorance model. In fact, it corresponds to a set of priors that generates vacuous prior probabilities and therefore, leading to an infimum and a supremum for the (prior) expectations of $\theta = P(X \leq Y)$ respectively equal to 0 and 1. This collection of priors leads to a collection of posterior probabilities for H_0 and H_1 , given the dataset, whose bounds we will respectively denoted by $\underline{P}(H_0 | (\vec{x}, \vec{y}))$ and $\overline{P}(H_0 | (\vec{x}, \vec{y}))$. To perform the hypothesis test $H_0 : \theta \leq 0.5$ against $H_1 : \theta > 0.5$, they compare each of these bounds with $\gamma = \frac{K_0}{K_0 + K_1}$ and consider the following decision rule:

- $d_I(\vec{x}, \vec{y}) = 1$ if $\overline{P}(H_0 : (\vec{x}, \vec{y})) < \frac{K_0}{K_0 + K_1}$
- $d_I(\vec{x}, \vec{y}) = 0$ if $\underline{P}(H_0 : (\vec{x}, \vec{y})) > \frac{K_0}{K_0 + K_1}$
- $d_I(\vec{x}, \vec{y}) = ?$ otherwise,

where “0”, “1” and “?” respectively denote “accept H_0 ”, “reject H_0 ” and “no decision”.

3. Relations between p-value, posterior probability and upper and lower posterior probabilities

3.1 Formal relations between p-value and Bayesian posterior probability

As mentioned in the Introduction, different authors have studied the relations between the frequentist p-value and the Bayesian posterior probability of the null hypothesis. Casella and Berger (1987) studied this relation for one-sided tests under some additional conditions: In particular, when the underlying distribution is assumed to be symmetric and it satisfies the property of monotone likelihood ratio (MLR), then the p-value coincides with the infimum of the set of posterior probabilities for H_0 , for several reasonable collections of priors.

We can prove an additional result relating the p-value and the posterior probability of the null hypothesis derived from any prior. It requires a MLR condition but it does not require any symmetry about the underlying distribution.

Definition 1 *The set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T if we can represent the likelihood ratio as*

$$\frac{L(\vec{x}, \vec{y}; \theta_1)}{L(\vec{x}, \vec{y}; \theta_2)} = g_{\theta_1, \theta_2}(T(\vec{x}, \vec{y})),$$

where g_{θ_1, θ_2} is strictly increasing for every pair $\theta_1 > \theta_2$.

Now we will prove that, when the family of distributions satisfies the MLR property with respect to some statistic T , the posterior probability associated to a one-sided-test is increasing wrt T :

Lemma 2 *Let us suppose that the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T and let us consider the test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then the posterior probability $P(H_0 | \vec{x})$ can be expressed as an increasing function of $T(\vec{x})$, i.e.:*

$$T(\vec{x}) < T(\vec{x}') \Rightarrow P(H_0 | \vec{x}) < P(H_0 | \vec{x}').$$

The following result is well known in the literature:

Theorem 3 *Assume the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T . Let us consider the one-sided test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then the test $\delta : \mathbb{R}^n \rightarrow \{0, 1\}$ defined as follows:*

$$\delta(\vec{x}) = \begin{cases} 0 & \text{if } T(\vec{x}) \leq c \\ 1 & \text{if } T(\vec{x}) > c \end{cases}$$

is a uniformly most powerful (UMP) test (among all the tests of size $\alpha_c = P_{\theta_0}(T > c)$).

We deduce the following result:

Theorem 4 *Let us suppose that the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T and that the cdf of T is strictly increasing for some θ_0 . Let us consider the test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Let us consider the family of UMP tests associated to the rejection regions $\{R_\alpha : \alpha \in (0, 1)\}$, each of them defined as $R_\alpha = \{\vec{x} :$*

$T(\vec{x}) > c_\alpha\}$, with $P_{\theta_0}(T > c_\alpha) = \alpha$. Let us consider the p-value associated to this family of tests as follows:

$$p(\vec{x}) = \inf\{\alpha : \vec{x} \in R_\alpha\} = P_{\theta_0}(T > T(\vec{x})), \forall \vec{x}. \quad (1)$$

Let us consider an arbitrary prior over Θ . Then there exists a (one-to-one) strictly increasing function $h : [0, 1] \rightarrow [0, 1]$ linking the posterior probability of H_0 and the p-value as $P(H_0|x) = g(p(\vec{x}))$, $\forall \vec{x}$, and therefore

$$p(\vec{x}) < p(\vec{x}') \Leftrightarrow P(H_0|\vec{x}) < P(H_0|\vec{x}').$$

As a consequence we can state the following corollary:

Corollary 5 Let us suppose that the set of distributions $\{P_\theta : \theta \in \Theta\}$ has a monotone likelihood ratio (MLR) with respect to a statistic T and that the cdf of T is strictly increasing for some θ_0 . Let us consider the test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Let us consider the family of UMP tests associated to the rejection regions $\{R_\alpha : \alpha \in (0, 1)\}$, each of them defined as $R_\alpha = \{\vec{x} : T(\vec{x}) > c_\alpha\}$, with $P_{\theta_0}(T > c_\alpha) = \alpha$. Let us consider an arbitrary prior over Θ , an arbitrary pair of loss values K_0 and K_1 , and the Bayesian test associated to it. Then there exists a UMP frequentist test that coincides with it, the size of it being an increasing function of $\gamma = \frac{K_0}{K_0 + K_1}$.

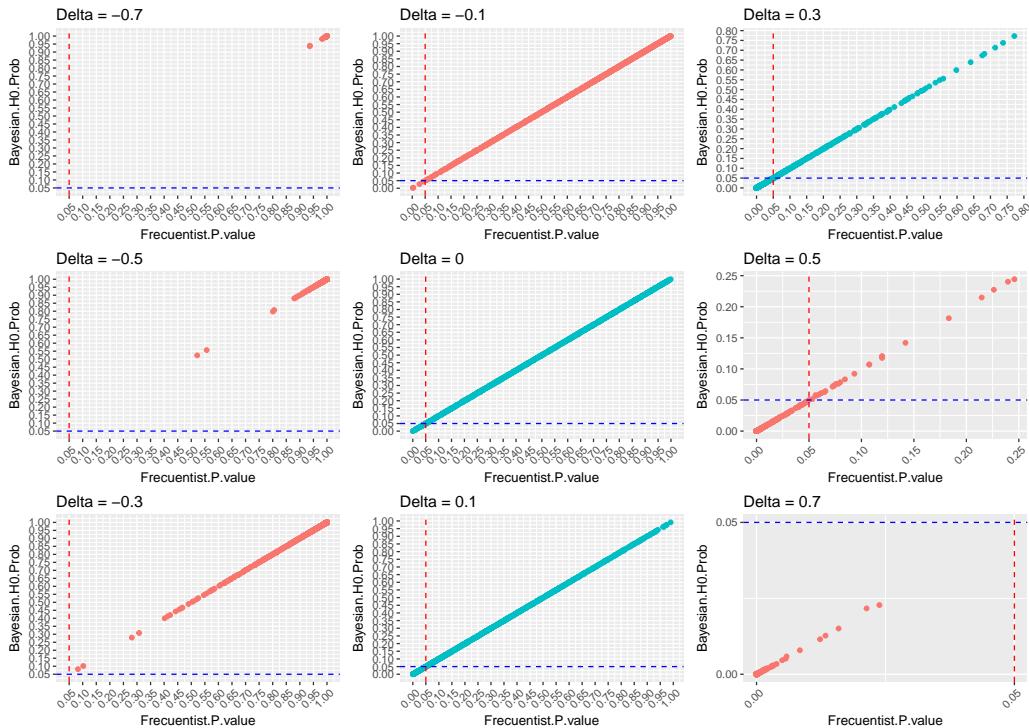
According to the above result, under the condition of MLR, and regardless the prior distribution we select, there exists a one-to-one correspondence between γ and α . This is to say, if we set an arbitrary prior, there exists a bijection $h : [0, 1] \rightarrow [0, 1]$ such that the Bayesian test associated to $\gamma = \frac{K_0}{K_0 + K_1}$ coincides with the UMP test of size $\alpha = h(\gamma)$. The next section deals with the particular case of the MWYW and its variations considered in [Benavoli et al. \(2015\)](#). In that particular case, this one-to-one correspondence is the identity, i.e., the p-value coincides with the posterior probability of the null hypothesis. Furthermore, we empirically show that the upper and lower posteriors can be also calculated as strictly increasing functions of the p-value.

3.2 Relations between the p-value and the pair of upper and lower posterior probabilities: an empirical study

[Benavoli et al. \(2015\)](#) have developed an empirical study in order to compare their IDP-based test with the MWYW frequentist test and the DP-based test obtained as the prior strength goes to zero (called the Bayesian Bootstrap Dirichlet Process test -the BB-DP test, for short-). They have considered a Monte Carlo experiment in which n_1, n_2 observations from X , and Y respectively are generated, where $X \equiv N(0, 1)$ and $Y \equiv N(\Delta, 1)$, and Δ ranges from -1.5 to 1.5 . For each value of Δ , they have performed 20000 Monte Carlo runs. They first compare the performance of the IDP test and the BB-DP test. They consider three different options for the loss quotient $\gamma = 1, \gamma = 0.1$ and $\gamma = 0.05$. They conclude that, in all those cases where the first of them is determinate, both of them return the same answer, the difference between them focussing only on those samples for which the first one is indeterminate. In a second round of experiments, they compare the IDP test with the frequentist MWYW test. They select the significance level $\alpha = 0.05$ in order to construct the frequentist test, and $\gamma = 0.05$ in order to define the IDP test. Again, the frequentist test returns the same answer when the IDP test is determinate. They also compute the proportion of rejections of the MWYW among those samples for which the IDP test is indeterminate. They observe that such

a proportion increases with respect to Δ . As an example, for $n_1 = n_2 = 10$ and $\Delta = 0.9$, the IDP is indeterminate in 30% of the runs, and the MWW test rejects the null hypothesis 50% of them. As it returns the same proportion of rejections as a 50/50 randomized test derived from the IDP test and they conclude that the MWW test “guesses at random” 30% of the times. Let us nevertheless notice that the MWW does not return a random answer from a given sample.

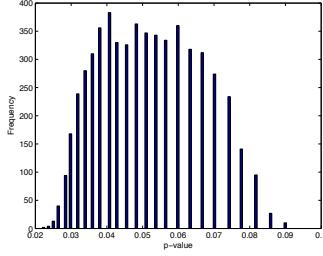
In this section, we deepen this study, with the aim of providing further insight about the behavior of the three tests (BB-DP, IDP and MWW) in practice. On one side, the p-value of the (frequentist) MWW test coincides with the posterior probability of the null hypothesis for the BB-DP, as we empirically show:



The posterior probability of the BB test depends on a bootstrap-based computation, and therefore small differences between the values of the posterior probability may occur if we launch the algorithm repeated times for the same sample (\vec{x}, \vec{y}) , the average of those posterior probabilities being the p-value. Consequently, the MWW test of size $\alpha = 0.05$ coincides with the BB-DP test for $\gamma = 0.05$.

Let us now examine the relation between the frequentist test and the IDP test. Since the p-value of the MWW coincides with the posterior probability of the BB test, we know that it is always bounded by the upper and lower posterior probabilities associated to the IDP test, $\bar{P}(H_0|(\vec{x}, \vec{y}))$ and $\underline{P}(H_0|(\vec{x}, \vec{y}))$. Therefore, we can write $\bar{P}(H_0|(\vec{x}, \vec{y})) = p(\vec{x}, \vec{y}) + \delta(\vec{x}, \vec{y})$ and $\underline{P}(H_0|(\vec{x}, \vec{y})) = p(\vec{x}, \vec{y}) - \delta'(\vec{x}, \vec{y})$, with $\delta(\vec{x}, \vec{y}) > 0$ and $\delta'(\vec{x}, \vec{y}) > 0$ for every pair of samples (\vec{x}, \vec{y}) .

Let us now recall an empirical result from [Benavoli et al. \(2015\)](#) about the distribution of the p-values over the collection of samples for which the IDP is indeterminate, i.e., those pairs of samples (\vec{x}, \vec{y}) satisfying the inequalities $\underline{P}(H_0|(\vec{x}, \vec{y})) < \gamma < \bar{P}(H_0|(\vec{x}, \vec{y}))$. The figure illustrates the distribution of the p-values for $\Delta = 0.5$ and $n_1 = n_2 = 20$ and $\gamma = 0.05$:



According to the above notation, these are the samples satisfying the following inequalities:

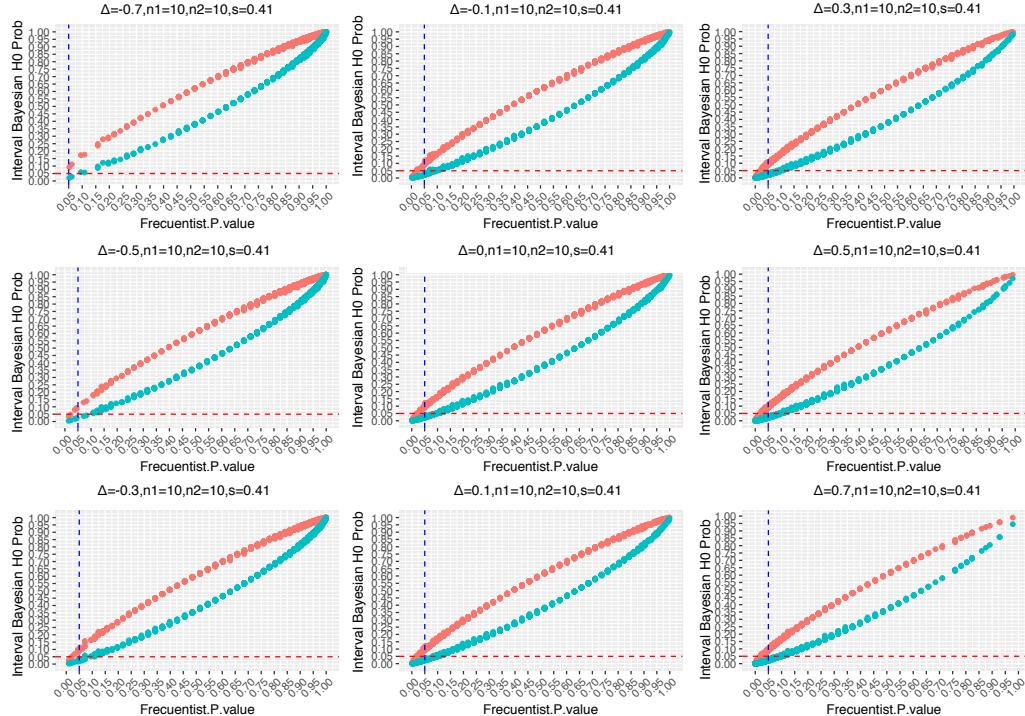
$$p(\vec{x}, \vec{y}) - \delta'(\vec{x}, \vec{y}) < 0.05 < p(\vec{x}, \vec{y}) + \delta(\vec{x}, \vec{y})$$

or, equivalently

$$0.05 - \delta(\vec{x}, \vec{y}) < p(\vec{x}, \vec{y}) < 0.05 + \delta'(\vec{x}, \vec{y}).$$

According to the above graph, we observe that the p-values are all of them in a neighbourhood of 0.05, and therefore δ and δ' take small values.

In order to get further information, we have computed and plotted, for every sample (\vec{x}, \vec{y}) , the upper and lower posterior probabilities from the IDP against the corresponding p-value, for different values of Δ and different sample sizes. Due to length restrictions, we just include the graphs for a specific choice of both sample sizes. In particular we have selected $n_1 = n_2 = 10$:



According to the above simulations, $\bar{P}(H_0 | (\vec{x}, \vec{y}))$ and $\underline{P}(H_0 | (\vec{x}, \vec{y}))$ can be written as functions of the p-value. For other sample sizes we have observed a similar shape of the graph, although the

difference $\bar{P}(H_0|(\vec{x}, \vec{y})) - \underline{P}(H_0|(\vec{x}, \vec{y}))$ is well known to decrease with respect to both sample sizes.

Furthermore, these two functions do not depend on the particular choice of Δ . Notwithstanding, it is well known that the p-value follows a uniform distribution over the unit interval when $\Delta = 0$, and as far as we get far away from $\Delta = 0$, the distribution of the p-values tends to concentrate over an extreme of the interval (the left side extreme for positive values of Δ and the right extreme for negative valued of Δ). The pair of upper and lower posterior probabilities also concentrate over the same extremes of the intervals for big values of Δ .

Let us now analyse some features of this functional relation. The p-value (which coincides with the posterior probability of the BB-DP test) is always bounded by $\bar{P}(H_0|(\vec{x}, \vec{y}))$ and $\underline{P}(H_0|(\vec{x}, \vec{y}))$, but it does not coincide with their half sum in general. On the other hand, when we plot their difference against the p-value, we observe that it increases from 0 to 0.5 and decreases from 0.5 to 1.

According to Equation 1, the p-value of a pair of samples (\vec{x}, \vec{y}) can be expressed $p(\vec{x}, \vec{y}) = G_0(U(\vec{x}, \vec{y}))$, with $G_0 = 1 - F_0$, where U denotes the MWW statistic and F_0 denotes the cdf of U under the assumption $\theta = 0.5$. The cdf F_0 corresponds to a unimodal distribution, and symmetric around μ_0 . In other words, the density function f_0 is increasing on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞) . Let us now consider $g_1(u) = G_0(u - \epsilon) - G_0(u)$ and $g_2(u) = G_0(u) - G_0(u + \epsilon)$. Both functions are increasing on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞) . Therefore we easily deduce that:

If either $U(\vec{x}, \vec{y}) < U(\vec{x}', \vec{y}') < \mu_0$ or $U(\vec{x}, \vec{y}) > U(\vec{x}', \vec{y}') > \mu_0$, then

$$g_1(U(\vec{x}, \vec{y})) < g_1(U(\vec{x}', \vec{y}')) \text{ and } g_2(U(\vec{x}, \vec{y})) < g_2(U(\vec{x}', \vec{y}')).$$

According to our Monte Carlo simulations, this is exactly what happens with the differences $\bar{P}(H_0|(\vec{x}, \vec{y})) - P(H_0|(\vec{x}, \vec{y}))$ and $P(H_0|(\vec{x}, \vec{y})) - \underline{P}(H_0|(\vec{x}, \vec{y}))$, i.e.:

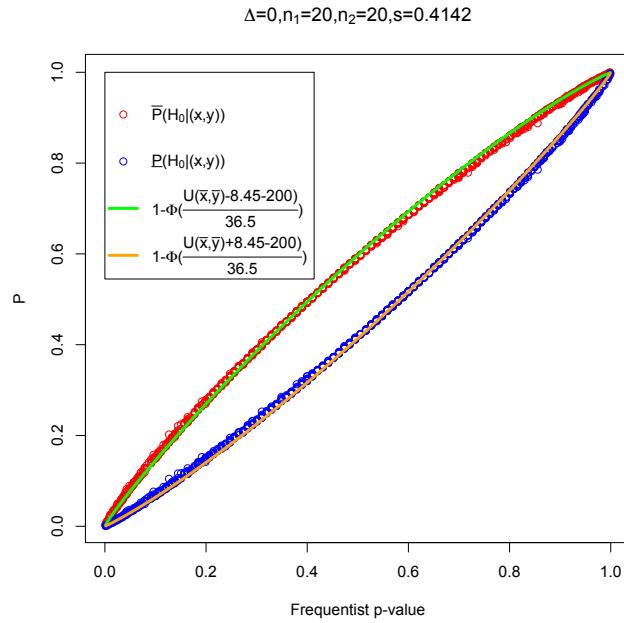
If $U(\vec{x}, \vec{y}) < U(\vec{x}', \vec{y}') < \mu_0$ or $U(\vec{x}, \vec{y}) > U(\vec{x}', \vec{y}') > \mu_0$, then

$$\bar{P}(H_0|(\vec{x}, \vec{y})) - P(H_0|(\vec{x}, \vec{y})) < \bar{P}(H_0|(\vec{x}', \vec{y}')) - P(H_0|(\vec{x}', \vec{y}')) \text{ and}$$

$$P(H_0|(\vec{x}, \vec{y})) - \underline{P}(H_0|(\vec{x}, \vec{y})) < P(H_0|(\vec{x}', \vec{y}')) - \underline{P}(H_0|(\vec{x}', \vec{y}')).$$

Therefore, it seems that the difference $\bar{P}(H_0|(\vec{x}, \vec{y})) - P(H_0|(\vec{x}, \vec{y}))$ is increasing with respect to $U(\vec{x}, \vec{y})$ on $(-\infty, \mu_0)$ and decreasing on (μ_0, ∞) . Something similar happens with the difference $P(H_0|(\vec{x}, \vec{y})) - \underline{P}(H_0|(\vec{x}, \vec{y}))$.

Since $P(H_0|(\vec{x}, \vec{y}))$ coincides with $p(\vec{x}, \vec{y}) = G_0(U(\vec{x}, \vec{y}))$ then we can deduce that there is a strictly increasing relation between $\bar{P}(H_0|(\vec{x}, \vec{y}))$ and $P(U(\vec{x}, \vec{y}) - \epsilon)$, for an arbitrary but fixed ϵ and the same happens with $\underline{P}(H_0|(\vec{x}, \vec{y}))$ and $P(U(\vec{x}, \vec{y}) + \epsilon)$. We have examined the nature of this strictly increasing (one-to-one) correspondence, and we have observed that it is in fact the identity.



This opens a door to the reconciliation between the Bayesian and the frequentist approaches also in the imprecise framework, following the path of Casella and Berger (1987) for the precise case. On one hand, there is a one-to-one correspondence between the upper (resp. the lower) posterior probability of the IDP test and the p-value of the MWW test. Furthermore, we can easily construct an imprecise test that relies on the MWW U-statistic and that mimics the behavior of the IDP test. Let us take an arbitrary α and let us define the new imprecise test as follows:

$$\delta(\vec{x}, \vec{y}) = \begin{cases} 0 & \text{if } U(\vec{x}, \vec{y}) \leq c_\alpha - \epsilon \\ 1 & \text{if } U(\vec{x}, \vec{y}) > c_\alpha + \epsilon \\ ? & \text{otherwise,} \end{cases} \quad (2)$$

where c_α is such that $P_{\theta_0}(U > c_\alpha) = \alpha$.

According to our simulations, for a specific choice of $\gamma = \frac{K_0}{K_0+K_1}$ and the triple (s, n_1, n_2) , there exists $\epsilon = g(s, n_1, n_2)$ such that the above imprecise test for $\alpha = \gamma$ coincides with the IDP test. Furthermore, the upper and lower posterior probabilities of the null hypothesis do respectively coincide with $G_0(U(\vec{x}, \vec{y}) + \epsilon)$ and $G_0(U(\vec{x}, \vec{y}) - \epsilon)$.

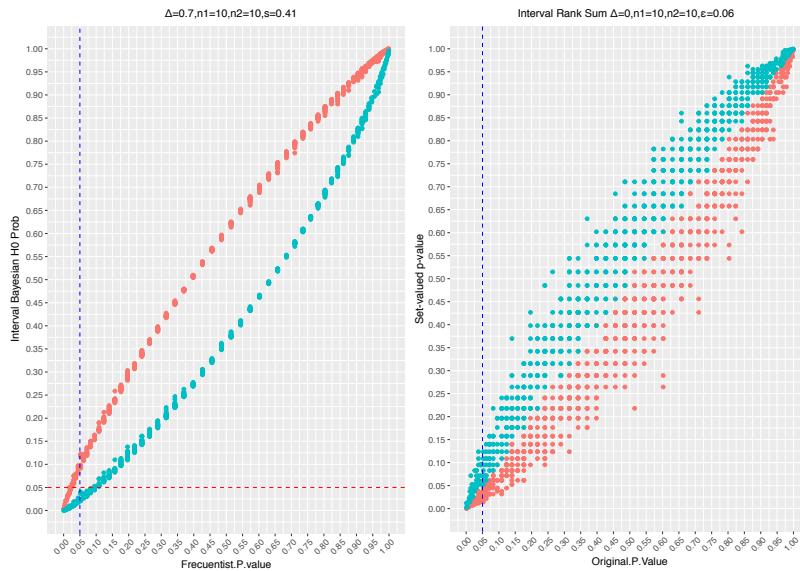
3.3 Conclusions and future directions

We have constructed an imprecise “frequentist” test that mimics the behavior of the so-called IDP test. It basically works as follows: it calculates the interval of values $(U(\vec{x}, \vec{y}) - \epsilon, U(\vec{x}, \vec{y}) + \epsilon)$ and it inherently considers the collection of samples (\vec{x}', \vec{y}') such that $U(\vec{x}, \vec{y}) - \epsilon < U(\vec{x}', \vec{y}') < U(\vec{x}, \vec{y}) + \epsilon$. If all of them are either in the rejection or the acceptance zone of the frequentist test, then the decision is clear. Otherwise, the outcome of the test is indeterminate. Thus, we conclude that, at least for the MWW test, the kind of “imprecision” over the set of priors considered in the IDP-based test may produce similar effects on the decision mechanism as an imprecision around the statistic values.

Let us notice that, in this specific case where the statistic is based on the ranks of the observations, but not on their numerical values, the statistic is not continuous with respect to those numerical values. Thus the following alternative test (see [Perolat et al. \(2015\)](#) for further discussion about it):

$$\begin{cases} 0 & \text{if } p(\vec{x}', \vec{y}') \leq \alpha, \forall (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ 1 & \text{if } p(\vec{x}', \vec{y}') > \alpha, \forall (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ ? & \text{otherwise.} \end{cases} \quad (3)$$

would produce different outcomes in practice, as we observe below:



Notwithstanding, for other frequentist tests, different from MWW, based on continuous statistics the variation proposed in Equation 3, could report similar results as the one provided in Equation 2, for adequate selections of ϵ and δ . For those cases, it seems that the kind of imprecision over the set of priors considered in the IDP-based test may produce similar effects on the decision mechanism as an imprecision around the sample values. Let us remind the reader that our empirical comparison in this paper refers to the case where the frequentist test completely coincides with the Bayesian one. But this may be not the case for other tests where the MLR condition is not satisfied.

In those cases we might directly compare the IDP-based test with an imprecise version of the Bayesian test as follows:

$$\begin{cases} 0 & \text{if } P(H_0 | \vec{x}', \vec{y}') \leq \gamma, \forall (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ 1 & \text{if } P(H_0 | \vec{x}', \vec{y}') > \gamma, \forall (\vec{x}', \vec{y}') \in B(\vec{x}, \vec{y}; \delta) \\ ? & \text{otherwise.} \end{cases} \quad (4)$$

Such a comparison could shed further light about the behaviour of IDP-based tests in practice. We conjecture that they could lead to similar decision rules. If our conjecture is true, this alternative procedure would lead to equivalent but computationally more efficient algorithms. On the other side, it would reflect that the kind of imprecision over the priors considered by this almost-ignorance model produces similar effects in the decision procedure as an imprecision around the sample values.

Acknowledgments

This paper has been partially supported by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias). We thank three anonymous reviewers for insightful comments on our manuscript.

References

- A. Benavoli, F. Mangili, F. Ruggeri, and M. Zaffalon. Imprecise Dirichlet process with application to the hypothesis test on the probability that $X \leq Y$. *Journal of Statistical Theory and Practice*, 9(3):658–684, 2015.
- J. Berger and T. Selke. Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–120, 1987.
- G. Casella and R. L. Berger. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397):106–111, 1987.
- I. Couso, S. Moral, and L. Sánchez. The behavioral meaning of the median. *Information Sciences*, 294:127–138, 2015.
- M. H. DeGroot. Doing what comes naturally: Interpreting a tail area as a posterior probability or likelihood ratio. *Journal of the American Statistical Association*, 68(344):966–969, 1973.
- B. Efron. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- H. Jeffreys. *Theory of Probability*. Oxford University Press., 1939.
- D. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.
- H. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.
- M. P. M.P. Fay. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010.
- J. Perolat, I. Couso, K. Loquin, and O. Strauss. Generalizing the Wilcoxon rank-sum test for interval data. *International Journal of Approximate Reasoning*, 56(A):108–121, 2015.
- J. Pratt. Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society, Ser. B*, 27(2):169–203, 1965.
- D. Rubin. Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.
- G. Shafer. Lindley’s Paradox. *Journal of the American Statistical Association*, 77(378):325–351, 1982.

Evenly Convex Credal Sets

Fabio Gagliardi Cozman

Escola Politecnica, Univ. de São Paulo - São Paulo (Brazil)

FGCOZMAN@USP.BR

Abstract

An evenly convex credal set is a set of probability measures that is evenly convex; that is, a set that is an intersection of open halfspaces. An evenly convex credal set can for instance encode preference judgments through strict and non-strict inequalities such as $\mathbb{P}(A) > 1/2$ and $\mathbb{P}(A) \leq 2/3$. This paper presents an axiomatization of evenly convex sets from preferences, where we introduce a new (and very weak) Archimedean condition.

Keywords: credal sets; sets of probability measures; preference axioms; convexity.

1. Introduction

The goal of this note is to show that relatively simple axioms on preference orderings can be used to characterize *evenly convex* sets of probability measures; that is, sets that are intersections of open halfspaces. Such sets allow assessments such as $\mathbb{P}(A) \geq 1/2$ and $1/4 < \mathbb{P}(B) \leq 3/4$; that is, strict and non-strict inequalities can be expressed on probability values.

A preference ordering is a binary relation \succ on *gambles*; a gamble is a function X that yields a real number $X(\omega)$ for each *state* ω , and $X \succ Y$ is understood as “ X is preferred to Y ”.

If a preference ordering is only a partial order, then, subject to a few additional conditions, it can be represented by a set of probability measures (Giron and Rios, 1980; Seidenfeld et al., 1990; Walley, 1991; Williams, 1975). Typically such axiomatizations of sets of probability measures focus on a single *maximal closed convex* set of probability measures. It seems that the only existing axiomatization that allows for open sets of probability measures sets has been given by Seidenfeld et al. (1995), using a more general setting where utilities are also derived, and a proof technique based on transfinite induction. Their representation result may require sets of state-dependent utilities to represent preferences; for this reason it may be a little difficult to grasp the geometric content of a preference profile. One wonders whether it is possible to capture assessments such as $\mathbb{P}(A) > 1/2$ with some intuitive construction.

Section 4 presents a concise axiomatization for *evenly convex* sets of probability measures. We use a new Archimedean condition, and emphasize the use of separating hyperplanes as much as possible, hopefully producing results that can be appreciated with moderate effort.

2. Preference orderings, sets of desirable gambles, and credal sets

In this section we present some basic concepts and results used throughout. Because some results here are in essence well-known, only very short proof sketches are mentioned for them.

Consider a finite set Ω containing n states $\{\omega_1, \dots, \omega_n\}$. An *event* is a subset of Ω ; a *gamble* is a function $X : \Omega \rightarrow \mathbb{R}$. A gamble can be viewed as a n -dimensional vector. A probability measure over Ω is entirely specified by a n -dimensional vector with non-negative elements that add up to one. Given such a vector p that induces a probability measure \mathbb{P} , and a gamble X , the expected value of X , denoted by $\mathbb{E}_{\mathbb{P}}[X]$, is simply the inner product $X \cdot p$.

All sets we consider are subsets of \Re^n ; throughout we assume the Euclidean topology. For a set \mathcal{A} , $\text{cl}\mathcal{A}$ is the closure of \mathcal{A} and $\text{relint}\mathcal{A}$ is the relative interior of \mathcal{A} . A *cone* \mathcal{A} is a set such that if $X \in \mathcal{A}$ then $\lambda X \in \mathcal{A}$ for $\lambda > 0$ (the origin may not be in \mathcal{A}). An *exposed ray* of a convex cone is an exposed face that is a half-line emanating from the origin (recall that an exposed face is a face that is equal to the set of points achieving the maximum of some linear function).

Most results in this paper deal with the representation of preferences:¹

Definition 1 A preference ordering \succ is a strict partial order over pairs of gambles.

Absence of preference between X and Y is indicated by $X \not\succ Y$. If $X \succ 0$, X is *desirable*; if $X \not\succ 0$, X is *neutral*.

We always assume two additional properties:

Monotonicity: If $X(\omega) > Y(\omega)$ for all $\omega \in \Omega$, then $X \succ Y$;

Cancellation: For all $\alpha \in (0, 1]$, $X \succ Y$ iff $\alpha X + (1 - \alpha)Z \succ \alpha Y + (1 - \alpha)Z$.

The following representation obtains:²

Proposition 2 If a preference ordering \succ satisfies monotonicity and cancellation, then there is a convex cone \mathcal{D} , not containing the origin but containing the interior of the positive octant, such that $X \succ Y$ iff $X - Y \in \mathcal{D}$.

Cones that encode preference orderings have received attention in the literature for some time (Giron and Rios, 1980; Seidenfeld et al., 1990; Williams, 1975; Walley, 1991). In fact, the literature on *sets of desirable gambles* (Miranda and Zaffalon, 2010; Quaeghebeur, 2014; Walley, 2000) employs cones of gambles to model preferences, often assuming *admissibility*: if $X(\omega) \geq 0$ for all ω and $X(\omega) > 0$ for some ω , then $X \succ 0$. We do not assume admissibility here; indeed, admissibility cannot be satisfied in general if preferences are to be encoded by expectation with respect to probability measures (when probability values may be equal to zero). In any case, we use the term *set of desirable gambles* to refer to a convex cone \mathcal{D} constructed as in Proposition 2. This proposition allows one to freely switch between preference orderings and sets of desirable gambles; we find the former to be more intuitive so we mostly employ them in the remainder of this paper.

One might think that any convex cone of gambles can be represented by a set of probability measures as follows: $X \in \mathcal{D}$ iff $\mathbb{E}_{\mathbb{P}}[X] > 0$ for all \mathbb{P} in some set \mathcal{K} of probability measures. This is not possible. Consider the set of desirable gambles depicted in Figure 1 (left). All gambles in the interior of \mathcal{D} satisfy $X(\omega_1)\mathbb{P}(\omega_1) + X(\omega_2)\mathbb{P}(\omega_2) > 0$ for $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = 1/2$. No other pair of probability values (or sets of pairs of probability values) can similarly represent the interior of \mathcal{D} . But even this probability measure cannot represent the fact that half-border is in \mathcal{D} ; for this half-border, $f(\omega_1)\mathbb{P}(\omega_1) + f(\omega_2)\mathbb{P}(\omega_2) = 0$. Thus some condition on boundaries is needed.

Conditions on boundaries of sets of desirable gambles inevitably focus on what “makes sense” concerning limiting behavior. For instance, Aumann (1962) has proposed the following condition:

1. A *strict partial order* is a binary relation that is irreflexive and transitive, an *equivalence* is a binary relation that is reflexive, transitive, and symmetric (a binary relation \diamond is *irreflexive* when $X \diamond X$ if false for every X ; it is *transitive* when $X \diamond X$ and $Y \diamond Z$ imply $X \diamond Z$; it is *symmetric* when $X \diamond Y$ implies $Y \diamond X$) (Fishburn, 1970, Section 2.3).
2. Proof sketch: Applying cancellation, $X \succ Y$ iff $X/2 - Y/2 \succ Y/2 - Y/2$ iff $X - Y \succ 0$. Now if $X \succ 0$ and $Y \succ 0$, then $0 \succ -Y$ (as $X \succ Y$ iff $-Y \succ -X$), and by transitivity we get $X + Y \succ 0$. For any $\lambda \in (0, 1)$, $X \succ 0$ iff $\lambda X \succ 0$ by cancellation. Finite induction leads to: $X \succ 0$ implies $\lambda X \succ 0$ for $\lambda > 0$, so we have the cone (monotonicity implies this cone contains every positive gamble; irreflexivity eliminates the origin).

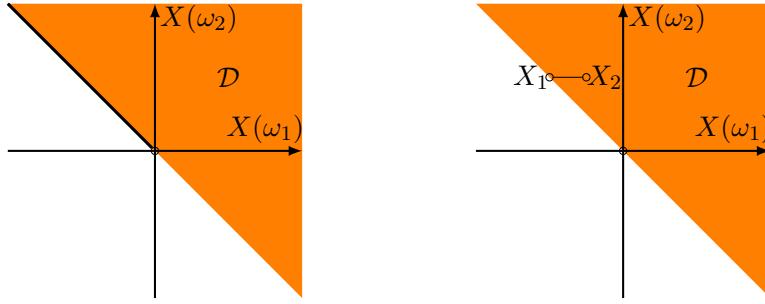


Figure 1: Left: a cone \mathcal{D} ; one bordering ray (thick line from the origin) belongs to \mathcal{D} , while the other bordering ray does not belong to \mathcal{D} . Right: to understand the effect of Aumann's continuity condition, take a similar cone \mathcal{D} ; the gamble X_2 is inside \mathcal{D} and the gamble X_1 is in the border, so the segment between X_1 and X_2 is in \mathcal{D} , implying that $X_1 \succ 0$ or $X_1 \not\succ 0$ by Aumann's continuity condition; the same reasoning could be repeated for $-X_1$, hence the border must be open because $X_1 \succ 0$ and $-X_1 \succ 0$ cannot happen.

Aumann's continuity: If $\alpha X + (1 - \alpha)Y \succ Z$ for all $\alpha > 0$, then either $Y \succ Z$ or $Y \not\succ Z$.

If the interior of the set of desirable gambles is an open halfspace, Aumann's continuity condition forces the set of desirable gambles to be open (see Figure 1 (right)). In general, if the interior of the set of desirable gambles is strictly smaller than a halfspace, Aumann's continuity condition does not imply that \mathcal{D} is entirely open; it only implies that each gamble in the boundary of \mathcal{D} is either desirable or neutral.

If the continuity condition is strengthened so that \mathcal{D} is assumed open (Seidenfeld et al., 1990; Walley, 1991), then it is possible to find a representation of preference orderings through probabilities. Walley imposes openness by basically requiring that $X \succ 0$ implies $X - \epsilon \succ 0$ for some ϵ (Walley, 1991, Section 3.7.8, D7). Another possibility could be to require that (note that limits of sequences of gambles are always assumed pointwise):

Open continuity If $X_i \succ 0$ is false for every i , and $X = \lim_i X_i$, then $X \succ 0$ is false.

Here is the representation result under the assumption of openness:³

Proposition 3 *If a set of desirable gambles \mathcal{D} is open, then it can be represented by a closed convex set \mathcal{K} of probability measures, in the sense that $X \in \mathcal{D}$ iff $\mathbb{E}_{\mathbb{P}}[X] > 0$ for all $\mathbb{P} \in \mathcal{K}$.*

A set of probability measures is called a *credal set* (Levi, 1980). There is a significant disadvantage in assuming that a set of desirable gambles is open; namely, the representing credal set is

3. Proof sketch: Copy the proof of Theorem 11, except Part 5 (note that in Part 1 one might choose to replace Theorem 7 by some appropriate separating hyperplane theorem (Klee Jr., 1955)). Now to show that \mathcal{K} is closed, show that the complement of the cone \mathcal{C} in the proof of Theorem 11 is open: If $p \notin \mathcal{C}$, then there is $X \in \mathcal{D}$ such that $X \cdot p \leq 0$, and also $X - \epsilon \in \mathcal{D}$ for some $\epsilon > 0$ (as \mathcal{D} is open by assumption). Consider the closed halfspace $\mathcal{H} = \{q : (X - \epsilon) \cdot q \leq 0\}$; this halfspace is disjoint from \mathcal{C} . Also, p is in \mathcal{H} but not in its boundary (there is a ball around p inside \mathcal{H} for any radius smaller than $|(X - \epsilon) \cdot p| / \|X - \epsilon\|$). So the complement of \mathcal{C} is open as desired.

necessarily closed. Hence one cannot say that a coin is biased simply by stating $\mathbb{P}(\text{Heads}) > 1/2$. It seems that the only existing condition in the literature that can accept such assessments has been proposed by Seidenfeld et al. (1995). Their condition has two parts, but only one is necessary:

SSK-continuity If $X_i \succ Y_i$ for every i , and $\lim_i Y_i \succ Z$, then $\lim_i X_i \succ Z$, whenever limits exist.

The other part of the original condition by Seidenfeld, Schervish and Kadane can actually be derived from the previous conditions:

Proposition 4 Suppose a preference ordering \succ satisfies cancellation and SSK-continuity. If $X_i \succ Y_i$ for every i , and $W \succ \lim_i X_i$, then $W \succ \lim_i Y_i$.

Proof The assumptions imply $X_i - Y_i \succ 0$ and then $-Y_i \succ -X_i$ for every i ; similarly, $-\lim X_i = \lim -X_i \succ -W$, so by SSK-continuity, $\lim -Y_i \succ -W$ and then $W \succ \lim Y_i$. ■

In fact we might simplify SSK-continuity even more in the presence of cancellation:

Proposition 5 Suppose \succ is a preference ordering satisfying cancellation. Suppose that if $X_i \succ Y_i$ and $\lim_i Y_i \succ 0$ then $\lim_i X_i \succ 0$. Then \succ satisfies SSK-continuity.

Proof If $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, $X_i \succ Y_i$ and $Y \succ Z$ then $\{X_i - Z\} \rightarrow X - Z$, $\{Y_i - Z\} \rightarrow Y - Z$, $X_i - Z \succ Y_i - Z$ and $Y - Z \succ 0$; if the property assumed in the statement is true, then $X - Z \succ 0$ so $X \succ Z$ as desired. ■

If a preference ordering satisfies SSK-continuity, and $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, and $X_i \succ Y_i$, then either $X \succ Y$ or $X \not\succ Y$ (for suppose otherwise that $Y \succ X$; SSK-continuity says that if $X_i \succ Y_i$ and $Y = \lim_i Y_i \succ X$ then $\lim_i X_i \succ X$, hence $X \succ X$, a contradiction). Thus we have that SSK-continuity conveys Aumann's continuity condition. We will return to SSK-continuity when we examine whether it implies even convexity (it does not).

3. Evenly convex sets and evenly convex cones

An *evenly convex* set \mathcal{A} is an intersection of open halfspaces (Fenchel, 1952). Hence an open convex set is evenly convex; also a closed convex set is evenly convex as it is an infinite intersection of halfspaces. For any set \mathcal{A} , its *evenly convex hull* $\text{eco}\mathcal{A}$ is the intersection of all evenly convex sets containing \mathcal{A} ; so $\text{eco}\mathcal{A}$ is the intersection of all open halfspaces that contain \mathcal{A} . Note that $\text{co}\mathcal{A} \subseteq \text{eco}\mathcal{A}$, where $\text{co}\mathcal{A}$ is the convex hull of \mathcal{A} .

There are many characterizations of evenly convex sets (Daniilidis and Martinez-Legaz, 2002; Goberna et al., 2003; Klee, 1968). In particular, we will use the following result in the proof of Theorem 9 (Daniilidis and Martinez-Legaz, 2002, Corollary 6): a convex set \mathcal{A} is evenly convex iff for every $X_0 \in \text{cl}\mathcal{A} \setminus \mathcal{A}$, and every $\{X_i\}_{i \geq 1} \subset \mathcal{A}$, and every $\{\lambda_i\}_{i \geq 1}$ such that $\lambda_i > 0$, we have $X_0 - \lim_i \lambda_i(X_i - X_0) \notin \mathcal{A}$ whenever the limit exists.

If \mathcal{A} is evenly convex, then if $X \in \mathcal{A}$ and $Y \in \text{cl}\mathcal{A}$ we have $\alpha X + (1 - \alpha)Y \in \mathcal{A}$ for $\alpha \in (0, 1)$ (Fenchel, 1952, Section 3.5). Consequently:

Lemma 6 Suppose \mathcal{A} is evenly convex and $0 \notin \mathcal{A}$. If X and $-X$ belong to $\text{cl}\mathcal{A}$, then neither is in \mathcal{A} .

Proof If $X \in \mathcal{A}$, then $-X \in \text{cl}\mathcal{A}$ implies $X/2 + (-X)/2 = 0 \in \mathcal{A}$, a contradiction; hence $X \notin \mathcal{A}$. By similar reasoning, $-X \notin \mathcal{A}$. \blacksquare

We then obtain the following separation property, that is used later:

Theorem 7 Suppose \mathcal{A} is an evenly convex cone such that $0 \notin \mathcal{A}$. If $X \notin \mathcal{A}$, then there is p such that $X \cdot p \leq 0$ and $Y \cdot p > 0$ for all $Y \in \mathcal{A}$.

Proof Part 1) Suppose $X \in \text{cl}\mathcal{A}$, but $X \notin \mathcal{A}$. Because \mathcal{A} is evenly convex, there is p and β such that $X \cdot p = \beta$ and $Y \cdot p > \beta$ for all $Y \in \mathcal{A}$ (Goberna et al., 2003, Proposition 3.1(ii)). If $\beta > 0$, then for any Y in a neighborhood of 0 we have $\epsilon Y \cdot p < \beta$ for some $\epsilon > 0$; this is a contradiction because some such Y is in \mathcal{A} , and for this Y we must have $\epsilon Y \cdot p > \beta$. Hence $\beta \leq 0$. We now show that actually $\beta = 0$.

For $Y \in \mathcal{A}$, $Y \cdot p > \beta = X \cdot p$, hence $(Y - X) \cdot p > 0$. Because X is in the boundary of \mathcal{A} , there is a gamble Y in a neighborhood of X that belongs to \mathcal{A} ; define $q = Y - X$, and note that the segment from Y to X (excluding X) is in \mathcal{A} (Fenchel, 1952, Section 3.5). That is, there is q such that $q \cdot p > 0$ and $(X + \epsilon q) \cdot p > \beta$ for $\epsilon > 0$ in a neighborhood of 0. Now for any $\lambda > 0$ we have $\lambda(X + \epsilon q) \in \mathcal{A}$. That is, $\lambda(X + \epsilon q) \cdot p > \beta$, so $X \cdot p > \beta/\lambda - \epsilon q \cdot p$. Again use $X \cdot p = \beta$, to obtain $\beta > \beta/\lambda - \epsilon q \cdot p$. Consequently, we have both $\beta \leq 0$ and $\beta > -\epsilon q \cdot p/(1 - 1/\lambda)$; take say $\lambda = 2$ to obtain the constraint $\beta > -\epsilon(2q \cdot p)$. These conditions can only be satisfied for $\epsilon > 0$ if $\beta = 0$.

Part 2) Now suppose instead that $X \notin \text{cl}\mathcal{A}$. Consider the cone $\mathcal{B} = \{\lambda X : \lambda \geq 0\}$. Using an appropriate separation result (Klee Jr., 1955, Theorem 2.5), we know that there is p such that $Y \cdot p > 0$ for $Y \in \text{cl}\mathcal{A} \setminus (\text{cl}\mathcal{A} \cap -\text{cl}\mathcal{A})$, $Y' \cdot p = 0$ for $Y' \in (\text{cl}\mathcal{A} \cap -\text{cl}\mathcal{A}) \cup (\mathcal{B} \cap -\mathcal{B})$, $Y'' \cdot p \leq 0$ for $Y'' \in \mathcal{B} \setminus (\mathcal{B} \cap -\mathcal{B})$. Clearly $\mathcal{B} \cap -\mathcal{B}$ contains just the zero gamble. Now note that $\text{cl}\mathcal{A} \cap -\text{cl}\mathcal{A}$ does not intersect \mathcal{A} (if $Y \in \text{cl}\mathcal{A} \cap -\text{cl}\mathcal{A}$, then $Y \in \text{cl}\mathcal{A}$ and $-Y \in \text{cl}\mathcal{A}$, so both are not in \mathcal{A} by Lemma 6). Hence there is p such that $X \cdot p \leq 0$ and $Y \cdot p > 0$ for $Y \in \mathcal{A}$. \blacksquare

4. Evenly convex sets of desirable gambles and evenly convex credal sets

In this section we consider preference orderings that can be represented by evenly convex sets of desirable gambles; such preference orderings can also be represented by evenly convex credal sets. This will allow us to consider assessments such as $1/4 \leq \mathbb{P}(\text{Heads}) < 1/2$.

4.1 Evenly convex sets of desirable gambles

We introduce the following condition:

Even continuity If $X_i \succ 0$ for every i , and $Y \succ 0$ is false, then $\lim_i (\lambda_i Y - X_i) \succ 0$ is false for any sequence of $\lambda_i > 0$ such that the limit exists.

Even though the condition is somewhat long, it is quite reasonable: one cannot take an undesirable gamble Y and make it desirable, not even in the limit, by multiplying it by a positive number and subtracting from it a desirable gamble.⁴

4. One might consider a weaker condition (as suggested by a reviewer): If $X_i \succ 0$ and not $Y \succ 0$, then not $\lim_i (Y - X_i) \succ 0$. But this is implied by SSK-continuity: if $X_i \succ 0$, then if $Y \succ \lim_i X_i$ then $Y \succ 0$ by SSK-continuity, implying that if $X_i \succ 0$, then if not $Y \succ 0$ then not $Y \succ \lim_i X_i$.

To make later results more concise, we introduce the following definition:

Definition 8 A preference ordering \succ is coherent when it satisfies monotonicity, cancellation, and even continuity.

We then obtain:

Theorem 9 If a preference ordering \succ is coherent, then there is an evenly convex cone \mathcal{D} of gambles, not containing the origin but containing the interior of the positive octant, such that $X \succ Y$ iff $X - Y \in \mathcal{D}$.

Proof Take the set of desirable gambles produced by Proposition 2.

For a fixed $Y \in \text{cl}\mathcal{D} \setminus \mathcal{D}$ (hence $Y \notin \mathcal{D}$) and $X_i \in \mathcal{D}$ for every i , and $\lambda_i > 0$, compute $\lambda'_i = 1 + \lambda_i$ and $X'_i = \lambda'_i X_i$. Clearly $\lambda'_i > 0$ and $X'_i \in \mathcal{D}$. By even continuity, $\lim_i (\lambda'_i Y - X'_i) \notin \mathcal{D}$; hence $\lim_i ((1 + \lambda_i)Y - \lambda_i X_i) \notin \mathcal{D}$, and then $Y - \lim_i \lambda_i (X_i - Y) \notin \mathcal{D}$. Thus \mathcal{D} is evenly convex (Daniilidis and Martinez-Legaz, 2002, Corollary 6). ■

Note that coherence implies Aumann's continuity condition:

Proposition 10 Suppose a preference ordering \succ is coherent. If $\alpha X + (1 - \alpha)Y \succ Z$ for all $\alpha > 0$, then either $Y \succ Z$ or $Y \not\succ Z$.

Proof If $X_i \succ 0$ for every i , then the fact that $\neg(0 \succ 0)$ and even continuity imply $\neg(-X \succ 0)$ for $X = \lim_i X_i$. Now, if $\alpha X + (1 - \alpha)Y \succ Z$, then take $\alpha_i = 1/2^i$ and $X_i = \alpha_i(X - Z) + (1 - \alpha_i)(Y - Z)$; hence $X_i \succ 0$, implying that $\neg(Z - Y \succ 0)$, so either $Y \succ Z$ or $Y \not\succ Z$. ■

4.2 Evenly convex credal sets

Evenly convex sets of desirable gambles can be nicely represented by evenly convex sets of probability measures, as described by the next theorem. In the next proof and later we use the nonempty cone

$$\mathcal{C} = \{p : X \cdot p > 0, \forall X \in \mathcal{D}\}.$$

Theorem 11 If a preference ordering \succ is coherent, then there is a unique maximal evenly convex credal set \mathcal{K} such that $X \succ Y$ iff for all $\mathbb{P} \in \mathcal{K}$ we have $\mathbb{E}_{\mathbb{P}}[X] > \mathbb{E}_{\mathbb{P}}[Y]$.

Proof Part 1) For any $X \notin \mathcal{D}$, there is p such that $X \cdot p \leq 0$ and $Y \cdot p > 0$ for all $Y \in \mathcal{D}$ by Theorem 7. So \mathcal{C} is nonempty, and in fact it is a cone (if p' and p'' satisfy the constraints, then so does $\lambda p'$ for $\lambda > 0$ and $p' + p''$). Hence if $X \notin \mathcal{D}$ then $\exists p \in \mathcal{C} : X \cdot p \leq 0$; equivalently, if $\forall p \in \mathcal{C} : X \cdot p > 0$, then $X \in \mathcal{D}$.

Part 2) By construction, if $X \in \mathcal{D}$ then $X \cdot p > 0$ for all $p \in \mathcal{C}$; using this and Part 1, $X \in \mathcal{D} \Leftrightarrow \forall p \in \mathcal{C} : X \cdot p > 0$.

Part 3) We now show that \mathcal{C} is equivalent to a set of probability measures \mathcal{K} . Denote by $\mathbf{1}$ a vector of ones, and $\mathbf{1}_i$ a vector whose i th element is 1 and all other elements are zero. By monotonicity, $\mathbf{1} \cdot p > 0$ for all $p \in \mathcal{C}$, so $\sum_i p_i > 0$. Also, for every $p \in \mathcal{C}$: $(\mathbf{1}_i + \epsilon) \cdot p > 0$ for every $\epsilon > 0$; hence $p_i + \epsilon \sum_j p_j > 0$ for every ϵ , implying that $p_i \geq 0$ (if $p_i < 0$ then for $\epsilon < -p_i / \sum_j p_j$ we

have $p_i + \epsilon \sum_j p_j < 0$, a contradiction). Hence we can normalize each p in \mathcal{C} , thus obtaining a set of probability measures \mathcal{K} that is a representation for \mathcal{D} : $X \in \mathcal{D} \Leftrightarrow \forall \mathbb{P} \in \mathcal{K} : \mathbb{E}_{\mathbb{P}}[X] > 0$.

Part 4) Take the set \mathcal{K} that is equal to the intersection of \mathcal{C} and the unitary simplex $\sum_i p_i = 1$: If p belongs to this intersection, it is normalized so $p \in \mathcal{K}$; and if $p \in \mathcal{K}$, then $p \in \mathcal{C}$ and also it is normalized so it belongs to the unitary simplex. Hence \mathcal{K} is the intersection of two convex sets, so \mathcal{K} is convex.

Part 5) The cone \mathcal{C} is defined as the intersection of open halfspaces, hence by definition it is evenly convex. And \mathcal{K} is the intersection of those open halfspaces and the unitary simplex (itself the intersection of open halfspaces), hence \mathcal{K} is evenly convex.

Part 6) To show that \mathcal{K} is the unique maximal credal set that represents \succ , suppose there is \mathcal{K}' that represents \succ , and $\mathbb{P}' \in \mathcal{K}'$ but $\mathbb{P}' \notin \mathcal{K}$. If $\mathbb{P}' \notin \mathcal{K}$, then by the definition of \mathcal{K} we must have some $X \in \mathcal{D}$ such that $\mathbb{E}_{\mathbb{P}'}[X] \leq 0$. However, because \mathcal{K}' represents \succ , for any $X \in \mathcal{D}$ we must have $\mathbb{E}_{\mathbb{P}'}[X] > 0$ for all $\mathbb{P} \in \mathcal{K}'$; that is, $\mathbb{E}_{\mathbb{P}'}[X] > 0$. Hence we get a contradiction, implying that no representing credal set can contain probability measures outside of \mathcal{K} . ■

In fact *many* sets of probability measures may encode the same ordering. For instance, if a representing \mathcal{K} is a closed set, then the set of its extreme points $\text{ext}\mathcal{K}$ is an equivalent representation for \succ ; that is, $X \succ Y \Leftrightarrow \forall \mathbb{P} \in \text{ext}\mathcal{K} : \mathbb{E}_{\mathbb{P}}[X] > \mathbb{E}_{\mathbb{P}}[Y]$.

Theorem 12 Suppose \succ is a coherent preference ordering, and the credal set \mathcal{K} has been built as in the proof of Theorem 11. A credal set \mathcal{K}' represents \succ iff $\text{eco}\mathcal{K}' = \mathcal{K}$.

Proof We need only to consider preferences with respect to the zero gamble.

Take a credal set \mathcal{K}' such that $\text{eco}\mathcal{K}' = \mathcal{K}$. Clearly if $X \succ 0$ then $\forall \mathbb{P} \in \mathcal{K} : \mathbb{E}_{\mathbb{P}}[X] > 0$ then $\forall \mathbb{P} \in \mathcal{K}' : \mathbb{E}_{\mathbb{P}}[f] > 0$ as $\mathcal{K}' \subseteq \text{eco}\mathcal{K}'$. Now suppose $\forall \mathbb{P} \in \mathcal{K}' : \mathbb{E}_{\mathbb{P}}[X] > 0$. Consider that $\text{eco}\mathcal{K}'$ is the set of all p such that $Y \cdot p > 0$ for all Y such that for all $q \in \mathcal{K}'$ we have $Y \cdot q > 0$. As X satisfies the last set of inequalities, then $X \cdot p > 0$ for all $p \in \text{eco}\mathcal{K}'$, hence $\mathbb{E}_{\mathbb{P}}[X] > 0$ for all $\mathbb{P} \in \mathcal{K}$, and then $X \succ 0$. Hence \mathcal{K}' represents \succ .

Now suppose \mathcal{K}' represents \succ . Then its elements must satisfy the constraints $X \cdot p > 0$ for all $X \in \mathcal{D}$. Suppose \mathcal{K}' also satisfies a nontrivial constraint $Y \cdot p > \alpha$ for some Y and α ; that is, there is p' that satisfies all other constraints but such that $Y \cdot p' \leq \alpha$. Because every p is a probability measure, $(Y - \alpha) \cdot p > 0$ is an equivalent constraint. Hence $(Y - \alpha) \cdot p > 0$ for all $p \in \mathcal{K}'$; because \mathcal{K}' represents \succ , $Y - \alpha$ is a desirable gamble. However there is $p' \in \mathcal{K}$ such that $(Y - \alpha) \cdot p' \leq 0$, implying $Y - \alpha \notin \mathcal{D}$, a contradiction. So there is no additional nontrivial strict linear inequality that distinguishes \mathcal{K}' and \mathcal{K} , and consequently they share the same evenly convex hull. ■

This theorem shows that if two evenly convex sets are different, then they represent distinct preference orderings. Figure 2 shows several different credal sets that have the same evenly convex hull, and hence represent the same coherent preference ordering.

4.3 A bit of duality

Additional insight can be obtained by investigating the duality between $\text{cl}\mathcal{D}$ and $\text{cl}\mathcal{C}$. As \mathcal{C} is nonempty, $\text{cl}\mathcal{C} = \{p : \forall X \in \mathcal{D} : X \cdot p \geq 0\}$; hence $\text{cl}\mathcal{C}$ is by definition the dual cone⁵ of \mathcal{D} ,

5. Given a convex set \mathcal{A} , its polar set is $\mathcal{A}^\circ = \{p : \forall X \in \mathcal{A} : X \cdot p \leq 1\}$ (Brondsted, 83); if \mathcal{A} is a convex cone its polar set is equal to its polar cone, defined as $\{p : \forall X \in \mathcal{A} : X \cdot p \leq 0\}$ (because any inequality with right hand



Figure 2: Five credal sets with the same evenly convex hull (the first credal set is evenly convex). Filled dots, thick lines and darker (orange) regions are in the credal sets.

denoted by \mathcal{D}^* (Boyd and Vandenberghe, 2004, Section 2.6). Then $(\text{cl}\mathcal{C})^*$ is just the closure of \mathcal{D} , as $\text{cl}\mathcal{D} = \mathcal{D}^{**}$ (Brondsted, 83, Theorem 6.2). Also, if a cone $\mathcal{F} \subset \mathcal{D}$ (say a proper face of \mathcal{D}), then $\mathcal{D}^* \subset \mathcal{F}^*$, and if we have several cones $\{\mathcal{D}_i\}_i$, then $(\cup_i \mathcal{D}_i)^* = \cap \mathcal{D}_i^*$ (Lay, 1982, Theorem 23.3).

It is also possible to establish a connection between the faces of $\text{cl}\mathcal{D}$ and $\text{cl}\mathcal{C}$. The following definition is necessary: for any face \mathcal{F} of a closed convex cone \mathcal{A} , define its dual face $\mathcal{F}^\Delta = \mathcal{A}^* \cap \mathcal{F}^\perp$ (Stoer and Witzgall, 1970, Section 2.13), where the superscript \perp denotes orthogonal complement (that is, $\mathcal{B}^\perp = \{p : \forall X \in \mathcal{B} : X \cdot p = 0\}$). If for two faces \mathcal{F}_1 and \mathcal{F}_2 of \mathcal{A} we have that \mathcal{F}_1 is a face of \mathcal{F}_2 , then \mathcal{F}_2^Δ is face of \mathcal{F}_1^Δ (Tam, 1985, Proposition 2.4). In fact, if all faces of \mathcal{A} are exposed, then the mapping between faces of \mathcal{A} and its dual is one-to-one and onto, in such a way that \mathcal{F}_1 is a face of \mathcal{F}_2 iff \mathcal{F}_2^Δ is face of \mathcal{F}_1^Δ (Tam, 1985, Corollary 2.6). In particular if $\text{cl}\mathcal{D}$ is generated by a finite number of gambles, then all its faces are exposed and the mapping is indeed one-to-one and onto the faces of $\text{cl}\mathcal{C}$ (Stoer and Witzgall, 1970, Theorem 2.13.2). Of course, this applies similarly to faces of $\text{cl}\mathcal{C}$ and its dual.

We can further refine these connections between \mathcal{D} and \mathcal{C} . For instance, if a face of $\text{cl}\mathcal{D}$ does intersect \mathcal{D} , its dual face does not intersect \mathcal{C} :

Theorem 13 *If \mathcal{F} is a face of $\text{cl}\mathcal{D}$, and $\mathcal{F} \cap \mathcal{D} \neq \emptyset$, then $\mathcal{F}^\Delta \cap \mathcal{C} = \emptyset$.*

Proof Suppose $\mathcal{F} \cap \mathcal{D} \neq \emptyset$. Pick $X \in \mathcal{F} \cap \mathcal{D}$. For any $p \in \mathcal{F}^\perp$ we must have $X \cdot p = 0$, so p cannot be in \mathcal{C} ; hence $\mathcal{F}^\perp \cap \mathcal{C} = \emptyset$ and consequently $\mathcal{F}^\Delta \cap \mathcal{C} = \mathcal{D}^* \cap \mathcal{F}^\perp \cap \mathcal{C} = \emptyset$. ■

The converse can be shown for finitely generated faces:⁶

Theorem 14 *If \mathcal{F} is a finitely generated face of $\text{cl}\mathcal{D}$, and $\mathcal{F} \cap \mathcal{D} = \emptyset$, then $\mathcal{F}^\Delta \cap \mathcal{C} \neq \emptyset$.*

Proof We have that \mathcal{F} is the conic hull of a finite set of gambles $\{X_1, \dots, X_n\}$. Suppose that no element of $\mathcal{F}^\perp = \{p : \forall X \in \mathcal{F} : X \cdot p = 0\}$ belongs to \mathcal{C} . Then for each $p \in \mathcal{C}$ there is at least a $X \in \mathcal{F}$ such that $X \cdot p > 0$. Write X as $\sum_i \alpha_i X_i$ (where all $\alpha_i \geq 0$) to obtain that $\sum_i \alpha_i X_i \cdot p > 0$; if we have $X_i \cdot p \geq 0$ for all X_i , then it must be that $X_i \cdot p > 0$ for at least one X_i . (To conclude that $X_i \cdot p \geq 0$ for all X_i , reason as follows. As any $Y \in \mathcal{F}$ is in the boundary of \mathcal{D} , for all such Y we have, for all $p \in \mathcal{C}$ and all $\epsilon > 0$, that $(Y + \epsilon) \cdot p > 0$. So for all $Y \in \mathcal{F}$ and all $p \in \mathcal{C}$ we must have $Y \cdot p \geq 0$ to satisfy $Y \cdot p > -\epsilon \sum_i p_i$ for all $\epsilon > 0$.) Consequently the convex combination $Z = \sum_{i=1}^n X_i / n$ must satisfy $Z \cdot p > 0$ for all $p \in \mathcal{C}$, and then $Z \in \mathcal{D}$. But Z must belong to \mathcal{F} , so Z cannot be in \mathcal{D} by assumption. Hence there must be an element of \mathcal{F}^\perp in \mathcal{C} ; this proves the theorem as $\mathcal{F}^\perp \cap \mathcal{C} = \mathcal{F}^\perp \cap \mathcal{C} \cap \text{cl}\mathcal{C} = \mathcal{F}^\perp \cap \mathcal{C} \cap \mathcal{D}^* = \mathcal{F}^\Delta \cap \mathcal{C}$. ■

side larger than zero is redundant). The dual cone is simply the mirror image of the polar cone: $\mathcal{A}^* = -\mathcal{A}^\circ$. Also, $\mathcal{A}^{\circ\circ} = \{X : \forall p \in \mathcal{A}^\circ : X \cdot p \leq 0\} = \{X : \forall -p \in -\mathcal{A}^\circ : X \cdot p \leq 0\} = \{X : \forall p \in \mathcal{A}^* : X \cdot p \geq 0\} = \mathcal{A}^{**}$.

6. Whether or not Theorem 14 holds for general faces is an open question.

4.4 Back to SSK-continuity

Note that SSK-continuity is satisfied by coherent preference orderings:

Proposition 15 *If \succ is a coherent preference ordering, then SSK-continuity holds.*

Proof Take $\{X_i\} \rightarrow X$ and $\{Y_i\} \rightarrow Y$ such that $X_i \succ Y_i$. Take the representing credal set \mathcal{K} ; any probability measure $\mathbb{P} \in \mathcal{K}$ satisfies $\mathbb{E}_{\mathbb{P}}[X_i - Y_i] > 0$, so $\lim_i \mathbb{E}_{\mathbb{P}}[X_i - Y_i] \geq 0$; then $\mathbb{E}_{\mathbb{P}}[\lim_i X_i] \geq \mathbb{E}_{\mathbb{P}}[\lim_i Y_i]$ as the state space is finite, hence $\mathbb{E}_{\mathbb{P}}[X] \geq \mathbb{E}_{\mathbb{P}}[Y]$. If additionally $Y \succ Z$, then $\mathbb{E}_{\mathbb{P}}[Y] > \mathbb{E}_{\mathbb{P}}[Z]$ for every $\mathbb{P} \in \mathcal{K}$, so $\mathbb{E}_{\mathbb{P}}[X] > \mathbb{E}_{\mathbb{P}}[Z]$ for every $\mathbb{P} \in \mathcal{K}$, and then $X \succ Z$ as desired. \blacksquare

The natural question is whether SSK-continuity implies even continuity. It does not; but to appreciate the matter, it is interesting to note that SSK-continuity implies even continuity in an important case. Start by considering a consequence of SSK-continuity that is quite reasonable as a property of preferences:

Proposition 16 *Suppose \succ is a preference ordering satisfying monotonicity and SSK-continuity. If $\alpha W + (1 - \alpha)X \succ Y \succ 0$ for $\alpha \in (0, 1]$, then $X \succ 0$.*

Proof Take $\alpha_i = 1/2^i$, $X_i = \alpha_i W + (1 - \alpha_i)X$ and $Y_i = Y$. As $X_i \succ Y_i$, $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, and $Y \succ 0$, SSK-continuity implies $X \succ 0$ as desired. \blacksquare

This result leads to:

Proposition 17 *Suppose \succ is a preference ordering satisfying monotonicity, cancellation, and SSK-continuity, with representing set of desirable gambles \mathcal{D} . If $X \in \mathcal{D}$ and $Y \in \text{cl}\mathcal{D}$, then $\alpha X + (1 - \alpha)Y \in \mathcal{D}$ for $\alpha \in (0, 1)$.*

Proof Take $X \in \mathcal{D}$, $Y \in \text{cl}\mathcal{D}$, $\alpha \in (0, 1)$, and $Z = \alpha X + (1 - \alpha)Y$. For some $\delta > 0$ we have $Y + \delta \in \text{relint}\mathcal{D}$ by monotonicity; hence $\beta(Y + \delta) + (1 - \beta)Y \in \mathcal{D}$ for $\beta \in (0, 1]$ (Rockafellar, 1970, Theorem 6.1). Note that $Y = \gamma Z - \alpha\gamma X$ where $\gamma = (1 - \alpha)^{-1}$; thus $\beta(\gamma Z - \alpha\gamma X + \delta) + (1 - \beta)(\gamma Z - \alpha\gamma X) \succ 0$. Hence $\beta(\gamma Z + \delta) + (1 - \beta)(\gamma Z) \succ \alpha\gamma X$ for $\beta \in (0, 1]$. By assumption $X \succ 0$, so $\alpha\gamma X \succ 0$; by Proposition 16, we obtain $\gamma Z \succ 0$, hence $Z \in \mathcal{D}$ as desired. \blacksquare

As noted by (Fenchel, 1952, Section 3.5), a cone \mathcal{A} whose closure is the intersection of finitely many closed halfspaces is evenly convex iff it satisfies: if $X \in \mathcal{A}$ and $Y \in \text{cl}\mathcal{A}$, then the segment between X and Y is in \mathcal{A} . Hence:

Theorem 18 *Suppose \succ is a preference ordering satisfying monotonicity, cancellation, and SSK-continuity, with representing set of desirable gambles \mathcal{D} . If the closure of \mathcal{D} is the intersection of finitely many closed halfspaces, then \mathcal{D} is evenly convex.*

That is, SSK-continuity produces even convexity of the set of desirable gambles, and therefore of the representing credal set, when only finitely many assessments affect preferences. However, in general SSK-continuity does not enforce even convexity of sets of desirable gambles.

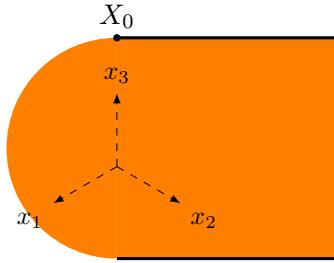


Figure 3: The set \mathcal{B} in Example 1, viewed from point $(1, 1, 1)$.

To understand how this is possible, take a coherent preference ordering \succ' and its representing set of desirable gambles \mathcal{D}' . Suppose \mathcal{D}' contains a *non-exposed* but extreme ray R_0 that goes through gamble X_0 (that is, $R_0 = \{\lambda X_0 : \lambda > 0\}$). Define $\mathcal{D}'' = \mathcal{D}' \setminus R_0$; this is still a convex set (hence a convex cone) containing the positive octant. We can then define a preference ordering \succ'' as $X \succ'' Y$ iff $X - Y \in \mathcal{D}''$. Note that \mathcal{D}'' is not an evenly convex set, but \succ'' built as described satisfies SSK-continuity as we argue in the remainder of this section. However, before we plunge into the arguments, consider a concrete example:

Example 1 Suppose $\Omega = \{\omega_1, \omega_2, \omega_3\}$; a gamble is a triple of numbers (x_1, x_2, x_3) , meaning $(X(\omega_1), X(\omega_2), X(\omega_3))$. Consider \mathcal{B} as the union of the open circle with center $(1/4, 1/4, 1/2)$ and radius $\sqrt{3}/2$ drawn on the simplex consisting of $x_1 + x_2 + x_3 = 1$, and the closed polygon with four vertices $(3/4, 3/4, -1/2)$, $(-1/4, -1/4, 3/2)$, $(-2, 3/2, 3/2)$, $(-1, 5/2, -1/2)$, and take $X_0 = (-1/4, -1/4, 3/2)$, a non-exposed extreme point of \mathcal{B} . Figure 3 depicts the set \mathcal{B} . Take the cone \mathcal{D}'' as the set of all rays emanating from the origin and going through points of \mathcal{B} except X_0 . This cone \mathcal{D}'' produces a preference ordering that satisfies SSK-continuity. \square

We now show that the preference ordering \succ'' induced by \mathcal{D}'' satisfies SSK-continuity.

As the cone \mathcal{D}' is evenly convex, we can build its representing credal set \mathcal{K}' . By construction $X \succ'' 0$ implies that for all $\mathbb{P} \in \mathcal{K}'$ we have $\mathbb{E}_{\mathbb{P}}[X] > 0$; also by construction $X \succ'' 0$ implies $X \notin R_0$. Also, if for all $\mathbb{P} \in \mathcal{K}'$ we have $\mathbb{E}_{\mathbb{P}}[X] > 0$ and $X \notin R_0$, then $X \succ'' 0$. That is, we have the representation: $X \succ'' 0 \Leftrightarrow (X \notin R_0) \wedge (\forall \mathbb{P} \in \mathcal{K}' : \mathbb{E}_{\mathbb{P}}[X] > 0)$.

By Proposition 5, we need to show that $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, $X_i \succ'' Y_i$, $Y \succ'' 0$ imply $X \succ'' 0$. If $Y \in R_0$, then $Y \succ'' 0$ is false and there is nothing to prove; hence assume that $Y \notin R_0$. We distinguish two cases: $X \notin R_0$ and $X \in R_0$.

Take $X \notin R_0$. To prove that $X \succ'' 0$, note that $\mathbb{E}_{\mathbb{P}}[X_i - Y_i] > 0$ for every $\mathbb{P} \in \mathcal{K}'$, so $\lim_i \mathbb{E}_{\mathbb{P}}[X_i - Y_i] \geq 0$ and therefore $\mathbb{E}_{\mathbb{P}}[X] \geq \mathbb{E}_{\mathbb{P}}[Y]$ for $\mathbb{P} \in \mathcal{K}'$. Thus $\mathbb{E}_{\mathbb{P}}[X] \geq \mathbb{E}_{\mathbb{P}}[Y] > \mathbb{E}_{\mathbb{P}}[0] = 0$ and then $\mathbb{E}_{\mathbb{P}}[X] > 0$ for every $\mathbb{P} \in \mathcal{K}'$, implying $X \succ'' 0$ as desired.

Now take $X \in R_0$; note that X is in an extreme ray of $\text{cl}\mathcal{D}'$. In the next paragraph we show that if $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, $X_i \succ'' Y_i$, then $Y \succ'' 0$ must be false. Hence it is irrelevant to consider $X \in R_0$ as the premise of SSK-continuity is never satisfied in this case, and the proof is finished.

To conclude we show that, if \mathcal{A} is a convex cone, $\{X_i\} \rightarrow X$, $\{Y_i\} \rightarrow Y$, $X_i - Y_i \in \mathcal{A}$, and X belongs to an extreme ray of $\text{cl}\mathcal{A}$ but $X \notin \mathcal{A}$, then $Y \notin \mathcal{A}$. We have that $Y \in \text{cl}\mathcal{A}$ and, as $X_i - Y_i \in \mathcal{A}$ for every i , $X - Y \in \text{cl}\mathcal{A}$ (the closure is the set of limiting points). So we have both Y and $X - Y$ in $\text{cl}\mathcal{A}$. If $Y \neq \lambda X$, then $X/2$ is the convex combination $Y/2 + (X - Y)/2$ of two

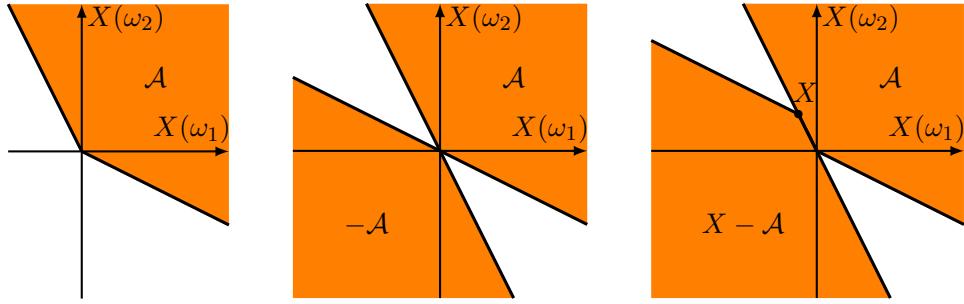


Figure 4: A closed convex cone \mathcal{A} (left), the cones \mathcal{A} and $-\mathcal{A}$ (middle), and the cones \mathcal{A} and $X - \mathcal{A}$ for X in an extreme ray of \mathcal{A} (right).

points not in the ray containing X , a contradiction with the assumption that X is in an extreme ray of $\text{cl}\mathcal{A}$. So $Y = \lambda X$ for some λ , and then $Y \notin \mathcal{A}$. (This result is illustrated by Figure 4: Y must belong to the closure of \mathcal{A} and to the closure of $X - \mathcal{A}$, so it belongs to the line from the origin through X .)

5. Conclusion

We have presented a few axioms on preference orderings that, together, imply a representation through evenly convex credal sets. This representation lets one handle assessments of strict inequality for probabilities, and go beyond what can be done with closed convex credal sets. The main idea is to adopt a novel Archimedean condition (even continuity) that implies even convexity. A similar representation can be obtained using SSK-continuity in many, but not all, cases.

Future work should look at natural and similar extensions, as well as to conditioning and independence. It should also be possible to use our proposed Archimedean condition to obtain *general* sets of probabilities, mimicking results by [Seidenfeld et al. \(2010\)](#).

Acknowledgement

The author is partially supported by the CNPq grant # 308433/2014-9 (PQ). This paper was partially funded by FAPESP grant #2015/21880-4 (project Proverbs).

References

- R. J. Aumann. Utility theory without the completeness axiom. *Econometrica*, 30(3):445–462, 1962.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- A. Brondsted. *An Introduction to Convex Polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer-Verlag Inc., New York, 83.
- A. Daniilidis and J.-E. Martínez-Legaz. Characterizations of evenly convex sets and evenly quasi-convex functions. *Journal of Mathematical Analysis and Applications*, 273:58–66, 2002.

- W. Fenchel. A remark on convex sets and polarity. *Meddelanden Lunds Universitets Matematiska, Tome Supplementaire*:82–89, 1952.
- P. C. Fishburn. *Utility Theory for Decision Making*. John Wiley and Sons, Inc., New York, 1970.
- F. J. Giron and S. Rios. Quasi-Bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 17–38. University Press, Valencia, Spain, 1980.
- M. A. Goberna, V. Jornet, and M. M. L. Rodriguez. On linear systems containing strict inequalities. *Linear Algebra and its Applications*, 360:151–171, 2003.
- V. L. Klee Jr. Separation properties of convex cones. *Proceedings of the American Mathematical Society*, 6(2):313–318, 1955.
- V. Klee. Maximal separation theorems for convex sets. *Transactions of the American Mathematical Society*, 134(1):133–147, 1968.
- S. R. Lay. *Convex Sets and Their Applications*. Wiley, New York, 1982.
- I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.
- E. Miranda and M. Zaffalon. Notes on desirability and conditional lower previsions. *Annals of Mathematics and Artificial Intelligence*, 60(3-4):251–309.
- E. Quaeghebeur. Desirability. In T. Augustin, F. P. A. Coolen, G. De Cooman, M. C. M. Troffaes (editors), *Introduction to Imprecise Probabilities*, pages 1–27, Wiley, 2014.
- R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Decisions without ordering. In W. Sieg, editor, *Acting and Reflecting*, pages 143–170. Kluwer Academic Publishers, 1990.
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. A representation of partially ordered preferences. *Annals of Statistics*, 23(6):2168–2217, 1995.
- T. Seidenfeld, M. Schervish, and J. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172(1):157–176, 2010.
- J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions I*. Springer-Verlag, 1970.
- B.-S. Tam. On the duality operator of a convex cone. *Linear Algebra and its Applications*, 64: 33–56, 1985.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.
- P. M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., University of Sussex, 1975.

Independent Natural Extension for Infinite Spaces: Williams-Coherence to the Rescue

Jasper De Bock

JASPER.DEBOCK@UGENT.BE

Ghent University - imec, IDLab, ELIS

Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde (Belgium)

Abstract

We define the independent natural extension of two local models for the general case of infinite spaces, using both sets of desirable gambles and conditional lower previsions. In contrast to [Miranda and Zaffalon \(2015\)](#), we adopt Williams-coherence instead of Walley-coherence. We show that our notion of independent natural extension always exists—whereas theirs does not—and that it satisfies various convenient properties, including factorisation and external additivity.

Keywords: independent natural extension; epistemic independence; Williams-coherence; infinite spaces; external additivity; factorisation; sets of desirable gambles; conditional lower previsions.

1. Introduction

When probabilities are imprecise, in the sense that they are only partially specified, it is no longer clear what it means for two variables to be independent ([Couso et al., 1999](#)). One approach is to apply the standard notion of independence to every element of some set of probability measures. The alternative, called epistemic independence, is to define independence as mutual irrelevance, in the sense that receiving information about one of the variables will not effect our uncertainty model for the other. The advantage of this intuitive alternative is that it has a much wider scope: since epistemic independence is expressed in terms of uncertainty models instead of probabilities, it can easily be applied to a variety of such models, including non-probabilistic ones; we here consider sets of desirable gambles and conditional lower previsions.

When an assessment of epistemic independence is combined with local uncertainty models, it leads to a unique corresponding joint uncertainty model that is called the independent natural extension. If the variables involved can take only a finite number of values, this independent natural extension always exists, and it then satisfies various convenient properties that allow for the design of efficient algorithms ([de Cooman et al., 2011](#); [de Cooman and Miranda, 2012](#)). If the variables involved take values in an infinite set, the situation becomes more complicated. On the one hand, for the specific case of lower probabilities, [Vicig \(2000\)](#) managed to obtain results that resemble the finite case. On the other hand, for the more general case of lower previsions, [Miranda and Zaffalon \(2015\)](#) recently found that the independent natural extension may not even exist.

Our present contribution generalises the results of [Vicig \(2000\)](#) to the case of conditional lower previsions, using sets of desirable gambles as an intermediate step. The key technical difference with [Miranda and Zaffalon \(2015\)](#) is that we use Williams-coherence instead of Walley-coherence. This difference turns out to be crucial because our notion of independent natural extension always exists. Furthermore, as we will see, it satisfies the same convenient properties that are known to hold in the finite case, including factorisation and external additivity.

Proofs are provided in the appendix of the arXiv version of this paper (De Bock, 2017), which had to be omitted from the published version because of the page limit constraint.

2. Preliminaries and Notation

We use \mathbb{N} to denote the natural numbers without zero and let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. \mathbb{R} is the set of real numbers and \mathbb{Q} is the set of rational numbers. Sign restrictions are imposed with subscripts. For example, we let $\mathbb{R}_{>0}$ be the set of positive real numbers and let $\mathbb{Q}_{\geq 0}$ be the set of non-negative rational numbers. The extended real numbers are denoted by $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.

For any non-empty set \mathcal{X} , the power set of \mathcal{X} —the set of all subsets of \mathcal{X} —is denoted by $\mathcal{P}(\mathcal{X})$, and we let $\mathcal{P}_0(\mathcal{X}) := \mathcal{P}(\mathcal{X}) \setminus \{\emptyset\}$ be the set of all non-empty subsets of \mathcal{X} . Elements of $\mathcal{P}(\mathcal{X})$ are called events. A set of events $\mathcal{B} \subseteq \mathcal{P}(\mathcal{X})$ is called a field if it is non-empty and closed with respect to complements and finite intersections and unions. If it is also closed with respect to countable intersections and unions, it is called a sigma field. A partition of \mathcal{X} is a set $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$ of pairwise disjoint non-empty subsets of \mathcal{X} whose union is equal to \mathcal{X} . We also adopt the notational trick of identifying \mathcal{X} with the set of atoms $\{\{x\} : x \in \mathcal{X}\}$, which allows us to regard \mathcal{X} as a partition of \mathcal{X} .

A bounded real-valued function on \mathcal{X} will be called a gamble on \mathcal{X} . The set of all gambles on \mathcal{X} is denoted by $\mathcal{G}(\mathcal{X})$, the set of all non-negative gambles on \mathcal{X} is denoted by $\mathcal{G}_{\geq 0}(\mathcal{X})$, and we let $\mathcal{G}_{>0}(\mathcal{X}) := \mathcal{G}_{\geq 0}(\mathcal{X}) \setminus \{0\}$ be the set of all non-negative non-zero gambles. For any set of gambles $\mathcal{A} \subseteq \mathcal{G}(\mathcal{X})$, we let

$$\text{posi}(\mathcal{A}) := \left\{ \sum_{i=1}^n \lambda_i f_i : n \in \mathbb{N}, \lambda_i \in \mathbb{R}_{>0}, f_i \in \mathcal{A} \right\} \quad (1)$$

and

$$\mathcal{E}(\mathcal{A}) := \text{posi}(\mathcal{A} \cup \mathcal{G}_{>0}(\mathcal{X})). \quad (2)$$

Indicators are a particular type of gamble. For any $A \in \mathcal{P}(\mathcal{X})$, the corresponding indicator \mathbb{I}_A of A is a gamble in $\mathcal{G}(\mathcal{X})$, defined for all $x \in \mathcal{X}$ by $\mathbb{I}_A(x) := 1$ if $x \in A$ and $\mathbb{I}_A(x) := 0$ otherwise.

Finally, for any $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$, we will also require the notion of a non-negative \mathcal{B} -measurable gamble, which we define as a uniform limit of simple \mathcal{B} -measurable gambles.

Definition 1 Let $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$. We call $g \in \mathcal{G}_{\geq 0}(\mathcal{X})$ a simple \mathcal{B} -measurable gamble if there are $c_0 \in \mathbb{R}_{\geq 0}$, $n \in \mathbb{N}_0$ and, for all $i \in \{1, \dots, n\}$, $c_i \in \mathbb{R}_{\geq 0}$ and $B_i \in \mathcal{B}$, such that $g = c_0 + \sum_{i=1}^n c_i \mathbb{I}_{B_i}$.

Definition 2 Let $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$. A gamble $g \in \mathcal{G}_{\geq 0}(\mathcal{X})$ is \mathcal{B} -measurable if it is a uniform limit of non-negative simple \mathcal{B} -measurable gambles, in the sense that there is a sequence $\{g_n\}_{n \in \mathbb{N}}$ of simple \mathcal{B} -measurable gambles in $\mathcal{G}_{\geq 0}(\mathcal{X})$ such that $\lim_{n \rightarrow +\infty} \sup |g - g_n| = 0$.

Readers that are familiar with the concepts of simple and measurable functions that are common in measure theory will observe some similarities. However, there are also some important differences. On the one hand, our definitions are more restrictive: we only consider bounded non-negative functions, Definition 1 requires that the coefficients c_i are non-negative, and Definition 2 considers uniform limits instead of pointwise limits. On the other hand, our definitions are more general because we allow for \mathcal{B} to be any subset of $\mathcal{P}_0(\mathcal{X})$. Nevertheless, if $\mathcal{B} \cup \{\emptyset\}$ is a sigma field, we have the following equivalence.

Proposition 3 Consider any $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$ such that $\mathcal{B}^* := \mathcal{B} \cup \{\emptyset\}$ is a sigma field. Then for any $g \in \mathcal{G}_{\geq 0}(\mathcal{X})$, g is \mathcal{B}^* -measurable in the measure-theoretic sense ([Nielsen, 1997](#), Definition 10.1) if and only if it is \mathcal{B} -measurable in the sense of Definition 2.

The proof of this result is based on the following sufficient condition for \mathcal{B} -measurability, which provides a convenient tool for establishing the \mathcal{B} -measurability of a given function. In particular, it implies that every non-negative gamble is $\mathcal{P}_0(\mathcal{X})$ -measurable.

Proposition 4 Let $\mathcal{B} \subseteq \mathcal{P}_0(\mathcal{X})$ and $g \in \mathcal{G}_{\geq 0}(\mathcal{X})$. If, for all $r \in \mathbb{Q}_{\geq 0}$, the set $\{x \in \mathcal{X} : g(x) \geq r\}$ is a finite union of pairwise disjoint events in $\mathcal{B} \cup \{\mathcal{X}, \emptyset\}$, then g is \mathcal{B} -measurable.

Corollary 5 Every $g \in \mathcal{G}_{\geq 0}(\mathcal{X})$ is $\mathcal{P}_0(\mathcal{X})$ -measurable.

3. Modelling Uncertainty

A subject's uncertainty about a variable X that takes values x in some non-empty set \mathcal{X} can be mathematically represented in various ways. The most popular such method is perhaps probability theory, but it is by no means the only one, nor is it the most general one. In order for our results to have a broader scope, we here adopt the frameworks of sets of desirable gambles and conditional lower previsions.

The main aim of this section is to provide an overview of the basic technical aspects of these frameworks, as these will be essential to the rest of the paper. Notably, we do not impose any constraints on the cardinality of \mathcal{X} : it may be finite, countably infinite or uncountably infinite. Connections with other—perhaps better known—models for uncertainty, including probability theory, will be discussed briefly at the end.

The basic idea behind *sets of desirable gambles* is to model a subject's uncertainty about X by considering his attitude towards gambles—bets—on \mathcal{X} . In particular, we consider the gambles $f \in \mathcal{G}(\mathcal{X})$ that he finds *desirable*, in the sense that he is willing to engage in a transaction where, once the actual value $x \in \mathcal{X}$ of X is known, he will receive a—possibly negative—reward $f(x)$ in some linear utility scale. Even more so, he prefers these desirable gambles over the status quo, that is, over not conducting any transaction at all. A set of desirable gambles is called coherent if it satisfies the following rationality requirements.

Definition 6 A coherent set of desirable gambles \mathcal{D} on \mathcal{X} is a subset of $\mathcal{G}(\mathcal{X})$ such that, for any two gambles $f, g \in \mathcal{G}(\mathcal{X})$ and any non-negative real number $\lambda \in \mathbb{R}_{>0}$:

D1: if $f \geq 0$ and $f \neq 0$, then $f \in \mathcal{D}$;

D2: if $f \in \mathcal{D}$ then $\lambda f \in \mathcal{D}$;

D3: if $f, g \in \mathcal{D}$, then $f + g \in \mathcal{D}$;

D4: if $f \leq 0$, then $f \notin \mathcal{D}$.

Despite their simplicity, sets of desirable gambles offer a surprisingly powerful framework for modelling uncertainty; see for example ([Walley, 2000](#)) and ([Quaeghebeur, 2014](#)). For our present purposes though, all we need for now is Definition 6.

Conditional lower previsions also model a subject's uncertainty about X by considering his attitude towards gambles on \mathcal{X} . However, in this case, instead of considering sets of gambles, we consider the prices at which a subject is willing to buy these gambles. Let

$$\mathcal{C}(\mathcal{X}) := \mathcal{G}(\mathcal{X}) \times \mathcal{P}_\emptyset(\mathcal{X})$$

be the set of all pairs (f, B) , where f is a gamble on \mathcal{X} and B is a non-empty subset of \mathcal{X} —an event. A conditional lower prevision is then defined as follows.

Definition 7 A conditional lower prevision \underline{P} on $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X})$ is a map

$$\underline{P}: \mathcal{C} \rightarrow \overline{\mathbb{R}}: (f, B) \mapsto \underline{P}(f|B).$$

For any (f, B) in the domain \mathcal{C} , the lower prevision $\underline{P}(f|B)$ of f conditional on B is interpreted as a subject's supremum price μ for buying f , under the condition that the transaction is called off when B does not happen—if $x \notin B$. In other words, $\underline{P}(f|B)$ is the supremum value of μ for which he is willing to engage in a transaction where he receives $f(x) - \mu$ if $x \in B$ and zero otherwise, and furthermore prefers this transaction to the status quo.

It is also possible to consider conditional upper previsions $\overline{P}(f|B)$, which are interpreted as infimum selling prices. However, since selling f for μ is equivalent to buying $-f$ for $-\mu$, we have that $\overline{P}(f|B) = -\underline{P}(-f|B)$. For that reason, we will mainly focus on conditional lower previsions. Unconditional lower previsions correspond to the special case where $B = \mathcal{X}$ for all $(f, B) \in \mathcal{C}$; we then use the shorthand notation $\underline{P}(f) := \underline{P}(f|\mathcal{X})$ and call $\underline{P}(f)$ the lower prevision of f . Similarly, we refer to $\overline{P}(f) := \overline{P}(f|\mathcal{X})$ as the upper prevision of f .

Because of their interpretation in terms of buying prices for gambles, a particularly intuitive way to obtain a conditional lower prevision \underline{P} is to derive it from a set of gambles \mathcal{D} . In particular, for every $\mathcal{D} \subseteq \mathcal{G}(\mathcal{X})$, we let

$$\underline{P}_{\mathcal{D}}(f|B) := \sup\{\mu \in \mathbb{R}: [f - \mu]\mathbb{I}_B \in \mathcal{D}\} \text{ for all } (f, B) \in \mathcal{C}(\mathcal{X}). \quad (3)$$

A conditional lower prevision is then called coherent if it can be derived from a coherent set of desirable gambles in this way.

Definition 8 A conditional lower prevision \underline{P} on a domain $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X})$ is coherent if there is a coherent set of desirable gambles \mathcal{D} on \mathcal{X} such that \underline{P} coincides with $\underline{P}_{\mathcal{D}}$ on \mathcal{C} .

This definition of coherence is heavily inspired by the work of Williams (1975, 2007). The only two minor differences are that our rationality axioms on \mathcal{D} are slightly different from his, and that we do not impose any structure on the domain \mathcal{C} . Nevertheless, when the domain \mathcal{C} satisfies the structural constraints in (Williams, 2007), Definition 8 is equivalent to that of Williams. More generally, as the following result establishes, it is equivalent to the structure-free notion of Williams-coherence that was developed by Pelessoni and Vicig (2009).

Proposition 9 A conditional lower prevision \underline{P} on $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X})$ is coherent if and only if it is real-valued and, for all $n \in \mathbb{N}_0$ and all choices of $\lambda_0, \dots, \lambda_n \in \mathbb{R}_{\geq 0}$ and $(f_0, B_0), \dots, (f_n, B_n) \in \mathcal{C}$:

$$\sup_{x \in B} \left(\sum_{i=1}^n \lambda_i \mathbb{I}_{B_i}(x) [f_i(x) - \underline{P}(f_i|B_i)] - \lambda_0 \mathbb{I}_{B_0}(x) [f_0(x) - \underline{P}(f_0|B_0)] \right) \geq 0, \quad (4)$$

where we let $B := \bigcup_{i=0}^n B_i$.

The advantage of this alternative characterisation is that it is expressed directly in terms of lower previsions. Nevertheless, we consider Equation (4) to be less intuitive than Definition 8, which is why we prefer the latter.

From a mathematical point of view, Definition 8 also has the advantage that it allows for simple and elegant proofs of some well-known results. For example, it follows trivially from our definition of coherence that the domain of a coherent conditional lower prevision can be arbitrarily extended while preserving coherence, whereas deriving this result directly from Equation 4 is substantially more involved; see for example the proof of (Pelessoni and Vicig, 2009, Proposition 1). Furthermore, our definition also allows for a very natural derivation of the so-called *natural extension* of \underline{P} , that is, the most conservative extension of \underline{P} to $\mathcal{C}(\mathcal{X})$. In particular, instead of having to derive this natural extension directly, Definition 8 allows us to rephrase this problem into a closely related yet simpler question: what is the smallest coherent set of desirable gambles \mathcal{D} on \mathcal{X} such that $\underline{P}_{\mathcal{D}}$ coincides with \underline{P} on \mathcal{C} ? The answer turns out to be surprisingly simple.

Proposition 10 *Consider a coherent conditional lower prevision \underline{P} on $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X})$ and let*

$$\mathcal{A}_{\underline{P}} := \{[f - \mu]\mathbb{I}_B : (f, B) \in \mathcal{C}, \mu < \underline{P}(f|B)\} \text{ and } \mathcal{E}(\underline{P}) := \mathcal{E}(\mathcal{A}_{\underline{P}}). \quad (5)$$

Then $\mathcal{E}(\underline{P})$ is a coherent set of desirable gambles on \mathcal{X} and $\underline{P}_{\mathcal{E}(\underline{P})}$ coincides with \underline{P} on \mathcal{C} . Furthermore, for any other coherent set of desirable gambles \mathcal{D} on \mathcal{X} such that $\underline{P}_{\mathcal{D}}$ coincides with \underline{P} on \mathcal{C} , we have that $\mathcal{E}(\underline{P}) \subseteq \mathcal{D}$.

Abstracting away some technical details, the reason why this result holds should be intuitively clear. First, since conditional lower previsions are interpreted as called-off supremum buying prices, we see that the gambles in $\mathcal{A}_{\underline{P}}$ should be desirable. Combined with D1–D3, the desirability of the gambles in $\mathcal{E}(\underline{P})$ then follows.

Since smaller sets of desirable gambles lead to more conservative—pointwise smaller—lower previsions, we conclude that the natural extension of \underline{P} is given by

$$\underline{E}(f|B) := \underline{P}_{\mathcal{E}(\underline{P})}(f|B) \text{ for all } (f, B) \in \mathcal{C}(\mathcal{X}). \quad (6)$$

The following proposition provides a formal statement of this result.

Proposition 11 *Let \underline{P} be a coherent conditional lower prevision on $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X})$. Then \underline{E} , as defined by Equation (6), is the pointwise smallest coherent conditional lower prevision on $\mathcal{C}(\mathcal{X})$ that coincides with \underline{P} on \mathcal{C} .*

All in all, we conclude that Definition 8 provides an intuitive as well as mathematically convenient characterisation of Williams-coherence that is furthermore equivalent to the structure-free version of Pelessoni and Vicig (2009). From a technical point of view, this equivalence will not be important further on, since all of our arguments will be based on the connection with sets of desirable gambles. From a practical point of view though, this equivalence is highly important, because the Williams-coherent conditional lower previsions that are considered in (Pelessoni and Vicig, 2009) are well-known to include as special cases a variety of other uncertainty models, including expectations, lower expectations, probabilities, lower probabilities and belief functions; lower probabilities, for example, can be obtained by restricting the domain of \underline{P} to indicators. For that reason, all of our results can be immediately applied to these special cases as well. A detailed treatment of these special cases, however, does not fit within the page constraints of this contribution, and therefore falls beyond the scope of our present work.

4. Epistemic Independence

Having introduced our main tools for modelling uncertainty, the next step towards developing a notion of independent natural extension is to agree on what we mean by independence. Within the context of lower previsions, there are basically two main options.

The first approach, which we will not consider here, is to interpret lower previsions as lower expectations, that is, as tight lower bounds on the expectations that correspond to some set of probability measures, and to then impose the usual notion of independence on each of the probability measures in that set. This approach has the advantage of being familiar, but is restricted in scope because it can only be applied to uncertainty models that are expressed in terms of probabilities.

The second approach, which is the one that we will adopt here, is to regard independence as an assessment of mutual irrelevance. In particular, we say that X_1 and X_2 are independent if our uncertainty model for X_1 is not affected by conditioning on information about X_2 , and vice versa. This definition can easily be applied to a probability measure, and then yields the usual notion of independence. However, and that is what makes this approach powerful and intuitive, it can just as easily be applied to lower previsions, sets of desirable gambles, or any other type of uncertainty model. This type of independence is usually referred to as epistemic independence. The aim of this section is to formalize this concept for the case of two variables, in terms of sets of desirable gambles and conditional lower previsions.

Consider two variables X_1 and X_2 where, for every $i \in \{1, 2\}$, X_i takes values x_i in a non-empty set \mathcal{X}_i that may be uncountably infinite, and let $X := (X_1, X_2)$ be the corresponding joint variable that takes values $x := (x_1, x_2)$ in $\mathcal{X}_1 \times \mathcal{X}_2$. In this context, whenever convenient, we will identify $B \in \mathcal{P}_0(\mathcal{X}_1)$ with $B \times \mathcal{X}_2$ and $B \in \mathcal{P}_0(\mathcal{X}_2)$ with $\mathcal{X}_1 \times B$. Similarly, for any $i \in \{1, 2\}$, we will identify $f \in \mathcal{G}(\mathcal{X}_i)$ with its cylindrical extension to $\mathcal{G}(\mathcal{X}_1 \times \mathcal{X}_2)$, defined by

$$f(x_1, x_2) := f(x_i) \text{ for all } x = (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2.$$

In order to make this explicit, we will then often denote this cylindrical extension by $f(X_i)$. In this way, for example, for any $f \in \mathcal{G}(\mathcal{X}_2)$ and $B \in \mathcal{P}(\mathcal{X}_1)$, we can write $f(X_2)\mathbb{I}_B(X_1)$ to denote a gamble in $\mathcal{G}(\mathcal{X}_1 \times \mathcal{X}_2)$ whose value in (x_1, x_2) is equal to $f(x_2)$ if $x_1 \in B$ and equal to zero otherwise. Using these conventions, for any set of gambles \mathcal{D} on $\mathcal{X}_1 \times \mathcal{X}_2$, we define the marginal models

$$\text{marg}_1(\mathcal{D}) := \{f \in \mathcal{G}(\mathcal{X}_1) : f(X_1) \in \mathcal{D}\} \text{ and } \text{marg}_2(\mathcal{D}) := \{f \in \mathcal{G}(\mathcal{X}_2) : f(X_2) \in \mathcal{D}\}$$

and, for any events $B_1 \in \mathcal{P}_0(\mathcal{X}_1)$ and $B_2 \in \mathcal{P}_0(\mathcal{X}_2)$, the conditional models

$$\text{marg}_1(\mathcal{D}|B_2) := \{f \in \mathcal{G}(\mathcal{X}_1) : f(X_1)\mathbb{I}_{B_2}(X_2) \in \mathcal{D}\}$$

and

$$\text{marg}_2(\mathcal{D}|B_1) := \{f \in \mathcal{G}(\mathcal{X}_2) : f(X_2)\mathbb{I}_{B_1}(X_1) \in \mathcal{D}\}.$$

Conditioning and marginalisation both preserve coherence: if \mathcal{D} is a coherent set of desirable gambles on $\mathcal{X}_1 \times \mathcal{X}_2$, then $\text{marg}_1(\mathcal{D})$ and $\text{marg}_1(\mathcal{D}|B_2)$ are coherent sets of desirable gambles on \mathcal{X}_1 , and $\text{marg}_2(\mathcal{D})$ and $\text{marg}_2(\mathcal{D}|B_1)$ are coherent sets of desirable gambles on \mathcal{X}_2 .

That being said, let us now recall our informal definition of epistemic independence, which was that the uncertainty model for X_1 is not affected by conditioning on information about X_2 , and vice versa. In the context of sets of desirable gambles, this can now be formalized as follows:

$$\text{marg}_1(\mathcal{D}|B_2) = \text{marg}_1(\mathcal{D}) \text{ and } \text{marg}_2(\mathcal{D}|B_1) = \text{marg}_2(\mathcal{D}).$$

The only thing that is left to specify are the conditioning events B_1 and B_2 for which we want this condition to hold. We think that the most intuitive approach is to impose this for every $B_1 \in \mathcal{P}_0(\mathcal{X}_1)$ and $B_2 \in \mathcal{P}_0(\mathcal{X}_2)$, and will call this epistemic subset-independence. However, this is not what is usually done. The conventional approach, which we will refer to as epistemic value-independence, is to focus on singleton events of the type $B_1 = \{x_1\}$ and $B_2 = \{x_2\}$; see for example (Walley, 1991) and (de Cooman and Miranda, 2012). We believe this conventional approach to be flawed and will argue against it further on. Until then, we postpone this debate by adopting a very general approach that subsumes the former two as special cases. In particular, for every $i \in \{1, 2\}$, we simply fix a generic set of conditioning events $\mathcal{B}_i \subseteq \mathcal{P}_0(\mathcal{X}_i)$. Epistemic value-independence corresponds to choosing $\mathcal{B}_i = \mathcal{X}_i$, whereas epistemic subset-independence corresponds to choosing $\mathcal{B}_i = \mathcal{P}_0(\mathcal{X}_i)$.

For sets of desirable gambles, this leads us to the following definition.

Definition 12 Let \mathcal{D} be a coherent set of desirable gambles on $\mathcal{X}_1 \times \mathcal{X}_2$. Then \mathcal{D} is epistemically independent if, for any i and j such that $\{i, j\} = \{1, 2\}$:

$$\text{marg}_i(\mathcal{D}|B_j) = \text{marg}_i(\mathcal{D}) \text{ for all } B_j \in \mathcal{B}_j.$$

For coherent lower previsions, as a prerequisite for defining epistemic independence, we require that the domain $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X}_1 \times \mathcal{X}_2)$ is independent, by which we mean that for any i and j such that $\{i, j\} = \{1, 2\}$, any pair $(f_i, B_i) \in \mathcal{C}(\mathcal{X}_i)$ and any event $B_j \in \mathcal{B}_j$:

$$(f_i, B_i) \in \mathcal{C} \Leftrightarrow (f_i, B_i \cap B_j) \in \mathcal{C}. \quad (7)$$

Other than that, we impose no restrictions on \mathcal{C} ; its elements $(f, B) \in \mathcal{C}$ are for example not restricted to the types that appear in Equation (7). As a result, the following definition of epistemic independence is applicable beyond the context of lower previsions. For example, by restricting the domain to indicators, we obtain a notion of epistemic independence that applies to conditional lower probabilities. A detailed discussion of these special cases, however, is left as future work.

Definition 13 Let $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X}_1 \times \mathcal{X}_2)$ be an independent domain. A coherent conditional lower prevision \underline{P} on \mathcal{C} is then epistemically independent if, for any i and j such that $\{i, j\} = \{1, 2\}$:

$$\underline{P}(f_i|B_i) = \underline{P}(f_i|B_i \cap B_j) \text{ for all } (f_i, B_i) \in \mathcal{C} \text{ and } B_j \in \mathcal{B}_j.$$

Another important feature of this definition is that B_j is not only irrelevant to unconditional local lower previsions of the form $\underline{P}(f_i)$ —in the sense that $\underline{P}(f_i) = \underline{P}(f_i|B_j)$ —but also to conditional local lower previsions such as $\underline{P}(f_i|B_i)$ —in the sense that $\underline{P}(f_i|B_i) = \underline{P}(f_i|B_i \cap B_j)$. This type of irrelevance is called h-irrelevance; see Cozman (2013) and De Bock (2015). Note, however, that this feature is optional within our framework; it only appears when \mathcal{C} is sufficiently large. If $B_i = \mathcal{X}_i$ for all $(f_i, B_i) \in \mathcal{C}$, our definition reduces to the simple requirement that $\underline{P}(f_i) = \underline{P}(f_i|B_j)$.

5. The Independent Natural Extension

All of that being said, we are now finally ready to introduce our central object of interest, which is the *independent natural extension*. Basically, the question to which this concept provides an answer is always the same: given two local uncertainty models and an assessment of epistemic

independence, what then should be the corresponding joint model? The answer, however, depends on the specific framework that is being considered.

Within the framework of sets of desirable gambles, the local uncertainty models are coherent sets of desirable gambles. In particular, for each $i \in \{1, 2\}$, we are given a coherent set of desirable gambles \mathcal{D}_i on \mathcal{X}_i . The aim is to combine these local models with an assessment of epistemic independence to obtain a coherent set of desirable gambles \mathcal{D} on $\mathcal{X}_1 \times \mathcal{X}_2$. The first requirement on \mathcal{D} , therefore, is that it should have \mathcal{D}_1 and \mathcal{D}_2 as its marginals, in the sense that $\text{marg}_i(\mathcal{D}) = \mathcal{D}_i$ for all $i \in \{1, 2\}$. The second is that \mathcal{D} should be epistemically independent. If both requirements are met, \mathcal{D} is called an independent product of \mathcal{D}_1 and \mathcal{D}_2 . The most conservative among these independent products is called the independent natural extension.

Definition 14 *An independent product of \mathcal{D}_1 and \mathcal{D}_2 is an epistemically independent coherent set of desirable gambles \mathcal{D} on $\mathcal{X}_1 \times \mathcal{X}_2$ that has \mathcal{D}_1 and \mathcal{D}_2 as its marginals.*

Definition 15 *The independent natural extension of \mathcal{D}_1 and \mathcal{D}_2 is the smallest independent product of \mathcal{D}_1 and \mathcal{D}_2 .*

If all we know is that \mathcal{D} is epistemically independent and has \mathcal{D}_1 and \mathcal{D}_2 as its marginal models, then the safest choice for \mathcal{D} —the only choice that does not require any additional assessments—is their independent natural extension, provided of course that it exists. In order to show that it always does, we let

$$\mathcal{D}_1 \otimes \mathcal{D}_2 := \mathcal{E}(\mathcal{A}_{1 \rightarrow 2} \cup \mathcal{A}_{2 \rightarrow 1}), \quad (8)$$

with

$$\mathcal{A}_{1 \rightarrow 2} := \{f_2(X_2) \mathbb{I}_{B_1}(X_1) : f_2 \in \mathcal{D}_2, B_1 \in \mathcal{B}_1 \cup \{\mathcal{X}_1\}\} \quad (9)$$

and

$$\mathcal{A}_{2 \rightarrow 1} := \{f_1(X_1) \mathbb{I}_{B_2}(X_2) : f_1 \in \mathcal{D}_1, B_2 \in \mathcal{B}_2 \cup \{\mathcal{X}_2\}\}. \quad (10)$$

The following result establishes that $\mathcal{D}_1 \otimes \mathcal{D}_2$ is the independent natural extension of \mathcal{D}_1 and \mathcal{D}_2 .

Theorem 16 *$\mathcal{D}_1 \otimes \mathcal{D}_2$ is the independent natural extension of \mathcal{D}_1 and \mathcal{D}_2 .*

Similar concepts can be defined for conditional lower previsions as well. In that case, the local uncertainty models are coherent conditional lower previsions. In particular, for every $i \in \{1, 2\}$, we are given a coherent conditional lower revision \underline{P}_i on some freely chosen local domain $\mathcal{C}_i \subseteq \mathcal{C}(\mathcal{X}_i)$. The aim is now to construct an epistemically independent coherent conditional lower revision \underline{P} on $\mathcal{C} \subseteq \mathcal{C}(\mathcal{X}_1 \times \mathcal{X}_2)$ that has \underline{P}_1 and \underline{P}_2 as its marginals, in the sense that \underline{P} coincides with \underline{P}_1 and \underline{P}_2 on their local domain: $\underline{P}(f_i | B_i) = \underline{P}_i(f_i | B_i)$ for all $i \in \{1, 2\}$ and $(f_i, B_i) \in \mathcal{C}_i$. As before, a model that meets these criteria is then called an independent product, and the most conservative among them is called the independent natural extension. Clearly, in order for these notions to make sense, the global domain \mathcal{C} must at least include the local domains \mathcal{C}_1 and \mathcal{C}_2 and must furthermore be independent in the sense of Equation (7). The definitions and results below take this for granted.

Definition 17 *An independent product of \underline{P}_1 and \underline{P}_2 is an epistemically independent coherent conditional lower revision on \mathcal{C} that has \underline{P}_1 and \underline{P}_2 as its marginals.*

Definition 18 *The independent natural extension of \underline{P}_1 and \underline{P}_2 is the point-wise smallest independent product of \underline{P}_1 and \underline{P}_2 .*

Here too, if all we know is that \underline{P} is epistemically independent and has \underline{P}_1 and \underline{P}_2 as its marginal models, then the safest choice for \underline{P} —the only choice that does not require any additional assessments—is the independent natural extension, provided that it exists. The following result establishes that it does, by showing that it is a restriction of the operator $\underline{P}_1 \otimes \underline{P}_2$, defined by

$$(\underline{P}_1 \otimes \underline{P}_2)(f|B) := \underline{P}_{\mathcal{D}}(f|B) \text{ for all } (f, B) \in \mathcal{C}(\mathcal{X}_1 \times \mathcal{X}_2), \text{ with } \mathcal{D} = \mathcal{E}(\underline{P}_1) \otimes \mathcal{E}(\underline{P}_2). \quad (11)$$

Theorem 19 *The independent natural extension of \underline{P}_1 and \underline{P}_2 is the restriction of $\underline{P}_1 \otimes \underline{P}_2$ to \mathcal{C} .*

Interestingly, as can be seen from this result, the choice of the joint domain \mathcal{C} does not affect the resulting independent natural extension, in the sense that any \mathcal{C} that includes (f, B) will lead to the same value of $(\underline{P}_1 \otimes \underline{P}_2)(f|B)$. For that reason, we will henceforth assume without loss of generality that $\mathcal{C} = \mathcal{C}(\mathcal{X}_1 \times \mathcal{X}_2)$.

6. On the Choice of Conditioning Events

The fact that the existence results in the previous section are valid regardless of the choice of \mathcal{B}_1 and \mathcal{B}_2 should not be taken to mean that this choice does not affect the model. In some cases, it most definitely does. In the remainder of this contribution, we will study the extend to which it does, and how it affects the properties of the resulting notion of independent natural extension.

As a first observation, we note that larger sets of conditioning events correspond to stronger assessments of epistemic independence, and therefore lead to more informative joint models. For example, as can be seen from Equations (8)–(10), adding events to \mathcal{B}_1 and \mathcal{B}_2 leads to a larger—more informative—set of desirable gambles $\mathcal{D}_1 \otimes \mathcal{D}_2$. Similarly, as can be seen from Equation (11), it leads to a joint lower prevision that is higher—and therefore again more informative. There is one important exception to this observation though, which occurs when we add conditioning events that are a finite disjoint union of other conditioning events. In that case, the resulting notion of independent natural extension does not change.

Proposition 20 *For each $i \in \{1, 2\}$, let \mathcal{B}'_i be a superset of \mathcal{B}_i that consists of finite disjoint unions of events in \mathcal{B}_i . Replacing \mathcal{B}_1 by \mathcal{B}'_1 and \mathcal{B}_2 by \mathcal{B}'_2 then has no effect on the resulting independent natural extension $\mathcal{D}_1 \otimes \mathcal{D}_2$ or $\underline{P}_1 \otimes \underline{P}_2$.*

As a particular case of this result, it follows that if \mathcal{B}_i is a finite partition of \mathcal{X}_i , we can replace it by the generated algebra—minus the empty event. As an even more particular case, if \mathcal{X}_1 and \mathcal{X}_2 are finite, we find that epistemic value- and subset-independence lead to the same notion of independent natural extension. For that reason, in the finite case, it does not really matter which of these two types of epistemic independence is adopted.

In the infinite case though, the difference does matter, and the debate between epistemic value- and subset-independence remains open. For lower previsions, [Miranda and Zaffalon \(2015\)](#) recently adopted epistemic value-independence in combination with Walley-coherence. Unfortunately, they found that the corresponding notion of independent natural extension does not always exist. They also considered the combination of epistemic value-independence with Williams-coherence, and argued that the resulting model was too weak. For the case of lower probabilities, [Vicig \(2000\)](#) adopted epistemic subset-independence in combination with Williams-coherence, showed that the corresponding independent natural extension always exists, and proved that it satisfies factorisation properties. Our results so far can be regarded as a generalisation of the existence results of [Vicig \(2000\)](#). As we are about to show, his factorisation results can be generalised as well.

7. Factorisation and External Additivity

When \mathcal{X}_1 and \mathcal{X}_2 are finite, the independent natural extension of two lower previsions \underline{P}_1 and \underline{P}_2 is well-known to satisfy the properties of factorisation and external additivity (de Cooman et al., 2011). Factorisation, on the one hand, states that

$$(\underline{P}_1 \otimes \underline{P}_2)(gh) = \underline{P}_1(g\underline{P}_2(h)) = \begin{cases} \underline{P}_1(g)\underline{P}_2(h) & \text{if } \underline{P}_2(h) \geq 0 \\ \bar{\underline{P}}_1(g)\underline{P}_2(h) & \text{if } \underline{P}_2(h) \leq 0, \end{cases} \quad (12)$$

where g is a non-negative gamble on \mathcal{X}_1 , h is a gamble on \mathcal{X}_2 and $\bar{\underline{P}}_1(g) := -\underline{P}_1(-g)$. By symmetry, the role of 1 and 2 can of course be reversed. External additivity, on the other hand, states that

$$(\underline{P}_1 \otimes \underline{P}_2)(f+h) = \underline{P}_1(f) + \underline{P}_2(h) \quad (13)$$

where f and h are gambles on \mathcal{X}_1 and \mathcal{X}_2 , respectively.

Compared to the properties that are satisfied by the joint expectation of a product measure of two precise probability measures, these notions of factorisation and external additivity are rather weak. For example, for a precise product measure, additivity is not ‘external’, in the sense that f and h do not have to be defined on separate variables, nor does factorisation require g to be non-negative. Nevertheless, even in this weaker form, these properties remain of crucial practical importance. For example, in the context of credal networks—Bayesian networks whose local models are imprecise—they turned out to be the key to the development of efficient inference algorithms; see for example de Cooman et al. (2010), De Bock and de Cooman (2014) and De Bock (2015). Any notion of independent natural extension that aims to extend such algorithms to infinite spaces, therefore, should preserve some suitable version of Equations (12) and (13).

The aim of this section is to study the extent to which these equations are satisfied by the notion of independent natural extension that was developed in this paper. As we will see, the answer ends up being surprisingly positive.

For all $i \in \{1, 2\}$, let \underline{P}_i be a coherent conditional lower prevision on $\mathcal{C}_i \subseteq \mathcal{C}(\mathcal{X}_i)$, let \underline{E}_i be its natural extension to $\mathcal{C}(\mathcal{X}_i)$, and let \mathcal{B}_i be a subset of $\mathcal{P}_0(\mathcal{X}_i)$. The independent natural extension of \underline{P}_1 and \underline{P}_2 then satisfies the following three properties, the first of which implies the other two as special cases.

Theorem 21 *Let $\{i, j\} = \{1, 2\}$. For any $f \in \mathcal{G}(\mathcal{X}_i)$, $h \in \mathcal{G}(\mathcal{X}_j)$ and \mathcal{B}_i -measurable $g \in \mathcal{G}_{\geq 0}(\mathcal{X}_i)$, we then have that*

$$(\underline{P}_1 \otimes \underline{P}_2)(f+gh) = \underline{E}_i(f+g\underline{E}_j(h)).$$

Corollary 22 (Factorisation) *Let $\{i, j\} = \{1, 2\}$. For any $h \in \mathcal{G}(\mathcal{X}_j)$ and any $g \in \mathcal{G}_{\geq 0}(\mathcal{X}_i)$ that is \mathcal{B}_i -measurable, we then have that*

$$(\underline{P}_1 \otimes \underline{P}_2)(gh) = \underline{E}_i(g\underline{E}_j(h)) = \begin{cases} \underline{E}_i(g)\underline{E}_j(h) & \text{if } \underline{E}_j(h) \geq 0; \\ \bar{\underline{E}}_i(g)\underline{E}_j(h) & \text{if } \underline{E}_j(h) \leq 0. \end{cases}$$

Corollary 23 (External additivity) *For any $f \in \mathcal{G}(\mathcal{X}_1)$ and $h \in \mathcal{G}(\mathcal{X}_2)$, we have that*

$$(\underline{P}_1 \otimes \underline{P}_2)(f+h) = \underline{E}_1(f) + \underline{E}_2(h).$$

In each of these results, if the local domains \mathcal{C}_1 and \mathcal{C}_2 are sufficiently large—that is, if they include the gambles that appear in the statement of the results—it follows from Proposition 11 that \underline{E}_i and \underline{E}_j can be replaced by \underline{P}_i and \underline{P}_j , respectively, and similarly for \bar{E}_i and \bar{P}_i .

That being said, let us now go back to the question of whether or not Equations (12) and (13) can be generalised to the case of infinite spaces. For the case of external additivity, it clearly follows from Corollary 23 that the answer is fully positive. Furthermore, this conclusion holds regardless of our choice for \mathcal{B}_1 and \mathcal{B}_2 ; they can even be empty. For factorisation, the answer does depend on \mathcal{B}_1 and \mathcal{B}_2 . If we adopt epistemic subset-independence—that is, if we choose $\mathcal{B}_1 = \mathcal{P}_0(\mathcal{X}_1)$ and $\mathcal{B}_2 = \mathcal{P}_0(\mathcal{X}_2)$ —it follows from Corollaries 5 and 22 that the answer is again fully positive, because $\mathcal{P}_0(\mathcal{X}_i)$ -measurability then holds trivially. If $\mathcal{B}_1 \cup \{\emptyset\}$ and $\mathcal{B}_2 \cup \{\emptyset\}$ are sigma fields, the answer remains fairly positive as well, because Proposition 3 then implies that it suffices for g to be measurable in the usual, measure-theoretic sense. If we adopt epistemic value-independence—that is, if we choose $\mathcal{B}_1 = \mathcal{X}_1$ and $\mathcal{B}_2 = \mathcal{X}_2$ —it is necessary for g to be \mathcal{X}_i -measurable, which is a rather strong requirement that easily fails. For that reason, we think that for the case of infinite spaces, when it comes to choosing between epistemic value- and subset-independence, the latter should be preferred over the former.

8. Conclusions and Future Work

The main conclusion of this work is that by combining Williams-coherence with epistemic subset-independence, we obtain a notion of independent natural extension that always exists, and that furthermore satisfies factorisation and external additivity. For weaker types of epistemic independence, including epistemic value-independence, the existence result and the external additivity property remain valid, but factorisation then requires measurability conditions.

We foresee several lines of future research. The first, which we expect to be rather straightforward, is to extend our results from the case of two variables to that of any finite number of variables. Next, these extended versions of our results could then be used to develop efficient algorithms for credal networks whose variables take values in infinite spaces, by suitably adapting existing algorithms for the finite case. On the more technical side, it would be useful to see whether our results can be extended to the case of unbounded functions. Finally, for variables that take values in Euclidean space, \mathcal{B}_1 and \mathcal{B}_2 could be restricted to the Lebesgue measurable events. Combined with an assessment of continuity, we think that this could lead to the development of a notion of independent natural extension that includes sigma additive product measures as a special case.

Acknowledgments

I am a Postdoctoral Fellow of the Research Foundation - Flanders (FWO) and wish to acknowledge its financial support. The research that lead to this paper was conducted during a research visit—funded by an FWO travel grant—to the Imprecise Probability Group of IDSIA (Institute Dalle Molle for Artificial Intelligence), the members of which I would like to thank for their warm hospitality. Finally, I would also like to thank two anonymous reviewers, for their generous constructive comments, and Enrique Miranda, for commenting on a preliminary version of this paper and for suggesting the idea of adopting a general notion of epistemic independence where \mathcal{B}_1 and \mathcal{B}_2 are allowed to be arbitrary.

References

- I. Couso, S. Moral, and P. Walley. Examples of Independence for Imprecise Probabilities. In *ISIPTA '99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*, pages 121–130. 1999.
- F. G. Cozman. Independence for sets of full conditional probabilities, sets of lexicographic probabilities, and sets of desirable gambles. In *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theory and Applications*, pages 87–97. 2013.
- J. De Bock. *Credal networks under epistemic irrelevance: theory and algorithms*. PhD thesis, Ghent University, 2015.
- J. De Bock. *Independent Natural Extension for Infinite Spaces: Williams-Coherence to the Rescue*. ArXiv report 1701.07295, 2017.
- 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.
- G. de Cooman and E. Miranda. Irrelevant and independent natural extension for sets of desirable gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012.
- G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052, 2010.
- G. de Cooman, E. Miranda, and M. Zaffalon. Independent natural extension. *Artificial Intelligence*, 175(12):1911–1950, 2011.
- E. Miranda and M. Zaffalon. Independent products in infinite spaces. *Journal of Mathematical Analysis and Applications*, 425(1):460 – 488, 2015.
- O. A. Nielsen. *An introduction to integration and measure theory*. Wiley, 1997.
- R. Pelessoni and P. Vicig. Williams coherence and beyond. *International Journal of Approximate Reasoning*, 50(4):612–626, 2009.
- E. Quaeghebeur. Desirability. In T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 1–27. John Wiley & Sons, 2014.
- P. Vicig. Epistemic independence for imprecise probabilities. *International Journal of Approximate Reasoning*, 24(2-3):235–250, 2000.
- P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.
- P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148, 2000.
- P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, 1975.
- P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007.

Computable Randomness is Inherently Imprecise

Gert de Cooman

Jasper De Bock

Ghent University, IDLab

Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde (Belgium)

GERT.DECOOMAN@UGENT.BE

JASPER.DEBOCK@UGENT.BE

Abstract

We use the martingale-theoretic approach of game-theoretic probability to incorporate imprecision into the study of randomness. In particular, we define a notion of computable randomness associated with interval, rather than precise, forecasting systems, and study its properties. The richer mathematical structure that thus arises lets us better understand and place existing results for the precise limit. When we focus on constant interval forecasts, we find that every infinite sequence of zeroes and ones has an associated filter of intervals with respect to which it is computably random. It may happen that none of these intervals is precise, which justifies the title of this paper. We illustrate this by showing that computable randomness associated with non-stationary precise forecasting systems can be captured by a stationary interval forecast, which must then be less precise: a gain in model simplicity is thus paid for by a loss in precision.

Keywords: computable randomness; imprecise probabilities; game-theoretic probability; interval forecast; supermartingale; computability.

1. Introduction

This paper documents the first steps in our attempt to incorporate indecision and imprecision into the study of randomness. Consider a infinite sequence $\omega = (z_1, \dots, z_n, \dots)$ of zeroes and ones; when do we call it *random*? There are many notions of randomness, and many of them have a number of equivalent definitions (Ambos-Spies and Kucera, 2000; Bienvenu et al., 2009). We focus here on *computable randomness*, mainly because its focus on computability—rather than, say, the weaker lower semicomputability—has allowed us in this first attempt to keep the mathematical nitpicking at arm’s length. Randomness of a sequence ω is typically associated with a probability measure on the sample space of all infinite sequences, or—what is equivalent—with a *forecasting system* γ that associates with each finite sequence of outcomes (x_1, \dots, x_n) the (conditional) expectation $\gamma(x_1, \dots, x_n)$ for the next (as yet unknown) outcome X_{n+1} . The sequence ω is then called *computably* random when it passes a (countable) number of *computable* tests of randomness, where the collection of randomness tests depends of the forecasting system γ . An alternative but equivalent definition, going back to Ville (1939), sees each forecast $\gamma(x_1, \dots, x_n)$ as a fair price for—and therefore a commitment to bet on—the as yet unknown next outcome X_{n+1} . The sequence ω is then computably random when there is no computable strategy for getting infinitely rich by exploiting the bets made available by the forecasting system γ along the sequence, without borrowing. Technically speaking, all computable non-negative supermartingales should remain bounded on ω , and the forecasting system γ determines what a supermartingale is.

It is this last, martingale-theoretic approach which seems to lend itself most easily to allowing for imprecision in the forecasts, and therefore in the definition of randomness. As we explain in Sections 2 and 3, an ‘imprecise’ forecasting system γ associates with each finite sequence of outcomes

(x_1, \dots, x_n) a (conditional) expectation interval $\gamma(x_1, \dots, x_n)$ for the next (as yet unknown) outcome X_{n+1} , whose lower bound represents a supremum acceptable buying price, and whose upper bound a infimum acceptable selling price for X_{n+1} . This idea rests firmly on the common ground between Walley's (1991) theory of coherent lower previsions and Shafer and Vovk's (2001) game-theoretic approach to probability that we have established in recent years, through our research on imprecise stochastic processes (De Cooman and Hermans, 2008; De Cooman et al., 2016). This allows us to associate supermartingales with an imprecise forecasting system, and therefore in Section 5 to extend the existing notion of computable randomness to allow for interval, rather than precise, forecasts—we discuss computability in Section 4. We show in Section 6 that our approach allows us to extend some of Dawid's (1982) well-known work on calibration, as well as an interesting ‘limiting frequencies’ or computable stochasticity result.

We believe the discussion becomes really interesting in Section 7, where we look at stationary interval forecasts to extend the classical account of randomness. That classical account typically considers a forecasting system with stationary expectation forecast $1/2$ —corresponding to flipping a fair coin. As we have by now come to expect from our experience with imprecise probability models, a much more interesting mathematical picture appears when allowing for interval forecasts than the rather simple case of precise forecasts would lead us to suspect. In the precise case, a given sequence may not be (computably) random for any stationary forecast, but in the imprecise case there is always a set filter of intervals that a given sequence is computably random for. Furthermore, as we show in Section 8, this filter may not have a smallest element, and even when it does, this smallest element may be a non-vanishing interval: randomness may be inherently imprecise.

In order to comply with the page limit, proofs are omitted; we refer the reader to the appendix of (De Cooman and De Bock, 2017), an extended version of this paper that is available on arXiv.

2. A single interval forecast

The dynamics of making a single forecast can be made very clear by considering a simple game, with three players, namely Forecaster, Sceptic and Reality.

Game: single forecast of an outcome X

In a first step, Forecaster specifies an interval bound $I = [\underline{p}, \bar{p}]$ for the expectation of an as yet unknown outcome X in $\{0, 1\}$ —or equivalently, for the probability that $X = 1$. We interpret this *interval forecast* I as a commitment, on the part of Forecaster, to adopt \underline{p} as a *supremum buying price* and \bar{p} as a *infimum selling price* for the gamble (with reward function) X . This is taken to mean that the second player, *Sceptic*, can now in a second step take Forecaster up on any (combination) of the following commitments:

- (i) for any $p \in [0, 1]$ such that $p \leq \underline{p}$, and any $\alpha \geq 0$ Forecaster must accept the gamble $\alpha[X - p]$, leading to an uncertain reward $-\alpha[X - p]$ for Sceptic;¹
- (ii) for any $q \in [0, 1]$ such that $q \geq \bar{p}$, and any $\beta \geq 0$ Forecaster accepts the gamble $\beta[q - X]$, leading to an uncertain reward $-\beta[q - X]$ for Sceptic.

Finally, in a third step, the third player, *Reality*, determines the value x of X in $\{0, 1\}$. \square

Elements x of $\{0, 1\}$ are called *outcomes*, and elements p of the real unit interval $[0, 1]$ are called (precise) *forecasts*. We denote by \mathcal{C} the set of non-empty closed subintervals of the real unit

1. Because we allow $p \leq \underline{p}$ rather than $p < \underline{p}$, we actually see \underline{p} as a *maximum* buying price, rather than a supremum one. We do this because it does not affect the conclusions, but simplifies the mathematics. Similarly for $q \geq \bar{p}$.

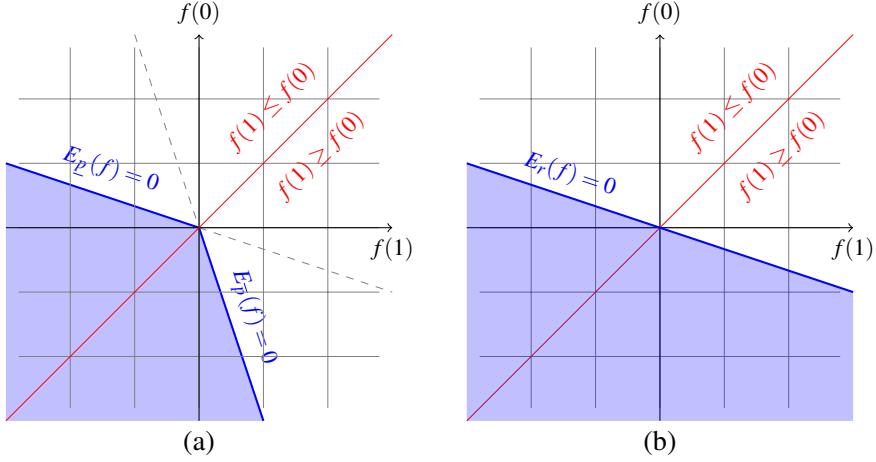


Figure 1: Gambles f available to Sceptic when (a) Forecaster announces $I \in \mathcal{C}$ with $\underline{p} < \bar{p}$; and when (b) Forecaster announces $I \in \mathcal{C}$ with $\underline{p} = \bar{p} =: r$.

interval $[0, 1]$. Any element I of \mathcal{C} is called an *interval forecast*. It has a smallest element $\min I$ and a greatest element $\max I$, so $I = [\min I, \max I]$. We will use the generic notation I for such an interval, and $\underline{p} := \min I$ and $\bar{p} := \max I$ for its lower and upper bounds, respectively.

After Forecaster announces a forecast interval I , what Sceptic can do is essentially to try and increase his capital by taking a gamble on the outcome X . Any such gamble can be considered as a map $f: \{0, 1\} \rightarrow \mathbb{R}$, and can therefore be represented as a vector $(f(1), f(0))$ in the two-dimensional vector space \mathbb{R}^2 ; see also Figure 1. $f(X)$ is then the increase in Sceptic's capital after the game has been played, as a function of the outcome variable X . Of course, not every gamble $f(X)$ on the outcome X will be available to Sceptic: which gambles he can take is determined by Forecaster's interval forecast I . In their most general form, they are given by $f(X) = -\alpha[X - \underline{p}] - \beta[q - X]$, where α and β are non-negative real numbers, $\underline{p} \leq p$ and $q \geq \bar{p}$. If we consider the so-called *lower expectation* (functional) \underline{E}_I associated with an interval forecast I , defined by

$$\underline{E}_I(f) = \min_{p \in I} E_p(f) = \min_{p \in I} [pf(1) + (1-p)f(0)] = \begin{cases} E_{\underline{p}}(f) & \text{if } f(1) \geq f(0) \\ E_{\bar{p}}(f) & \text{if } f(1) \leq f(0) \end{cases} \quad (1)$$

for any gamble $f: \{0, 1\} \rightarrow \mathbb{R}$, and similarly, the *upper expectation* (functional) \bar{E}_I , defined by

$$\bar{E}_I(f) = \max_{p \in I} E_p(f) = \begin{cases} E_{\bar{p}}(f) & \text{if } f(1) \geq f(0) \\ E_{\underline{p}}(f) & \text{if } f(1) \leq f(0) \end{cases} = -\underline{E}_I(-f), \quad (2)$$

then it is not difficult to see that *the cone of gambles $f(X)$ that are available to Sceptic after Forecaster announces an interval forecast I is completely determined by the condition $\bar{E}_I(f) \leq 0$* , as depicted by the blue regions in Figure 1. The functionals \underline{E}_I and \bar{E}_I are easily shown to have the following properties, typical for the more general lower and upper expectation operators defined on more general gamble spaces (Walley, 1991; Troffaes and De Cooman, 2014):

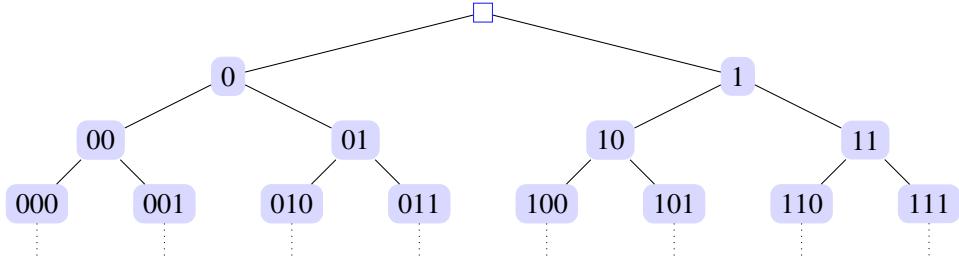
Proposition 1 Consider any forecast interval $I \in \mathcal{C}$. Then for all gambles f, g on $\{0, 1\}$, $\mu \in \mathbb{R}$ and non-negative $\lambda \in \mathbb{R}$:

- C1. $\min f \leq \underline{E}_I(f) \leq \bar{E}_I(f) \leq \max f$; [bounds]
- C2. $\underline{E}_I(\lambda f) = \lambda \underline{E}_I(f)$ and $\bar{E}_I(\lambda f) = \lambda \bar{E}_I(f)$; [non-negative homogeneity]
- C3. $\underline{E}_I(f+g) \geq \underline{E}_I(f) + \underline{E}_I(g)$ and $\bar{E}_I(f+g) \leq \bar{E}_I(f) + \bar{E}_I(g)$; [super/subadditivity]
- C4. $\underline{E}_I(f+\mu) = \underline{E}_I(f) + \mu$ and $\bar{E}_I(f+\mu) = \bar{E}_I(f) + \mu$. [constant additivity]

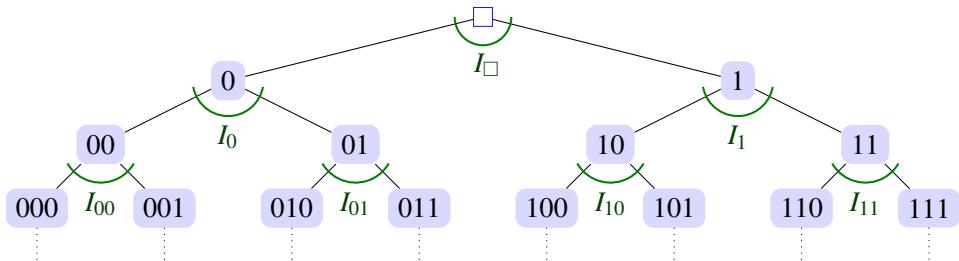
3. Interval forecasting systems and imprecise probability trees

We now consider a sequence of repeated versions of the forecast game in the previous section, where at each stage $k \in \mathbb{N}$, Forecaster presents an interval forecast $I_k = [\underline{p}_k, \bar{p}_k]$ for the unknown outcome variable X_k . This effectively allows Sceptic to choose any gamble $f_k(X_k)$ such that $\bar{E}_{I_k}(f_k) \leq 0$. Reality then chooses a value x_k for X_k , resulting in a gain, or increase in capital, $f_k(x_k)$ for Sceptic.

We call $(x_1, x_2, \dots, x_n, \dots)$ an *outcome sequence*, and collect all possible outcome sequences in the set $\Omega := \{0, 1\}^{\mathbb{N}}$. We collect the finite outcome sequences (x_1, \dots, x_n) in the set $\Omega^\diamond := \{0, 1\}^* = \bigcup_{n \in \mathbb{N}_0} \{0, 1\}^n$. Finite sequences s in Ω^\diamond and infinite sequences ω in Ω are the nodes—called *situations*—and paths in an event tree with unbounded horizon, part of which is depicted below.



In this repeated game, Forecaster will only provide interval forecasts I_k after observing the actual sequence (x_1, \dots, x_{k-1}) that Reality has chosen. This is the essence of so-called *prequential forecasting* (Dawid, 1982, 1984; Dawid and Vovk, 1999). But for technical reasons, it will be useful to consider the more involved setting where a forecast I_s is specified in each of the possible situations $s \in \Omega^\diamond$; see the figure below.



Indeed, we can use this idea to generalise the notion of a forecasting system (Vovk and Shen, 2010).

Definition 2 (Forecasting system) A forecasting system is a map $\gamma: \Omega^\diamond \rightarrow \mathcal{C}$, that associates with any situation s in the event tree a forecast $\gamma(s) \in \mathcal{C}$. With any forecasting system γ we can associate two real-valued maps $\underline{\gamma}$ and $\bar{\gamma}$ on Ω^\diamond , defined by $\underline{\gamma}(s) := \min \gamma(s)$ and $\bar{\gamma}(s) := \max \gamma(s)$ for all $s \in \Omega^\diamond$. A forecasting system γ is called precise if $\underline{\gamma} = \bar{\gamma}$. Γ denotes the set $\mathcal{C}^{\Omega^\diamond}$ of all forecasting systems.

Specifying such a forecasting system requires imagining in advance all moves that Reality could make, and devising in advance what forecasts to give in each imaginable situation s . In the precise case, that is typically what one does when specifying a probability measure on the so-called *sample space* Ω —the set Ω of all paths.

Since in each situation s the interval forecast $I_s = \gamma(s)$ corresponds to a local lower expectation \underline{E}_{I_s} , we can use the argumentation in our earlier papers (De Cooman and Hermans, 2008; De Cooman et al., 2016) on stochastic processes to let the forecasting system γ turn the event tree into a so-called *imprecise probability tree*, with an associated global lower expectation, and a corresponding notion of ‘(strictly) almost surely’. In what follows, we briefly recall how to do this; for more context, we also refer to the seminal work by Shafer and Vovk (2001).

For any path $\omega \in \Omega$, the initial sequence that consists of its first n elements is a situation in $\{0, 1\}^n$ that is denoted by ω^n . Its n -th element belongs to $\{0, 1\}$ and is denoted by ω_n . As a convention, we let its 0-th element be the *initial* situation $\omega^0 = \omega_0 = \square$. We write that $s \sqsubseteq t$, and say that the situation s *precedes* the situation t , when every path that goes through t also goes through s —so s is a precursor of t .

A *process* F is a map defined on Ω^\diamond . A *real process* is a real-valued process: it associates a real number $F(s) \in \mathbb{R}$ with every situation $s \in \Omega^\diamond$. With any real process F , we can always associate a process ΔF , called the *process difference*. For every situation (x_1, \dots, x_n) with $n \in \mathbb{N}_0$, $\Delta F(x_1, \dots, x_n)$ is a gamble on $\{0, 1\}$ defined by $\Delta F(x_1, \dots, x_n)(x_{n+1}) := F(x_1, \dots, x_{n+1}) - F(x_1, \dots, x_n)$ for all $x_{n+1} \in \{0, 1\}$. In the imprecise probability tree associated with a *given* forecasting system γ , a *submartingale* M for γ is a real process such that $\underline{E}_{\gamma(x_1, \dots, x_n)}(\Delta M(x_1, \dots, x_n)) \geq 0$ for all $n \in \mathbb{N}_0$ and $(x_1, \dots, x_n) \in \{0, 1\}^n$. A real process M is a *supermartingale* for γ if $-M$ is a submartingale, meaning that $\overline{E}_{\gamma(x_1, \dots, x_n)}(\Delta M(x_1, \dots, x_n)) \leq 0$ for all $n \in \mathbb{N}_0$ and $(x_1, \dots, x_n) \in \{0, 1\}^n$: all supermartingale differences have non-positive upper expectation, so supermartingales are real processes that Forecaster expects to decrease. We denote the set of all submartingales for a given forecasting system γ by $\underline{\mathbb{M}}^\gamma$ —whether a real process is a submartingale depends of course on the forecasts in the situations. Similarly, the set $\overline{\mathbb{M}}^\gamma := -\underline{\mathbb{M}}^\gamma$ is the set of all supermartingales for γ .

It is clear from the discussion in Section 2 that the supermartingales are effectively all the possible capital processes \mathcal{K} for a Sceptic who starts with an initial capital $\mathcal{K}(\square)$, and in each possible subsequent situation s selects a gamble $f_s = \Delta \mathcal{K}(s)$ that is available there because Forecaster specifies the interval forecast $I_s = \gamma(s)$ and because $\overline{E}_{I_s}(f_s) = \overline{E}_{\gamma(s)}(\Delta \mathcal{K}(s)) \leq 0$. If Reality chooses outcomes $s = (x_1, \dots, x_n)$, then Sceptic ends up with capital $\mathcal{K}(x_1, \dots, x_n) = \mathcal{K}(\square) + \sum_{k=0}^{n-1} \Delta \mathcal{K}(x_1, \dots, x_k)(x_{k+1})$. A *non-negative* supermartingale M is non-negative in all situations, which corresponds to Sceptic never borrowing any money. We call *test supermartingale* any non-negative supermartingale M that starts with unit capital $M(\square) = 1$. We collect all test supermartingales for γ in the set $\overline{\mathbb{T}}^\gamma$.

In the context of probability trees, we call *variable* any function defined on the sample space Ω . When this variable is real-valued and bounded, we call it a *gamble* on Ω . 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. The following expressions define lower and upper expectations on such gambles g on Ω :

$$\underline{E}^\gamma(g) := \sup \left\{ M(\square) : M \in \underline{\mathbb{M}}^\gamma \text{ and } \limsup_{n \rightarrow +\infty} M(\omega^n) \leq g(\omega) \text{ for all } \omega \in \Omega \right\} \quad (3)$$

$$\overline{E}^\gamma(g) := \inf \left\{ M(\square) : M \in \overline{\mathbb{M}}^\gamma \text{ and } \liminf_{n \rightarrow +\infty} M(\omega^n) \geq g(\omega) \text{ for all } \omega \in \Omega \right\} = -\underline{E}^\gamma(g). \quad (4)$$

They satisfy coherence properties similar to those in Proposition 1. We refer to extensive discussions elsewhere (De Cooman et al., 2016; Shafer and Vovk, 2001) about why these expressions are interesting and useful. For our present purposes, it may suffice to mention that for precise forecasts, they lead to models that coincide with the ones found in measure-theoretic probability theory (Shafer and Vovk, 2001, Chapter 8). In particular, when all $I_s = \{1/2\}$, they coincide with the usual uniform (Lebesgue) expectations on measurable gambles.

We call an event $A \subseteq \Omega$ *null* if $\bar{P}^\gamma(A) := \bar{E}^\gamma(\mathbb{I}_A) = 0$, or equivalently $\underline{P}^\gamma(A^c) := \underline{E}^\gamma(\mathbb{I}_{A^c}) = 1$, and *strictly null* if there is some test supermartingale $T \in \bar{\mathbb{T}}^\gamma$ that converges to $+\infty$ on A , meaning that $\lim_{n \rightarrow +\infty} T(\omega^n) = +\infty$ for all $\omega \in A$. Any strictly null event is null, but null events need not be strictly null (Vovk and Shafer, 2014; De Cooman et al., 2016). Because it is easily checked that $\bar{P}^\gamma(\emptyset) = \underline{P}^\gamma(\emptyset) = 0$, the complement A^c of a (strictly) null event A is never empty. As usual, any property that holds, except perhaps on a (strictly) null event, is said to hold (strictly) *almost surely*.

4. Basic computability notions

We recall a few notions and results from computability theory that are relevant to the discussion. For a much more extensive treatment, we refer for instance to the books by Pour-El and Richards (1989) and Li and Vitányi (1993).

A *computable* function $\phi: \mathbb{N}_0 \rightarrow \mathbb{N}_0$ is a function that can be computed by a Turing machine. All further notions of computability that we will need, build on this basic notion. It is clear that it in this definition, we can replace any of the \mathbb{N}_0 with any other countable set.

We start with the definition of a computable real number. We call a sequence of rational numbers r_n *computable* if there are three computable functions a, b, σ from \mathbb{N}_0 to \mathbb{N}_0 such that $b(n) > 0$ and $r_n = (-1)^{\sigma(n)} \frac{a(n)}{b(n)}$ for all $n \in \mathbb{N}_0$, and we say that it *converges effectively* to a real number x if there is some computable function $e: \mathbb{N}_0 \rightarrow \mathbb{N}_0$ such that $n \geq e(N) \Rightarrow |r_n - x| \leq 2^{-N}$ for all $n, N \in \mathbb{N}_0$. A real number is then called *computable* if there is a computable sequence of rational numbers that converges effectively to it. Of course, every rational number is a computable real.

We also need a notion of computable real processes, or in other words, computable real-valued maps $F: \Omega^\diamond \rightarrow \mathbb{R}$ defined on the set Ω^\diamond of all situations. Because there is an obvious computable bijection between \mathbb{N}_0 and Ω^\diamond , whose inverse is also computable, we can in fact identify real processes and real sequences, and simply import, *mutatis mutandis*, the definitions for computable real sequences common in the literature (Li and Vitányi, 1993, Chapter 0). Indeed, we call a net of rational numbers $r_{s,n}$ *computable* if there are three computable functions a, b, s from $\Omega^\diamond \times \mathbb{N}_0$ to \mathbb{N}_0 such that $b(s, n) > 0$ and $r_{s,n} = (-1)^{\sigma(s,n)} \frac{a(s,n)}{b(s,n)}$ for all $s \in \Omega^\diamond$ and $n \in \mathbb{N}_0$. We call a real process $F: \Omega^\diamond \rightarrow \mathbb{R}$ *computable* if there is a computable net of rational numbers $r_{s,n}$ and a computable function $e: \Omega^\diamond \times \mathbb{N}_0 \rightarrow \mathbb{N}_0$ such that $n \geq e(s, N) \Rightarrow |r_{s,n} - F(s)| \leq 2^{-N}$ for all $s \in \Omega^\diamond$ and $n, N \in \mathbb{N}_0$. Obviously, it follows from this definition that in particular $F(t)$ is a computable real number for any $t \in \Omega^\diamond$: fix $s = t$ and consider the sequence $r_{t,n}$ that converges to $F(s)$ as $n \rightarrow +\infty$. Also, a constant real process is computable if and only if its constant value is.

The following definitions are now obvious. A gamble f on $\{0, 1\}$ is called *computable* if both its values $f(0)$ and $f(1)$ are computable real numbers. An interval forecast $I = [\underline{p}, \bar{p}] \in \mathcal{C}$ is called *computable* if both its lower bound \underline{p} and upper bound \bar{p} are computable real numbers. A forecasting system γ is called *computable* if the associated real processes $\underline{\gamma}$ and $\bar{\gamma}$ are.

5. Random sequences in an imprecise probability tree

We will now associate a notion of randomness with a forecasting system γ —or in other words, with an imprecise probability tree. In what follows, we will often consider computable test supermartingales. These computable test supermartingales for a forecasting system are countable in number, because the computable processes are (Li and Vitányi, 1993; Vovk and Shen, 2010).

Definition 3 (Computable randomness) Consider any forecasting system $\gamma: \Omega^\diamond \rightarrow \mathcal{C}$. We call an outcome sequence ω computably random for γ if all computable test supermartingales T remain bounded above on ω , meaning that there is some $B \in \mathbb{R}$ such that $T(\omega^n) \leq B$ for all $n \in \mathbb{N}$, or equivalently, that $\sup_{n \in \mathbb{N}} T(\omega^n) < +\infty$. We then also say that the forecasting system γ makes ω computably random. We denote by $\Gamma_C(\omega) := \{\gamma \in \Gamma: \omega \text{ is computably random for } \gamma\}$ the set of all forecasting systems for which the outcome sequence ω is computably random.

Computable randomness of an outcome sequence means that there is no computable strategy that starts with capital 1 and avoids borrowing, and allows Sceptic to increase his capital without bounds by exploiting the bets on these outcomes that are made available to him by Forecaster’s specification of the forecasting system γ . When the forecasting system γ is precise and computable, our notion of computable randomness reduces to the classical notion of computable randomness (Ambos-Spies and Kucera, 2000; Bienvenu et al., 2009).

The (computable) vacuous forecasting system γ_v assigns the vacuous forecast $\gamma_v(s) := [0, 1]$ to all situations $s \in \Omega^\diamond$. The following proposition implies that no $\Gamma_C(\omega)$ is empty.

Proposition 4 All paths are computably random for the vacuous forecasting system: $\gamma_v \in \Gamma_C(\omega)$ for all $\omega \in \Omega$.

More conservative (or imprecise) forecasting systems have more computably random sequences.

Proposition 5 Let ω be computably random for a forecasting system γ . Then ω is also computably random for any forecasting system γ^* such that $\gamma \subseteq \gamma^*$, meaning that $\gamma(s) \subseteq \gamma^*(s)$ for all $s \in \Omega^\diamond$.

6. Consistency results

We first show that any Forecaster who specifies a forecasting system is consistent in the sense that he believes himself to be *well calibrated*: in the imprecise probability tree generated by his own forecasts, (strictly) almost all paths will be computably random, so he is sure that Sceptic will not be able to become infinitely rich at his expense, by exploiting his—Forecaster’s—forecasts. This also generalises the arguments and conclusions in a paper by Dawid (1982).

Theorem 6 Consider any forecasting system $\gamma: \Omega^\diamond \rightarrow \mathcal{C}$. Then (strictly) almost all outcome sequences are computably random for γ in the imprecise probability tree that corresponds to γ .

This result is quite powerful, and it guarantees in particular that:

Corollary 7 For any sequence of interval forecasts (I_1, \dots, I_n, \dots) there is a forecasting system given by $\gamma(x_1, \dots, x_n) := I_{n+1}$ for all $(x_1, \dots, x_n) \in \{0, 1\}^n$ and all $n \in \mathbb{N}_0$, and associated imprecise probability tree such that (strictly) almost all—and therefore definitely at least one—outcome sequences are computably random for γ in the associated imprecise probability tree.

The following weaker consistency result deals with limits (inferior and superior) of relative frequencies, taken with respect to a so-called *selection process* $S: \Omega^\Diamond \rightarrow \{0, 1\}$. It is a counterpart in our more general context of the notions of *computable stochasticity* or *Church randomness* in the precise case with $I = \{1/2\}$ (Ambos-Spies and Kucera, 2000).

Theorem 8 (Church randomness) *Let $\gamma: \Omega^\Diamond \rightarrow \mathcal{C}$ be any computable forecasting system, let $\omega = (x_1, \dots, x_n, \dots) \in \Omega$ be any outcome sequence that is computably random for γ , and let f be any computable gamble on $\{0, 1\}$. If $S: \Omega^\Diamond \rightarrow \{0, 1\}$ is any computable selection process such that $\sum_{k=0}^n S(x_1, \dots, x_k) \rightarrow +\infty$, then also*

$$\liminf_{n \rightarrow +\infty} \frac{\sum_{k=0}^{n-1} S(x_1, \dots, x_k) [f(x_{k+1}) - \underline{E}_{\gamma(x_1, \dots, x_k)}(f)]}{\sum_{k=0}^{n-1} S(x_1, \dots, x_k)} \geq 0.$$

7. Constant interval forecasts

We now introduce a significant simplification. For any interval $I \in \mathcal{C}$, we let γ_I be the corresponding *stationary forecasting system* that assigns the same interval forecast I to all nodes: $\gamma_I(s) := I$ for all $s \in \Omega^\Diamond$. In this way, with any outcome sequence ω , we can associate the collection of all interval forecasts for which the corresponding stationary forecasting system makes ω computably random:

$$\mathcal{C}_C(\omega) := \{I \in \mathcal{C} : \gamma_I \in \Gamma_C(\omega)\} = \{I \in \mathcal{C} : \gamma_I \text{ makes } \omega \text{ computably random}\}.$$

As an immediate consequence of Propositions 4 and 5, we find that this set of intervals is non-empty and increasing.

Proposition 9 (Non-emptiness) *For all $\omega \in \Omega$, $[0, 1] \in \mathcal{C}_C(\omega)$, so any sequence of outcomes ω has at least one stationary forecast that makes it computably random: $\mathcal{C}_C(\omega) \neq \emptyset$.*

Proposition 10 (Increasingness) *Consider any $\omega \in \Omega$ and any $I, J \in \mathcal{C}$. If $I \in \mathcal{C}_C(\omega)$ and $I \subseteq J$, then also $J \in \mathcal{C}_C(\omega)$.*

Theorem 8 implies the following property. However, quite remarkably, and seemingly in contrast with Theorem 8, it does not require any computability assumptions on the (stationary) forecasts.

Corollary 11 (Church randomness) *Consider any outcome sequence $\omega = (x_1, \dots, x_n, \dots)$ in Ω and any stationary interval forecast $I = [\underline{p}, \bar{p}] \in \mathcal{C}_C(\omega)$ that makes ω computably random. Then for any computable selection process $S: \Omega^\Diamond \rightarrow \{0, 1\}$ such that $\sum_{k=0}^n S(x_1, \dots, x_k) \rightarrow +\infty$:*

$$\underline{p} \leq \liminf_{n \rightarrow +\infty} \frac{\sum_{k=0}^{n-1} S(x_1, \dots, x_k) x_{k+1}}{\sum_{k=0}^{n-1} S(x_1, \dots, x_k)} \leq \limsup_{n \rightarrow +\infty} \frac{\sum_{k=0}^{n-1} S(x_1, \dots, x_k) x_{k+1}}{\sum_{k=0}^{n-1} S(x_1, \dots, x_k)} \leq \bar{p}.$$

The following proposition can of course be straightforwardly extended to any finite number of interval forecasts, and guarantees, together with Proposition 10, that $\mathcal{C}_C(\omega)$ is a *set filter*.

Proposition 12 *For any $\omega \in \Omega$ and any two interval forecasts I and J : if $I \in \mathcal{C}_C(\omega)$ and $J \in \mathcal{C}_C(\omega)$ then $I \cap J \neq \emptyset$, and $I \cap J \in \mathcal{C}_C(\omega)$.*

This result also tells us that the collection $\mathcal{C}_C(\omega)$ of closed subsets of the compact set $[0, 1]$ has the finite intersection property, and its intersection is therefore a non-empty closed interval: $\bigcap \mathcal{C}_C(\omega) = [\underline{p}_C(\omega), \bar{p}_C(\omega)]$. Propositions 10 and 12 guarantee that all intervals $[\underline{p}_C(\omega) - \varepsilon_1, \bar{p}_C(\omega) + \varepsilon_2]$ in \mathcal{C} with $\varepsilon_1, \varepsilon_2 > 0$ belong to $\mathcal{C}_C(\omega)$. But we will see in the next section that this does not generally hold for $\varepsilon_1 = 0$ and/or $\varepsilon_2 = 0$. For this reason, we now define the following two subsets of $[0, 1]$:

$$L_C(\omega) := \{\min I : I \in \mathcal{C}_C(\omega)\} \text{ and } U_C(\omega) := \{\max I : I \in \mathcal{C}_C(\omega)\}.$$

Then Proposition 10 guarantees that $L_C(\omega)$ is a decreasing set, and that $U_C(\omega)$ is increasing. They are therefore both subintervals of $[0, 1]$. Obviously, $\underline{p}_C(\omega) = \sup L_C(\omega)$ and $\bar{p}_C(\omega) = \inf U_C(\omega)$. On the one hand clearly $L_C(\omega) = [0, \underline{p}_C(\omega)]$ or $L_C(\omega) = [0, \underline{p}_C(\omega)]$, and on the other hand $U_C(\omega) = (\bar{p}_C(\omega), 1]$ or $U_C(\omega) = [\bar{p}_C(\omega), 1]$. Proposition 12 easily allows us to give the following simple description of the set $\mathcal{C}_C(\omega)$ in terms of $L_C(\omega)$ and $U_C(\omega)$:

$$I \in \mathcal{C}_C(\omega) \Leftrightarrow \left(\min I \in L_C(\omega) \text{ and } \max I \in U_C(\omega) \right).$$

A trivial example is given by:

Proposition 13 *If the sequence ω is computable with infinitely many zeroes and ones, then $\mathcal{C}_C(\omega) = \{[0, 1]\}$, and therefore $L_C(\omega) = \{0\}$, $U_C(\omega) = \{1\}$, $\underline{p}_C(\omega) = 0$ and $\bar{p}_C(\omega) = 1$.*

At the other extreme, there are the sequences ω that are computably random for some *precise* stationary forecasting system $\gamma_{\{p\}}$, with $p \in [0, 1]$. They are amongst the random sequences that have received most attention in the literature, thus far. For any such sequence, $\mathcal{C}_C(\omega) = \{I \in \mathcal{C} : p \in I\}$, $L_C(\omega) = [0, p]$ and $U_C(\omega) = [p, 0]$, and therefore also $\underline{p}_C(\omega) = \bar{p}_C(\omega) = p$.

We show in the next section that, in between these extremes of total imprecision and maximal precision, there lies a—to the best of our knowledge—previously uncharted realm of sequences, with similar (and even in some sense ‘larger’) unpredictability than the ones traditionally called ‘computably random’, for which $L_C(\omega)$ and $U_C(\omega)$ need not always be closed, and more importantly, for which $0 < \underline{p}_C(\omega) < \bar{p}_C(\omega) < 1$. This is what we mean when we claim that ‘computable randomness is inherently imprecise’.

8. Randomness is inherently imprecise

Our work on imprecise Markov chains (De Cooman et al., 2016) has taught us that in some cases, we can very efficiently compute tight bounds on expectations in non-stationary precise Markov chains, by replacing them with their stationary imprecise versions. Similarly, in statistical modelling, when learning from data sampled from a distribution with a varying (non-stationary) parameter, it seems hard to estimate the exact time sequence of its values. But we may be more successful in learning about its (stationary) interval *range*. This idea was also considered earlier by Fierens et al. (2009), when they argued for a frequentist interpretation of imprecise probability models based on non-stationarity.

In this section, we exploit this idea, by showing that randomness associated with non-stationary precise forecasting systems can be captured by a stationary forecasting system, which must then be less precise: we gain simplicity of representation, but pay for it by losing precision.

We begin with a simple example. Consider any p and q in $[0, 1]$ with $p \leq q$, and any outcome sequence $\omega = (x_1, \dots, x_n, \dots)$ that is computably random for the forecasting system $\gamma_{p,q}$ that is

defined by

$$\gamma_{p,q}(z_1, \dots, z_n) := \begin{cases} p & \text{if } n \text{ is odd} \\ q & \text{if } n \text{ is even} \end{cases} \quad \text{for all } (z_1, \dots, z_n) \in \Omega^\diamond.$$

We know from Corollary 7 that there is at least one such outcome sequence. It turns out that the stationary forecasting systems that make such ω computably random have a simple characterisation:

Proposition 14 *Consider any ω that is computably random for the forecasting system $\gamma_{p,q}$. Then for all $I \in \mathcal{C}$, $I \in \mathcal{C}_C(\omega) \Leftrightarrow [p, q] \subseteq I$.*

Its proof relies on a very simple argument involving Corollary 11. This result implies in particular also that $L_C(\omega) = [0, p]$, $U_C(\omega) = [q, 1]$, $\underline{p}_C(\omega) = p$ and $\bar{p}_C(\omega) = q$.

Next, we turn to a more complicated example, where we look at sequences that are ‘nearly’ computably random for the stationary precise forecast $1/2$, but not quite. This example was inspired by the ideas involving Hellinger-like divergences in a beautiful paper by Vovk (2009).

Consider the following sequence $\{p_n\}_{n \in \mathbb{N}}$ of precise forecasts:

$$p_n := \frac{1}{2} + (-1)^n \delta_n, \text{ with } \delta_n := e^{-\frac{1}{n+1}} \sqrt{e^{\frac{1}{n+1}} - 1} \text{ for all } n \in \mathbb{N},$$

converging to $1/2$. Observe that the sequence δ_n is decreasing towards its limit 0 and that $\delta_n \in (0, 1/2)$ and $p_n \in (0, 1)$, for all $n \in \mathbb{N}$. Now consider any outcome sequence $\omega = (x_1, \dots, x_n, \dots)$ that is computably random for the precise forecasting system $\gamma_{\sim 1/2}$ that is defined by

$$\gamma_{\sim 1/2}(z_1, \dots, z_{n-1}) := p_n \text{ for all } n \in \mathbb{N} \text{ and } (z_1, \dots, z_{n-1}) \in \Omega^\diamond.$$

We know from Corollary 7 that there is at least one such outcome sequence. It turns out that the stationary forecasting systems that make such ω computably random have a simple characterisation:

Proposition 15 *Consider any ω that is computably random for the forecasting system $\gamma_{\sim 1/2}$. Then for all $I \in \mathcal{C}$, $I \in \mathcal{C}_C(\omega)$ if and only if $\min I < 1/2$ and $\max I > 1/2$.*

This result implies in particular that $L_C(\omega) = [0, 1/2]$, $U_C(\omega) = (1/2, 1]$ and $\underline{p}_C(\omega) = \bar{p}_C(\omega) = 1/2$.

9. Conclusion

Even with the limited number of examples we have been able to examine in this paper, it becomes apparent that incorporating imprecision in the study of randomness allows for much more mathematical structure to arise, which we would argue lets us better understand and place the existing results in the precise limit.

In our argumentation that ‘randomness is inherently imprecise’, we are well aware that we are restricting ourselves to stationary forecasts. Our examples in Section 8 all involve sequences that are computably random for a precise non-stationary forecasting system, but no longer computably random for any stationary precise variant. To make our claim irrefutable, we would have to show that there are sequences that are computably random for forecasting systems more precise than the vacuous one, but not for any (computable) precise forecasting system. Or in other words, that there is ‘randomness’ or ‘unpredictability’ that cannot be ‘explained’ by any non-stationary (computable) precise forecasting system. We will of course keep this challenge foremost in our minds.

Nevertheless, the examples in Section 8 do indicate that it is in some ways possible to replace an ‘explanation’ by a complex non-stationary precise forecasting model by a(n infinite filter of) more imprecise stationary one(s).

This work may seem promising, but we are well aware that it is only a humble beginning. We see many extensions in many directions. First of all, we want to find out if our approach can also be used to find interval versions of *Martin-Löf* and *Schnorr randomness* (Ambos-Spies and Kucera, 2000; Bienvenu et al., 2009) with similarly interesting properties and conclusions. Secondly, our preliminary exploration suggests that it will be possible to formulate equivalent randomness definitions in terms of *randomness tests*, rather than supermartingales, but this needs to be checked in much more detail. Thirdly, the approach we follow here is not prequential: we assume that our Forecaster specifies an entire forecasting system γ , or in other words an interval forecast in all possible situations (x_1, \dots, x_n) , rather than only interval forecasts in those situations z_1, \dots, z_n of the sequence $\omega = (z_1, \dots, z_n, \dots)$ whose potential randomness we are considering. The *prequential approach*, which we eventually will want to come to, looks at the randomness of a sequence of interval forecasts and outcomes $(I_1, z_1, I_2, z_2, \dots, I_n, z_n, \dots)$, where each I_k is an interval forecast for the as yet unknown X_k , which is afterwards revealed to be z_k , without the need of specifying forecasts in situations that are never reached; see the paper by Vovk and Shen (2010) for an account of how this works for precise forecasts. Fourthly, we need to connect our work with earlier approaches to associating imprecision with randomness (Walley and Fine, 1982; Fierens et al., 2009; Fierens, 2009; Gorban, 2016). And finally, and perhaps most importantly, we believe this research could be a very early starting point for an approach to statistics that takes imprecise or set-valued parameters more seriously, when learning from finite amounts of data.

Acknowledgments

This research started with discussions between Gert and Philip Dawid about what prequential interval forecasting would look like, during a joint stay at Durham University in late 2014. Gert, and Jasper who joined in late 2015, wrote an early prequential version of the present paper during a joint research visit to the Universities of Strathclyde and Durham in May 2016, trying to extend the results by Volodya Vovk (Vovk, 1987, 2009; Vovk and Shen, 2010) to make them allow for interval forecasts. In an email exchange, Volodya pointed out a number of difficulties with our approach, which we were able to resolve by letting go of its prequential emphasis, at least for the time being. This was done during research visits of Gert to Jasper at IDSIA in late 2016 and early 2017.

We are grateful to Philip and Volodya for their inspiring and helpful comments and guidance. Gert’s research and travel were partly funded through project number G012512N of the Research Foundation – Flanders (FWO), Jasper is a Postdoctoral Fellow of the FWO and wishes to acknowledge its financial support.

References

- K. Ambos-Spies and A. Kucera. Randomness in computability theory. *Contemporary Mathematics*, 257:1–14, 2000.
- L. Bienvenu, G. Shafer, and A. Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5, 2009.

- A. P. Dawid. The well-calibrated Bayesian. *Journal of The American Statistical Association*, 77(379):605–610, 1982.
- A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:278–292, 1984.
- A. P. Dawid and V. G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5:125–162, 1999.
- G. de Cooman and F. 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](https://doi.org/10.1016/j.artint.2008.03.001).
- G. de Cooman and J. De Bock. Computable randomness is inherently imprecise. ArXiv report 1703.00931 [math.PR], 2017. <https://arxiv.org/abs/1703.00931>.
- G. de Cooman, J. De Bock, and S. Lopatatzidis. Imprecise stochastic processes in discrete time: global models, imprecise Markov chains, and ergodic theorems. *International Journal of Approximate Reasoning*, 76:18–46, 2016.
- P. I. Fierens. An extension of chaotic probability models to real-valued variables. *International Journal of Approximate Reasoning*, 50(4):627–641, 2009.
- P. I. Fierens, L. C. Rego, and T. L. Fine. A frequentist understanding of sets of measures. *Journal of Statistical Planning and Inference*, 139(6):1879–1892, 2009.
- I. I. Gorban. *The Statistical Stability Phenomenon*. Springer, 2016.
- M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 1993.
- M. B. Pour-El and J. I. Richards. *Computability in Analysis and Physics*. Springer Verlag, 1989.
- G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.
- M. C. M. Troffaes and G. de Cooman. *Lower Previsions*. Wiley, 2014.
- J. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.
- V. Vovk and G. Shafer. Game-theoretic probability. In T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*. John Wiley & Sons, 2014.
- V. G. Vovk. On a criterion of randomness. *Doklady Akademii Nauk SSSR*, 294(6):1298–1302, 1987.
- V. G. Vovk. Merging of opinions in game-theoretic probability. *Annals of the Institute of Statistical Mathematics*, 61(4):969–993, 2009.
- V. G. Vovk and A. Shen. Prequential randomness and probability. *Theoretical Computer Science*, 411(29-30):2632–2646, 2010.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- P. Walley and T. L. Fine. Towards a frequentist theory of upper and lower probability. *Annals of Statistics*, 10:741–761, 1982.

Imprecise Continuous-Time Markov Chains: Efficient Computational Methods with Guaranteed Error Bounds

Alexander Erreygers

Ghent University, SMACS Research Group (Belgium)

ALEXANDER.ERREYGERS@UGENT.BE

Jasper De Bock

Ghent University - imec, IDLab, ELIS (Belgium)

JASPER.DEBOCK@UGENT.BE

Abstract

Imprecise continuous-time Markov chains are a robust type of continuous-time Markov chains that allow for partially specified time-dependent parameters. Computing inferences for them requires the solution of a non-linear differential equation. As there is no general analytical expression for this solution, efficient numerical approximation methods are essential to the applicability of this model. We here improve the uniform approximation method of [Krak et al. \(2016\)](#) in two ways and propose a novel and more efficient adaptive approximation method. For ergodic chains, we also provide a method that allows us to approximate stationary distributions up to any desired maximal error.

Keywords: imprecise continuous-time Markov chain; lower transition operator; lower transition rate operator; approximation method; ergodicity; coefficient of ergodicity.

1. Introduction

Markov chains are a popular type of stochastic processes that can be used to model a variety of systems with uncertain dynamics, both in discrete and continuous time. In many applications, however, the core assumption of a Markov chain—i.e., the Markov property—is not entirely justified. Moreover, it is often difficult to exactly determine the parameters that characterise the Markov chain. In an effort to handle these modelling errors in an elegant manner, several authors have recently turned to imprecise probabilities ([Škulj and Hable, 2013](#); [Hermans and de Cooman, 2012](#); [Škulj, 2015](#); [Krak et al., 2016](#); [De Bock, 2017](#)).

As [Krak et al. \(2016\)](#) thoroughly demonstrate, making inferences about an imprecise continuous-time Markov chain—determining lower and upper expectations or probabilities—requires the solution of a non-linear vector differential equation. To the best of our knowledge, this differential equation cannot be solved analytically, at least not in general. [Krak et al. \(2016\)](#) proposed a method to numerically approximate the solution of the differential equation, and argued that it outperforms the approximation method that [Škulj \(2015\)](#) previously introduced. One of the main results of this contribution is a novel approximation method that outperforms that of [Krak et al. \(2016\)](#).

An important property—both theoretically and practically—of continuous-time Markov chains is the behaviour of the solution of the differential equation as the time parameter recedes to infinity. If regardless of the initial condition the solution converges, we say that the chain is ergodic. We show that in this case the approximation is guaranteed to converge as well. This constitutes the second main result of this contribution and serves as a motivation behind the novel approximation method. Furthermore, we also quantify a worst-case convergence rate for the approximation. This unites the work of [Škulj \(2015\)](#), who studied the rate of convergence for discrete-time Markov chains,

and [De Bock \(2017\)](#), who studied the ergodic behaviour of continuous-time Markov chains from a qualitative point of view. One of the uses of our worst-case convergence rate is that it allows us to approximate the limit value of the solution up to a guaranteed error.

In order to comply with the page limit, we do not provide any proofs for our statements. We refer the interested reader to the appendix of ([Erreygers and De Bock, 2017](#)), an extended version of this contribution that is available on arXiv.

2. Mathematical Preliminaries

Throughout this contribution, we denote the set of real, non-negative real and strictly positive real numbers by \mathbb{R} , $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$, respectively. The set of natural numbers is denoted by \mathbb{N} , if we include zero we write $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For any set S , we let $|S|$ denote its cardinality. If a and b are two real numbers, we say that a is lower (greater) than b if $a \leq b$ ($a \geq b$), and that a is strictly lower (greater) than b if $a < b$ ($a > b$).

2.1 Gambles and Norms

We consider a finite *state space* \mathcal{X} , and are mainly concerned with real-valued functions on \mathcal{X} . All of these real-valued functions on \mathcal{X} are collected in the set $\mathcal{L}(\mathcal{X})$, which is a vector space. If we identify the state space \mathcal{X} with $\{1, \dots, |\mathcal{X}|\}$, then any function $f \in \mathcal{L}(\mathcal{X})$ can be identified with a vector: for all $x \in \mathcal{X}$, the x -component of this vector is $f(x)$. A special function on \mathcal{X} is the indicator \mathbb{I}_A of an event A . For any $A \subseteq \mathcal{X}$, it is defined for all $x \in \mathcal{X}$ as $\mathbb{I}_A(x) = 1$ if $x \in A$ and $\mathbb{I}_A(x) = 0$ otherwise. In order not to obfuscate the notation too much, for any $y \in \mathcal{X}$ we write \mathbb{I}_y instead of $\mathbb{I}_{\{y\}}$. If it is required from the context, we will also identify the real number $\gamma \in \mathbb{R}$ with the map γ from \mathcal{X} to \mathbb{R} , defined as $\gamma(x) = \gamma$ for all $x \in \mathcal{X}$.

We provide the set $\mathcal{L}(\mathcal{X})$ of functions with the standard maximum norm $\|\cdot\|$, defined for all $f \in \mathcal{L}(\mathcal{X})$ as $\|f\| := \max \{|f(x)| : x \in \mathcal{X}\}$. A seminorm that captures the variation of $f \in \mathcal{L}(\mathcal{X})$ will also be of use; we therefore define the variation seminorm $\|f\|_v := \max f - \min f$. Since the value $\|f\|_v / 2$ occurs often in formulas, we introduce the shorthand notation $\|f\|_c := \|f\|_v / 2$.

2.2 Non-Negatively Homogeneous Operators

An operator A that maps $\mathcal{L}(\mathcal{X})$ to $\mathcal{L}(\mathcal{X})$ is *non-negatively homogeneous* if for all $\mu \in \mathbb{R}_{\geq 0}$ and all $f \in \mathcal{L}(\mathcal{X})$, $A(\mu f) = \mu Af$. The maximum norm $\|\cdot\|$ for functions induces an operator norm:

$$\|A\| := \sup\{\|Af\| : f \in \mathcal{L}(\mathcal{X}), \|f\| = 1\}.$$

If for all $\mu \in \mathbb{R}$ and all $f, g \in \mathcal{L}(\mathcal{X})$, $A(\mu f + g) = \mu Af + Ag$, then the operator A is *linear*. In that case, it can be identified with a matrix of dimension $|\mathcal{X}| \times |\mathcal{X}|$, the (x, y) -component of which is $[A\mathbb{I}_y](x)$. The identity operator I is an important special case, defined for all $f \in \mathcal{L}(\mathcal{X})$ as $If := f$.

Two types of non-negatively homogeneous operators play a vital role in the theory of imprecise Markov chains: lower transition operators and lower transition rate operators.

Definition 1 An operator \underline{T} from $\mathcal{L}(\mathcal{X})$ to $\mathcal{L}(\mathcal{X})$ is called a *lower transition operator* if for all $f \in \mathcal{L}(\mathcal{X})$ and all $\mu \in \mathbb{R}_{\geq 0}$:

$$L1: \underline{T}f \geq \min f; \quad L2: \underline{T}(f + g) \geq \underline{T}f + \underline{T}g; \quad L3: \underline{T}(\mu f) = \mu \underline{T}f.$$

Every lower transition operator \underline{T} has a conjugate upper transition operator \overline{T} , defined for all $f \in \mathcal{L}(\mathcal{X})$ as $\overline{T}f := -\underline{T}(-f)$.

Definition 2 An operator \underline{Q} from $\mathcal{L}(\mathcal{X})$ to $\mathcal{L}(\mathcal{X})$ is called a lower transition rate operator if for any $f, g \in \mathcal{L}(\mathcal{X})$, any $\mu \in \mathbb{R}_{\geq 0}$, any $\gamma \in \mathbb{R}$ and any $x, y \in \mathcal{X}$ such that $x \neq y$:

$$R1: \underline{Q}\gamma = 0; \quad R2: \underline{Q}(f + g) \geq \underline{Q}f + \underline{Q}g; \quad R3: \underline{Q}(\mu f) = \mu \underline{Q}f; \quad R4: [\underline{Q}\mathbb{I}_x](y) \geq 0.$$

The conjugate lower transition rate operator \overline{Q} is defined for all $f \in \mathcal{L}(\mathcal{X})$ as $\overline{Q}f := -\underline{Q}(-f)$.

As will become clear in Section 3, lower transition operators and lower transition rate operators are tightly linked. For instance, we can use a lower transition rate operator to construct a lower transition operator. One way is to use Eqn. (1) further on. Another one is given in the following proposition, which is a strengthened version of (De Bock, 2017, Proposition 5).

Proposition 3 Consider any lower transition rate operator \underline{Q} and any $\delta \in \mathbb{R}_{\geq 0}$. Then the operator $(I + \delta \underline{Q})$ is a lower transition operator if and only if $\delta \|\underline{Q}\| \leq 2$.

We end this section with the first—although minor—novel result of this contribution. The norm of a lower transition rate operator is essential for all the approximation methods that we will discuss. The following proposition supplies us with an easy formula for determining it.

Proposition 4 Let \underline{Q} be a lower transition rate operator. Then $\|\underline{Q}\| = 2 \max\{|[\underline{Q}\mathbb{I}_x](x)| : x \in \mathcal{X}\}$.

Example 1 Consider a binary state space $\mathcal{X} = \{0, 1\}$ and two closed intervals $[\underline{q}_0, \bar{q}_0] \subset \mathbb{R}_{\geq 0}$ and $[\underline{q}_1, \bar{q}_1] \subset \mathbb{R}_{\geq 0}$. Let

$$\underline{Q}f := \min \left\{ \begin{bmatrix} q_0(f(1) - f(0)) \\ q_1(f(0) - f(1)) \end{bmatrix} : q_0 \in [\underline{q}_0, \bar{q}_0], q_1 \in [\underline{q}_1, \bar{q}_1] \right\} \text{ for all } f \in \mathcal{L}(\mathcal{X}).$$

Then one can easily verify that \underline{Q} is a lower transition rate operator.

Krak et al. (2016) also consider a running example with a binary state space, but they let $\mathcal{X} := \{\text{healthy}, \text{sick}\}$. We here identify healthy with 0 and sick with 1. In (Krak et al., 2016, Example 18), they propose the following values for the transition rates: $[\underline{q}_0, \bar{q}_0] := [1/52, 3/52]$ and $[\underline{q}_1, \bar{q}_1] := [1/2, 2]$. It takes Krak et al. a lot of work to determine the exact value of the norm of \underline{Q} , see (Krak et al., 2016, Example 19). We simply use Proposition 4: $\|\underline{Q}\| = 2 \max\{3/52, 2\} = 4$.

3. Imprecise Continuous-Time Markov Chains

For any lower transition rate operator \underline{Q} and any $f \in \mathcal{L}(\mathcal{X})$, Škulj (2015) has shown that the differential equation

$$\frac{d}{dt} \underline{T}_t f = \underline{Q} \underline{T}_t f. \quad (1)$$

with initial condition $\underline{T}_0 f := f$ has a unique solution for all $t \in \mathbb{R}_{\geq 0}$. Later, De Bock (2017) proved that the time-dependent operator \underline{T}_t itself satisfies a similar differential equation, and that it is a lower transition operator. Finding the unique solution of Eqn. (1) is non-trivial. Fortunately, we can approximate this solution, as by (De Bock, 2017, Proposition 10)

$$\underline{T}_t = \lim_{n \rightarrow \infty} \left(I + \frac{t}{n} \underline{Q} \right)^n. \quad (2)$$

Example 2 In the simple case of Example 1, we can use Eqn. (2) to obtain analytical expressions for the solution of Eqn. (1). Assume that $\underline{q}_0 + \bar{q}_1 > 0$ and fix some $t \in \mathbb{R}_{\geq 0}$. Then

$$[\underline{T}_t f](0) = f(0) + \underline{q}_0 h(t) \text{ and } [\underline{T}_t f](1) = f(1) - \bar{q}_1 h(t) \text{ for all } f \in \mathcal{L}(\mathcal{X}) \text{ with } f(0) \leq f(1),$$

where $h(t) := \|f\|_v (\underline{q}_0 + \bar{q}_1)^{-1} (1 - e^{-t(\underline{q}_0 + \bar{q}_1)})$. The case $f(0) \geq f(1)$ yields similar expressions.

For a linear lower transition rate operator \underline{Q} —i.e., if it is a transition rate matrix Q —Eqn. (2) reduces to the definition of the matrix exponential. It is well-known—see (Anderson, 1991)—that this matrix exponential $T_t = e^{tQ}$ can be interpreted as the transition matrix at time t of a time-homogeneous or stationary continuous-time Markov chain: the (x, y) -component of T_t is the probability of being in state y at time t if the chain started in state x at time 0. Therefore, it follows that the expectation of the function $f \in \mathcal{L}(\mathcal{X})$ at time $t \in \mathbb{R}_{\geq 0}$ conditional on the initial state $x \in \mathcal{X}$, denoted by $E(f(X_t)|X_0 = x)$, is equal to $[\underline{T}_t f](x)$.

As Eqn. (2) is a non-linear generalisation of the definition of the matrix exponential, we can interpret \underline{T}_t as the non-linear generalisation of the matrix exponential $T_t = e^{tQ}$. Extending this parallel, we might interpret \underline{T}_t as the non-linear generalisation of the transition matrix—i.e., as the lower transition operator—at time t of a generalised continuous-time Markov chain. In fact, Krak et al. (2016) prove that this is the case. They show that—under some conditions on \underline{Q} — $[\underline{T}_t f](x)$ can be interpreted as the tightest lower bound for $E(f(X_t)|X_0 = x)$ with respect to a set of—not necessarily Markovian—stochastic processes that are consistent with \underline{Q} . Krak et al. (2016) argue that, just like a transition rate matrix Q characterises a (precise) continuous-time Markov chain, a lower transition rate operator \underline{Q} characterises a so-called imprecise continuous-time Markov chain.

The main objective of this contribution is to determine $\underline{T}_t f$ for some $f \in \mathcal{L}(\mathcal{X})$ and some $t \in \mathbb{R}_{>0}$. Our motivation is that, from an applied point of view on imprecise continuous-time Markov chains, what one is most interested in are tight lower and upper bounds on expectations of the form $E(f(X_t)|X_0 = x)$. As explained above, the lower bound is given by $\underline{E}(f(X_t)|X_0 = x) = [\underline{T}_t f](x)$. Similarly, the upper bound is given by $\overline{E}(f(X_t)|X_0 = x) = -[\underline{T}_t(-f)](x)$. Note that the lower (or upper) probability of an event $A \subseteq \mathcal{X}$ conditional on the initial state x is a special case of a lower (or upper) expectation: $\underline{P}(X_t \in A|X_0 = x) = \underline{E}(\mathbb{I}_A(X_t)|X_0 = x)$ and similarly for the upper probability. Hence, for the sake of generality we can focus on $\underline{T}_t f$ and forget about its interpretation. As in most cases analytically solving Eqn. (1) is infeasible or even impossible, we resort to methods that yield an approximation up to some guaranteed maximal error.

4. Approximation Methods

Škulj (2015) was, to the best of our knowledge, the first to propose methods that approximate the solution $\underline{T}_t f$ of Eqn. (1). He proposes three methods: one with a uniform grid, a second with an adaptive grid and a third that is a combination of the previous two. In essence, he determines a step size δ and then approximates $\underline{T}_{t+\delta} f$ with $e^{\delta Q} \underline{T}_t f$, where Q is a transition rate matrix determined from \underline{Q} and $\underline{T}_t f$. One drawback of this method is that it needs the matrix exponential $e^{\delta Q}$, which—in general—needs to be approximated as well. Škulj (2015) mentions that his methods turn out to be quite computationally heavy, even if the uniform and adaptive methods are combined.

We consider two alternative approximation methods—one with a uniform grid and one with an adaptive grid—that both work in the same way. First, we pick a small step $\delta_1 \in \mathbb{R}_{\geq 0}$ and apply the operator $(I + \delta_1 \underline{Q})$ to the function $g_0 = f$, resulting in a function $g_1 := (I + \delta_1 \underline{Q})f$. Recall from

Proposition 3 that if we want $(I + \delta_1 \underline{Q})$ to be a lower transition operator, then we need to demand that $\delta_1 \|\underline{Q}\| \leq 2$. Next, we pick a (possibly different) step $\delta_2 \in \mathbb{R}_{\geq 0}$ such that $\delta_2 \|\underline{Q}\| \leq 2$ and apply the lower transition operator $(I + \delta_2 \underline{Q})$ to the function g_1 , resulting in a function $g_2 := (I + \delta_2 \underline{Q})g_1$. If we continue this process until the sum of all the small steps is equal to t , then we end up with an approximation for $\underline{T}_t f$. More formally, let $s := (\delta_1, \dots, \delta_k)$ denote a sequence in $\mathbb{R}_{\geq 0}$ such that, for all $i \in \{1, \dots, k\}$, $\delta_i \|\underline{Q}\| \leq 2$. Using this sequence s we define the *approximating lower transition operator*

$$\Phi(s) := (I + \delta_k \underline{Q}) \cdots (I + \delta_1 \underline{Q}).$$

What we are looking for is a convenient way to determine the sequence s such that the error $\|\underline{T}_t f - \Phi(s)f\|$ is guaranteed to be lower than some desired maximal error $\epsilon \in \mathbb{R}_{>0}$.

4.1 Using a Uniform Grid

Krak et al. (2016) provide one way to determine the sequence s . They assume a uniform grid, in the sense that all elements of the sequence s are equal to δ . The step size δ is completely determined by the desired maximal error ϵ , the time t , the variation norm of the function f and the norm of \underline{Q} ; (Krak et al., 2016, Proposition 8.5) guarantees that the actual error is lower than ϵ . Algorithm 1 provides a slightly improved version of (Krak et al., 2016, Algorithm 1). The improvement is due to Proposition 3: we demand that $n \geq t \|\underline{Q}\| / 2$ instead of $n \geq t \|\underline{Q}\|$.

Algorithm 1: Uniform approximation

Data: A lower transition rate operator \underline{Q} , a function $f \in \mathcal{L}(\mathcal{X})$, a maximal error $\epsilon \in \mathbb{R}_{>0}$, and a time point $t \in \mathbb{R}_{\geq 0}$.

Result: $\underline{T}_t f \pm \epsilon$

- 1 $g_0 \leftarrow f$
 - 2 **if** $\|f\|_c = 0$ **or** $\|\underline{Q}\| = 0$ **or** $t = 0$ **then** $(n, \delta) \leftarrow (0, 0)$
 - 3 **else**
 - 4 $n \leftarrow \lceil \max\{t \|\underline{Q}\| / 2, t^2 \|\underline{Q}\|^2 \|f\|_c / \epsilon\} \rceil$
 - 5 $\delta \leftarrow t/n$
 - 6 **for** $i = 0, \dots, n - 1$ **do**
 - 7 $g_{i+1} \leftarrow g_i + \delta \underline{Q} g_i$
 - 8 **return** g_n
-

More formally, for any $t \in \mathbb{R}_{\geq 0}$ and any $n \in \mathbb{N}$ such that $t \|\underline{Q}\| \leq 2n$, we consider the *uniformly approximating lower transition operator*

$$\Psi_t(n) := \left(I + \frac{t}{n} \underline{Q} \right)^n.$$

As a special case, we define $\Psi_t(0) := I$. The following theorem then guarantees that the choice of n in Algorithm 1 results in an error $\|\underline{T}_t f - \Psi_t(n)f\|$ that is lower than the desired maximal error ϵ .

Theorem 5 Let \underline{Q} be a lower transition rate operator and fix some $f \in \mathcal{L}(\mathcal{X})$, $t \in \mathbb{R}_{\geq 0}$ and $\epsilon \in \mathbb{R}_{>0}$. If we use Algorithm 1 to determine n , δ and g_0, \dots, g_n , then we are guaranteed that

$$\|\underline{T}_t f - \Psi_t(n)f\| = \|\underline{T}_t f - g_n\| \leq \epsilon' := \delta^2 \|\underline{Q}\|^2 \sum_{i=0}^{n-1} \|g_i\|_c \leq \epsilon.$$

Table 1: Comparison of the presented approximation methods, obtained using a naive, unoptimised implementation of the algorithms in Python. N is the total number of iterations, D_ϵ ($D_{\epsilon'}$) is the average duration—in seconds, averaged over 50 independent runs—without (with) keeping track of ϵ' , and ϵ_a is the actual error. The Python code is made available at github.com/alexander-e/ictmc.

Method	N	D_ϵ	$D_{\epsilon'}$	$\epsilon' \times 10^3$	$\epsilon_a \times 10^3$
Uniform	8,000	0.0345	0.0574	0.430	0.0335
Uniform	250	0.00171	0.0264	13.8	1.07
Adaptive with $m = 1$	3,437	0.0371	0.0428	1.000	0.108
Adaptive with $m = 20$	3,456	0.0143	0.0254	0.992	0.107
Uniform ergodic with $m = 1$	6,133	0.0264	0.0449	0.560	0.0437

Theorem 5 is an extension of (Krak et al., 2016, Proposition 8.5). We already mentioned that the demand $n \geq t \|Q\|$ can be relaxed to $n \geq t \|Q\| / 2$. Furthermore, it turns out that we can compute an upper bound $\overline{\epsilon'}$ on the error that is (possibly) lower than the desired maximal error ϵ . If we want to determine this ϵ' while running Algorithm 1, we simply need to add $\epsilon' \leftarrow 0$ to line 1 and insert $\epsilon' \leftarrow \epsilon' + \delta^2 \|Q\|^2 \|g_i\|_c$ just before line 7.

Example 3 We again consider the simple case of Example 1 and illustrate the use of Theorem 5 with a numerical example based on (Krak et al., 2016, Example 20). Krak et al. (2016) use Algorithm 1 to approximate $\underline{T}_1 \mathbb{I}_1$, and find that $n = 8,000$ guarantees an error lower than the desired maximal error $\epsilon := 1 \times 10^{-3}$. As reported in Table 1, we use Theorem 5 to compute ϵ' . We find that $\epsilon' \approx 0.430 \times 10^{-3}$, which is approximately a factor two smaller than the desired maximal error ϵ .

In this case, since we know the analytical expression for $\underline{T}_1 \mathbb{I}_1$ from Example 2, we can determine the actual error $\epsilon_a = \|\underline{T}_1 \mathbb{I}_1 - \Psi_1(8000) \mathbb{I}_1\|$. Quite remarkably, the actual error is approximately 3.35×10^{-5} , which is roughly 30 times smaller than the desired maximal error. This leads us to think that the number of iterations used by the uniform method is too high. In fact, we find that using as few as 250 iterations—roughly $8,000/30$ —already results in an actual error that is approximately equal to the desired one: $\|\underline{T}_1 \mathbb{I}_1 - \Psi_1(250) \mathbb{I}_1\| \approx 1.07 \times 10^{-3}$.

4.2 Using an Adaptive Grid

In Example 3, we noticed that the maximal desired error was already satisfied for a uniform grid that was much coarser than that constructed by Algorithm 1. Because of this, we are led to believe that we can find a better approximation method than the uniform method of Algorithm 1.

To this end, we now consider grids where, for some integer m , every m consecutive time steps in the grid are equal. In particular, we consider a sequence $\delta_1, \dots, \delta_n$ in $\mathbb{R}_{\geq 0}$ and some $k \in \mathbb{N}$ such that $1 \leq k \leq m$ and, for all $i \in \{1, \dots, n\}$, $\delta_i \|Q\| \leq 2$. From such a sequence, we then construct the m -fold approximating lower transition operator:

$$\Phi_{m,k}(\delta_1, \dots, \delta_n) := (I + \delta_n \underline{Q})^k (I + \delta_{n-1} \underline{Q})^{m-k} \cdots (I + \delta_1 \underline{Q})^m,$$

where if $n = 1$ only $(I + \delta_1 \underline{Q})^k$ remains and if $n = 2$ only $(I + \delta_2 \underline{Q})^k (I + \delta_1 \underline{Q})^{m-k}$ remains.

The uniform approximation method of before is a special case of the m -fold approximating lower transition operator; a more interesting method to construct an m -fold approximation is Algorithm 2. In this algorithm, we re-evaluate the time step every m iterations, possibly increasing its length.

Algorithm 2: Adaptive approximation

Data: A lower transition rate operator \underline{Q} , a gamble $f \in \mathcal{L}(\mathcal{X})$, an integer $m \in \mathbb{N}$, a tolerance $\epsilon \in \mathbb{R}_{>0}$, and a time period $t \in \mathbb{R}_{\geq 0}$.

Result: $\underline{T}_t f \pm \epsilon$

```

1  $(g_{(0,m)}, \Delta, i) \leftarrow (f, t, 0)$ 
2 if  $\|f\|_c = 0$  or  $\|\underline{Q}\| = 0$  or  $t = 0$  then  $(n, k) \leftarrow (0, m)$ 
3 else
4   while  $\Delta > 0$  and  $\|g_{(i,m)}\|_c > 0$  do
5      $i \leftarrow i + 1$ 
6      $\delta_i \leftarrow \min\{\Delta, 2/\|\underline{Q}\|, \epsilon/(t \|\underline{Q}\|^2 \|g_{(i-1,m)}\|_c)\}$ 
7     if  $m\delta_i > \Delta$  then
8        $k_i \leftarrow \lceil \Delta/\delta_i \rceil$ 
9        $\delta_i \leftarrow \Delta/k_i$ 
10    else  $k_i \leftarrow m$ 
11     $g_{(i,0)} \leftarrow g_{(i-1,m)}, \Delta \leftarrow \Delta - k_i \delta_i$ 
12    for  $j = 0, \dots, k_i - 1$  do
13       $g_{(i,j+1)} \leftarrow g_{(i,j)} + \delta_i \underline{Q} g_{(i,j)}$ 
14   $(n, k) \leftarrow (i, k_i)$ 
15 return  $g_{(n,k)}$ 

```

From the properties of lower transition operators, it follows that for all $i \in \{2, \dots, n-1\}$, $\|g_{(i-1,m)}\|_c \leq \|g_{(i-2,m)}\|_c$. Hence, the re-evaluated step size δ_i is indeed larger than (or equal to) the previous step size δ_{i-1} . The only exception to this is the final step size δ_n : it might be that the remaining time Δ is smaller than $m\delta_n$, in which case we need to choose k and δ_n such that $k\delta_n = \Delta$.

Theorem 6 guarantees that the adaptive approximation of Algorithm 2 indeed results in an actual error lower than the desired maximal error ϵ . Even more, it provides a method to compute an upper bound ϵ' of the actual error that is lower than the desired maximal error. Finally, it also states that the adaptive method of Algorithm 2 needs at most an equal number of iterations than the uniform method of Algorithm 1.

Theorem 6 Let \underline{Q} be a lower transition rate operator, $f \in \mathcal{L}(\mathcal{X})$, $t \in \mathbb{R}_{\geq 0}$, $\epsilon \in \mathbb{R}_{>0}$ and $m \in \mathbb{N}$. We use Algorithm 2 to determine n and k , and if applicable also k_i , δ_i and $g_{(i,j)}$. If $\|f\|_c = 0$, $\|\underline{Q}\| = 0$ or $t = 0$, then $\|\underline{T}_t f - g_{(n,k)}\| = 0$. Otherwise, we are guaranteed that

$$\|\underline{T}_t f - \Phi_{m,k}(\delta_1, \dots, \delta_n) f\| = \|\underline{T}_t f - g_{(n,k)}\| \leq \epsilon' := \sum_{i=1}^n \delta_i^2 \|\underline{Q}\|^2 \sum_{j=0}^{k_i-1} \|g_{(i,j)}\|_c \leq \epsilon$$

and that the total number of iterations has an upper bound:

$$\sum_{i=1}^n k_i = (n-1)m + k \leq \left\lceil \max \left\{ \|\underline{Q}\| t/2, t^2 \|\underline{Q}\|^2 \|f\|_c / \epsilon \right\} \right\rceil.$$

Again, we can determine ϵ' while running Algorithm 2. An alternate—less tight—version of ϵ' can be obtained by replacing the sum of $\|g_{(i,j)}\|_c$ for j from 0 to $k_i - 1$ by $k_i \|g_{(i,0)}\|_c = k_i \|g_{(i-1,m)}\|_c$. Determining this alternative ϵ' while running Algorithm 2 adds negligible computational overhead compared to the ϵ' of Theorem 6, as $\|g_{(i-1,m)}\|_c$ is needed to re-evaluate the step size anyway.

The reason why we only re-evaluate the step size δ after every m iterations is twofold. First and foremost, all we currently know for sure is that for all $\delta \in \mathbb{R}_{\geq 0}$ such that $\delta \|\underline{Q}\| \leq 2$, all $m \in \mathbb{N}$ and all $f \in \mathcal{L}(\mathcal{X})$, $\|(I + \delta \underline{Q})^m f\|_c \leq \|f\|_c$. Re-evaluating the step size every m iterations is therefore only justified if a priori we are certain that $\|(I + \delta_i \underline{Q})^m g_{(i-1,m)}\|_c < \|g_{(i-1,m)}\|_c$. We come back to this in Section 5. A second reason is that there might be a trade-off between the time it takes to re-evaluate the step size and the time that is gained by the resulting reduction of the number of iterations. The following numerical example illustrates this trade off.

Example 4 Recall that in Example 3 we wanted to approximate $\underline{T}_1 \mathbb{I}_1$ up to a maximal desired error $\epsilon = 1 \times 10^{-3}$. Instead of using the uniform method of Algorithm 1, we now use the adaptive method of Algorithm 2 with $m = 1$. The initial step size is the same as that of the uniform method, but because we re-evaluate the step size we only need 3,437 iterations, as reported in Table 1. We also find that in this case $\epsilon' = 1.00 \times 10^{-3}$, which is a coincidence. Nevertheless, the actual error of the approximation is 0.108×10^{-3} , which is about ten times smaller than what we were aiming for.

However, fewer iterations do not necessarily imply a shorter duration of the computations. Qualitatively, we can conclude the following from Table 1. First, keeping track of ϵ' increases the duration, as expected. Second, the adaptive method is faster than the uniform method, at least if we choose m large enough. And third, both methods yield an actual error that is at least an order of magnitude lower than the desired maximal error.

5. Ergodicity

Let $\Phi_{m,k}(\delta_1, \dots, \delta_n)f$ be an approximation constructed using the adaptive method of Algorithm 2. Re-evaluating the step size is then only justified if a priori we are sure that

$$\frac{1}{2} \|(I + \delta_i \underline{Q})^m \Phi_{i-1} f\|_v = \|g_{(i,m)}\|_c < \|g_{(i-1,m)}\|_c = \frac{1}{2} \|\Phi_{i-1} f\|_v \text{ for all } i \in \{1, \dots, n-1\},$$

where $\Phi_0 := I$ and $\Phi_i := (I + \delta_i \underline{Q})^m \Phi_{i-1}$. As $(\Phi_{i-1} f) \in \mathcal{L}(\mathcal{X})$, this is definitely true if we require that

$$(\forall \delta \in \{\delta_1, \dots, \delta_{n-1}\})(\forall f \in \mathcal{L}(\mathcal{X})) \quad \|(I + \delta \underline{Q})^m f\|_v < \|f\|_v. \quad (3)$$

In fact, since this inequality is invariant under translation or positive scaling of f , it suffices if

$$(\forall \delta \in \{\delta_1, \dots, \delta_{n-1}\})(\forall f \in \mathcal{L}(\mathcal{X}): 0 \leq f \leq 1) \quad \|(I + \delta \underline{Q})^m f\|_v < 1.$$

Readers that are familiar with (the ergodicity of) imprecise discrete-time Markov chains—see (Hermans and de Cooman, 2012) or (Škulj and Hable, 2013)—will probably recognise this condition, as it states that the (weak) coefficient of ergodicity of $(I + \delta \underline{Q})^m$ should be strictly smaller than 1. For all lower transition operators \underline{T} , Škulj and Hable (2013) define this (weak) *coefficient of ergodicity* as

$$\rho(\underline{T}) := \max \{\|\underline{T}f\|_v : f \in \mathcal{L}(\mathcal{X}), 0 \leq f \leq 1\}. \quad (4)$$

5.1 Ergodicity of Lower Transition Rate Operators

As will become apparent, whether or not combinations of $m \in \mathbb{N}$ and $\delta \in \mathbb{R}_{\geq 0}$ exist such that $\delta \|\underline{Q}\| \leq 2$ and $\rho((I + \delta \underline{Q})^m) < 1$ is tightly connected with the behaviour of $\underline{T}_t f$ for large t . De Bock (2017) proved that for all lower transition rate operator \underline{Q} and all $f \in \mathcal{L}(\mathcal{X})$, the limit $\lim_{t \rightarrow \infty} \underline{T}_t f$ exists. An important case is when this limit is a constant function for all f .

Definition 7 (Definition 2 of (De Bock, 2017)) *The lower transition rate operator \underline{Q} is ergodic if for all $f \in \mathcal{L}(\mathcal{X})$, $\lim_{t \rightarrow \infty} \underline{T}_t f$ exists and is a constant function.*

As shown by De Bock (2017), ergodicity is easily verified in practice: it is completely determined by the signs of $[\underline{Q}\mathbb{I}_x](y)$ and $[\underline{Q}\mathbb{I}_A](z)$, for all $x, y \in \mathcal{X}$ and certain combinations of $z \in \mathcal{X}$ and $A \subset \mathcal{X}$. It turns out that an ergodic lower transition rate operator \underline{Q} does not only induce a lower transition operator \underline{T}_t that converges, it also induces discrete approximations—of the form $(I + \delta_k \underline{Q}) \cdots (I + \delta_1 \underline{Q})$ —with special properties. The following theorem, which we consider to be one of the main results of this contribution, highlights this.

Theorem 8 *The lower transition rate operator \underline{Q} is ergodic if and only if there is some $n < |\mathcal{X}|$ such that $\rho(\Phi(\delta_1, \dots, \delta_k)) < 1$ for one (and then all) $k \geq n$ and one (and then all) sequence(s) $\delta_1, \dots, \delta_k$ in $\mathbb{R}_{>0}$ such that $\delta_i \|\underline{Q}\| < 2$ for all $i \in \{1, \dots, k\}$.*

5.2 Ergodicity and the Uniform Approximation Method

Theorem 8 guarantees that the conditions that were discussed at the beginning of this section are satisfied. In particular, if the lower transition rate operator is ergodic, then there is some $n < |\mathcal{X}|$ such that $\rho((I + \delta \underline{Q})^m) < 1$ for all $m \geq n$ and all $\delta \in \mathbb{R}_{>0}$ such that $\delta \|\underline{Q}\| < 2$. Consequently, if we choose $m \geq |\mathcal{X}| - 1$ then re-evaluating the step size δ will—except maybe for the last re-evaluation—result in a new step size that is strictly greater than the previous one. Therefore, we conclude that if the lower transition rate operator is ergodic, then using the adaptive method of Algorithm 2 is certainly justified; it will result in fewer iterations, provided we choose a large enough m .

Another nice consequence of the ergodicity of a lower transition rate operator \underline{Q} is that we can prove an alternate a priori guaranteed upper bound for the error of uniform approximations.

Proposition 9 *Let \underline{Q} be a lower transition rate operator and fix some $f \in \mathcal{L}(\mathcal{X})$, $m, n \in \mathbb{N}$ and $\delta \in \mathbb{R}_{>0}$ such that $\delta \|\underline{Q}\| < 2$. If $\beta := \rho((I + \delta \underline{Q})^m) < 1$, then*

$$\|\underline{T}_t f - \Psi_t(n)\| \leq \epsilon_e := m\delta^2 \|\underline{Q}\|^2 \|f\|_c \frac{1 - \beta^k}{1 - \beta} \leq \epsilon_d := \frac{m\delta^2 \|\underline{Q}\|^2 \|f\|_c}{1 - \beta},$$

where $t := n\delta$ and $k := \lceil n/m \rceil$. The same is true for $\beta = \rho(\underline{T}_{m\delta})$.

Interestingly enough, the upper bound ϵ_d is not dependent on t (or n) at all! This is a significant improvement on the upper bound of Theorem 5, as that upper bound is proportional to t^2 .

By Theorem 8, there always is an $m < |\mathcal{X}|$ such that $\rho((I + \delta \underline{Q})^m) < 1$ for all $\delta \in \mathbb{R}_{>0}$ such that $\delta \|\underline{Q}\| < 2$. Thus, given such an m , we can easily improve Algorithm 1. After we have determined n and δ with Algorithm 1, we can simply determine the upper bound of Proposition 9. If $m(1 - \beta^k) < n(1 - \beta)$ (or $m < n(1 - \beta)$), then this upper bound is smaller than the desired maximal error ϵ , and we have found a tighter upper bound on the actual error. We can even go the extra mile and replace line 4 with a method that looks for the smallest possible $n \in \mathbb{N}$ that yields

$$m\delta^2 \|\underline{Q}\|^2 \|f\|_c (1 - \beta^k) \leq (1 - \beta)\epsilon,$$

where $k = \lceil n/m \rceil$ and $\delta = t/n$ —and therefore also β —are dependent of n . This method could yield a smaller n , but the time we gain by having to execute fewer iterations does not necessarily compensate the time lost by looking for a smaller n . In any case, to actually implement these improvements we need to be able to compute $\beta := \rho((I + \delta\underline{Q})^m)$.

Example 5 For the simple case of Example 1, we can derive an analytical expression for $\rho((I + \delta\underline{Q}))$ that is valid for all $\delta \in \mathbb{R}_{\geq 0}$ such that $\delta \|\underline{Q}\| \leq 2$. Therefore, we can use Proposition 9 to a priori determine an upper bound for the error. If we choose $m = 1$, then $\epsilon_e = 0.767 \times 10^{-3}$ and $\epsilon_d = 1.79 \times 10^{-3}$. Note that $\epsilon_e < \epsilon$, so we can probably decrease the number of iterations n . As reported in Table 1, we find that $n = 6,133$ still suffices, and that this results in an approximation correct up to $\epsilon' = 0.560 \times 10^{-3}$, roughly two times smaller than the desired maximal error ϵ . The actual error is 0.0437×10^{-3} , roughly ten times smaller than ϵ .

5.3 Approximating the Coefficient of Ergodicity

Unfortunately, determining the exact value of $\rho((I + \delta\underline{Q})^m)$ —and of $\rho(\underline{T})$ in general—turns out to be non-trivial and is often even impossible. Nevertheless, the following theorem gives some—actually computable—lower and upper bounds for the coefficient of ergodicity.

Theorem 10 Let \underline{T} be a lower transition operator. Then

$$\rho(\underline{T}) \leq \max \left\{ \max \{ [\underline{T}\mathbb{I}_A](x) - [\underline{T}\mathbb{I}_A](y) : x, y \in \mathcal{X} \} : \emptyset \neq A \subset \mathcal{X} \right\}, \quad (5)$$

$$\rho(\underline{T}) \geq \max \left\{ \max \{ [\underline{T}\mathbb{I}_A](x) - [\underline{T}\mathbb{I}_A](y) : x, y \in \mathcal{X} \} : \emptyset \neq A \subset \mathcal{X} \right\}. \quad (6)$$

The upper bound in Theorem 10 is particularly useful in combination with Proposition 9, as it allows us to replace $\beta := \rho((I + \delta\underline{Q})^m)$ with a guaranteed upper bound. Of course, this only makes sense if this upper bound is strictly smaller than one. The following proposition guarantees that, for ergodic lower transition rate operators \underline{Q} , this is always the case.

Proposition 11 Let \underline{Q} be an ergodic lower transition rate operator. Then there is some $n < |\mathcal{X}|$ such that, for all $k \geq n$ and $\delta_1, \dots, \delta_k$ in $\mathbb{R}_{>0}$ such that $\delta_i \|\underline{Q}\| < 2$ for all $i \in \{1, \dots, k\}$, the upper bound for $\rho(\Phi(\delta_1, \dots, \delta_k))$ that is given by Eqn. (5) is strictly smaller than one.

5.4 Approximating Limit Values

The results that we have obtained earlier in this section naturally lead to a method to approximate $\underline{T}_\infty f := \lim_{t \rightarrow \infty} \underline{T}_t f$ up to some maximal error. This is an important problem in applications; for instance, Troffaes et al. (2015) try to determine $\underline{T}_\infty f$ for an ergodic lower transition rate operator that arises in their specific reliability analysis application. The method they use is rather ad hoc: they pick some t and n and then determine the uniform approximation $\Psi_t(n)f$. As $\|\Psi_t(n)f\|_v$ is small, they suspect that they are close to the actual limit value. They also observe that $\Psi_{2t}(4n)f$ only differs from $\Psi_t(n)f$ after the fourth significant digit, which they regard as further empirical evidence for the correctness of their approximation. While this ad hoc method seemingly works, the initial values for t and n have to be chosen somewhat arbitrarily. Also, this method provides no guarantee that the actual error is lower than some desired maximal error.

Theorem 8, Proposition 9, Theorem 10 and the following stopping criterion allow us to propose a method that corrects these two shortcomings.

Proposition 12 Let \underline{Q} be an ergodic lower transition rate operator and let $f \in \mathcal{L}(\mathcal{X})$, $t \in \mathbb{R}_{\geq 0}$ and $\epsilon \in \mathbb{R}_{>0}$. Let s denote a sequence $\delta_1, \dots, \delta_k$ in $\mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^k \delta_i = t$ and, for all $i \in \{1, \dots, k\}$, $\delta_i \|\underline{Q}\| \leq 2$. If $\|\underline{T}_t f - \Phi(s)f\| \leq \epsilon/2$ and $\|\Phi(s)f\|_c \leq \epsilon/2$, then for all $\Delta \in \mathbb{R}_{\geq 0}$:

$$\left| \underline{T}_{t+\Delta} f - \frac{\max \Phi(s) + \min \Phi(s)}{2} \right| \leq \epsilon \quad \text{and} \quad \left| \underline{T}_\infty f - \frac{\max \Phi(s) + \min \Phi(s)}{2} \right| \leq \epsilon.$$

Without actually stating it, we mention that a similar—though less useful—stopping criterion can be proved for non-ergodic transition rate matrices as well.

Our method for determining $\underline{T}_\infty f$ is now relatively straightforward. Let \underline{Q} be an ergodic lower transition rate operator and fix some $f \in \mathcal{L}(\mathcal{X})$. We can then approximate $\underline{T}_\infty f$ up to any desired maximal error $\epsilon \in \mathbb{R}_{>0}$ as follows. First, we look for some $m \in \mathbb{N}$ and some—preferably large— $\delta \in \mathbb{R}_{>0}$ such that $\delta \|\underline{Q}\| < 2$ and

$$2m\delta^2 \|\underline{Q}\|^2 \|f\|_c \leq (1 - \beta)\epsilon,$$

where $\beta := \rho((I + \delta\underline{Q})^m)$. From Theorem 8, we know that a possible starting point for m is $|\mathcal{X}| - 1$. If we do not have an analytical expression for $\rho((I + \delta\underline{Q})^m)$, then we know from Proposition 11 that we can instead use the guaranteed upper bound of Theorem 10. If no such m and δ exist—for instance because the guaranteed upper bound on β is too conservative—then this method does not work. If on the other hand we do find such an m and δ , then we can keep on running the iterative step (line 7) of Algorithm 1 until we reach the first index $i \in \mathbb{N}$ such that $\|g_i\|_c \leq \epsilon/2$. By Propositions 9 and 12, we are now guaranteed that $(\max g_i + \min g_i)/2$ is an approximation of $\underline{T}_\infty f$ up to a maximal error ϵ .

Alternatively, we can fix a step size δ ourselves and use the method of Theorem 5 to compute ϵ' . In that case, we simply need to run the iterative scheme until we reach the first index i such that $\|g_i\|_c \leq \epsilon'$. By Proposition 12, we are then guaranteed that the error $(\max g_i + \min g_i)/2$ is an approximation of $\underline{T}_\infty f$ up to a maximal error $\epsilon = 2\epsilon'$. The same is true if we replace ϵ' by the error ϵ_e that is used in Proposition 9.

Example 6 Using the analytical expressions of Example 2, we obtain $\underline{T}_\infty \mathbb{I}_1 \approx 9.5238095 \times 10^{-3}$.

We want to approximate $\underline{T}_\infty \mathbb{I}_1$ up to a maximum error $\epsilon := 1 \times 10^{-6}$. We observe that $m = 1$ and $\delta \approx 3.485 \times 10^{-8}$ yield an ϵ_d that is lower than $\epsilon/2$. After 196,293,685 iterations, the norm of the approximation is sufficiently small, resulting in the approximation $\underline{T}_\infty \mathbb{I}_1 = (9.524 \pm 0.001) \times 10^{-3}$. Alternatively, choosing $\delta = 1 \times 10^{-7}$ and continuing until $\|g_i\|_c \leq \epsilon'$ yields the approximation $\underline{T}_\infty \mathbb{I}_1 = (9.5242 \pm 0.0008) \times 10^{-3}$ after only 69,572,154 iterations.

Mimicking Troffaes et al. (2015), we also tried the heuristic method of increasing t and n until we observe empirical convergence. After some trying, we find that $t = 7$ and $n = 7 \cdot 250 = 1750$ already yield an approximation with sufficiently small error: $\|\underline{T}_\infty \mathbb{I}_1 - \Psi_7(1750)\mathbb{I}_1\| \approx 7 \times 10^{-7} < \epsilon$. Note however that for non-binary examples, where $\underline{T}_\infty f$ cannot be computed analytically, this heuristic approach is unable to provide a guaranteed bound.

6. Conclusion

We have improved an existing method and proposed a novel method to approximate $\underline{T}_t f$ up to any desired maximal error, where $\underline{T}_t f$ is the solution of the non-linear differential equation (1) that plays an essential role in the theory of imprecise continuous-time Markov chains. As guaranteed by our

theoretical results, and as verified by our numerical examples, our methods outperform the existing method by Krak et al. (2016), especially if the lower transition rate operator is ergodic. For these ergodic lower transition rate operators, we also proposed a method to approximate $\lim_{t \rightarrow \infty} \underline{T}_t f$ up to any desired maximal error.

For the simple case of a binary state space, we observed in numerical examples that there is a rather large difference between the theoretically required number of iterations and the number of iterations that are empirically found to be sufficient. Similar differences can—although this falls beyond the scope of our present contribution—also be observed for the lower transition rate operator that is studied in (Troffaes et al., 2015). The underlying reason for these observed differences remains unclear so far. On the one hand, it could be that our methods are still on the conservative side, and that further improvements are possible. On the other hand, it might be that these differences are unavoidable, in the sense that guaranteed theoretical bounds come at the price of conservatism. We leave this as an interesting line of future research. Additionally, the performance of our proposed methods for systems with a larger state space deserves further inquiry.

Acknowledgments

Jasper De Bock is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO) and wishes to acknowledge its financial support. His work was also partially supported by the H2020-MSCA-ITN-2016 UTOPIAE, grant agreement 722734. Finally, the authors would like to express their gratitude to three anonymous reviewers, for their time, effort and constructive feedback.

References

- W. J. Anderson. *Continuous-Time Markov Chains*. Springer-Verlag New York, 1991.
- J. De Bock. The limit behaviour of imprecise continuous-time Markov chains. *Journal of Nonlinear Science*, 27(1):159–196, 2017.
- A. Erreygers and J. De Bock. Imprecise continuous-time Markov chains: Efficient computational methods with guaranteed error bounds. 2017. arXiv Report 1702.07150 [math.PR].
- F. Hermans and G. de Cooman. Characterisation of ergodic upper transition operators. *International Journal of Approximate Reasoning*, 53(4):573 – 583, 2012.
- T. Krak, J. De Bock, and A. Siebes. Imprecise continuous-time Markov chains. 2016. arXiv Report 1611.05796 [math.PR].
- D. Škulj. Efficient computation of the bounds of continuous time imprecise Markov chains. *Applied Mathematics and Computation*, 250:165–180, 2015.
- M. C. M. Troffaes, J. Gledhill, D. Škulj, and S. Blake. Using imprecise continuous time Markov chains for assessing the reliability of power networks with common cause failure and non-immediate repair. In *Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pages 287–294, 2015.
- D. Škulj and R. Hable. Coefficients of ergodicity for Markov chains with uncertain parameters. *Metrika*, 76(1):107–133, 2013.

(Generalized) Linear Regression on Microaggregated Data – From Nuisance Parameter Optimization to Partial Identification

Paul Fink

Thomas Augustin

Department of Statistics, Ludwig-Maximilians-Universität München (LMU Munich)

Munich (Germany)

PAUL.FINK@STAT.UNI-MUENCHEN.DE

AUGUSTIN@STAT.UNI-MUENCHEN.DE

Abstract

Protecting sensitive micro data prior to publishing or passing the data itself on is a crucial aspect: A trade-off between sufficient disclosure control and analyzability needs to be found. This paper presents a starting point to evaluate the effect of k -anonymity microaggregated data in (generalized) linear regression. Taking a rigorous imprecision perspective, microaggregated data are understood inducing a set \mathbb{X} of potentially true data. Based on this representation two conceptually different approaches deriving estimations from the *ideal* likelihood are discussed. The first one picks a single element of \mathbb{X} , for instance by naively treating the microaggregated data as true ones or by introducing a maximax approach taking the elements of \mathbb{X} as nuisance parameters to be optimized. The second one seeks, in the spirit of Partial Identification, the set of all maximum likelihood estimators compatible with the elements of \mathbb{X} , thus creating cautious estimators. As the simulation study corroborates, the obtained sets of estimators of the latter approach are still precise enough to be practically relevant.

Keywords: maximum likelihood estimation; generalized linear regression; microaggregation; anonymization; partial identification.

1. Introduction

In recent years, more data are made available, for instance in the context of websites for marketing purposes or also by institutions of Official Statistics¹. These micro data usually contain sensitive information of the units involved. As the combination of different disjoint data sources requires lesser effort now, data obtained in one isolated context may be not revealing, however, the join of multiple data sources may make the unit identifiable. Therefore, powerful anonymization techniques for disclosure control, inducing an information reduction, are essential to protect the privacy and strengthen the collected data quality. Scientific researchers have a diametrically opposite aim: They desire a deeper understanding of the unit's action and/or the social or economic processes involved, hence the availability of minute information about the units under consideration is essential. As data sources they may rely either on self-collected data or on such from a different source, for instance Official Statistics or private companies. Henceforth a trade-off between granting sufficient privacy while still ensuring data utility is required. In cases when the anonymization technique does not provide sufficient information after its deployment, the availability of the data itself is reduced to absurdity as the data provision might be scrapped entirely.

A well known concept in this setting is the so-called k -*anonymity* proposed by Sweeney (2002): It guarantees that each value of each anonymized variable occurs at least k times. Hence, even if

1. The European Statistics Code of Practice explicitly encourages to make the collected data publicly available (cf. Eurostat and European Statistical System (2011, principle 15))

attackers know identifying aspects of one record within the micro data, they are unable to deduce the actual value of the sensitive variable in question. One specific concept ensuring k -anonymity is *microaggregation*, which belongs to the family of perturbative methods in statistical disclosure control (e.g. [Willenborg and de Waal \(2001\)](#)), replacing individual records by a representative substitute, e.g. a group average. Initially developed to deal with just continuous variables, microaggregation is nowadays applicable to any measurement scale of the variable(s) to anonymize. However, in this paper the focus is on continuous variables. Microaggregation by a given technique is herein understood as a mapping m , which by design maps several situations of the values to be microaggregated \mathbf{x} onto the same microaggregation result in the image $\tilde{\mathbf{x}}$. However, when analyzing microaggregated data, one is interested in findings on the original data and therefore in the reverse mapping which induces a set of compatible underlying data situations

$$\mathbb{X}(\tilde{\mathbf{x}}) = \{\mathbf{x} \mid m(\mathbf{x}) = \tilde{\mathbf{x}}\}.$$

In literature, microaggregation techniques have been evaluated mostly in the light of disclosure control, but few on the potential for analyses: To which extent are structures in the original data inferable by looking at the anonymized data? This paper deals with linear regression, a standard statistical model strategy which is commonly employed in econometrics, biometrics and social sciences. As basis for further research it is considered in the broader framework of generalized linear regression. In the standard case for precise observations the estimators for the structural parameters are obtained when maximizing the likelihood with respect to those. In order to obtain concise estimators in the case of microaggregated data, one could just ignore the nature of the data and use them as-is or employ a maximax-like approach by maximizing the likelihood with respect to the structure parameters and all compatible underlying data situations, suggested herein as first way to proceed. By taking only such an optimistic instantiation of \mathbb{X} the introduced imprecision is not taken seriously. Therefore, in the spirit of Partial Identification ([Manski, 2003](#)) a more cautious estimation approach is introduced by reporting the set of all maximum likelihood estimators based on the elements of \mathbb{X} . All findings are derived in case of a generic microaggregation and thus are suitable for any microaggregation technique. However, it is also demonstrated how the estimation is further improved when considering the specific masking microaggregation technique.

This paper is structured as follows: In section 2 a short overview of microaggregation as anonymization technique and of generalized linear regression is given. In section 3 approaches to obtain concise estimators are presented, while in section 4 the partial identification view is taken resulting in sets-valued estimators. A simulation study in section 5 evaluates the theoretically obtained results. The paper concludes with some remarks and an outlook for further research in section 6.

2. Microaggregation and (Generalized) Linear Regression

2.1 Basics of Microaggregation

The general idea of microaggregation is to replace the individual records by a representative substitute of at least k individuals, in turn microaggregation in this sense then satisfies k -anonymity. Any microaggregation technique may be represented as a two-step process.

Grouping: The individual records of the micro data are partitioned into clusters in a certain way such that records within a cluster are similar and each cluster contains at least $k \geq 3$ records.

Aggregation: Each individual record within a cluster is replaced by the cluster's characteristic value, e.g. mean or median.

The choice on how to define similarity of observations and how to deal with multiple variables allows for a variety of actual techniques. As minimal requirement the previously mentioned concept of k -anonymity (Sweeney, 2002) needs to be fulfilled, guaranteeing that each value of each anonymized variable occurs at least k times. In the simplest case of a single variable, neighboring observations are grouped together and their values are replaced by their group mean. Thus without knowledge about the original data, the membership to the employed groups in the first microaggregation step are deducible as same values in the microaggregated data indicate membership to the same group.

Microaggregation techniques relying on a sorting of variable(s) include *Single-axis Sorting*, where the data are globally ordered according to single (external) sorting variable, and *Individual Ranking*, in which for every variable to be microaggregated an independent Single-Axis sorting is applied according to the ranks of itself. From the perspective of disclosure control it is seen critically that regions for the underlying true records are deducible for both Individual Ranking and Single-axis Sorting in cases when one of the variables to anonymize acts as the sorting variable.² From the analyst's view those regions are exploitable when estimating statistical models as will be seen in the following sections. Other microaggregation techniques do not rely on the concept of an underlying sorting variable, but employ directly a multivariate clustering, e.g. *Maximum Distance to Average Vector* (MDAV) by Domingo-Ferrer and Mateo-Sanz (2002), also providing natural regions.

As desired, the grouping and aggregation introduce imprecision, which means that several different data sets of the underlying true data will lead to the same microaggregated data set. In case of inference on structures in the underlying true data one needs to account for this imprecision. In this paper two conceptually different views are presented: The first, often implicitly found in the literature without embedding into a formal framework, will prove characterizable as a *maximax*-like approach, where the most plausible data situation(s) in the light of the estimation function are used to obtain the parameter estimate(s). The second is the Partial Identification view, where instead of an optimistic estimator, a set-valued one is reported, reflecting the imprecision in the input data more accurately.

2.2 (Generalized) Linear Regression in a Nutshell

The two views are presented for the case of classical linear regression, in order to prepare its extension in the formulation of the generalized linear regression framework, which uses a maximum likelihood approach for parameter estimation instead of the ordinary least squares. The basic settings of the likelihood approach are briefly sketched now³: The aim of (generalized) linear regression is to model the dependency of p independent covariates $\mathbf{X} = (X_1, \dots, X_q, \dots, X_p)$ on a response variable Y , without claiming a causal relation in either direction. For each unit $i = 1, \dots, n$ the response y_i and the covariates \mathbf{x}_i are observed, densely written as $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)^T$ and $\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_i^T, \dots, \mathbf{x}_n^T)^T$. The dependency is modeled by means of a linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} = (1, \mathbf{x}_i)\boldsymbol{\beta}$, which is in the context of generalized regression transformed by a so-called *response* function⁴ h to model the conditional expectation $\mathbb{E}(Y|\mathbf{X})$, where

2. Rosemann et al. (2004) argues that in some practically relevant contexts it is negligible.

3. For further details see Fahrmeir et al. (2013, p. 301ff).

4. h^{-1} is the so-called *link function*.

the form of the conditional distribution is to be specified as a certain one from the exponential family. Common choices lead to well-known models, e.g. a normal distribution for classical linear regression or a Bernoulli distribution for logistic regression; in the general formulation several models are dealt with.

The parameters of interest β are estimated by maximizing the (log-)likelihood induced by the modeling assumptions or equivalently by solving the equation system when setting the *score function* (derivate of the log-likelihood with respect to the parameters of interest) to zero. The score function in classical linear regression for β_q is obtainable to

$$s(\beta_q) = \frac{1}{\phi} \sum_{i=1}^n x_{iq} (y_i - (1, \mathbf{x}_i)\beta), \quad (1)$$

where in the case $q = 0$ the value of x_{i0} is set to one.⁵

For a start only microaggregation of all of the covariates is covered in this work; a situation common in social sciences, when the effect of several covariates, requiring anonymization, on a response variable, for which anonymization is not required, is to be estimated. An extension of this situation to mixtures of aggregation techniques employed for covariates of different scale is subject to further research. It is believed that the findings presented herein may be adapted straightforwardly to the case when only some covariates are microaggregated and the others are left as-is. Considering a microaggregated response variable appears more difficult as conditional independence of the now microaggregated y_i 's in the likelihood does no longer hold. Moreover, it is believed that the actual choice of microaggregation technique has a more severe impact on the quality of the estimation.

2.3 Common Structural Implications of Microaggregation

Before discussing microaggregation in the context of (generalized) linear regression, the structure of microaggregation is recalled. To draw the distinction between the original values and its microaggregated counterparts, as briefly sketched in the introduction, the first are denoted by \mathbf{x} , whereas for the microaggregated values a tilde is placed above: $\tilde{\mathbf{x}}$. A technique suitable for k -anonymity is assumed with k as the fixed group size. For simplicity reasons within this paper it is further assumed that the number of observation n is a multiple of k . Nonetheless, the proposed approach may be straightforwardly generalized to the case when k is only a minimal group size. As a result there are $G = n/k$ distinct groups for each variable under microaggregation. As the microaggregation process involves grouping and averaging, depending on the actual technique chosen even per variable, the notation for generic microaggregation described herein is severely affected. In order to index an observation, a two level indexing in the superscript is utilized: The first place is taken by the membership in a specific group and in the second place the index within this group is given. To denote the grouping according to a specific microaggregated covariate, the group label is further indexed by it. This means, $\tilde{x}_q^{g_r,j}$ corresponds to the microaggregated value of the q^{th} covariate for the j^{th} observation in the g^{th} group, when grouping is induced by the microaggregation of the r^{th} covariate. Please recall that $\tilde{x}_q^{g_q,j}$ has the same value for $j = 1, \dots, k$.

In the above notation the original values are representable by adding a further parameter level:

$$x_q^{g_r,j} = \tilde{x}_q^{g_r,j} + \Delta_q^{g_r,j}, \quad (2)$$

5. The estimation equation for generalized linear regression with canonical link function takes a similar form, only replacing $(1, \mathbf{x}_i)$ by $h(1, \mathbf{x}_i)$

where $\Delta_q^{g_r,j}$ is the corresponding deviation of the individual record from its corresponding group mean. By just looking at the microaggregated values, one is able to deduce the group membership used in the microaggregation's aggregation step. As the mean value within each group is already known, there is a restriction on those deviations per group, which can be formulated by means of the underlying true values associated with a each group

$$\sum_{j=1}^k x_q^{g_q,j} = k \cdot \tilde{x}_q^{g_q,j} \quad \forall q, g , \quad (3)$$

or expressed in terms of the deviations within each group

$$\sum_{j=1}^k \Delta_q^{g_q,j} = 0 \iff \Delta_q^{g_q,k} = - \sum_{j=1}^{k-1} \Delta_q^{g_q,j} \quad \forall q, g . \quad (4)$$

Depending on the employed microaggregation technique, further information on the regions in the data space in which the true underlying values are lying are deducible. Those regions are especially straightforward obtainable in case of Individual Ranking. Also in case of multivariate clustering one could identify such regions.

In the following two different approaches on obtaining meaningful estimators for the regression coefficients are presented.

3. A Nuisance Parameter Optimization Approach

This section discusses the optimistic estimation of the structural parameter when only considering favorable data situation. What is actually deemed as favorable is depending on the view taken.

A naive estimation approach just substitutes the true underlying data x with the microaggregated data \tilde{x} , treating the microaggregated data as independent observations, resulting in the so-called *naive estimator* for β . However, this estimator entirely neglects the nature of the data, it even explicitly rules out the imprecision. In the literature the properties of this naive estimator have been studied in the context of OLS estimation in order to improve it if necessary: [Schmid and Schneeweiss \(2008\)](#) have investigated the effects on the regression coefficient estimates in a general case when Individual Ranking was used, while in [Schmid et al. \(2007\)](#) Single-axis Sorting was considered with the response variable as the sorting variable. Further situations, including also external sorting variables, were discussed in [Schmid \(2007\)](#). [Schmid and Schneeweiss \(2005\)](#) conducted simulation studies, looking at the effect of different microaggregation techniques on the bias of the estimators. They prove that in some situations a bias correction is necessary and actually derive it for either the coefficient estimator or the error variance or even both. They also demonstrate that the corrected estimators are consistent and sometimes even the naive estimator has this property with respect to the true underlying data. However, they rely heavily on the fact that the OLS estimator is obtainable in a closed form, necessarily limiting their investigations to linear models.

3.1 Implications on the Score Function

When looking at the same task from the generalized linear regression perspective, a closed form is also obtainable for the classical linear regression as the estimation equations coincide, however the framework provides means for uniformly modeling different types of models for which neither

an OLS estimation is appropriate nor a closed form expression for the estimates is obtainable. Additionally, the fitting of the proposed corrections into the likelihood approach to obtain consistent estimators are tedious and not straightforwardly applicable; especially when thinking a step ahead when dealing with generalized linear regression. Yet again to obtain a concise estimate only one specific data constellation is to be deemed favorable, which is implicitly assessed by the likelihood.

In the following an extension is elaborated on stating this implicit assumption explicitly: One includes the underlying true values $\mathbf{x} \in \mathbb{X}$, or equivalently their deviation Δ from the respective group mean, as nuisance parameters into the ideal likelihood and estimates them alongside:

$$\hat{\beta} : \ell(\beta, \mathbf{x}; \mathbf{y}) \longrightarrow \max_{\beta, \mathbf{x} \in \mathbb{X}}$$

The restrictions in (3) or (4) introduce a well perceived loss of freedom which in turn results in the fact that within each group there are $k - 1$ nuisance parameters to estimate. For simplicity reason the restrictions as on the right in (4) are employed.

One should note that in the case of a classical linear model the resulting likelihood function to maximize is already a polynomial of grade 4 in the parameters, while taking the same approach in the logistic regression setting the function is not even polynomial.

As seen in (1) the score function consists of all observations' contributions. This overall summation may be re-ordered as one desires, and additionally even separately for the different score function parts; yet this comes at the cost that the score function is no longer directly expressible in the straightforward matrix notation as in (1). The general trick when looking at the score function is to exploit the grouping structure: For any $q = 1, \dots, p$ the grouping with respect to the q^{th} covariate is employed when considering the score function part of β_q .

In the following the contribution $s(\beta_q)_g$ of such a group g to the score function with respect to β_q is given by including (4) into the ideal likelihood and then taking the respective derivatives:

$$\begin{aligned} s(\beta_q)_g &= \frac{1}{\phi} \sum_{j=1}^k \left(\tilde{x}_q^{g_q,j} \left(y^{g_q,j} - \left(\beta_0 + \beta_q \tilde{x}_q^{g_q,j} + \sum_{\substack{r=1 \\ r \neq q}}^p \beta_r (\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j}) \right) \right) \right) \\ &\quad + \frac{1}{\phi} \sum_{j=1}^{k-1} \Delta_r^{g_q,j} \left[(y^{g_q,j} - y^{g_q,k}) - \beta_q \left(\Delta_q^{g_q,j} + \sum_{l=1}^{k-1} \Delta_q^{g_q,l} \right) \right. \\ &\quad \left. - \sum_{\substack{r=1 \\ r \neq q}}^p \beta_r (\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j} - \tilde{x}_r^{g_q,k} - \Delta_r^{g_q,k}) \right]. \end{aligned} \quad (5)$$

As each deviation is group specific for its respective covariate, the global score function $s(\Delta_q^{g_q,j})$ for the deviation $\Delta_q^{g_q,j}$ takes the following form:

$$\begin{aligned} s(\Delta_q^{g_q,j}) &= \frac{\beta_q}{\phi} \left[(y^{g_q,j} - y^{g_q,k}) - \beta_q \left(\Delta_q^{g_q,j} + \sum_{l=1}^{k-1} \Delta_q^{g_q,l} \right) \right. \\ &\quad \left. - \sum_{\substack{r=1 \\ r \neq q}}^p \beta_r (\tilde{x}_r^{g_q,j} + \Delta_r^{g_q,j} - \tilde{x}_r^{g_q,k} - \Delta_r^{g_q,k}) \right]. \end{aligned} \quad (6)$$

As a necessary condition for the optimum the term in (6) needs to be zero. Under the assumption $\beta_q \neq 0$ it follows that the term in square brackets needs to equal zero. Furthermore, noting that the terms in square brackets are the same in (5) and (6), the score function $s(\beta_q)$ for β_q simplifies to:

$$s(\beta_q) = \frac{1}{\phi} \sum_{g=1}^{n/k} \sum_{j=1}^k \tilde{x}_q^{g_q,j} \left(y^{g_q,j} - \left(\beta_0 + \beta_q \tilde{x}_q^{g_q,j} + \sum_{\substack{r=1 \\ r \neq q}}^p \beta_r \tilde{x}_r^{g_q,j} \right) \right) \\ - \frac{1}{\phi} \sum_{g=1}^{n/k} \sum_{j=1}^k \left(\tilde{x}_q^{g_q,j} \sum_{\substack{r=1 \\ r \neq q}}^p \beta_r \Delta_r^{g_q,j} \right). \quad (7)$$

With Equation (7) it becomes obvious that, in the general case the estimate of β_q will be different to the one obtained by the naive estimator. However, when the grouping is the same for all covariates, the second line in (7) equals zero and the function is identical to the one obtained for the naive estimator. When looking at the respective equations for the intercept β_0 they are same in this approach and for the naive estimation. Nonetheless, as they are dependent on the other coefficients, the resulting estimates will still differ.

3.2 Notes on Consistency

When considering the limit behavior of the estimators in case of infinitely many observations, in particular their consistency, two situations are to be covered. In any case the number of observations prior to microaggregation increases, however the outcome will be dependent on whether the maximal group size is still fixed, i.e. k is independent of n , or if we increase k along with n , such that the ratio k/n is fixed. In the first case, as the number of observations increases and the group size does not change, the value of observations within a group converges to its group mean and therefore the deviations converge to zero. This then implies that the score estimation equation converges to the naive estimator equations, which converge to the ones obtain on the original data. Hence the nuisance parameters vanish in the converging point and the consistency of the one proposed in this section follows from the consistency of the general maximum likelihood estimator on the original data. However, such a property is in practice irrelevant, or even not desirable at all as it means that the privacy preserving effect of microaggregation vanishes in the limit. Hence the second case of a fixed ratio should be considered: In this case the observations will not converge towards the group means and equivalently the deviations will not vanish in the limit. This allows to conclude that the naive estimator is generally not a consistent one, as it neglects the non-vanishing deviations.

3.3 Maximax Optimization

The maximization of the likelihood as presented in the above subsection was formulated as an unconstrained optimization task with respect to β and Δ . Nonetheless, the task may also be formulated as a constrained optimization problem with respect to β and either the true underlying values x or the deviations Δ , with the microaggregation as constraint condition. As it appears more natural, in the following the choice of the restriction is switched to the ones on x , as presented in (3). The previous formulation in terms of the Δ already implements the microaggregation equality constraints into the target function by so-called *elimination*.

Furthermore, there might be additional restrictions on the original values which are specific for certain microaggregation techniques. For instance considering Individual Ranking: Here each variable is ordered and grouped individually and therefore additional bounds of the underlying true values for each variable are inferable as each underlying true value within a group must lie between the respective means of the neighboring groups. Also for MDAV, based on the Euclidean distance, one can add the heuristic restrictions that the true values are within a ball around each group mean with a radius of the minimal distance to any other group mean. By taking this optimization view, other – even external – constraints are implementable to restrict the task further.

In the here presented case the target function is still polynomial, while for generalized linear regression this does not necessarily hold. Therefore algorithms which can deal with such optimization tasks are required. As there already exist powerful algorithms to solve the maximization task in a generalized linear regression setting, instead of the direct maximization of the (log-)likelihood, the popular hill climbing strategy is employed. It consists of two major steps, which are to be repeated until a specified stopping criterion is reached:

1. Maximization of the likelihood for given $\hat{\beta}^{(v-1)}$ estimates, such that all constraints on \boldsymbol{x} are satisfied (not only microaggregation ones):

$$\hat{\boldsymbol{x}}^{(v)} = \arg \max_{\boldsymbol{x}} \ell(\boldsymbol{x}; \hat{\beta}^{(v-1)}, \boldsymbol{y})$$

2. Maximization of the likelihood for given $\hat{\boldsymbol{x}}^{(v)}$ estimates:

$$\hat{\beta}^{(v)} = \arg \max_{\beta} \ell(\beta; \hat{\boldsymbol{x}}^{(v)}, \boldsymbol{y}).$$

$\hat{\boldsymbol{x}}^{(v)}$ and $\hat{\beta}^{(v)}$ denote the obtained estimates for \boldsymbol{x} and β in the v -th iteration step. By this approach the complex optimization task is split into two different sub-tasks, which are in themselves easier to solve. The first is in general a non-convex optimization problem with linear constraints, while the second step is a simple estimation of a generalized linear regression for which standard software may be employed. One should note that the number of parameters to estimate is in general greater than the number of available observations which might result in more than a single solution. Furthermore, extensive care needs to be taken for choosing the initial values $\hat{\beta}^{(0)}$: if they are set poorly the algorithm might converge only to a local maximum instead of the global one or might not converge at all, especially without the region constraints. Their introduction does make a difference to avoid some local maxima entirely, as it became visible in the simulation study in section 5.

4. A Partial Identification View

The previous sections rely implicitly on the assumption that the functions presented actually deserve their name as score function. As the number of nuisance parameters increases with the sample size, the situation is comparable to an example in [Neyman and Scott \(1948\)](#), therefore violating standard regularity conditions of maximum likelihood theory. They derive that in their situation the obtained estimators may not be consistent.

Instead of trying to correct the above functions and derive valid score functions, an ideological break is taken by looking at the task from a different angle, stressing the points of interest for generalized linear regression. The structural parameters β are of primary interest, while the underlying

values of the covariates are of minor interest or do not matter at all. In the spirit of partial identification only the available information is to be exploited, for instance the region constraints, while questionable assumptions are dropped, e.g. taking only the most favorable covariate constellation(s) into account in the model estimation. The partial identification approach in general leads to a set of estimates compatible with the available data. A natural approach is to calculate the so-called *collection regions*⁶, or an outer approximation of it in form of a hypercube by calculating bounds on the coefficient estimates component wise. Those obtained may still be informative enough in practice. The collection region collects all such coefficients $\hat{\beta}$ which are obtained as maximum likelihood estimates, or equivalently as zeros of the score function $s(\beta; \mathbf{x}_0)$, for at least one feasible $\mathbf{x} \in \mathbb{X}$:

$$\hat{\mathcal{B}} := \{\hat{\beta} \mid \exists \mathbf{x}_0 \in \mathbb{X} : \ell(\hat{\beta}; \mathbf{x}_0, \mathbf{y}) \geq \ell(\beta; \mathbf{x}_0, \mathbf{y}), \forall \beta \in \mathbb{R}^{p+1}\} \stackrel{7}{=} \{\hat{\beta} \mid \exists \mathbf{x}_0 \in \mathbb{X} : s(\hat{\beta}; \mathbf{x}_0) = 0\}.$$

With this approach only the actually present information of the covariates is employed in the estimation of coefficients, it is guaranteed that the estimator of the unknown underlying micro data prior to microaggregation is contained within the resulting set, as well as the naive estimator. Furthermore, additional information on the covariates like marginal distributions are also includable.

When actually estimating the collection region $\hat{\mathcal{B}}$ or an appropriate approximation, an optimization perspective on the tasks proves fruitful once again. In order to obtain the component wise lower and upper bounds on $\hat{\beta}$, the target function to minimize or maximize takes then a rather simple form:

$$\hat{\beta}_q \longrightarrow \min / \max .$$

Additionally to the constraints introduced by the microaggregation, namely the mean and region constraints, the score function constraint needs to be taken into account:

$$s_r(\hat{\beta}; \mathbf{x}) = 0 \quad \forall r \in \{0, \dots, p\},$$

i.e. each score function part evaluated at the coefficient vector $\hat{\beta}$ and a feasible \mathbf{x} .

Please note that in contrast to the previous section, the \mathbf{x} are only indirectly subject to the optimization. They are allowed to vary freely within their respective bounds, as long as the summation restrictions induced by the microaggregation and any other constraints on them are satisfied. There is no further plausibility assessment on their actual values, as it was employed when optimizing them in the light of the log-likelihood concomitantly or when constructing corrected estimators.

As in general the constraint on the score function has a not negligible complexity, especially as it is not linear in the parameters, the equality constraint may be better incorporated into the target function in terms of a penalty. This leads to the following target function:

$$\hat{\beta}_q \pm \sum_{r=0}^p \lambda_r (s_r(\hat{\beta}; \mathbf{x}))^2 \longrightarrow \min / \max ,$$

where the sign before the sum is chosen appropriately⁸ and λ_r are the so-called *penalty parameters*. If λ_r increases each deviation of the evaluated score function from zero is penalized to a greater extent. Therefore by sufficiently large enough λ_r the deviation is numerically forced onto zero.

Another benefit of those views is the ability to check for any given vector β^* if it is included in the feasible region $\hat{\mathcal{B}}$ of regression coefficients, which is an optimization task in \mathbf{x} only.

6. cf. [Schollmeyer and Augustin \(2015, sec. 3.2\)](#) for an overview of other types of identification regions.

7. This equivalence holds for generalized linear regression as the log-likelihood is concave; cf. [Wedderburn \(1976\)](#).

8. The addition is chosen when minimizing and subtraction when maximizing the component.

<i>n</i>	Average RMSE			
	Truth	IR	SaS	MDAV
100	0.453	0.667 (0.459)	0.797 (0.797)	0.461 (0.462)
250	0.268	0.352 (0.269)	0.469 (0.469)	0.270 (0.270)

Table 1: Average root mean squared error (RMSE) for the maximax-like approach in different microaggregation settings; values in parentheses are the respective value for the naive estimator

5. Simulation Results

Besides the theoretical investigations, a simulation study is conducted with the aim to visualize the effects of microaggregation and display the adequacy of the proposed methods.⁹ The simulation was carried out with the statistical software *R* ([R Core Team, 2016](#))¹⁰.

The classical linear model in the generalized linear regression setting is considered, for simplicity reasons with two independent covariates. Exemplary, the microaggregation techniques of Single-axis Sorting (SaS), Individual Ranking (IR) and MDAV are employed onto the setting of two independent covariates; the dependent variable is left un-aggregated. The regression coefficients are estimated by means of the hill climbing algorithm with all available constraints in the sense of section 3 and for the component wise bounds by taking the partial identification view as presented in section 4. The covariates are each drawn from a uniform distribution on [0, 10], while the response variable is obtained by inducing a dependency structure based on $\beta = (1, 0.5, 1.9)^T$ and adding white noise with variance $\sigma^2 = 4$. To study the behavior when more observations are available, the number of observations varied between $n = 100$ and 250, while the aggregation size was kept fixed with $k = 5$. Each setting was repeated 1000 times.

Under consideration were the different proposed methods: optimization with only microaggregation equality constraints, optimization with microaggregation equality constraints and region inequality constraints and finally component wise coefficient min/max estimation of the collection region. However, in a preliminary run, it turned out that the first method, which ignores de-facto available information, is also numerically unstable as the results are very highly dependent on the choice of the initial value, which is a well known difficulty when applying a hill climbing strategy.¹¹ However as soon as the additional region constraints entered the task, this was no longer the case and the obtained solution was a reliable one. Therefore in the here presented simulation the optimization with just the microaggregation equality constraints is left out.

In Table 1 some results of the simulation are summarized: For the concise estimates their average root mean squared error (RMSE) is reported. As can be seen for the average RMSE, in the case of Individual Ranking the maximax-like approach performs poorer in comparison to the naive estimator. For Single-axis Sorting the estimator coincide, as theoretically shown in section 3, which

9. The code and supplementary files are available on request.

10. The employed optimizer *SLSQP* ([Kraft, 1994](#)) is provided by the R package *NLopt* ([Johnson, 2014](#)).

11. It was found that most of the times in cases when the initial values for $\beta^{(0)}$ were considerably far away from the true ones, a solution was returned, which was easily improvable.

practically also holds for the MDAV approach¹². Furthermore, with higher n the average RMSE is smaller, reflecting the considerations in section 3.2. When comparing the average RMSE obtained on the aggregated data to the one on the original, one finds that Single-axis Sorting produced the highest discrepancy, while the others yield quite comparable results. This is mainly due to the fact that Single-axis Sorting destroys any multivariate structure within the data. In general it is found that the maximax-like approach is too optimistic resulting in poorer performance, and that the naive estimator should be preferred if the user insists on a guaranteed precise solution. The simulation essentially confirms the analytical results of section 3 and the corresponding ones in Schmid and Schneeweiss (2005).

For the outer approximation of the identification region the volume of the coefficient box was calculated. For those the results are also in line with the findings for average RMSE, as in case of Single-Axis Sorting the boxes are considerably large ($n = 100$: 38.031; $n = 250$: 1.530), while for Individual Ranking and MDAV the estimators are numerically point identified. Nonetheless, for Single-Axis Sorting the box volume shrinks as n increases. For any methods in any repetition the naive estimator is indeed included within the box or coincides with the point identified estimators.¹³ The tight boxes are mainly due to the exploitation of the guessable region constraints for Individual Ranking and MDAV.

6. Concluding Remarks

In the light of protecting sensitive data, especially micro data, and availability of suitable methods for protection, the usability of such protected data should no longer be neglected. In this paper a general investigation of the effect of different microaggregation methods on the outcome of regression coefficient estimation in generalized linear regression was started. For generic microaggregation it was demonstrated how the ideal likelihood as the core of generalized linear models can be used as basis in estimation: on the one hand by introducing nuisance parameters concerning the true underlying values and on the other hand by a partial identification view resulting in a set of reachable values. In the present paper for the partial identification approach an outer approximation of the actual set was given, yet it may be further refined. Furthermore, the basic ideas were demonstrated on a classical linear regression, expressed in the framework of generalized linear regression. However, the developed concepts are general and therefore may be employed to other models, for instance the logistic regression.

Acknowledgments

We are grateful for the remarks of three anonymous reviewers, also stimulating further research.

References

- J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002. doi:[10.1109/69.979982](https://doi.org/10.1109/69.979982).

12. The minor difference for MDAV is subject to the heuristic region constraint used.

13. Interestingly the estimator on the original data was not always contained within the calculated box, which might be subject to the employed optimizer. This clearly needs further investigations as theory guarantees the inclusion.

- Eurostat and European Statistical System. The European Statistics Code of Practice, 2011. doi:[10.2785/18474](https://doi.org/10.2785/18474).
- L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer, Berlin, 2013.
- S. G. Johnson. *The NLOpt Nonlinear-optimization Package*, 2014. URL <http://ab-initio.mit.edu/nlopt>.
- D. Kraft. Algorithm 733: TOMP–Fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software*, 20(3):262–281, 1994. doi:[10.1145/192115.192124](https://doi.org/10.1145/192115.192124).
- C. F. Manski. *Partial Identification of Probability Distributions*. Springer, Berlin, 2003.
- J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948. doi:[10.2307/1914288](https://doi.org/10.2307/1914288).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL <https://www.R-project.org>.
- M. Rosemann, D. Vorgrimler, and R. Lenz. Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten. *Allgemeines Statistisches Archiv*, 88(1):73–99, 2004. doi:[10.1007/s101820400160](https://doi.org/10.1007/s101820400160).
- M. Schmid. *Estimation of a Linear Regression with Microaggregated Data*. Verlag Dr. Hut, Munich, 2007.
- M. Schmid and H. Schneeweiss. The effect of microaggregation procedures on the estimation of linear models: A simulation study. Technical Report 443, Institut für Statistik, Sonderforschungsbereich 386, München, 2005. URL <https://epub.ub.uni-muenchen.de/1831>.
- M. Schmid and H. Schneeweiss. Estimation of a linear model in transformed variables under microaggregation by individual ranking. *AStA Advances in Statistical Analysis*, 92(4):359–374, 2008. doi:[10.1007/s10182-008-0087-9](https://doi.org/10.1007/s10182-008-0087-9).
- M. Schmid, H. Schneeweiss, and H. Küchenhoff. Estimation of a linear regression under microaggregation with the response variable as a sorting variable. *Statistica Neerlandica*, 61(4):407–431, 2007. doi:[10.1111/j.1467-9574.2007.00366.x](https://doi.org/10.1111/j.1467-9574.2007.00366.x).
- G. Schollmeyer and T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning*, 56:224–248, 2015. doi:[10.1016/j.ijar.2014.07.003](https://doi.org/10.1016/j.ijar.2014.07.003).
- L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002. doi:[10.1142/S0218488502001648](https://doi.org/10.1142/S0218488502001648).
- R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):22–32, 1976. doi:[10.2307/2335080](https://doi.org/10.2307/2335080).
- L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Springer, New York, 2001.

Maximum Likelihood with Coarse Data based on Robust Optimisation

Romain Guillaume

IRIT, University of Toulouse, Toulouse (France)

ROMAIN.GUILLAUME@IRIT.FR

Inés Couso

Universidad de Oviedo, Gijon (Spain)

COUSO@UNIOVI.ES

Didier Dubois

IRIT, CNRS and University of Toulouse, Toulouse (France)

DUBOIS@IRIT.FR

Abstract

This paper deals with the problem of probability estimation in the context of coarse data. Probabilities are estimated using the maximum likelihood principle. 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. As a consequence, the likelihood function of the ill-observed sample is set-valued. In this paper, we apply a robust optimization method to find a safe plausible estimate of the probabilities of elementary events on finite state spaces. More precisely we use a maximin criterion on the imprecise likelihood function. We show that there is a close connection between the robust maximum likelihood strategy and the maximization of entropy among empirical distributions compatible with the incomplete data. A mathematical model in terms of maximal flow on graphs, based on duality theory, is proposed. It results in a linear objective function and convex constraints. This result is somewhat surprising since maximum entropy problems are known to be complex due to the maximization of a concave function on a convex set.

Keywords: maximum likelihood; incomplete information; robust optimization; entropy.

1. Introduction

Interval observations, and more generally, set-valued ones, do not always reflect the same phenomenon (Couso and Dubois, 2014). Sets, e.g. intervals, may either represent exact observations of items taking the form of sets (for instance, the daily min-max temperature ranges across one year), or, on the contrary, imprecise observations of precise quantities. In the later case, we speak of coarse data (Heitjan and Rubin, 1991). In the first situation, set data are a special kind of functional data where observations lie in a space of characteristic functions equipped with a suitable metric structure, enabling precise statistical parameters to be derived, e.g., (González-Rodríguez et al., 2012). In this paper we are interested in the statistical analysis of data when observations are imprecise, or coarse, more specifically, when we only know that the precise values of observations are restricted by sets of possible outcomes of a random variable of interest. In this kind of representation, sets model epistemic states (or states of knowledge) in the sense that no value outside the set is possibly the true observed value (unreachable for the observer). Under the epistemic approach, the expected value and the variance of a collection of intervals are themselves intervals (Kruse and Meyer, 2012).

This paper addresses the problem of statistical inference in the presence of epistemic set-valued data using the maximum likelihood principle. Under imprecise observations, the likelihood function itself becomes imprecisely appraised too and is thus set-valued. There are several possible ways of defining a scalar likelihood function in this situation (Couso and Dubois, 2016a). In this paper we adopt a robust optimisation point of view and maximize the lower bound of the imprecise likelihood

function, with a view to obtain a probability density that accounts for the potential variability of the random variable observed via sets of possible outcomes. We give an interpretation of the robust solution in terms of entropy maximization, and propose algorithms for computing robust maximum likelihood distributions in the discrete (finite) case of coarse nominal data, based on a maximal flow approach.

The paper is organized as follows. In Section 2, we recall a general framework for maximum likelihood estimation under coarse data due to [Couso and Dubois \(2016a\)](#), and situate our robust optimization strategy in this framework. It consists in maximizing the minimal likelihood function in agreement with the coarse data. We discuss the difference between our approach and the optimistic maximax strategy. Section 3 proposes a methodology for solving the robust optimisation problem in the discrete case, based on max-flow formulation and duality. Section 4 shows that the optimal estimate corresponds to maximizing entropy among empirical distributions of all possible samples in agreement with the coarse data. In section 5, we propose a new method for solving the maximin likelihood estimation problem and discuss an illustrative example.

2. General framework

A likelihood function is proportional to the probability of obtaining the observed data given a hypothesis, according to a probability model. Observed data are considered as outcomes, i.e., elementary events. If this point of view is accepted, what becomes of the likelihood function under coarse observations? If coarse observations are considered as results, we can construct the likelihood function for set-valued outcomes, and compute a random set. However, coarse observations being set-valued, they do not directly inform us about the underlying random variable. In order to properly exploit such incomplete information, we must first decide what to model ([Couso and Dubois, 2016a](#)): (1) the random phenomenon *despite* the deficiencies its measurement process; or (2) the random phenomenon *as known via* its measurement process.

In the first case, authors have proposed several ways of restoring a distribution for the underlying random phenomenon. The most traditional approach constructs a virtual sample of the ill-observed random variable in agreement with the imprecise data, by minimizing divergence from a parametric model, and maximizing likelihood wrt this sample, so as to update this parametric model. This is often carried out by means of EM algorithm ([Dempster et al., 1977](#)). The problem with this approach is that there may be several optimal distributions, hence virtual samples, especially when the connection between the hidden random variable and its observation process is loose ([Couso and Dubois, 2016b](#)). The result of an iterative algorithm such as EM may depend on the initial parameter value. Moreover the EM algorithm assumes that observed coarse data form a partition of the domain of the random variable of interest (see the introduction of ([Dempster et al., 1977](#))).

In this paper, we take the other point of view, the one of ill-observed outcomes. Then, there are as many likelihood functions as precise datasets in agreement with the coarse observations, and it is not clear which one to maximize. We pursue our study of a methodology based on a robust maximin optimisation approach applied to a set-valued likelihood [Guillaume and Dubois \(2015\)](#). Note that here, we do not consider the issue of modelling imprecision due to too small a number of precise observations (see for instance ([Serrurier and Prade, 2013](#))). Let us recall the formal setting for statistics with coarse data proposed by [Couso and Dubois \(2016a\)](#), then we study the meaning of the solution to the maximin approach, and finally propose an algorithm to solve it in the case of nominal outcome sets.

2.1 The random phenomenon and its measurement process

Let a random variable $X : \Omega \rightarrow \mathcal{X}$ represent the outcome of a certain random experiment. For the sake of simplicity, let us assume that $\mathcal{X} = \{a_1, \dots, a_m\}$ is finite. Suppose that there is a measurement tool driven by a random variable Y that provides an incomplete report of observations. Namely, there is a set-valued random variable $Y : \Omega \rightarrow \wp(\mathcal{X})$ that models the reports of a measurement device, where $\wp(\mathcal{X})$ is the set of subsets of \mathcal{X} . Y is thus a multimapping which represents our (imprecise) perception of X , in the sense that we assume that X is a selection of Y , i.e. $X(\omega) \in Y(\omega)$, $\forall \omega \in \Omega$, in agreement with the setting of imprecise probabilities proposed by Dempster (1967). X is often called the latent variable. Let $\mathcal{Y} = \{A_1, \dots, A_r\}$ denote the set of possible set-valued outcomes of Y , where $A_j \in \wp(\mathcal{X})$.

The information about the joint distribution of the random vector (X, Y) modeling the random variable X and its measurement process can be represented by a joint probability on $\mathcal{X} \times \mathcal{Y}$ defined by means of $m \times r$ coefficients $p_{ij} = P(X = a_i, Y = A_j)$. Some knowledge may be available about this probability matrix. For instance, in the case when \mathcal{Y} is a partition of \mathcal{X} , we have

$$p_{ij} = P(Y = A_j | X = a_i) \cdot P(X = a_i) = \begin{cases} P(X = a_i) & \text{if } a_i \in A_j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Sometimes assumptions are made about the conditional probability $P(Y = A_j | X = a_i)$ describing the imprecise measurement process, like the superset assumption (Hüllermeier and Cheng, 2015) considering that $\mathcal{Y} = \wp(\mathcal{X})$ and stating that this probability is a constant c_i over all sets containing a_i , i.e. $c_i = 1/2^{m-1}$ that does not depend on i . Another less restrictive assumption is called “coarse at random” (CAR) whereby $P(Y = A_j | X = a_i)$ does not depend on the value $a_i \in A_j$ (Heitjan and Rubin, 1991). In this paper, we shall just ignore the measurement process.

2.2 Different likelihood functions

The respective marginals on \mathcal{X} and \mathcal{Y} are denoted as follows:

- $p_{\cdot j} = \sum_{k=1}^m p_{kj}$ denotes the mass of $Y = A_j$, $j = 1, \dots, r$, and
- $p_{k \cdot} = \sum_{j=1}^r p_{kj}$ denotes the mass of $X = a_k$, $k = 1, \dots, m$.

Now, let us assume that the above joint distribution is characterized by means of a (vector of) parameter(s) $\theta \in \Theta$ that determines a joint distribution on $\mathcal{X} \times \mathcal{Y}$.

For a sequence of N iid copies of $Z = (X, Y)$, $\mathbf{Z} = ((X_1, Y_1), \dots, (X_N, Y_N))$, we denote by $\mathbf{z} = ((x_1, y_1), \dots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ a specific sample of the vector (X, Y) . Thus, $\mathbf{y} = (y_1, \dots, y_N)$ will denote the observed sample (an observation of the set-valued vector $\mathbf{Y} = (Y_1, \dots, Y_N)$), and $\mathbf{x} = (x_1, \dots, x_N)$ will denote an arbitrary artificial sample from \mathcal{X} for the unobservable latent variable X , that we shall vary in \mathcal{X}^N . The samples \mathbf{x} are chosen such that the number of repetitions n_{kj} of each pair $(a_k, A_j) \in \mathcal{X} \times \mathcal{Y}$ in the sample are in agreement with the number q_j of actual observations A_j . We denote by $\mathcal{X}^{\mathbf{y}}$ (resp. $\mathcal{Z}^{\mathbf{y}}$), the set of samples \mathbf{x} (resp. complete joint samples \mathbf{z}) respecting this condition. We assume that the measurements are reliable in the sense that, observing $y = G \subseteq \mathcal{X}$, we can be sure that the actual outcome $X = x \in G$. If we let n_k be the number of appearances of a_k in the virtual sample \mathbf{x} , we have that any $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$

satisfies:

$$\begin{cases} \sum_{k=1,\dots,r} n_k = \sum_{j=1,\dots,r} q_j = N \\ n_k = \sum_{j=1}^r n_{kj}, \forall k = 1, \dots, m \\ q_j = \sum_{k=1}^m n_{kj}, \forall j = 1, \dots, r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \quad (2)$$

For a complete sample \mathbf{z} to be compatible with the observation \mathbf{y} , we have that any $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ satisfies:

$$\begin{cases} \sum_{k=1,\dots,r} \sum_{j=1,\dots,r} n_{kj} = N \\ q_j = \sum_{k=1}^m n_{kj}, \forall j = 1, \dots, r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \quad (3)$$

As pointed out by [Couso and Dubois \(2016a\)](#), we may consider three different log-likelihood functions depending on whether we refer to

1. *the observed sample in Y :* $L^{\mathbf{y}}(\theta) = \log \prod_{i=1}^N p(y_i; \theta) = \sum_{j=1}^r q_j \log p_j^\theta$.
2. *the hidden sample in X :* $L^{\mathbf{x}}(\theta) = \log \prod_{i=1}^N p(x_i; \theta) = \sum_{k=1}^m n_k \log p_k^\theta$.
3. *the complete sample in $X \times Y$:* $L^{\mathbf{z}}(\theta) = \log \prod_{i=1}^N p(z_i; \theta) = \sum_{k=1}^m \sum_{j=1}^r n_{kj} \log p_{kj}^\theta$

The two last ones are ill-known. The choice of one likelihood function vs. another depends upon what problem we are interested to solve. Maximizing $L^{\mathbf{y}}(\theta)$ means that we are interested in modeling our perception of the random variable only. It is the standard maximum likelihood estimation (MLE) that computes the argument of the maximum of $L^{\mathbf{y}}$ considered as a mapping defined on Θ , i.e.: $\hat{\theta} = \arg \max_{\theta \in \Theta} L^{\mathbf{y}}(\theta) = \arg \max_{\theta \in \Theta} \prod_{j=1}^r (p_j^\theta)^{q_j}$. The result is a mass assignment on $2^{\mathcal{X}}$ if there is no constraint relating the distributions of X and Y via the parameter θ . It computes a belief function on \mathcal{X} with focal sets in \mathcal{Y} .

The EM algorithm ([Dempster et al., 1977](#)) is an iterative technique maximizing this likelihood function via the use of the latent variable X and a virtual sample for X in order to achieve a local maximum of $L^{\mathbf{y}}$. This procedure makes sense if the observed sample \mathbf{y} provides enough information on X (via suitable assumptions on the model parameters θ) to guarantee the convergence of the iterative procedure to a solution that minimizes the distance between the empirical distribution of the final virtual sample in agreement with \mathbf{y} , and the resulting parametric distribution on X ([Couso and Dubois, 2016a](#)).

Maximizing $L^{\mathbf{z}}(\theta)$ enables to take into account the knowledge we may have about the measurement process, and allows for a fine-grained modeling. On the contrary, maximizing $L^{\mathbf{x}}(\theta)$ means that we completely give up modeling the measurement process and try to extract information about X based on information about Y , assuming complete ignorance about the measurement process. The difficulty with $L^{\mathbf{z}}(\theta)$ and $L^{\mathbf{x}}(\theta)$ is that they are ill-known, namely we must consider for all values of θ , the sets $\mathbb{L}^{\mathbf{z}}(\theta) = \{L^{\mathbf{z}}(\theta) : \mathbf{z} \in \mathcal{Z}^{\mathbf{y}}\}$ and $\mathbb{L}^{\mathbf{x}}(\theta) = \{L^{\mathbf{x}}(\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$, respectively. In the paper we shall deal with $L^{\mathbf{x}}(\theta)$, i.e., try to find results independently of the measurement process.

Applying the maximum likelihood principle when the likelihood function is ill-known requires the choice of a representative likelihood function from $\mathbb{L}^{\mathbf{x}}(\theta)$. Obvious natural choices are $\bar{L}(\theta) = \max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} \mathbb{L}^{\mathbf{x}}(\theta)$ and $\underline{L}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} \mathbb{L}^{\mathbf{x}}(\theta)$.

On this basis, there are two strategies of likelihood maximization, based on a sequence of imprecise observations $\mathbf{y} = (y_1, \dots, y_N) \in \mathcal{Y}^N$:

1. *The maximax strategy* (Hüllermeier, 2014): it aims at finding the pair $(\mathbf{x}^*, \theta^*) \in \mathcal{X}^\mathbf{y} \times \Omega$ that maximizes $L^\mathbf{x}(\theta)$. In other words, compute $(\mathbf{x}^*, \theta^*) = \arg \max_{\mathbf{x} \in \mathcal{X}^\mathbf{y}, \theta \in \Theta} L^\mathbf{x}(\theta)$.
2. *The maximin strategy* (Guillaume and Dubois, 2015): it aims at finding $\theta_* \in \Theta$ that maximizes $\underline{L}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^\mathbf{y}} L^\mathbf{x}(\theta)$. It is a robust optimization approach that takes a pessimistic view on likelihood maximization.

The maximax strategy tries to disambiguate the coarse data by choosing a virtual sample \mathbf{x} that makes the parametric model maximally in agreement with the data. In the case of the maximin strategy, it is pessimistic in the sense that it tends to select distributions with large variability as we shall show.

3. The robust approach to discrete probability estimation with coarse data

In this section, we try to estimate the probability $P(X = a_k), k = 1, \dots, m$ when the reports of N observations of X are imperfect and take the form of an imprecise sample \mathbf{y} containing q_j copies of subsets $A_j \in \mathcal{Y}$ of values of X , for $j = 1 \dots r$. To determine the parameter we adopt the usual approach based on likelihood maximization, which in the case of precise observations takes the form:

$$\text{maximize} : L^\mathbf{x}(\theta) = \log p(\mathbf{x}; \theta) = \sum_{i=1}^m n_k \log p(X = a_k | \theta) \quad (4)$$

Note that in our context the numbers $n_k, k = 1, \dots, m$ are ill-known, because we did not fully observe the outcomes. All we know is that $\mathbf{x} \in \mathcal{X}^\mathbf{y}$. Hence, the vector $\mathbf{n} = (n_k)_{k=1, \dots, m} \in \mathcal{N}^\mathbf{y}$, where $\mathcal{N}^\mathbf{y}$ is the set of possible statistics in agreement with the imprecise observations \mathbf{y} , that is, respecting equation (2). We call an assignment $\mathbf{n} \in \mathcal{N}^\mathbf{y}$ a virtual sample. To manage the uncertainty on \mathbf{n} we use the pessimistic maxmin strategy. Namely, we will maximize the minimal value of likelihood function for the hidden sample \mathbf{x} :

$$\max_{\theta} \min_{\mathbf{n} \in \mathcal{N}^\mathbf{y}} \sum_{i=1}^m n_k \log p(X = a_k | \theta) \quad (5)$$

Note that by using the likelihood based on the hidden sample \mathbf{x} , we make no assumption on the measurement process that from observing $X \in \mathcal{X}$, yields a subset of \mathcal{X} . We only know that if $Y = A_j$ is observed, some $x_k \in A_j$ has been produced. One can see that Equation (5) is then equivalent to the more explicit mathematical formulation:

$$\begin{aligned} & \max_{\mathbf{p}} \min_{\mathbf{n}} \sum_{k=1, \dots, m} n_k \cdot \log p_k \\ & \text{s.t.} \\ & (a) \quad \sum_{k=1, \dots, m} n_k = \sum_{j=1, \dots, r} q_j = N \\ & (b) \quad n_k = \sum_{j: (j, k) \in \mathcal{E}^\mathbf{y}} n_{j, k}, \quad \forall k = 1, \dots, m \\ & (c) \quad q_j = \sum_{k: (j, k) \in \mathcal{E}^\mathbf{y}} n_{j, k}, \quad \forall j = 1, \dots, r \\ & (d) \quad \sum_{k=1, \dots, m} p_k = 1 \\ & (e) \quad n_k, n_{j, k} \in \mathbb{N}^+, p_k > 0, \quad \forall k = 1, \dots, m, \end{aligned} \quad (6)$$

where

- the value $q_j, j = 1, \dots, r$ is the number of actual observations of Y of the form A_j ,
- the decision variables $(p_k)_{k=1,\dots,m}$ stand for the ill-known model probabilities $p(X = a_k|\theta)$, $k = 1, \dots, m$ on \mathcal{X} ; in the loosest situation, there is no constraint relating the p_k via an explicit parameter θ .
- $(n_k)_{k=1,\dots,m}$ are the ill-known numbers of occurrences of values $a_k, k = 1, \dots, m$ of X ,
- $\mathbb{E}^y = \{(j, k) : a_k \in A_j, \forall k = 1, \dots, m\}$. Indeed, since coarse observations are supposed to be faithful, $n_{j,k} = 0$ if $a_k \notin A_j$.

The constraints (6(a)) guarantee that all observations are taken into account. The constraints (6(b)) and (6(c)) guarantee that the number of virtual samples $\mathbf{n} \in \mathcal{N}^y$ is in agreement with observations. Equation (6(d)) is a normalisation constraint. Moreover we add constraints (6(e)) since the observation is integer and $\log(x)$ is not defined for $x = 0$. Constraint (6(d)) expresses the reliability of imprecise observations. In particular, the set \mathcal{N} of feasible statistics \mathbf{n} for X is thus defined by the set of m -tuples of integers verifying constraints (6(a, b, c)), and such that $(j, k) \in \mathbb{E}^y$.

Remark 1 Note that the maximal size of \mathcal{Y} is a linear function of the number of observations and not exponential of the form $2^{|\mathcal{X}|}$. More precisely, it is $\min(2^{|\mathcal{X}|}, \sum_{k=1}^r q_r)$. In fact, in the case where $2^{|\mathcal{X}|} > \sum_{k=1}^r q_r$, observations could be different from one another, i.e., $q_k = 1, k = 1, \dots, r$.

4. The maxmin strategy maximizes entropy

Problems of the form (6) are well-known in the framework of game theory. The major issue is to find conditions under which the expression $\max_u \min_v f(u, v)$ is equal to $\min_u \max_v f(u, v)$ for (u, v) lying in a compact convex subset of \mathbb{R}^2 . In the general case, the following inequality always holds:

$$\max_u \min_v f(u, v) \leq \min_v \max_u f(u, v).$$

When there is a saddle point, that is a pair (u^*, v^*) such that

$$f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*), \forall u, v,$$

then the equality holds, and corresponds to the notion of Nash equilibrium in game theory. This is the case when the function f is convex-concave and continuous, that is when f is convex in u and concave in v (Von Neumann, 1928; Sion, 1958; Komiya, 1988).

Problem (6) can be written as $\max_\theta \min_{\mathbf{n} \in \mathcal{N}^y} f(\mathbf{n}, \theta)$, where function f has the form: $f(\mathbf{n}, \theta) = \sum_{i=1}^m n_k \log p(X = a_k|\theta)$. Provisionally, let us drop the assumption that \mathbf{n} is a vector of integers, and assume it is a set of reals obeying (6(a, f)). It is easy to see that $f(\mathbf{n}, \theta)$ is increasing and linear in \mathbf{n} , while is concave and continuous with respect to $\theta = (p_k)_{k=1,\dots,m}$. The optimisation domain is clearly a compact and convex set. So, f is convex concave, and the above known result then applies:

Proposition 1 Assuming \mathbf{n} is not restricted to being integer-valued, the equality $\max_p \min_{\mathbf{n}} \sum_{k=1,\dots,m} n_k \cdot \log p_k = \min_{\mathbf{n}} \max_p \sum_{k=1,\dots,m} n_k \cdot \log p_k$ holds.

The solution to the minmax problem is easier to find. Indeed the problem $\max_{\mathbf{p}} \sum_{k=1,\dots,m} n_k \cdot \log p_k$ is a standard maximum likelihood problem with a fixed vector \mathbf{n} . The optimal solution is given by $p_k = n_k/N, k = 1, \dots, m$. Now we are led to find \mathbf{n} that maximizes an expression of the form $-n_k \cdot \log(n_k/N)$ which, divided by N , is clearly the entropy of $(n_1/N, \dots, n_m/N)$. We can thus conclude that:

Corollary 1 *The optimal solution to the maxmin likelihood problem (6) is the solution with maximal entropy, namely the solution to: $\max_{\mathbf{n}} -\sum_{k=1,\dots,m} \frac{n_k}{N} \cdot \log \frac{n_k}{N}$ under conditions (6(a, b, c)), and $n_k \in \mathbb{R}^+$, i.e. \mathbf{n} in the convex hull of \mathcal{N}^y .*

In fact, it is easy to see that the observed data $(q_j, A_j), j = 1 \dots r$ defines a belief function Bel with mass function $\mu(A_j) = \frac{q_j}{N}, j = 1 \dots r$, and that the convex set of probabilities $\mathcal{P} = \{P : P \geq Bel\}$ is nothing but the credal set defined by the set of probability assignments $\mathbf{p} = (n_1/N, \dots, n_m/N)$ (Zaffalon, 2002), where $\sum_{k=1,\dots,m} n_k = \sum_{j=1,\dots,r} q_j = N$ plus conditions (6(b, c)) and $n_i \geq 0$ are reals. So the maxmin likelihood problem (6) comes down to a finding the maximum entropy probability in the credal set \mathcal{P} , a problem already addressed in the past by Abellán and Moral (2003).

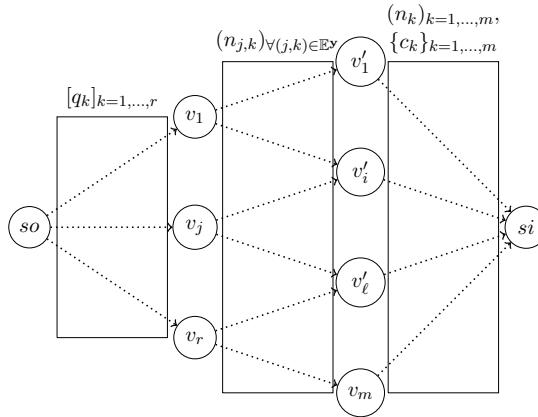


Figure 1: Graph representation of the problem

It remains to be checked whether the optimal solution \mathbf{n}^* is integer or not. To this end, we focus on the problem of minimizing $\sum_{k=1,\dots,m} n_k \cdot \log p_k$ for a given probability distribution \mathbf{p} . The decision variables form the vector \mathbf{n} in the convex hull of \mathcal{N}^y . The problem considered is :

$$\min_{\mathbf{n}} \sum_{k=1,\dots,m} n_k c_k \quad (7)$$

s.t.

- (a) $\sum_{k=1,\dots,m} n_k = \sum_{k=1,\dots,r} q_k = N,$
- (b) $n_k - \sum_{j:(j,k) \in \mathcal{E}^y} n_{j,k} = 0, \quad \forall k = 1, \dots, m$
- (c) $\sum_{k:(j,k) \in \mathcal{E}^y} n_{j,k} = q_j, \quad \forall j = 1, \dots, r$
- (d) $n_k \in \mathbb{N}^+, \quad \forall k = 1, \dots, m$

where $c_k = \log p_k, k = 1, \dots, m$ are constant. The problem (7) can be modeled by a bipartite transportation graph, as done by Zaffalon (2002). The graph is (V, \mathcal{E}) where the vertices V include

a source node so related to r vertices corresponding to elements of \mathcal{Y} , themselves related to m nodes corresponding to the elements of \mathcal{X} , and finally a sink node si . Edges in \mathcal{E} are of the form (so, v_j) , (v_j, v'_k) if $(j, k) \in \mathbb{E}^y$, and (v'_k, si) (see Fig. 1). The values in brackets provide the flow along these edges.

Proposition 2 *The problem (7) is a maximum flow minimum cost problem.*

Proof: The constraint (7(a)) is the equality constraint between the source flow and the sink flow. The constraints (7(b)) and (7(c)) are flow conservation constraints. In our case, the maximum flow is equal to $\sum_{k=1,\dots,r} q_k$. \square

From Proposition 2, we know that this problem has a totally unimodular structure, i.e., it is a linear problem with a totally unimodular constraint matrix. Therefore, the linear program relaxation of the model (7), letting $n_k \in \mathbb{R}^+$ yields an integral solution, which is thus the one of problem (7). So, the maximal entropy solution we have defined above for the maximization of the lower hidden likelihood is indeed of the form $(n_1/N, \dots, n_m/N)$ for integer values of n_k .

Remark 2 *The solution of the maximization of the upper hidden likelihood, that is $\max_p \max_n \sum_{k=1,\dots,m} n_k \cdot \log p_k$ under constraints (7(a-d)), is trivially equivalent to $\max_n \max_p \sum_{k=1,\dots,m} n_k \cdot \log p_k$. It corresponds to minimizing the entropy of the vector $(n_1/N, \dots, n_m/N) \in \mathcal{N}^y$ in the credal set induced by y , i.e. looking for the minimally uncertain frequency tuples compatible with observations, which corresponds to the idea of disambiguation put forward by [Hüllermeier \(2014\)](#).*

The above results shed light on the significance of the maximin and the maximax strategies and are useful to understand when to apply one or the other.

- The maximin strategy makes sense if we know that the process generating the variable X is genuinely non-deterministic, and that the imprecision of the observation may hide some variability (for instance the pace of variability of X is higher than the one of the observation process, so that X may vary during the making of one observation). Consider the case of reporting daily the temperature of the outside air based on a device that records the temperature variation within each day. This information is representative of the “average daily temperature”, which may lead to their modelling as epistemic intervals containing this average value. Then it is reasonable to interpret the coarseness of A_i in terms of underlying variability and to go for a maximal entropy solution to the maximum likelihood problem.
- The maximax strategy makes sense if it is assumed that the underlying phenomenon is deterministic but the observations are noisy and coarse. If we try to learn a best model taken from a class of models and we have some good reason to think that the phenomenon under study can be represented by one of these models, then it is natural to try and choose one of them. In particular, it is clear that if $A = \cap_{j=1} A_j \neq \emptyset$, then the maximax strategy yields any Dirac function on \mathcal{X} such that $P(A) = 1$ (it picks any element in A). For instance, consider a linear regression problem with interval observations, an example from [Hüllermeier \(2014\)](#). If the studied phenomenon is known to be affine, then one may choose the straight line that achieves a best fit with respect to the intervals. Especially any linear model that would be consistent with all interval observations will be preferred. The maximin strategy clearly yields a very different result due to the link with maximal entropy laid bare above.

5. Resolution method and an example

In this section, we propose a mathematical programming approach to solving problem (6), which comes down to optimizing a linear objective function under convex constraints. From the duality theorem, we know that the cost value of an optimal solution of the original (primal) model is equal to the cost value of the optimal solution of its dual. Let α be the dual variable associated to constraint (7(a)), $\beta_k, k = 1, \dots, m$ the dual variables for constraints (7(b)) and $\gamma_k, k = 1, \dots, r$ the dual variables for constraints (7(c)). The dual form of problem (7) is:

$$\begin{aligned} & \max_{\alpha, \beta, \gamma} -(\alpha \sum_{k=1}^r q_k + \sum_{k=1}^r \gamma_k q_k) \\ \text{s.t. } & \alpha + \beta_k \geq -\log(p_k), \quad \forall k = 1, \dots, m \\ & -\beta_j + \gamma_k \geq 0, \quad \forall (j, k) \in \mathbb{E}^y \\ & \alpha, \beta_j, \gamma_k \in \mathbb{R}, \quad \forall j = 1, \dots, r, k = 1, \dots, m \end{aligned} \tag{8}$$

Let us now return to the initial problem (eq.6) where the probability distribution is a decision variable. Its dual problem can be now written as a maximax problem:

$$\begin{aligned} & \max_{\mathbf{p}} \max_{\alpha, \beta, \gamma} -(\alpha \sum_{k=1}^r q_k + \sum_{k=1}^r \gamma_k q_k) \\ \text{s.t. } & (a) \quad \alpha + \beta_k \geq -\log(p_k), \quad \forall k = 1, \dots, m \\ & (b) \quad -\beta_j + \gamma_k \geq 0, \quad \forall (j, k) \in \mathbb{E}^y \\ & (c) \quad \sum_{k=1, \dots, m} p_k = 1 \\ & (d) \quad p_k > 0, \quad \forall k = 1, \dots, m \\ & (e) \quad \alpha, \beta_j, \gamma_k \in \mathbb{R}, \quad \forall j = 1, \dots, r, k = 1, \dots, m \end{aligned} \tag{9}$$

One can reformulate the problem (9) as follows with $\epsilon \rightarrow 0$:

$$\begin{aligned} & \min_{P, \alpha, \beta, \gamma} \alpha \sum_{k=1}^r q_k + \sum_{k=1}^r \gamma_k q_k \\ \text{s.t. } & (a) \quad \alpha + \beta_k \geq -\log(p_k), \quad \forall k = 1, \dots, m \\ & (b) \quad -\beta_j + \gamma_k \geq 0, \quad \forall (j, k) \in \mathbb{E}^y \\ & (c) \quad \sum_{k=1, \dots, m} p_k = 1 \\ & (d) \quad p_k + \epsilon \geq 0, \quad \forall k = 1, \dots, m \\ & (e) \quad p_k, \alpha, \beta_j, \gamma_k \in \mathbb{R}, \quad \forall j = 1, \dots, r, k = 1, \dots, m \end{aligned} \tag{10}$$

The problem (10) has a linear objective function to minimize, m convex constraints 10.(a) plus linear constraints. Hence this problem can be efficiently solved using a nonlinear solver.

Example

We want to estimate the probability that a type of car is present in some parking lot. The custodian provides some characteristics of cars (color and the number of doors) in a data base. For simplicity we consider three colors: red (r), blue(b), grey (g) and two situations for doors: 3 doors (3) and 5 doors (5). There are 6 possible types of cars: $\{r3, r5, b3, b5, g3, g5\}$. The information reported by the custodian can be both the color and the number of doors or only the color or only the number

\mathcal{Y}	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
q	9	167	120	199	164	188
\mathcal{Y}	$\{r3, b3, g3\}$	$\{r5, b5, g5\}$	$\{r3, r5\}$	$\{b3, b5\}$	$\{g3, g5\}$	
q	80	80	18	107	100	

Table 1: Distribution of Coarse Observations

of doors. So, we have $\mathcal{Y} = \{\{r3\}, \{r5\}, \{b3\}, \{b5\}, \{g3\}, \{g5\}, \{r3, b3, g3\}, \{r5, b5, g5\}, \{r3, r5\}, \{b3, b5\}, \{g3, g5\}\}$. Table 1 provides the coarse dataset.

To estimate the maximin probability distribution on \mathcal{X} (noted p^{Mm}) we solve the mathematical formulation given in section 5 using the solver SQP of software Octave.¹ To discuss the result, we compare it with the probability distribution obtained using a maximax approach (noted p^{MM}). The results are given in table 2. Firstly, the maximin allow us to conclude that $\{r3\}$ is the least probable,

\mathcal{X}	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{Mm}(X = a_i) \approx$	0.141	0.171	0.171	0.171	0.173	0.173
$p^{MM}(X = a_i) \approx$	0.007	0.150	0.098	0.313	0.279	0.153

Table 2: Estimations of probability distributions on the latent variable

$\{r5\}$, $\{b3\}$, and $\{b5\}$ have the same probability to be present in this parking. Finally $\{g5\}$ and $\{g3\}$ are the most expected ones in this parking. The uncertainty on data prevents us from differentiating between $\{r5\}$, $\{b3\}$ or $\{b5\}$. In the same way, it is not possible to differentiate the probabilities of a car of types $\{g3\}$ or $\{g5\}$.

Let us compare both approaches on the resulting distributions pictured on Table 2. Both the maximin as the maximax approaches suggest that the cars of type $r3$ have the least probability to appear. But its probability in the maxmin approach is around two times the probability obtained by the maximax approach. In fact, in the maximax approach the observations $\{r3, b3, g3\}$ and $\{r3, r5\}$ are respectively interpreted as $\{g3\}$ and $\{r5\}$. It supposes that when the custodian just writes the characteristic “3 doors” in data base, the car is supposed to be grey. And when the custodian only writes the characteristic “red”, the car has 3 doors. One can see that the probability of $\{r5\}$, $\{b3\}$, and $\{b5\}$ are very different, like probability $\{g5\}$ and $\{g3\}$.

We focus now on the probability of $\{b3\}$, and $\{b5\}$. In the maximin approach they were equal but in the maximax approach the probability $\{b3\}$ is the second less probable while $\{b5\}$ is the most probable type of car. But one can see that around half of observations concerning $\{b3\}$ or $\{b5\}$ are imprecise. It is clear that the maximin approach favors uniform distributions over outcomes while the maximax approach tends to put more weights on some specific cars, namely those which have been already most often observed precisely (such as $\{b5\}$).

Let us swap observations $\{b5\}$ and $\{b3\}$, i.e., suppose there are 120 observations for $\{b5\}$ and 199 observations for $\{b3\}$. The probability distribution of maximin approach does not change since the number of imprecise observations $\{b3, b5\}$ is too high to separate the probabilities of $\{b3\}$ and $\{b5\}$. But the probability distribution of the maximax approach is very sensitive to this exchange (see Table 3). Of course, the probability of $\{b3\}$ becomes higher than that of $\{b5\}$. We point out to that the probability of $\{g3\}$ and $\{g5\}$ changes a lot. In fact, now the observations $\{r3, b3, g3\}$ are

1. <https://www.gnu.org/software/octave>.

interpreted as $\{b3\}$ and not $\{g3\}$ while the observations $\{g3, g5\}$ are interpreted as $\{g3\}$ and not $\{g5\}$.

\mathcal{X}	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{MM}(X = a_i) \approx$	0.008	0.150	0.313	0.097	0.133	0.299

Table 3: Maximax probability distribution with the modified dataset

In this example we show that the maximin approach is cautious compared to the maximax approach. More precisely, a high number of very coarse observations tends to equalize the probabilities of elementary outcomes while the maximax approach tends to select a best outcome consistent to coarse observations and this result can be completely altered by slightly changing the number of observations of each kind, which may lead to very different results.

6. Conclusion

This paper is a contribution to the study of maximum likelihood methods when data are coarse. The most popular approaches often assume some knowledge about the measurement process (as witnessed by the use of the superset of the CAR assumptions). These assumptions are strong and lead to work with the likelihood function of the complete joint sample involving both the observed and the latent variables. In our approach, we ignore the measurement process, and adopt a cautious approach involving robust optimisation and graph-theoretic methods. This approach, introduced previously ([Guillaume and Dubois, 2015](#)) for continuous parametric distributions and interval data, is here studied for finite sets of outcomes. The close connections between maximax and maximin strategies with entropy optimization shed light on the significance of each approach: the intuitive character of the resulting distribution depends on whether the observed phenomenon is genuinely random, or if it is deterministic, with a known class of models, and randomness comes from the measurement tool that is both imprecise and noisy: only in the latter case does the disambiguation strategy sound natural. On the contrary, the maximin approach interprets imprecision as the effect of the variability of the real outcomes. Moreover, we have proposed an efficient solving technique that can use existing non-linear optimization software. Further work is needed to test the approach on real data, and compare obtained results with other approaches that use belief functions ([Denoeux, 2013](#)), and also recent probabilistic maximum likelihood methods, which yield possibility distributions with fixed levels of specificity ([Haddad et al., 2016](#)).

References

- J. Abellán and S. Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(5):587–598, 2003.
- 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.
- I. Couso and D. Dubois. Maximum likelihood under incomplete information: Toward a comparison of criteria. In *Soft Methods for Data Science*, pages 141–148. Springer, 2016a.

- I. Couso and D. Dubois. Belief revision and the em algorithm. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 279–290. Springer, 2016b.
- A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38:325–339, 1967.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on knowledge and data engineering*, 25(1):119–130, 2013.
- G. González-Rodríguez, A. Colubi, and M. Á. Gil. Fuzzy data treated as functional data: A one-way anova test approach. *Computational Statistics & Data Analysis*, 56(4):943–955, 2012.
- R. Guillaume and D. Dubois. Robust parameter estimation of density functions under fuzzy interval observations. In *9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'15)*, pages 147–156, 2015.
- M. Haddad, P. Leray, and N. B. Amor. Possibilistic networks: Parameters learning from imprecise data and evaluation strategy. *CoRR*, abs/1607.03705, 2016.
- D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.
- E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.
- E. Hüllermeier and W. Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer, 2015.
- H. Komiya. Elementary proof for sion's minimax theorem. *Kodai mathematical journal*, 11(1):5–7, 1988.
- R. Kruse and K. D. Meyer. *Statistics with vague data*, volume 6. Springer Science & Business Media, 2012.
- M. Serrurier and H. Prade. An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. *International Journal of Approximate Reasoning*, 54(7):919–933, 2013.
- M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.
- J. Von Neumann. Zur theorie der gesellschaftsspiele. *Math. Annalen.*, 100:295–320, 1928.
- M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105:105–122, 2002.

Concepts for Decision Making under Severe Uncertainty with Partial Ordinal and Partial Cardinal Preferences

Christoph Jansen

Georg Schollmeyer

Thomas Augustin

CHRISTOPH.JANSEN@STAT.UNI-MUENCHEN.DE

GEORG.SCHOLLMAYER@STAT.UNI-MUENCHEN.DE

AUGUSTIN@STAT.UNI-MUENCHEN.DE

Department of Statistics, Ludwig-Maximilians-Universität München

Munich (Germany)

Abstract

We introduce three different approaches for decision making under uncertainty, if (I) there is only partial (both cardinal and ordinal) information on an agent's preferences and (II) the uncertainty about the states of nature is described by a credal set. Particularly, (I) is modeled by a pair of relations, one specifying the partial rank order of the alternatives and the other modeling partial information on the strength of preference. Our first approach relies on criteria that construct complete rankings of the acts based on generalized expectation intervals. Subsequently, we introduce different concepts of global admissibility that construct partial orders by comparing all acts simultaneously. Finally, we define criteria induced by suitable binary relations on the set of acts and, therefore, can be understood as concepts of local admissibility. Whenever suitable, we provide linear programming based algorithms for checking optimality/admissibility of acts.

Keywords: partial preferences; ordinality; cardinality; decision making under uncertainty; linear programming; decision criterion; stochastic dominance; utility representation; admissibility.

1. Introduction

One of the constantly recurring topics discussed in the *imprecise probabilities* community (and on ISIPTA conferences in particular) is defining meaningful criteria for decision making under complex uncertainty, finding persuading axiomatic justifications for them and providing efficient algorithms capable to deal with them. Examples ranging from early ISIPTA contributions by, e.g., [Jaffray \(1999\)](#) to (most) recent ones by, e.g., [Bradley \(2015\)](#). However, in the vast majority of works in this field, the complexity underlying the decision situation is assumed to solely arise from beliefs on the mechanism generating the states of nature that are expressed by an imprecise probabilistic model. In contrast, the cardinal utility function adequately describing the decision maker's preference structure is often unquestioned and assumed to be precisely given in advance.¹ Our paper generalizes the classical (generalized) setting to situations, in which this assumption is no longer justified. Particularly, we consider the case that the decision maker's preference structure is both partially ordinal and partially cardinal and, therefore, no longer can be characterized by (a set of positive linear transformations of) *one* cardinal utility function.

The paper is structured as follows: In Section 2, we give a brief overview on the background of our work and show how our approach naturally fits into this picture. In Section 3, we introduce the crucial concept of a *preference system* over a set of alternatives that allows for modeling partially ordinal and partially cardinal preference structures. Section 4 introduces three different approaches

1. Exceptions include [Montes \(2014\)](#), Section 4.2.1), who uses set-valued utility functions.

for decision making with acts taking values in a preference system by proposing decision criteria based on generalized expectation intervals (Section 4.2), on global comparisons of acts (Section 4.3) and on pairwise comparisons of acts (Section 4.4). Whenever suitable, we give linear programming driven algorithms for checking feasibility of acts in finite decision settings. Section 5 concludes.

2. Brief Overview on the Fundamentals underlying our Approach

In classical subjective expected utility theory (SEUT), the decision maker is assumed to be able to specify (I) a *cardinal* utility function (unique up to a positive linear transformation) representing his preferences on a set of alternatives and (II) a unique and *precise* subjective probability measure on the space of states of nature specifying his beliefs on the occurrence of the states. Once these ingredients are specified, according to SEUT, the decision maker should choose any act that maximizes expected utility with respect to his utility and his subjective probability measure. However, in practice both assumptions (I) and (II) often are systematically too restrictive. In particular, (I) demands the decision maker to act in accordance with the axioms of von Neumann and Morgenstern, i.e. to be able to specify a complete preference ranking of all simple lotteries that is both independent and continuous (see, e.g., [Fishburn, 1970](#), Ch. 8), whereas (II) requires that the decision maker can completely order the resulting utility-valued acts by preference in accordance with the axioms of de Finetti, i.e. continuous, additive and monotone (see, e.g., [Gilboa, 2009](#), Ch. 9).

Consequently, there exists plenty of literature relaxing these assumptions. If *only* (II) is violated in the sense that there is *partial* probabilistic information on the occurrence of the states of nature together with a cardinal preference structure, the common relaxation is to allow for *imprecise* probabilistic models in order to represent the probabilistic information. In this case, one can define optimality of acts in terms of some imprecise decision criterion such as Γ -*maximin*, Γ -*maximax*, *maximality* or *E-admissibility* that, each in its own way, takes into account the whole set of probabilities for constructing a ranking of the acts (see [Huntley et al. \(2014\)](#) for a survey and, e.g., [Kofler and Menges \(1976\)](#); [Levi \(1983\)](#); [Walley \(1991\)](#); [Gilboa and Schmeidler \(1989\)](#) for original sources). Accordingly, there exists a very well-investigated and established theory as well as efficient and powerful algorithms to deal with this kind of violation of the classical assumptions (see, e.g., [Utkin and Augustin, 2005](#); [Kikuti et al., 2011](#); [Hable and Troffaes, 2014](#)).

If (I) is violated in the sense that the decision maker has only complete *ordinal* preferences and (II) is violated in the sense that there is no probabilistic information *at all*, it is nearly unanimously favored to define optimality of acts in terms of Wald's classical *maximin criterion*: Choose whatever act receiving highest possible rank under the worst possible state of nature (see [Wald, 1949](#)). However, note that the completeness of the involved ordinal ranking is essential, since, otherwise, the worst consequences of two distinct acts might be incomparable and, therefore, an optimal act with respect to the maximin criterion simply does not exist. Even more severe, also the vacuousness assumption is crucial: Applying the minimax criterion in the presence of (partial) probabilistic information means willingly ignoring information. This seems not reasonable at all (cf. also Example 1 for an illustration). Finally, if *only* (I) is violated in the sense that there is no cardinal information at all and the available ordinal information is possibly incomplete, one commonly applies the concept of first order *stochastic dominance*: Dismiss an act X taking values in the partially ordered set, if there exists another act Y such that $u \circ Y$ dominates $u \circ X$ in expectation for every real-valued function u respecting the partial order (see, e.g., [Lehmann, 1955](#); [Kamae et al., 1977](#); [Mosler and Scarsini, 1991](#)).

3. Preference Systems

In this section we define the crucial concept of a *preference system*. The intuition behind this concept is simple: In many decision problems, the (available information on) the agent's preferences is incomplete. More precisely, it often is the case that some pairs of possible decision outcomes are incomparable, whereas others can be ordered by preference. For some pairs there might even be an idea of the strength of the preference. There are several situations that could lead to such incomplete preferences. For example, if a company wants to analyze the choice behavior of their customers, the information on the customer's preferences will often be given in form of observed choices and/or survey data. In this case, incompleteness is a missing data problem and originates in lacking information. However, also the agent herself might have incomplete preferences. Suppose she knows (e.g. from earlier experience) certain outcomes better than others. Then for pairs involving better known outcomes, she might be able to specify a preference ranking and even some intuition for the strength of the preference, whereas for pairs involving unfamiliar outcomes, she might be able to specify only a ranking or can't make a comparison at all. The following definition captures the intuition just described.

Definition 1 Let A be a non-empty set and let $R_1 \subset A \times A$ denote a preorder (i.e. reflexive and transitive) on A . Moreover, let $R_2 \subset R_1 \times R_1$ denote a preorder on R_1 . Then the triplet $\mathcal{A} = [A, R_1, R_2]$ is called a **preference system** on A .

Except from transitivity, Definition 1 makes no rationality and/or compatibility assumption on the relations R_1 and R_2 . Hence, a preference system in the sense of the above definition needs by no means to be reasonable or rational. In Krantz et al. (1971, Chapter 4), an axiomatic approach for characterizing consistent preference systems is provided for the case that the involved relations are complete. The corresponding axioms then imply the existence of a real valued function representing both relations simultaneously that is unique up to a positive linear transformation. Another axiomatization that uses quaternary relations instead of pairs of relations is established in Pivato (2013), where it is shown that under some quite strong conditions (like, e.g., *solvability*) there exists a *multitility characterization* of the corresponding quaternary relation. A weaker consistency condition that still applies to settings in which conditions like solvability no longer can be expected is given in the following definition, for which we need some further notation: If R is a preorder on A , we denote by I_R and P_R its *indifference* and its *strict part*, respectively. More precisely, for $(a, b) \in A \times A$, we have $(a, b) \in I_R : \Leftrightarrow ((a, b) \in R \wedge (b, a) \in R)$ and $(a, b) \in P_R : \Leftrightarrow ((a, b) \in R \wedge (b, a) \notin R)$.

Definition 2 Let $\mathcal{A} = [A, R_1, R_2]$ be a preference system. Then \mathcal{A} is said to be **consistent** if there exists a function $u : A \rightarrow [0, 1]$ such that for all $a, b, c, d \in A$ the following two properties hold:

- i) If $(a, b) \in R_1$, then $u(a) \geq u(b)$ with equality iff $(a, b) \in I_{R_1}$.
- ii) If $((a, b), (c, d)) \in R_2$, then $u(a) - u(b) \geq u(c) - u(d)$ with equality iff $((a, b), (c, d)) \in I_{R_2}$.

Every such function u is then said to (*weakly*²) **represent** the preference system \mathcal{A} . The set of all (weak) representations u of \mathcal{A} is denoted by $\mathcal{U}_{\mathcal{A}}$. The set of all $u \in \mathcal{U}_{\mathcal{A}}$ satisfying $\inf_{a \in A} u(a) = 0$ and $\sup_{a \in A} u(a) = 1$ is denoted by $\mathcal{N}_{\mathcal{A}}$.

2. Here, the term *weakly* refers to the fact that the representation is meant in the *if* and not the *iff* sense.

The idea behind the set $\mathcal{N}_{\mathcal{A}}$ in the above definition is the following: For the special case, that the preference system \mathcal{A} is in accordance with the axioms in Krantz et al. (1971, Chapter 4), the representation is unique up to a positive linear transformation. Hence, the conditions $\inf_a u(a) = 0$ and $\sup_a u(a) = 1$ guarantee a unique representation for that special case. For the general case of a consistent preference system \mathcal{A} with non complete relations R_1 and R_2 , restricting analysis to the set $\mathcal{N}_{\mathcal{A}}$ ensures that comparison will not be made with respect to equivalent representation which only measure utility on a different scale. Note that for finite A , the boundedness condition on the utility function implies the existence of alternatives in A with greatest and lowest utility value, but not necessarily of worst and best alternatives in A w.r.t. the relation R_1 . The restriction on $\mathcal{N}_{\mathcal{A}}$, together with the concept of *granularity* of Definition 3, will prove crucial when comparing acts by means of numerical representation in Section 4.2. Obviously, for a preference system $\mathcal{A} = [A, R_1, R_2]$ to be consistent, certain compatibility criteria between the relations R_1 and R_2 have to be satisfied. For example it cannot be the case that, for some elements $a, b, c \in A$, it simultaneously holds that $(c, a) \in P_{R_1}$ and $((a, b), (c, b)) \in R_2$, since any element $u \in \mathcal{U}_{\mathcal{A}}$ would have to satisfy $u(c) > u(a)$ and $u(a) - u(b) \geq u(c) - u(b)$. We now provide an algorithm for checking the consistency of a finite preference system. The proof is straightforward and therefore left out.

Proposition 1 *Let $\mathcal{A} = [A, R_1, R_2]$ be a preference system, where $A = \{a_1, \dots, a_n\}$ is a finite and non-empty set. Consider the linear optimization problem*

$$\varepsilon = \langle (0, \dots, 0, 1)', (u_1, \dots, u_n, \varepsilon)' \rangle \longrightarrow \max_{(u_1, \dots, u_n, \varepsilon) \in \mathbb{R}^{n+1}} \quad (1)$$

with constraints $0 \leq (u_1, \dots, u_n, \varepsilon) \leq 1$ and

- i) $u_p = u_q$ for all $(a_p, a_q) \in I_{R_1} \setminus \text{diag}(A)$
- ii) $u_q + \varepsilon \leq u_p$ for all $(a_p, a_q) \in P_{R_1}$
- iii) $u_p - u_q = u_r - u_s$ for all $((a_p, a_q), (a_r, a_s)) \in I_{R_2} \setminus \text{diag}(R_1)$
- iv) $u_r - u_s + \varepsilon \leq u_p - u_q$ for all $((a_p, a_q), (a_r, a_s)) \in P_{R_2}$

Then \mathcal{A} is consistent if and only if the optimal outcome of (1) is strictly positive.

The linear programming problem (1) possesses $|R_2| + n + 2$ constraints. Thus, the number of constraints increases with the preciseness of the available information on the agent's preferences. In applications, typically the relation R_2 will be rather sparse, whereas the relation R_1 will be rather dense. This is intuitive: While R_1 is directly observable in the choice behavior of the agent, edges in R_2 need to be gained by hypothetical comparisons in interviews and polls by asking questions like: "Imagine you have objects a and b . Would you rather be willing to accept the exchange of a by c or the exchange of b by d ?" In order to reduce the number of constraints of the problem, note that (weak) representability of a preference system $\mathcal{A} = [A, R_1, R_2]$ automatically implies transitivity of the relations R_1 and R_2 . Therefore, in the constraints of the above optimization problem it actually suffices to quantify only over the transitive reduction of the relations I_{R_1} , P_{R_1} , I_{R_2} and P_{R_2} . Before turning to decision theory with preference system valued acts, we need one further concept:

Definition 3 *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent preference system. Moreover, for $\delta \in (0, 1)$, let $\mathcal{N}_{\mathcal{A}}^{\delta}$ denote the set of all $u \in \mathcal{N}_{\mathcal{A}}$ satisfying $u(a) - u(b) \geq \delta$ for all $(a, b) \in P_{R_1}$ and $u(c) - u(d) - u(e) + u(f) \geq \delta$ for all $((c, d), (e, f)) \in P_{R_2}$. Then, $\mathcal{N}_{\mathcal{A}}^{\delta}$ is called the **(weak) representation set of granularity (at least) δ** .*

The granularity can be given a similar interpretation as the *just noticeable difference* in the context of psychophysics (see Luce (1956) for details): It is the minimal difference in utility that the specific decision maker under consideration is able to notice given that utility is measured on a $[0, 1]$ -scale. More practically, the restriction to utility functions that reflect the fact that utility differences below some threshold are not distinguishable empirically will play a crucial role when it comes to defining generalized expectations in Section 4.2. For now, it is sufficient to note that the algorithm given in Proposition 1 straightforwardly extends to checking whether the preference system is consistent for a decision maker with granularity $\delta > 0$: If $(u_1^*, \dots, u_n^*, \varepsilon^*)$ is an optimal solution to problem (1), then the system is δ -consistent if and only if it holds that $\delta \leq \varepsilon^*$.

4. Decision Theory with ps-valued Acts

Differently from axiomatic approaches followed in, e.g., Seidenfeld et al. (1995); Nau (2006); Galaabaatar and Karni (2013), where (multi-)utility and (imprecise) probability representations are obtained by preferences over acts, the aim of the present paper is to obtain preferences on acts given a preference system and some additional probabilistic information. Therefore, we now propose and discuss some first ideas on decision making under uncertainty with acts taking values in some preference system (short: *ps-valued acts*) and partial probabilistic information on the occurrence of the states available. Before turning to these ideas, let us briefly give some intuition why the standard criteria for decision making under uncertainty generally will fail (or at least produce counter-intuitive results) in our context: The classical *maximin criterion*, originally proposed by A. Wald (see Wald (1949)), is the prototypical criterion for decision under complete lack of information. However, applying this criterion in the presence of probabilistic information means willingly ignoring available information and will often lead to counter-intuitive decisions (see Example 1). On the other hand, the principle of *maximizing expected utility* requires both cardinal utility and precise probabilistic information and, therefore, obviously is not applicable in our situation. Moreover, the common imprecise decision criteria, while explicitly allowing to take into account the incompleteness of the probabilistic information, still require cardinal utility scale. Contrarily, stochastic dominance allows for dealing with non-cardinal utility scales, however, requires precise probabilistic information (for approaches generalizing stochastic dominance to credal sets, see Montes (2014, Section 4.1.1)).

4.1 Basic Setting

We start by defining the central concepts of the theory for the most general case. Let S denote some non-empty set equipped with some suitable σ -algebra $\sigma(S)$. The elements of S are interpreted as all possible states of nature about whose occurrence the decision maker is uncertain. Moreover, let \mathcal{M} denote the credal set on the measurable space $(S, \sigma(S))$, interpreted as the set of all probabilities that are compatible with the available (partial) probabilistic information and thus describing the uncertainty about the occurrence of the states. For a given consistent preference system \mathcal{A} , a state space S and a credal set \mathcal{M} , a *ps-valued act* is a mapping $X : S \rightarrow A$ assigning states of nature to values in the preference system. Define the set $\mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)} \subset A^S := \{f | f : S \rightarrow A\}$ by setting

$$\mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)} := \left\{ X \in A^S : u \circ X \text{ is } \sigma(S)\text{-}\mathcal{B}_{\mathbb{R}}\text{-measurable for all } u \in \mathcal{U}_{\mathcal{A}} \right\} \quad (2)$$

where $\mathcal{B}_{\mathbb{R}}$ denotes the Borel sigma field on \mathbb{R} . By construction, the space $\mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ consists of exactly those acts $X : S \rightarrow A$ whose expectation exists with respect to all pairs $(u, \pi) \in \mathcal{U}_{\mathcal{A}} \times \mathcal{M}$

of compatible probability measure and utility representation (since bounded and measurable random variables have finite expectation). Given this notation, we can now define our main object of study:

Definition 4 *In the situation above, call every subset $\mathcal{G} \subset \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ a **decision system** (with information base $(\mathcal{A}, \mathcal{M})$). Moreover, call a decision system \mathcal{G} finite, if both $|\mathcal{G}| < \infty$ and $|S| < \infty$.*

The elements of a decision system \mathcal{G} are interpreted as those elements of the space $\mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ that are available in the specific choice situation under consideration. Given a decision system \mathcal{G} , we are interested in the following question: How can we utilize the information base $(\mathcal{A}, \mathcal{M})$ best possibly in order to define meaningful and reasonable choice criteria on the set \mathcal{G} ? In the following sections, we propose three different classes of approach that address exactly this question.

4.2 Criteria based on Generalized Expectation Intervals

In this section, we consider decision criteria that are based on the analysis of generalized expectation intervals. Depending on the attitude towards ambiguity of the decision maker of interest, such intervals give rise to different criteria for decision making. Specifically, for a ps-valued act and a decision maker with granularity $\delta > 0$, the corresponding interval will range from the lowest to the highest possible expected value that choosing this act can lead to under some pair $(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}$. This leads to the definition of the basic quantity of this section.

Definition 5 *Let $X \in \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ and $\delta \in (0, 1)$. With $\mathcal{D}_{\delta} := \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}$, we call the quantity*

$$\mathbb{E}_{\mathcal{D}_{\delta}}(X) := [\underline{\mathbb{E}}_{\mathcal{D}_{\delta}}(X), \bar{\mathbb{E}}_{\mathcal{D}_{\delta}}(X)] := \left[\inf_{(u, \pi) \in \mathcal{D}_{\delta}} \mathbb{E}_{\pi}(u \circ X), \sup_{(u, \pi) \in \mathcal{D}_{\delta}} \mathbb{E}_{\pi}(u \circ X) \right] \quad (3)$$

the generalized interval expectation of X with respect to \mathcal{A}, \mathcal{M} and granularity δ .

In the spirit of the theory of imprecise probabilities, the set $\mathbb{E}_{\mathcal{D}_{\delta}}(X)$ can be given an *epistemic* or an *ontological* interpretation: If the imprecision/ambiguity in the sets arises from lack of information in the sense of e.g. partially observed choice behavior and/or partially known precise probabilities, the set $\mathbb{E}_{\mathcal{D}_{\delta}}(X)$ is the set of all expectations arising in at least one situation that is compatible with the data. In contrast, if both sets $\mathcal{N}_{\mathcal{A}}^{\delta}$ and \mathcal{M} have an ontological interpretation, i.e. are interpreted as holistic entities of their own, the same holds true for the set of expectations $\mathbb{E}_{\mathcal{D}_{\delta}}(X)$. Of course, all decision theory that is based on comparisons of the set $\mathbb{E}_{\mathcal{D}_{\delta}}(X_i)$ of different acts X_i should reflect the underlying interpretation. The following definition gives three criteria rather relying on an ontological interpretation of the set \mathcal{D}_{δ} . Note that all of them are straightforward generalizations of the (complete order inducing) decision criteria commonly used in the theory of imprecise probabilities and reviewed, e.g., in [Huntley et al. \(2014\)](#).

Definition 6 *Let $\mathcal{G} \subset \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ be a decision system and $\delta, \alpha \in (0, 1)$. An act $X \in \mathcal{G}$ is called*

- i) **\mathcal{D}_{δ} -maximin** : iff $\forall Y \in \mathcal{G} : \underline{\mathbb{E}}_{\mathcal{D}_{\delta}}(X) \geq \underline{\mathbb{E}}_{\mathcal{D}_{\delta}}(Y)$
- ii) **\mathcal{D}_{δ} -maximax** : iff $\forall Y \in \mathcal{G} : \bar{\mathbb{E}}_{\mathcal{D}_{\delta}}(X) \geq \bar{\mathbb{E}}_{\mathcal{D}_{\delta}}(Y)$
- iii) **$\mathcal{D}_{\delta}^{\alpha}$ -maximix** : iff $\forall Y \in \mathcal{G} : \alpha \underline{\mathbb{E}}_{\mathcal{D}_{\delta}}(X) + (1 - \alpha) \bar{\mathbb{E}}_{\mathcal{D}_{\delta}}(X) \geq \alpha \underline{\mathbb{E}}_{\mathcal{D}_{\delta}}(Y) + (1 - \alpha) \bar{\mathbb{E}}_{\mathcal{D}_{\delta}}(Y)$

We denote by $\underline{\mathcal{G}}_{\delta}$, $\bar{\mathcal{G}}_{\delta}$ and $\mathcal{G}_{\delta}^{\alpha}$ the sets of \mathcal{D}_{δ} -maximin, \mathcal{D}_{δ} -maximax and $\mathcal{D}_{\delta}^{\alpha}$ -maximix acts in \mathcal{G} .

Independent of its interpretation, we need ways for computing the set $\mathbb{E}_{\mathcal{D}_\delta}(X)$ in concrete situations. The following proposition gives a linear programming based algorithm for doing so in finite decision systems. However, note that applying the proposition requires the extreme points of the underlying credal set \mathcal{M} and, therefore, is ideal for situations where the number of extreme points is moderate and where closed formulas for computing the extreme points are available. For credal sets induced by 2-monotone lower/ 2-alternating upper probabilities such formulas exist (cf., Shapley, 1971, Theorem 3, p.19). While generally the number of extreme points could be very high (maximally $|S|!$ for lower probabilities), convenient cases exist where furthermore efficient enumeration procedures are available (such special cases include *ordinal probabilities* (cf., Kofler, 1989, p. 26), *comparative probabilities* (cf., Miranda and Destercke, 2015), *necessity measures* (cf., Schollmeyer, 2015) or *p-boxes* (cf., Montes and Destercke, 2017)).

Proposition 2 *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent preference system, where $A = \{a_1, \dots, a_n\}$ such that $(a_1, b), (b, a_n) \in R_1$ for all $b \in A$ and let ε^* denote the optimal outcome of problem (1). Moreover, let $S = \{s_1, \dots, s_m\}$ be finite, \mathcal{M} be some polyhedral credal set on $(S, 2^S)$ with extreme points $\mathcal{E}(\mathcal{M}) := \{\pi^{(1)}, \dots, \pi^{(T)}\}$ and let $X \in \mathcal{G}$. For $\varepsilon^* \geq \delta > 0$, consider the collection of linear programs $LP_1^\delta, \dots, LP_T^\delta$ given by:*

$$\sum_{i=1}^n u_i \cdot \pi^{(t)}(X^{-1}(\{a_i\})) \longrightarrow \min_{(u_1, \dots, u_n) \in \mathbb{R}^n} / \max_{(u_1, \dots, u_n) \in \mathbb{R}^n} \quad (\text{LP}_t^\delta)$$

with constraints $0 \leq (u_1, \dots, u_n) \leq 1$, $u_1 = 1$, $u_n = 0$ and i) to iv) as given in Proposition 1 (with $\varepsilon := \delta$ fixed). Let $\underline{v}(t, \delta)$ and $\bar{v}(t, \delta)$ denote the optimal outcomes of problem LP_t^δ in minimum and maximum form. Then, we have $\mathbb{E}_{\mathcal{D}_\delta}(X) = [\min_t \underline{v}(t, \delta), \max_t \bar{v}(t, \delta)]$.

Proof. Let $X \in \mathcal{G}$ and $\varepsilon^* \geq \delta > 0$. Then, $\mathcal{N}_{\mathcal{A}}^\delta$ is non-empty and we can define the function $f : \mathcal{D}_\delta \rightarrow \mathbb{R}, (u, \pi) \mapsto \mathbb{E}_\pi(u \circ X)$. For any $u \in \mathcal{N}_{\mathcal{A}}^\delta$ fixed, the function $\pi \mapsto f(u, \pi)$ is linear and, therefore, both convex and concave. By applying standard results on families of convex and concave functions, we know that the functions $\pi \mapsto \inf_u f(u, \pi)$ and $\pi \mapsto \sup_u f(u, \pi)$ have to be concave and convex, respectively. But concave functions on polyhedral set attain their minimum and convex functions on polyhedral set attain their maximum on the set of extreme points. Hence, in order to find global maximum and minimum of the function f , it suffices to check on the set $\mathcal{N}_{\mathcal{A}}^\delta \times \mathcal{E}(\mathcal{M})$.

Now, let (u_1^*, \dots, u_n^*) denote an optimal solution to problem LP_t^δ in maximum form for fixed $t \in \{1, \dots, T\}$. One easily verifies that the constraints imply $u^* \in \mathcal{N}_{\mathcal{A}}^\delta$, where $u^* : A \rightarrow [0, 1], u^*(a_i) := u_i^*$ and $\bar{v}(t, \delta) = \mathbb{E}_{\pi^{(t)}}(u^* \circ X) = \sup\{\mathbb{E}_{\pi^{(t)}}(u \circ X) : u \in \mathcal{N}_{\mathcal{A}}^\delta\}$. Analogous reasoning for the problem in minimum form yields $\underline{v}(t, \delta) = \inf_{u \in \mathcal{N}_{\mathcal{A}}^\delta} \mathbb{E}_{\pi^{(t)}}(u \circ X)$. Thus, applying our considerations from before yields $\mathbb{E}_{\mathcal{D}_\delta}(X) = [\min_t \underline{v}(t, \delta), \max_t \bar{v}(t, \delta)]$. \square

Another way to compute the bounds in (3) in the case of 2-monotone lower probabilities on a finite space A is to use the Choquet representation of the upper (lower) expectation (cf., e.g., Denneberg (1994, Proposition 10.3, p. 126)): For a fixed utility u and a 2-alternating upper probability ν with associated credal set \mathcal{M}_ν the corresponding expected upper utility can be written as $\bar{\mathbb{E}}_{\{u\} \times \mathcal{M}_\nu}(X) = \sum_{i=1}^n (u_{(i)} - u_{(i-1)}) \cdot \nu(\{s \in S \mid u(X(s)) \geq u_{(i)}\})$. If R_1 is complete then the expectation is a linear form in the utility u and the maximization $\max_{u \in \mathcal{N}_{\mathcal{A}}^\delta} \bar{\mathbb{E}}_{\{u\} \times \mathcal{M}_\nu}(X)$ translates to a simple linear program. If the relation R_1 is not complete then the ordering of the utility values u_i can change as u ranges in $\mathcal{N}_{\mathcal{A}}^\delta$ and one has to compute the expectation separately for every possible ordering of the utility values and then take the maximum. If there are totally comparable values u_i meaning that

for every u_j either $u_i \leq u_j$ or $u_i > u_j$, independently from the concrete $u \in \mathcal{N}_{\mathcal{A}}^{\delta}$ then one can split the sum in a part containing all utility values below u_i and a part containing all utility values above u_i and then analyze every subsum independently which would help in reducing the combinatorial complexity. The criteria from Definition 6 allow for comparing acts *given* the granularity δ of the specific decision maker of interest. However, note that knowing the granularity might be a strong assumption if R_1 and R_2 are *partial* orderings, since experimental settings in which this additional parameter could precisely be elicited are not as straightforward as in the complete case. Further possibilities to deal with these issues are treated in the next two sections, where we propose two approaches completely overcoming the choice of a granularity parameter.

4.3 Criteria based on Global Comparisons

The decision criteria defined in Section 4.2 all construct complete rankings on the set \mathcal{G} by comparing numerical representations of parts of the decision system and by somehow ignoring the inherent utility and probability structure. Therefore, when defining optimality of acts in terms of one of the criteria from Definition 6, it makes no difference if the ranking is constructed by pairwise or global comparisons. In the next sections, we turn to two approaches that explicitly take into account a global and local viewpoint for defining optimality of acts, respectively.³ We start with the global perspective in the sense that we try to find existing utilities (or probabilities, respectively) that can establish a form of global admissibility of a given act X over all other acts Y that is valid for every possible underlying probability (or utility, respectively). This is reflected in the fact that in the three admissibility concepts of Definition 7 a \forall quantifier can follow an \exists quantifier but not vice versa.

Definition 7 Let $\mathcal{G} \subset \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ denote a decision system. We call an act $X \in \mathcal{G}$

- i) $\mathcal{A}|\mathcal{M}$ -**admissible** : iff $\exists u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$
- ii) \mathcal{A} -**admissible** : iff $\exists u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$
- iii) \mathcal{M} -**admissible** : iff $\exists \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$
- iv) $\mathcal{A}|\mathcal{M}$ -**dominant** : iff $\forall u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$

Denote by $\mathcal{G}_{\mathcal{A}|\mathcal{M}}$, $\mathcal{G}_{\mathcal{A}}$, $\mathcal{G}_{\mathcal{M}}$ and $\mathcal{G}_{\mathcal{A}|\mathcal{M}}^d$ the sets of such acts, respectively.

All four act properties just defined rely on the idea that, if there *was* perfect information on both the state probabilities (i.e. $\mathcal{M} = \{\pi\}$ is a singleton) and the utility values (i.e. the utility representation u is unique up to a positive linear transformation), then an act X should be labeled optimal iff X has greater or equal expected utility than every other act $Y \in \mathcal{G}$ with respect to (u, π) . However, they differ in the way they handle the ambiguity underlying the involved sets \mathcal{M} and $\mathcal{U}_{\mathcal{A}}$: While $\mathcal{A}|\mathcal{M}$ -admissibility only demands the existence of at least *one* compatible combination (u, π) with respect to which X maximizes expected utility, $\mathcal{A}|\mathcal{M}$ -dominance requires this for *all* compatible combinations. \mathcal{M} - and \mathcal{A} -admissibility relax the \forall -assumption on probability and utility level, respectively. Clearly, it holds that $\mathcal{G}_{\mathcal{A}}, \mathcal{G}_{\mathcal{M}}, \mathcal{G}_{\mathcal{A}|\mathcal{M}}^d \subset \mathcal{G}_{\mathcal{A}|\mathcal{M}}$ and $\mathcal{G}_{\mathcal{A}|\mathcal{M}}^d \subset \mathcal{G}_{\mathcal{A}}$ and $\mathcal{G}_{\mathcal{A}|\mathcal{M}}^d \subset \mathcal{G}_{\mathcal{M}}$, but in general neither $\mathcal{G}_{\mathcal{A}} \subset \mathcal{G}_{\mathcal{M}}$ nor $\mathcal{G}_{\mathcal{M}} \subset \mathcal{G}_{\mathcal{A}}$. The following example demonstrates that ignoring the available information base and applying the maximin criterion instead leads to counter-intuitive decisions even in very simple situations.

3. Note that in the context of IP decision theory, fundamental differences between global criteria and criteria based on pairwise comparisons have already been discussed (Schervish et al., 2003).

Example 1 Let $A = \{a_1, a_2, a_3, a_4\}$, the (complete) relation R_1 induced by $a_2P_{R_1}a_3P_{R_1}a_4P_{R_1}a_1$ and $P_{R_2} = \{((a_2, a_4), (a_3, a_1))\}$ consists of one single edge. Consider the decision system $\mathcal{G} = \{X_1, X_2\}$, where the acts $X_1, X_2 : \{s_1, s_2\} \rightarrow A$ are defined by $(X_1(s_1), X_1(s_2)) = (a_1, a_2)$ and $(X_2(s_1), X_2(s_2)) = (a_3, a_4)$. Moreover, suppose our probabilistic information is given by the credal set $\mathcal{M} := \{\pi : \pi(\{s_1\}) \leq 0.5\}$. In this case, act X_1 is $\mathcal{A}|\mathcal{M}$ -dominant, since it maximizes expected utility w.r.t. every pair $(u, \pi) \in \mathcal{U}_{\mathcal{A}} \times \mathcal{M}$. In contrast, X_2 is not even $\mathcal{A}|\mathcal{M}$ -admissible, although it is the unique optimal act w.r.t. the maximin criterion!

To complete the section, we give a proposition containing a linear programming based approach for checking whether an act X is \mathcal{A} -admissible in finite decision settings.

Proposition 3 Consider again the situation of Proposition 2. Moreover, let $\mathcal{G} := \{X_1, \dots, X_k\} \subset \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ denote a finite decision system and let $X_z \in \mathcal{G}$. Consider again the linear optimization problem (1) with additional constraints

$$\sum_{i=1}^n u_i \cdot \pi^{(t)}(X_z^{-1}(\{a_i\})) \geq \sum_{i=1}^n u_i \cdot \pi^{(t)}(X_l^{-1}(\{a_i\})) \text{ for all } l = 1, \dots, k \quad (C_t)$$

for every $t = 1, \dots, T$. Then X_z is \mathcal{A} -admissible if and only if the optimal outcome of this optimization problem is strictly greater than 0.

Proof. A similar argument as in the proof of Proposition 1 guarantees the existence of an optimal solution $(u_1^*, \dots, u_n^*, \varepsilon^*)$ such that $u : A \rightarrow \mathbb{R}$, $u(a_i) := u_i^*$ for all $i \in \underline{n}$ (weakly) represents the preference system \mathcal{A} . Now, let $\pi \in \mathcal{M}$ be arbitrary. Choose $\alpha \in \Delta_{T-1}$ such that $\pi(\cdot) = \sum_{t=1}^T \alpha_t \cdot \pi^{(t)}(\cdot)$. Then, condition (C_t) additionally guarantees that for all $l = 1, \dots, k$ it holds

$$\begin{aligned} \mathbb{E}_{\pi}(u \circ X_z) &= \sum_{i=1}^n u_i^* \cdot \pi(X_z^{-1}(\{a_i\})) &= \sum_{i=1}^n u_i^* \cdot \left(\sum_{t=1}^T \alpha_t \cdot \pi^{(t)}(X_z^{-1}(\{a_i\})) \right) \\ &= \sum_{t=1}^T \alpha_t \left(\sum_{i=1}^n u_i^* \cdot \pi^{(t)}(X_z^{-1}(\{a_i\})) \right) &\geq \sum_{t=1}^T \alpha_t \left(\sum_{i=1}^n u_i^* \cdot \pi^{(t)}(X_l^{-1}(\{a_i\})) \right) \\ &= \sum_{i=1}^n u_i^* \cdot \left(\sum_{t=1}^T \alpha_t \cdot \pi^{(t)}(X_l^{-1}(\{a_i\})) \right) &= \mathbb{E}_{\pi}(u \circ X_l) \end{aligned}$$

Hence, X_z maximizes expected utility with respect to (u, π) . Since $\pi \in \mathcal{M}$ was chosen arbitrarily, this implies that X_z is \mathcal{A} -admissible. \square

Note that a similar algorithm as given in in Proposition 3 could be used for checking \mathcal{M} -admissibility of acts. However, this would require the set $\mathcal{E}(\mathcal{U}_{\mathcal{A}})$ of extreme points of the representation set to be known, which is way less straightforward than assuming $\mathcal{E}(\mathcal{M})$ to be known.

4.4 Criteria based on Pairwise Comparisons

While the criteria defined in Section 4.3 rather relied on global comparisons of acts in the sense that an act, in order to be labeled admissible, has to dominate all other acts in expectation for (at least one) fixed pair (π, u) , we now turn to criteria induced by pairwise expectation comparisons of acts (i.e. binary relations on the set of acts). Similarly as already seen in the global case, there are

several different ways to define such relations each of which reflecting a different attitude towards the underlying ambiguity. In particular, we define six binary relations $R_{\exists\exists}, R_{\exists\forall}^1, R_{\exists\forall}^2, R_{\forall\exists}^1, R_{\forall\exists}^2$ and $R_{\forall\forall}$ on $\mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$ by setting for all $X, Y \in \mathcal{F}_{(\mathcal{A}, \mathcal{M}, S)}$:

$$(X, Y) \in R_{\exists\exists} : \Leftrightarrow \exists u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (4)$$

$$(X, Y) \in R_{\exists\forall}^1 : \Leftrightarrow \exists u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (5)$$

$$(X, Y) \in R_{\exists\forall}^2 : \Leftrightarrow \exists \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (6)$$

$$(X, Y) \in R_{\forall\exists}^1 : \Leftrightarrow \forall u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (7)$$

$$(X, Y) \in R_{\forall\exists}^2 : \Leftrightarrow \forall \pi \in \mathcal{M} \exists u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (8)$$

$$(X, Y) \in R_{\forall\forall} : \Leftrightarrow \forall \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y) \quad (9)$$

Obviously, it holds that $R_{\forall\forall}$ is subset of all other relation, whereas $R_{\exists\exists}$ is a superset of them. For the remaining relations, in general, no sub- or superset relation has to be satisfied. Furthermore, transitivity is only guaranteed for $R_{\forall\forall}$ in general. Similarly as already discussed in the global case, each of the desirability relations just defined relies on the idea that, given perfect information on utilities and probabilities, maximizing expected utility should be the criterion of choice. Again, the relations differ only in the way they handle the ambiguity on the involved sets $\mathcal{U}_{\mathcal{A}}$ and \mathcal{M} . Naturally, each of the relations defined above induces a different criterion of (local) admissibility. These criteria are summarized in the following definition.

Definition 8 Let $R \in \{R_{\exists\exists}, R_{\exists\forall}^1, R_{\exists\forall}^2, R_{\forall\exists}^1, R_{\forall\exists}^2, R_{\forall\forall}\} =: \mathcal{R}_p$. We call an act $X \in \mathcal{G}$ **locally admissible** with respect to R , if it is an element of the set $\max_R(\mathcal{G}) := \{Y \in \mathcal{G} : \nexists Z \in \mathcal{G} \text{ s.t. } (Z, Y) \in P_R\}$, that is if it is a maximal element in \mathcal{G} with respect to the relation $R \cap (\mathcal{G} \times \mathcal{G})$.

So, which of the relations defined above are most important in our context? To address this question, we discuss some special cases: If the credal set \mathcal{M} is a singleton $\mathcal{M} = \{\pi\}$ and if $\mathcal{U}_{\mathcal{A}} = \{a \cdot u_0 + b \mid a > 0, b \in \mathbb{R}\}$ is unique up to a positive linear transformation then all relations $R \in \mathcal{R}_p$ coincide with the classical expected utility criterion. If \mathcal{M} is a singleton and $\mathcal{U}_{\mathcal{A}}$ is the class of all non-decreasing functions then the relations $R_{\forall\exists}^1$ and $R_{\forall\forall}$ essentially coincide with the classical concept of first order stochastic dominance (cf., e.g., [Mosler and Scarsini \(1991\)](#); [Lehmann \(1955\)](#); [Kamae et al. \(1977\)](#)) while second order stochastic dominance is obtained if $\mathcal{U}_{\mathcal{A}}$ is the set of all continuous concave non-decreasing utility functions that are related to the concept of *decreasing returns to scale*. An intermediate case would arise if one has information about decreasing returns to scale only for parts of the preference system. To compute the relations $R_{\exists\exists}$ and $R_{\forall\forall}$ in the general case one can use the same technique as in Proposition 2 by noting that $\mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$ is equivalent to $\mathbb{E}_{\pi}(u \circ X - u \circ Y) \geq 0$. The other relations $R \in \mathcal{R}_p$ do not appear to be manageable in such a straightforward manner. However, if \mathcal{M} is the core of a belief function then all $\pi \in \mathcal{M}$ can be understood as obtained from a mass transfer of probability mass to singleton sets of S . Since classical first order stochastic dominance can be checked via the solution of a mass transportation problem (cf., [Mosler and Scarsini \(1991, p. 269\)](#)), the computation of $R_{\exists\forall}^2$ can be done by solving a composite mass transportation problem. The most rigorous relation $R_{\forall\forall}$ is also discussed in [Montes \(2014, Ch. 4.1\)](#). Note that the locally $R_{\forall\forall}$ -admissible acts coincide with the $\mathcal{A}|\mathcal{M}$ -dominant acts. Note also that, in general, the other global concepts of admissibility from Definition 4.3 are not expressable as induced by one of the local criteria from Definition 4.4 (for the special case of a cardinal u this is discussed in [Schervish et al. \(2003\)](#)).

5. Summary and Outlook

We proposed three approaches for decision making under severe uncertainty if acts are ps-valued: The first is based on granularity-dependent expectation intervals, while the other two rely on local and global comparisons of specific expectations of acts. For selected criteria, we gave linear programs. Several challenges should be addressed in future research. Clearly, further algorithms for the remaining criteria need to be explored in order to make the theory computationally feasible and, therefore, applicable in practice. Further, it is certainly worth investigating in more detail how the criteria from the different approaches relate to each other. Finally, designing experimental settings for eliciting the parameter δ could help to receive a more canonical interpretation of granularity.

Acknowledgements

The authors would like to thank the three anonymous referees for their helpful comments and Jean Baccelli for stimulating discussions on the topic and hints to further relevant references.

References

- S. Bradley. How to choose among choice functions. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *Proc. of ISIPTA '15*, pages 57–66. Aracne, 2015.
- D. Denneberg. *Non-additive Measure and Integral*. Kluwer Academic Publishers, Dordrecht, Boston and London, 1994.
- P. Fishburn. *Utility Theory for Decision Making*. Wiley, London and New York, 1970.
- T. Galaabaatar and E. Karni. Subjective expected utility with incomplete preferences. *Econometrica*, 81:255–284, 2013.
- I. Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, New York, 2009.
- I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *J Math Econ*, 18: 141–153, 1989.
- R. Hable and M. Troffaes. Computation. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 329–337. Wiley, Chichester, 2014.
- N. Huntley, R. Hable, and M. Troffaes. Decision making. In T. Augustin, Coolen, Frank, G. de Cooman, and Matthias Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 190–206. Wiley, Chichester, 2014.
- J.-Y. Jaffray. Rational decision making with imprecise probabilities. In G. de Cooman, F. Cozman, S. Moral, and P. Walley, editors, *Proc. of ISIPTA '99*, pages 183–188. IPP, 1999.
- T. Kamae, U. Krengel, and G. O'Brien. Stochastic inequalities on partially ordered spaces. *Ann Probab*, pages 899–912, 1977.
- D. Kikuti, F. Cozman, and R. Filho. Sequential decision making with partially ordered preferences. *Artif Intel*, 175:1346 – 1365, 2011.

- E. Kofler. *Prognosen und Stabilität bei unvollständiger Information*. Campus, Frankfurt, 1989.
- E. Kofler and G. Menges. *Entscheiden bei unvollständiger Information*. Springer, Berlin, 1976.
- D. Krantz, R. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Academic Press, San Diego and London, 1971.
- E. Lehmann. Ordered families of distributions. *Ann Math Stat*, 26:399–419, 1955.
- I. Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press, Cambridge, Mass., 1983.
- R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.
- E. Miranda and S. Destercke. Extreme points of the credal sets generated by comparative probabilities. *J Math Psychol*, 64:44–57, 2015.
- I. Montes. *Comparison of Alternatives under Uncertainty and Imprecision*. PhD thesis, Oviedo, 2014. digibuo.uniovi.es/dspace/bitstream/10651/28953/6/TD_IgnacioMontesGutierrez.pdf.
- I. Montes and S. Destercke. On extreme points of p-boxes and belief functions. In M. Ferraro, P. Giordani, B. Vantaggi, M. Gagolewski, M. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *Soft Methods for Data Science*, pages 363–371. Springer, 2017.
- K. Mosler and M. Scarsini. Some theory of stochastic dominance. In K. Mosler and M. Scarsini, editors, *Stochastic Orders and Decision under Risk*, pages 203–212. Institute of Mathematical Statistics, Hayward, CA, 1991.
- R. Nau. The shape of incomplete preferences. *Ann Stat*, 34:2430–2448, 2006.
- M. Pivato. Multiutility representations for incomplete difference preorders. *Math Soc Sci*, 66:196–220, 2013.
- M. Schervish, T. Seidenfeld, J. Kadane, and I. Levi. Extensions of expected utility theory and some limitations of pairwise comparisons. In J.-M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *Proc. of ISIPTA '03*, pages 496–510. Carleton Scientific, 2003.
- G. Schollmeyer. On the number and characterization of the extreme points of the core of necessity measures on finite spaces. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *Proc. of ISIPTA '15*, pages 277–286. Aracne, 2015.
- T. Seidenfeld, J. Kadane, and M. Schervish. A representation of partially ordered preferences. *Ann Stat*, 23:2168–2217, 1995.
- L. Shapley. Cores of convex games. *Int J Game Theory*, 1:11–26, 1971.
- L. Utkin and T. Augustin. Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *Proc. of ISIPTA '05*, pages 349–358. SIPTA, 2005.
- A. Wald. Statistical decision functions. *Ann Math Stat*, 20:165–205, 1949.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Efficient Computation of Updated Lower Expectations for Imprecise Continuous-Time Hidden Markov Chains

Thomas Krak

T.E.KRAK@UU.NL

[†]*Department of Information and Computing Sciences, Utrecht University (The Netherlands)*

Jasper De Bock

JASPER.DEBOCK@UGENT.BE

Department of Electronics and Information Systems, imec, IDLab, Ghent University (Belgium)

Arno Siebes[†]

A.P.J.M.SIEBES@UU.NL

Abstract

We consider the problem of performing inference with *imprecise continuous-time hidden Markov chains*, that is, *imprecise continuous-time Markov chains* that are augmented with random *output* variables whose distribution depends on the hidden state of the chain. The prefix ‘imprecise’ refers to the fact that we do not consider a classical continuous-time Markov chain, but replace it with a robust extension that allows us to represent various types of model uncertainty, using the theory of *imprecise probabilities*. The inference problem amounts to computing lower expectations of functions on the state-space of the chain, given observations of the output variables. We develop and investigate this problem with very few assumptions on the output variables; in particular, they can be chosen to be either discrete or continuous random variables. Our main result is a polynomial runtime algorithm to compute the lower expectation of functions on the state-space at any given time-point, given a collection of observations of the output variables.

Keywords: continuous-time hidden Markov chains; imprecise probabilities; updating.

1. Introduction

A continuous-time Markov chain (CTMC) is a stochastic model that describes the evolution of a dynamical system under uncertainty. Specifically, it provides a probabilistic description of how such a system might move through a finite state-space, as time elapses in a continuous fashion. There are various ways in which this model class can be extended.

One such extension are continuous-time *hidden* Markov chains (CTHMC’s) (Wei et al., 2002). Such a CTHMC is a stochastic model that contains a continuous-time Markov chain as a latent variable—that is, the actual realised behaviour of the system cannot be directly observed. This model furthermore incorporates random *output* variables, which depend probabilistically on the current state of the system, and it is rather realisations of these variables that one observes. Through this stochastic dependency between the output variables and the states in which the system might be, one can perform inferences about quantities of interest that depend on these states—even though they have not been, or cannot be, observed directly.

Another extension of CTMC’s, arising from the theory of *imprecise probabilities* (Walley, 1991), are *imprecise continuous-time Markov chains* (ICTMC’s) (Škulj, 2015; Krak et al., 2016). This extension can be used to robustify against uncertain numerical parameter assessments, as well as the simplifying assumptions of time-homogeneity and that the model should satisfy the Markov property. Simply put, an ICTMC is a *set* of continuous-time stochastic processes, some of which are “traditional” time-homogeneous CTMC’s. However, this set also contains more complicated processes, which are non-homogeneous and do not satisfy the Markov property.

In this current work, we combine these two extensions by considering *imprecise continuous-time hidden Markov chains*—a stochastic model analogous to a CTHMC, but where the latent CTMC is replaced by an ICTMC. We will focus in particular on practical aspects of the corresponding inference problem. That is, we provide results on how to efficiently compute lower expectations of functions on the state-space, given observed realisations of the output variables.

Throughout, all results are stated without proof. We have made available an extended version of this work ([Krak et al., 2017](#)), which includes an appendix containing the proofs of all our results.

1.1 Related Work

As should be clear from the description of CTHMC’s in Section 1, this model class extends the well-known (discrete-time) *hidden Markov models* (HMM’s) to a continuous-time setting. In the same sense, the present subject of ICTHMC’s can be seen to extend previous work on *imprecise hidden Markov models* (iHMM’s) ([de Cooman et al., 2010](#)) to a continuous-time setting. Hence, the model under consideration should hopefully be intuitively clear to readers familiar with (i)HMM’s.

The main novelty of this present work is therefore not the (somewhat obvious) extension of iHMM’s to ICTHMC’s, but rather the application of recent results on ICTMC’s ([Krak et al., 2016](#)) to derive an efficient solution to the continuous-time analogue of inference in iHMM’s. The algorithm that we present is largely based on combining these results with the ideas behind the MePiCTIr algorithm ([de Cooman et al., 2010](#)) for inference in credal trees under epistemic irrelevance.

A second novelty of the present paper is that, contrary to most of the work in the literature on iHMM’s, we allow the output variables of the ICTHMC to be either discrete or continuous. This allows the model to be applied to a much broader range of problems. At the same time, it turns out that this does not negatively influence the efficiency of the inference algorithm.

2. Preliminaries

We denote the reals as \mathbb{R} , the non-negative reals as $\mathbb{R}_{\geq 0}$, and the positive reals as $\mathbb{R}_{> 0}$. The natural numbers are denoted by \mathbb{N} , and we define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Since we are working in a continuous-time setting, a *time-point* is an element of $\mathbb{R}_{\geq 0}$, and these are typically denoted by t or s . We also make extensive use of non-empty, finite sequences of time points $u \subset \mathbb{R}_{\geq 0}$. These are taken to be ordered, so that they may be written $u = t_0, \dots, t_n$, for some $n \in \mathbb{N}_0$, and such that then $t_i < t_j$ for all $i, j \in \{0, \dots, n\}$ for which $i < j$. Such sequences are usually denoted by u or v , and we let \mathcal{U} be the entire set of them.

Throughout, we consider some fixed, finite state space \mathcal{X} . A generic element of \mathcal{X} will be denoted by x . When considering the state-space at a specific time t , we write $\mathcal{X}_t := \mathcal{X}$, and x_t denotes a generic state-assignment at this time. When considering multiple time-points u simultaneously, we define the joint state-space as $\mathcal{X}_u := \prod_{t_i \in u} \mathcal{X}_{t_i}$, of which $x_u = (x_{t_0}, \dots, x_{t_n})$ is a generic element.

For any $u \in \mathcal{U}$, we let $\mathcal{L}(\mathcal{X}_u)$ be the set of all real-valued functions on \mathcal{X}_u .

2.1 Imprecise Continuous-Time Markov Chains

We here briefly recall the most important properties of imprecise continuous-time Markov chains (ICTMC’s), following the definitions and results of [Krak et al. \(2016\)](#). For reasons of brevity, we provide these definitions in a largely intuitive, non-rigorous manner, and refer the interested reader to this earlier work for an in-depth treatise on the subject.

An ICTMC will be defined below as a specific set of *continuous-time stochastic processes*. Simply put, a continuous-time stochastic process is a joint probability distribution over random variables X_t , for each time $t \in \mathbb{R}_{\geq 0}$, where each random variable X_t takes values in \mathcal{X} .

It will be convenient to have a way to numerically parameterise such a stochastic process P . For this, we require two different kinds of parameters. First, we need the specification of the initial distribution $P(X_0)$ over the state at time zero; this simply requires the specification of some probability mass function on \mathcal{X}_0 . Second, we need to parameterise the dynamic behaviour of the model.

In order to describe this dynamic behaviour, we require the concept of a *rate matrix*. Such a rate matrix Q is a real-valued $|\mathcal{X}| \times |\mathcal{X}|$ matrix, whose off-diagonal elements are non-negative, and whose every row sums to zero—thus, the diagonal elements are non-positive. Such a rate matrix may be interpreted as describing the “rate of change” of the conditional probability $P(X_s | X_t, X_u = x_u)$, when s is close to t . In this conditional probability, it is assumed that $u < t$, whence the state assignment x_u is called the *history*. For small enough $\Delta \in \mathbb{R}_{>0}$, we may now write that

$$P(X_{t+\Delta} | X_t, X_u = x_u) \approx [I + \Delta Q_{t,x_u}] (X_t, X_{t+\Delta}),$$

for some rate matrix Q_{t,x_u} , where I denotes the $|\mathcal{X}| \times |\mathcal{X}|$ identity matrix, and where the quantity $[I + \Delta Q_{t,x_u}] (X_t, X_{t+\Delta})$ denotes the element at the X_t -row and $X_{t+\Delta}$ -column of the matrix $I + \Delta Q_{t,x_u}$. Note that in general, this rate matrix Q_{t,x_u} may depend on the specific time t and history x_u at which this relationship is stated.

If these rate matrices only depend on the time t and not on the history x_u , i.e. if $Q_{t,x_u} = Q_t$ for all t and all x_u , then it can be shown that P satisfies the *Markov property*: $P(X_s | X_t, X_u) = P(X_s | X_t)$. In this case, P is called a *continuous-time Markov chain*.

Using this method of parameterisation, an *imprecise continuous-time Markov chain* (ICTMC) is similarly parameterised using a *set* of rate matrices \mathcal{Q} , and a *set* of initial distributions \mathcal{M} . The corresponding ICTMC, denoted by $\mathbb{P}_{\mathcal{Q},\mathcal{M}}$, is the set of all continuous-time stochastic processes whose dynamics can be described using the elements of \mathcal{Q} , and whose initial distributions are consistent with \mathcal{M} . That is, $\mathbb{P}_{\mathcal{Q},\mathcal{M}}$ is the set of stochastic processes P for which $P(X_0) \in \mathcal{M}$ and for which $Q_{t,x_u} \in \mathcal{Q}$ for every time t and history x_u .

The *lower expectation* with respect to this set $\mathbb{P}_{\mathcal{Q},\mathcal{M}}$ is then defined as

$$\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\cdot | \cdot] := \inf \{ \mathbb{E}_P[\cdot | \cdot] : P \in \mathbb{P}_{\mathcal{Q},\mathcal{M}} \},$$

where $\mathbb{E}_P[\cdot | \cdot]$ denotes the expectation with respect to the (precise) stochastic process P . The *upper expectation* $\overline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\cdot | \cdot]$ is defined similarly, and is derived through the well-known conjugacy property $\overline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\cdot | \cdot] = -\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[-\cdot | \cdot]$. Note that it suffices to focus on lower (or upper) expectations, and that *lower* (and *upper*) *probabilities* can be regarded as a special case; for example, for any $A \subseteq \mathcal{X}$, we have that $\underline{P}_{\mathcal{Q},\mathcal{M}}(X_s \in A | X_t) := \inf\{P(X_s \in A | X_t) : P \in \mathbb{P}_{\mathcal{Q},\mathcal{M}}\} = \underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\mathbb{I}_A(X_s) | X_t]$, where \mathbb{I}_A is the indicator of A , defined for all $x \in \mathcal{X}$ by $\mathbb{I}_A(x) := 1$ if $x \in A$ and $\mathbb{I}_A(x) := 0$ otherwise.

In the sequel, we will assume that \mathcal{M} is non-empty, and that \mathcal{Q} is non-empty, bounded,¹ convex, and has *separately specified rows*. This latter property states that \mathcal{Q} is closed under arbitrary recombination of rows from its elements; see (Krak et al., 2016, Definition 24) for a formal definition. Under these assumptions, $\mathbb{P}_{\mathcal{Q},\mathcal{M}}$ satisfies an *imprecise Markov property*, in the sense that $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[f(X_s) | X_t, X_u = x_u] = \overline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[f(X_s) | X_t]$. This property explains why we call this model an imprecise continuous-time “Markov” chain.

1. That is, that there exists a $c \in \mathbb{R}_{\geq 0}$ such that, for all $Q \in \mathcal{Q}$ and $x \in \mathcal{X}$, it holds that $|Q(x,x)| < c$.

2.2 Computing Lower Expectations for ICTMC's

Because we want to focus in this paper on providing efficient methods of computation, we here briefly recall some previous results from Krak et al. (2016) about how to compute lower expectations for ICTMC's. We focus in particular on how to do this for functions on a single time-point.

To this end, it is useful to introduce the *lower transition rate operator* \underline{Q} that corresponds to \mathcal{Q} . This operator is a map from $\mathcal{L}(\mathcal{X})$ to $\mathcal{L}(\mathcal{X})$, defined for every $f \in \mathcal{L}(\mathcal{X})$ by

$$[\underline{Q}f](x) := \inf \left\{ \sum_{x' \in \mathcal{X}} Q(x, x') f(x') : Q \in \mathcal{Q} \right\} \text{ for all } x \in \mathcal{X}. \quad (1)$$

Using this lower transition rate operator \underline{Q} , we can compute conditional lower expectations in the following way. For any $t, s \in \mathbb{R}_{\geq 0}$, with $t \leq s$, and any $f \in \mathcal{L}(\mathcal{X})$, it has been shown that

$$\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_s) | X_t] = \mathbb{E}_{\mathcal{Q}}[f(X_s) | X_t] := \lim_{n \rightarrow +\infty} \left[I + \frac{(s-t)}{n} \underline{Q} \right]^n f,$$

where I is the identity operator on $\mathcal{L}(\mathcal{X})$, in the sense that $Ig = g$ for every $g \in \mathcal{L}(\mathcal{X})$. The notation $\mathbb{E}_{\mathcal{Q}}$ is meant to indicate that this conditional lower expectation only depends on \mathcal{Q} , and not on \mathcal{M} . The above implies that for large enough $n \in \mathbb{N}$, and writing $\Delta := (s-t)/n$, we have

$$\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_s) | X_t] = \mathbb{E}_{\mathcal{Q}}[f(X_s) | X_t] \approx [I + \Delta \underline{Q}]^n f. \quad (2)$$

Concretely, this means that if one is able to solve the minimisation problem in Equation (1)—which is relatively straightforward for “nice enough” \mathcal{Q} , e.g., convex hulls of finite sets of rate matrices—then one can also compute conditional lower expectations using the expression in Equation 2. In practice, we do this by first computing $f'_1 := \underline{Q}f$ using Equation (1), and then computing $f_1 := f + \Delta f'_1$. Next, we compute $f'_2 := \underline{Q}f_1$, from which we obtain $f_2 := f_1 + \Delta f'_2$. Proceeding in this fashion, after n steps we then finally obtain $f_n := [I + \Delta \underline{Q}]f_{n-1} = [I + \Delta \underline{Q}]^n f$, which is roughly the quantity of interest $\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_s) | X_t]$ provided that n was taken large enough.²

As noted above, the conditional lower expectation $\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_s) | X_t]$ only depends on \mathcal{Q} . Similarly, and in contrast, the unconditional lower expectation at time zero only depends on \mathcal{M} . That is,

$$\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_0)] = \mathbb{E}_{\mathcal{M}}[f(X_0)] := \inf \left\{ \sum_{x \in \mathcal{X}} p(x) f(x) : p \in \mathcal{M} \right\}. \quad (3)$$

Furthermore, the unconditional lower expectation at an arbitrary time $t \in \mathbb{R}_{\geq 0}$, is given by

$$\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_t)] = \mathbb{E}_{\mathcal{M}}[\mathbb{E}_{\mathcal{Q}}[f(X_t) | X_0]], \quad (4)$$

which can therefore be computed by combining Equations (2) and (3). In particular, from a practical point of view, it suffices to first compute the conditional lower expectation $\mathbb{E}_{\mathcal{Q}}[f(X_t) | X_0]$, using Equation (2). Once this quantity is obtained, it remains to compute the right-hand side of Equation (3), which again is relatively straightforward when \mathcal{M} is “nice enough”, e.g., the convex hull of some finite set of probability mass functions.

2. We refer the reader to (Krak et al., 2016, Proposition 8.5) for a theoretical bound on the minimum such n that is required to ensure a given maximum error on the approximation in Equation (2). We here briefly note that this bound scales polynomially in every relevant parameter. This means that $\mathbb{E}_{\mathcal{Q}, \mathcal{M}}[f(X_s) | X_t]$ is numerically computable in polynomial time, provided that \mathcal{Q} is such that Equation (1) can also be solved in the same time-complexity order.

3. Imprecise Continuous-Time Hidden Markov Chains

In this section, we construct the *hidden* model that is the subject of this paper. Our aim is to augment the stochastic processes that were introduced in the previous section, by adding random *output* variables Y_t whose distribution depends on the state X_t at the same time point t .

We want to focus in this paper on the more practical aspect of solving the inference problem of interest, i.e., computing lower expectations on the state-space *given some observations*. Hence, we will assume that we are given some finite sequence of time points, and we then only consider these time points in augmenting the model. In order to disambiguate the notation, we will henceforth denote stochastic processes as $P_{\mathcal{X}}$, to emphasise that they are only concerned with the state-space.

3.1 Output Variables

We want to augment stochastic processes with random “output variables” Y_t , whose distribution depends on the state X_t . We here define the corresponding (conditional) distribution.

We want this definition to be fairly general, and in particular do not want to stipulate that Y_t should be either a discrete or a continuous random variable. To this end, we simply consider some set \mathcal{Y} to be the outcome space of the random variable. We then let Σ be some algebra on \mathcal{Y} . Finally, for each $x \in \mathcal{X}$, we consider some finitely (and possibly σ -)additive probability measure $P_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ on (\mathcal{Y}, Σ) , with respect to which the random variable Y_t can be defined.

Definition 1 An output model is a tuple $(\mathcal{Y}, \Sigma, P_{\mathcal{Y}|\mathcal{X}})$, where \mathcal{Y} is an outcome space, Σ is an algebra on \mathcal{Y} , and, for all $x \in \mathcal{X}$, $P_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is a finitely additive probability measure on (\mathcal{Y}, Σ) .

When considering (multiple) explicit time points, we use notation analogous to that used for states; so, $\mathcal{Y}_t := \mathcal{Y}$ for any time $t \in \mathbb{R}_{\geq 0}$, and for any $u \in \mathcal{U}$, we write $\mathcal{Y}_u := \prod_{t \in u} \mathcal{Y}_t$.

We let Σ_u denote the set of all events of the type $O_u = \times_{t \in u} O_t$, where, for all $t \in u$, $O_t \in \Sigma$. This set Σ_u lets us describe observations using assessments of the form ($Y_t \in O_t$ for all $t \in u$). For any $O_u \in \Sigma_u$ and $x_u \in \mathcal{X}_u$, we also adopt the shorthand notation $P_{\mathcal{Y}|\mathcal{X}}(O_u|x_u) := \prod_{t \in u} P_{\mathcal{Y}|\mathcal{X}}(O_t|x_t)$.

3.2 Augmented Stochastic Processes

We now use this notion of an output model to define the stochastic model P that corresponds to a—precise—continuous-time *hidden* stochastic process. So, consider some fixed output model $(\mathcal{Y}, \Sigma, P_{\mathcal{Y}|\mathcal{X}})$, some fixed continuous-time stochastic process $P_{\mathcal{X}}$ and some fixed, non-empty and finite sequence of time-points $u \in \mathcal{U}$ on which observations of the outputs may take place.

We assume that Y_t is conditionally independent of *all* other variables, given the state X_t . This means that the construction of the augmented process P is relatively straightforward; we can simply multiply $P_{\mathcal{Y}|\mathcal{X}}(\cdot|X_t)$ with any distribution $P_{\mathcal{X}}(X_t, \cdot)$ that includes X_t to obtain the joint distribution including Y_t : for any $t \in u$ and $v \in \mathcal{U}$ such that $t \notin v$, any $x_t \in \mathcal{X}_t$ and $x_v \in \mathcal{X}_v$, and any $O_t \in \Sigma$:

$$P(Y_t \in O_t, X_t = x_t, X_v = x_v) := P_{\mathcal{Y}|\mathcal{X}}(O_t|x_t)P_{\mathcal{X}}(X_t = x_t, X_v = x_v).$$

Similarly, when considering multiple output observations at once—say for the entire sequence u —then for any $v \in \mathcal{U}$ such that $u \cap v = \emptyset$, any $x_u \in \mathcal{X}_u$ and $x_v \in \mathcal{X}_v$, and any $O_u \in \Sigma_u$:

$$P(Y_u \in O_u, X_u = x_u, X_v = x_v) := P_{\mathcal{Y}|\mathcal{X}}(O_u|x_u)P_{\mathcal{X}}(X_u = x_u, X_v = x_v).$$

Other probabilities can be derived by appropriate marginalisation. We denote the resulting augmented stochastic process as $P = P_{\mathcal{Y}|\mathcal{X}} \otimes P_{\mathcal{X}}$, for the specific output model $P_{\mathcal{Y}|\mathcal{X}}$ and stochastic process $P_{\mathcal{X}}$ that were taken to be fixed in this section.

3.3 Imprecise Continuous-Time Hidden Markov Chains

An *imprecise continuous-time hidden Markov chain* (ICTHMC) is a set of augmented stochastic processes, obtained by augmenting all processes in an ICTMC with some given output model.

Definition 2 Consider any ICTMC $\mathbb{P}_{\mathcal{Q}, \mathcal{M}}$, and any output model $(\mathcal{Y}, \Sigma, P_{\mathcal{Y}|\mathcal{X}})$. Then, the corresponding imprecise continuous-time hidden Markov chain (ICTHMC) \mathcal{Z} is the set of augmented stochastic processes that is defined by $\mathcal{Z} := \{P_{\mathcal{Y}|\mathcal{X}} \otimes P_{\mathcal{X}} : P_{\mathcal{X}} \in \mathbb{P}_{\mathcal{Q}, \mathcal{M}}\}$. The lower expectation with respect to \mathcal{Z} will be denoted by $\underline{\mathbb{E}}_{\mathcal{Z}}$.

Note that we leave the parameters \mathcal{M} , \mathcal{Q} and $P_{\mathcal{Y}|\mathcal{X}}$ implicit in the notation of the ICTHMC \mathcal{Z} —we will henceforth take these parameters to be fixed.

Also, the output model is taken to be precise, and shared by all processes in the set. One further generalisation that we aim to make in the future is to allow for an imprecise specification of this output model. However, this would force us into choosing an appropriate notion of independence; e.g., whether to enforce the independence assumptions made in Section 3.2, leading to strong or complete independence, or to only enforce the lower envelopes to have these independence properties, leading to epistemic irrelevance. It is currently unclear which choice should be preferred, e.g. with regard to computability, so at present we prefer to focus on this simpler model.

4. Updating the Model

Suppose now that we have observed that some event $(Y_u \in O_u)$ has taken place, with $O_u \in \Sigma_u$. We here use the terminology that we *update* our model with these observations, after which the updated model reflects our revised beliefs about some quantity of interest. These updated beliefs, about some function $f \in \mathcal{L}(\mathcal{X}_v)$, say, are then denoted by $\mathbb{E}_P[f(X_v) | Y_u \in O_u]$ or $\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v) | Y_u \in O_u]$, depending on whether we are considering a precise or an imprecise model. In this section, we provide definitions and alternative expressions for such updated (lower) expectations.

4.1 Observations with Positive (Upper) Probability

When our assertion $(Y_u \in O_u)$ about an observation at time points u has positive probability, we can—in the precise case—update our model by application of Bayes’ rule. The following gives a convenient expression for the updated expectation $\mathbb{E}_P[f(X_v) | Y_u \in O_u]$, which makes use of the independence assumptions in Section 3.2 for augmented stochastic processes.

Proposition 3 Let P be an augmented stochastic process and consider any $u, v \in \mathcal{U}$, $O_u \in \Sigma_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. Then the updated expectation is given by

$$\mathbb{E}_P[f(X_v) | Y_u \in O_u] := \sum_{x_v \in \mathcal{X}_v} f(x_v) \frac{P(X_v = x_v, Y_u \in O_u)}{P(Y_u \in O_u)} = \frac{\mathbb{E}_{P_{\mathcal{Z}}}[f(X_v) P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)]}{\mathbb{E}_{P_{\mathcal{Z}}}[P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)]},$$

whenever $P(Y_u \in O_u) = \mathbb{E}_{P_{\mathcal{Z}}}[P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)] > 0$, and is left undefined, otherwise.

Having defined above how to update all the precise models $P \in \mathcal{Z}$, we will now update the imprecise model through *regular extension* (Walley, 1991). This corresponds to simply discarding from \mathcal{Z} those precise models that assign zero probability to $(Y_u \in O_u)$, updating the remaining models, and then computing their lower envelope.

Definition 4 Let \mathcal{Z} be an ICHMC and consider any $u, v \in \mathcal{U}$, $O_u \in \Sigma_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. Then the updated lower expectation is defined by

$$\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v) | Y_u \in O_u] := \inf\{\mathbb{E}_P[f(X_v) | Y_u \in O_u] : P \in \mathcal{Z}, P(Y_u \in O_u) > 0\},$$

whenever $\bar{P}_{\mathcal{Z}}(Y_u \in O_u) = \bar{\mathbb{E}}_{\mathcal{Z}, \mathcal{M}}[P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)] > 0$, and is left undefined, otherwise.

As is well known, the updated lower expectation that is obtained through regular extension satisfies Walley's *generalised Bayes' rule* (Walley, 1991). The following proposition gives an expression for this generalised Bayes' rule, rewritten using some of the independence properties of the model. We will shortly see why this expression is useful from a computational perspective.

Proposition 5 Let \mathcal{Z} be an ICHMC and consider any $u, v \in \mathcal{U}$, $O_u \in \Sigma_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. Then, if $\bar{P}_{\mathcal{Z}}(Y_u \in O_u) = \bar{\mathbb{E}}_{\mathcal{Z}, \mathcal{M}}[P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)] > 0$, the quantity $\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v) | Y_u \in O_u]$ satisfies

$$\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v) | Y_u \in O_u] = \max\{\mu \in \mathbb{R} : \underline{\mathbb{E}}_{\mathcal{Z}, \mathcal{M}}[P_{\mathcal{Y}|\mathcal{X}}(O_u | X_u)(f(X_v) - \mu)] \geq 0\}.$$

4.2 Uncountable Outcome Spaces, Point Observations, and Probability Zero

An important special case where observations have probability zero for all precise models, but where we can still make informative inferences, is when we have an uncountable outcome space \mathcal{Y} and the observations are points $y_u \in \mathcal{Y}_u$ —i.e., when Y_u is continuous. In this case, it is common practice to define the updated expectation $\mathbb{E}_P[f(X_v) | Y_u = y_u]$ as a limit of *conditional* expectations, where each conditioning event is an increasingly smaller region around this point y_u . We will start by formalising this idea in a relatively abstract way, but will shortly make this practicable. For the sake of intuition, note that we are working towards the introduction of probability density functions.

Fix any $P \in \mathcal{Z}$, consider any $y_u \in \mathcal{Y}_u$ and choose a sequence $\{O_u^i\}_{i \in \mathbb{N}}$ of events in Σ_u which shrink to y_u —i.e., such that $O_u^i \supseteq O_u^{i+1}$ for all $i \in \mathbb{N}$, and such that $\cap_{i \in \mathbb{N}} O_u^i = \{y_u\}$. We then define

$$\mathbb{E}_P[f(X_v) | Y_u = y_u] := \lim_{i \rightarrow +\infty} \mathbb{E}_P[f(X_v) | Y_u \in O_u^i]. \quad (5)$$

This limit exists if there is a sequence $\{\lambda_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ such that, for every $x_u \in \mathcal{X}_u$, the limit

$$\phi_u(y_u | x_u) := \lim_{i \rightarrow +\infty} \frac{P_{\mathcal{Y}|\mathcal{X}}(O_u^i | x_u)}{\lambda_i}$$

exists, is real-valued—in particular, finite—and satisfies $\mathbb{E}_{P_{\mathcal{Z}}}[\phi_u(y_u | X_u)] > 0$:

Proposition 6 Let P be an augmented stochastic process and consider any $u, v \in \mathcal{U}$, $y_u \in \mathcal{Y}_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. For any $\{O_u^i\}_{i \in \mathbb{N}}$ in Σ_u that shrinks to y_u , if for some $\{\lambda_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ the quantity $\phi_u(y_u | X_u)$ exists, is real-valued, and satisfies $\mathbb{E}_{P_{\mathcal{Z}}}[\phi_u(y_u | X_u)] > 0$, then

$$\mathbb{E}_P[f(X_v) | Y_u = y_u] := \lim_{i \rightarrow +\infty} \mathbb{E}_P[f(X_v) | Y_u \in O_u^i] = \frac{\mathbb{E}_{P_{\mathcal{Z}}}[f(X_v)\phi_u(y_u | X_u)]}{\mathbb{E}_{P_{\mathcal{Z}}}[\phi_u(y_u | X_u)]}. \quad (6)$$

Note that $\mathbb{E}_P[f(X_v) | Y_u = y_u]$ is clearly dependent on the exact sequence $\{O_u^i\}_{i \in \mathbb{N}}$. Unfortunately, this is the best we can hope for at the level of generality that we are currently dealing with. For brevity, we nevertheless omit from the notation the updated expectation's dependency on this sequence. However, as we will explain below, this should not be problematic for most practical applications.

It is also useful to note that $\phi_u(y_u|x_u)$ can often be constructed ‘‘piecewise’’. That is, if for every $t \in u$ there is a sequence $\{\lambda_{t,i}\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ such that, for all $x_t \in \mathcal{X}_t$,

$$\phi_t(y_t|x_t) := \lim_{i \rightarrow +\infty} \frac{P_{\mathcal{Y}|\mathcal{X}}(O_t^i|x_t)}{\lambda_{t,i}}$$

exists and is real-valued, then choosing $\{\lambda_i\}_{i \in \mathbb{N}}$ as $\lambda_i := \prod_{t \in u} \lambda_{t,i}$ yields $\phi_u(y_u|x_u) = \prod_{t \in u} \phi_t(y_t|x_t)$.

Now, to make the above practicable, we can for example assume that if \mathcal{Y} is uncountable, then it is the set $\mathcal{Y} = \mathbb{R}^d$, for some $d \in \mathbb{N}$, and that Σ is the Borel σ -algebra on \mathbb{R}^d . For each $x \in \mathcal{X}$, we then assume that the measure $P_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is induced by some given *probability density function*: a measurable function $\psi(\cdot|x) : \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ such that $\int_{\mathcal{Y}} \psi(y|x) dy = 1$ and, for every $O \in \Sigma$,

$$P_{\mathcal{Y}|\mathcal{X}}(O|x) := \int_O \psi(y|x) dy,$$

where the integrals are understood in the Lebesgue sense.

Then choose any $y_u \in \mathcal{Y}_u$, any $t \in u$, any sequence $\{O_t^i\}_{i \in \mathbb{N}}$ of open balls in \mathcal{Y}_t that are centred on, and shrink to, y_t , and fix any $x_u \in \mathcal{X}_u$. If $\psi(\cdot|x_t)$ is continuous at y_t , it can be shown that

$$\phi_t(y_t|x_t) = \lim_{i \rightarrow +\infty} \frac{P_{\mathcal{Y}|\mathcal{X}}(O_t^i|x_t)}{\lambda(O_t^i)} = \psi(y_t|x_t), \quad (7)$$

where $\lambda(O_t^i)$ denotes the Lebesgue measure of O_t^i . So, we can construct the sequence $\{O_u^i\}_{i \in \mathbb{N}}$ such that every $O_u^i := \prod_{t \in u} O_t^i$, with each O_t^i chosen as above. If we then choose the sequence $\{\lambda_i\}_{i \in \mathbb{N}}$ as $\lambda_i := \prod_{t \in u} \lambda(O_t^i)$ for each $i \in \mathbb{N}$, we find $\phi_u(y_u|x_u) = \prod_{t \in u} \phi_t(y_t|x_t) = \prod_{t \in u} \psi(y_t|x_t)$, provided that each $\phi_t(y_t|x_t)$ satisfies Equation (7). It can now be seen that, under these assumptions, the right-hand side of Equation (6) is simply the well-known Bayes’ rule for (finite) mixtures of densities.

In most practical applications, therefore, the function $\phi_u(\cdot|x_u)$ is known explicitly; one may assume, for example, that Y_t follows a Normal distribution with parameters depending on X_t , and the functions $\phi_t(\cdot|x_t)$ —and by extension, $\phi_u(\cdot|x_u)$ —then follow directly by identification with $\psi(\cdot|x_t)$. Furthermore, arguably, most of the density functions that one encounters in practice will be continuous and strictly positive at y_t . This guarantees that the limit in Equation (7) exists, and largely solves the interpretation issue mentioned above: when $\phi_u(y_u|X_u) = \prod_{t \in u} \psi(y_t|X_t)$ is continuous and positive at y_u , $\mathbb{E}_P[f(X_v)|Y_u = y_u]$ exists and is the same for almost³ all sequences $\{O_u^i\}_{i \in \mathbb{N}}$.

Moving on, note that if $\phi_u(y_u|X_u)$ exists and satisfies $\underline{\mathbb{E}}_{\mathcal{D}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$, then the updated expectation $\mathbb{E}_P[f(X_v)|Y_u = y_u]$ is well-defined for every $P \in \mathcal{Z}$. Hence, we can then update the imprecise model by updating each of the precise models that it consists of.

Definition 7 Let \mathcal{Z} be an ICHMC and consider any $u, v \in \mathcal{U}$, $y_u \in \mathcal{Y}_u$, and $f \in \mathcal{L}(\mathcal{X}_v)$. For any $\{O_u^i\}_{i \in \mathbb{N}}$ in Σ_u that shrinks to y_u , if for some $\{\lambda_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ the quantity $\phi_u(y_u|X_u)$ exists and is real-valued, the updated lower expectation is defined by

$$\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v)|Y_u = y_u] := \inf\{\mathbb{E}_P[f(X_v)|Y_u = y_u] : P \in \mathcal{Z}\},$$

whenever $\underline{\mathbb{E}}_{\mathcal{D}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$, and is left undefined, otherwise.

3. It suffices if, for all $t \in u$, there is a sequence of open balls $\{B_t^i\}_{i \in \mathbb{N}}$ in \mathcal{Y} that shrinks to y_t such that, for all $i \in \mathbb{N}$, O_t^i has positive Lebesgue measure and is contained in B_t^i .

Similar to the results in Section 4.1, this updated lower expectation satisfies a “generalised Bayes’ rule for mixtures of densities”, in the following sense.

Proposition 8 *Let \mathcal{Z} be an ICHMC and consider any $u, v \in \mathcal{U}$, $y_u \in \mathcal{Y}_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. For any $\{O_u^i\}_{i \in \mathbb{N}}$ in Σ_u that shrinks to y_u , if for some $\{\lambda_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ the quantity $\phi_u(y_u|X_u)$ exists, is real-valued, and satisfies $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$, then*

$$\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v)|Y_u = y_u] = \max \left\{ \mu \in \mathbb{R} : \underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)(f(X_v) - \mu)] \geq 0 \right\}. \quad (8)$$

Furthermore, this updated imprecise model can be given an intuitive limit interpretation.

Proposition 9 *Let \mathcal{Z} be an ICHMC and consider any $u, v \in \mathcal{U}$, $y_u \in \mathcal{Y}_u$ and $f \in \mathcal{L}(\mathcal{X}_v)$. For any $\{O_u^i\}_{i \in \mathbb{N}}$ in Σ_u that shrinks to y_u , if for some $\{\lambda_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ the quantity $\phi_u(y_u|X_u)$ exists, is real-valued, and satisfies $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$, then $\underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v)|Y_u = y_u] = \lim_{i \rightarrow +\infty} \underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v)|Y_u \in O_u^i]$.*

Now, recall that the requirement $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$ for updating the imprecise model is a sufficient condition to guarantee that *all* the precise updated models are well-defined. However, one may wonder whether it is also possible to update the imprecise model under weaker conditions. Indeed, one obvious idea would be to define the updated model more generally as

$$\underline{\mathbb{E}}_{\mathcal{Z}}^R[f(X_v)|Y_u = y_u] := \inf \left\{ \underline{\mathbb{E}}_P[f(X_v)|Y_u = y_u] : P \in \mathcal{Z}, \underline{\mathbb{E}}_{P_{\mathcal{Z}}}[\phi_u(y_u|X_u)] > 0 \right\},$$

whenever $\overline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$; this guarantees that *some* of the precise updated models are well-defined. This updated lower expectation satisfies the same generalised Bayes’ rule as above, i.e. the right-hand side of Equation (8) is equal to $\underline{\mathbb{E}}_{\mathcal{Z}}^R[f(X_v)|Y_u = y_u]$ whenever $\overline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$. However, the limit interpretation then fails to hold, in the sense that it is possible to construct an example where $\underline{\mathbb{E}}_{\mathcal{Z}}^R[f(X_v)|Y_u = y_u] \neq \lim_{i \rightarrow +\infty} \underline{\mathbb{E}}_{\mathcal{Z}}[f(X_v)|Y_u \in O_u^i]$, with $\overline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] > 0$ but $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\phi_u(y_u|X_u)] = 0$. We feel that this makes this more general updating scheme somewhat troublesome from an interpretation (and hence philosophical) point of view.

On the other hand, we recall that the existence of $\phi_u(y_u|X_u)$ and the positivity of $\underline{\mathbb{E}}_{P_{\mathcal{Z}}}[\phi_u(y_u|X_u)]$ are necessary and sufficient conditions for the limit in Equation (5) to exist and be computable using Equation (6). However, these conditions are sufficient but non-necessary for that limit to simply exist. Therefore, a different way to generalise the imprecise updating method would be

$$\underline{\mathbb{E}}_{\mathcal{Z}}^L[f(X_v)|Y_u = y_u] := \inf \left\{ \underline{\mathbb{E}}_P[f(X_v)|Y_u = y_u] : P \in \mathcal{Z}, \underline{\mathbb{E}}_P[f(X_v)|Y_u = y_u] \text{ exists} \right\},$$

whenever $\{P \in \mathcal{Z} : \underline{\mathbb{E}}_P[f(X_v)|Y_u = y_u] \text{ exists}\} \neq \emptyset$. We conjecture that this updated model *does* satisfy the limit interpretation, but on the other hand, it is possible to show that this, in turn, no longer satisfies the above generalised Bayes’ rule. That makes this updating scheme somewhat troublesome from a practical point of view because, as we discuss below, the expression in Equation (8) is crucial for our method of efficient computation of the updated lower expectation.

5. Inference Algorithms

In the previous section, we have seen that we can use the generalised Bayes’ rule for updating our ICHMC with some given observations. From a computational point of view, this is particularly useful because, rather than having to solve the non-linear optimisation problems in Definitions 4 or 7 directly, we can focus on evaluating the function $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[P_{\mathcal{Y}|\mathcal{X}}(O_u|X_u)(f(X_v) - \mu)]$, or its

density-analogue, for some fixed value of μ . Finding the updated lower expectation is then a matter of finding the maximum value of μ for which this quantity is non-negative. As we will discuss in Section 5.1, this is a relatively straightforward problem to solve numerically.

Therefore, in order for this approach to be computationally tractable, we require efficient algorithms that can evaluate this quantity for a given value of μ . In Section 5.2, we provide such an algorithm for the important case where the function f depends on a single time-point.

We first generalise the problem so that these results are applicable both for observations of the form ($Y_u \in O_u$), and for point-observations ($Y_u = y_u$) in an uncountable outcome space. Recall that

$$P_{\mathcal{Y}|\mathcal{X}}(O_u|X_u) = \prod_{t \in u} P_{\mathcal{Y}|\mathcal{X}}(O_t|X_t) \quad \text{and} \quad \phi_u(y_u|X_u) = \prod_{t \in u} \phi_t(y_t|X_t).$$

In both cases, we can rewrite this expression as $\prod_{t \in u} g_t(X_t)$, where, for all $t \in u$, $g_t \in \mathcal{L}(\mathcal{X}_t)$ and $g_t \geq 0$. The function of interest is then $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[(\prod_{t \in u} g_t(X_t))(f(X_v) - \mu)]$ and the sign conditions in Propositions 5 and 8 reduce to $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$ and $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$, respectively.

5.1 Solving the Generalised Bayes' Rule

Finding the maximum value of μ for which the function of interest in the generalised Bayes' rule is non-negative, is relatively straightforward numerically. This is because this function, parameterised in μ , is very well-behaved. The proposition below explicitly states some of its properties. These are essentially well-known, and can also be found in other work; see, e.g., (De Bock, 2015, Section 2.7.3). The statement below is therefore intended to briefly recall these properties, and is stated in a general form where we can also use it when working with densities.

Proposition 10 *Let $\mathbb{P}_{\mathcal{Q},\mathcal{M}}$ be an ICTMC and consider any $u, v \in \mathcal{U}$, any $f \in \mathcal{L}(\mathcal{X}_v)$ and, for all $t \in u$, any $g_t \in \mathcal{L}(\mathcal{X}_t)$ such that $g_t \geq 0$. Consider the function $G : \mathbb{R} \rightarrow \mathbb{R}$ that is given, for all $\mu \in \mathbb{R}$, by $G(\mu) := \underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[(\prod_{t \in u} g_t(X_t))(f(X_v) - \mu)]$. Then the following properties hold:*

- G1: *G is continuous, non-increasing, concave, and has a root, i.e. $\exists \mu \in \mathbb{R} : G(\mu) = 0$.*
- G2: *If $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$, then G is (strictly) decreasing, and has a unique root.*
- G3: *If $\underline{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] = 0$ but $\bar{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$, then G has a maximum root μ_* , satisfies $G(\mu) = 0$ for all $\mu \leq \mu_*$, and is (strictly) decreasing for $\mu > \mu_*$.*
- G4: *If $\bar{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] = 0$, then G is identically zero, i.e. $\forall \mu \in \mathbb{R} : G(\mu) = 0$.*

Note that the function G in Proposition 10 can behave in three essentially different ways. These correspond to the cases where the observed event has strictly positive probability(/density) for *all* processes in the set; to where it only has positive probability(/density) for *some* processes; and to where it has *zero* probability(/density) for *all* processes. In the first two cases—which are the important ones to apply the generalised Bayes' rule—the function is “well-behaved” enough to make finding its maximum root a fairly simple task. For instance, a standard bisection/bracketing algorithm can be applied here, known in this context as Lavine's algorithm (Cozman, 1997).

We sketch this method below. First, note that due to Propositions 5 and 8, the maximum root will always be found in the interval $[\min f, \max f]$. The properties above therefore provide us with a way to check the sign conditions for updating. That is, for any $\mu > \max f$, we see that $G(\mu) < 0$ if and only if $\bar{\mathbb{E}}_{\mathcal{Q},\mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$; similarly, for any $\mu < \min f$, we see that $G(\mu) > 0$ if and only

if $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\prod_{t \in u} g_t(X_t)] > 0$. Evaluating G at such values of μ is therefore sufficient to check the sign conditions in Propositions 5 and 8.

The algorithm now starts by setting $\mu_- := \min f$, and $\mu_+ := \max f$; if $G(\mu_+) = 0$, we know that μ_+ is the quantity of interest. Otherwise, proceed iteratively in the following way. Compute the half-way point $\mu := 1/2(\mu_+ - \mu_-)$; then, if $G(\mu) \geq 0$ set $\mu_- := \mu$, otherwise set $\mu_+ := \mu$; then repeat. Clearly, the interval $[\mu_-, \mu_+]$ still contains the maximum root after each step. The procedure can be terminated whenever $(\mu_+ - \mu_-) < \varepsilon$, for some desired numerical precision $\varepsilon > 0$. Since the width of the interval is halved at each iteration, the runtime of this procedure is $O(\log\{(\max f - \min f)\varepsilon^{-1}\})$. Methods for improving the numerical stability of this procedure can be found in (De Bock, 2015, Section 2.7.3).

5.2 Functions on a Single Time Point

Having discussed an efficient method to find the maximum root of the function $G(\mu)$ in Section 5.1, it now remains to provide an efficient method to numerically *evaluate* this function for a given value of μ . Clearly, any such method will depend on the choice of f .

We focus on a particularly useful special case, which can be used to compute the updated lower expectation of a function $f \in \mathcal{L}(\mathcal{X}_s)$ on a single time point s , given observations at time points u . If $s \notin u$, then it will be notationally convenient to define $g_s := f - \mu$, and to let $u' := u \cup \{s\}$. We can then simply focus on computing

$$\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}} \left[\left(\prod_{t \in u} g_t(X_t) \right) (f(X_s) - \mu) \right] = \underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}} \left[\prod_{t \in u'} g_t(X_t) \right].$$

On the other hand, if $s = t$ for some $t \in u$, we let $u' := u$ and replace g_t by $(f - \mu)g_t$. Clearly, the above equality then also holds; the point is simply to establish a uniform indexing notation over all time-points and functions. The right hand side of the above equality can now be computed using the following dynamic programming technique.

For all $t \in u'$, we define auxiliary functions $g_t^+, g_t^- \in \mathcal{L}(\mathcal{X}_t)$, as follows. Writing $u' = t_0, \dots, t_n$, let $g_{t_n}^+ := g_{t_n}^- := g_{t_n}$. Next, for all $i \in \{0, \dots, n-1\}$ and all $x_{t_i} \in \mathcal{X}_{t_i}$, let

$$g_{t_i}^+(x_{t_i}) := \begin{cases} g_{t_i}(x_{t_i}) \underline{\mathbb{E}}_{\mathcal{Q}}[g_{t_{i+1}}^+(X_{t_{i+1}}) | X_{t_i} = x_{t_i}] & \text{if } g_{t_i}(x_{t_i}) \geq 0, \\ g_{t_i}(x_{t_i}) \bar{\underline{\mathbb{E}}}_{\mathcal{Q}}[g_{t_{i+1}}^-(X_{t_{i+1}}) | X_{t_i} = x_{t_i}] & \text{if } g_{t_i}(x_{t_i}) < 0 \end{cases}$$

and

$$g_{t_i}^-(x_{t_i}) := \begin{cases} g_{t_i}(x_{t_i}) \bar{\underline{\mathbb{E}}}_{\mathcal{Q}}[g_{t_{i+1}}^-(X_{t_{i+1}}) | X_{t_i} = x_{t_i}] & \text{if } g_{t_i}(x_{t_i}) \geq 0, \\ g_{t_i}(x_{t_i}) \underline{\mathbb{E}}_{\mathcal{Q}}[g_{t_{i+1}}^+(X_{t_{i+1}}) | X_{t_i} = x_{t_i}] & \text{if } g_{t_i}(x_{t_i}) < 0. \end{cases}$$

Clearly, backward recursion allows us to compute all these functions in a time-complexity order that is linear in the number of time points in u' . Practically, at each step, computing the quantities $\underline{\mathbb{E}}_{\mathcal{Q}}[g_{t_{i+1}}^+(X_{t_{i+1}}) | X_{t_i} = x_{t_i}]$ and $\bar{\underline{\mathbb{E}}}_{\mathcal{Q}}[g_{t_{i+1}}^-(X_{t_{i+1}}) | X_{t_i} = x_{t_i}]$ can be done using Equation (2) and the method described in Section 2.2. Due to the results in (Krak et al., 2016), each of these quantities is computable in polynomial time. So, the total complexity of computing all these functions is clearly also polynomial. We now have the following result.

Proposition 11 *For all $t \in u'$, let g_t , g_t^+ and g_t^- be as defined above. Then the function of interest is given by $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[\prod_{t \in u'} g_t(X_t)] = \underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[g_{t_0}^+(X_{t_0})]$. Also, $\bar{\underline{\mathbb{E}}}_{\mathcal{Q}, \mathcal{M}}[\prod_{t \in u'} g_t(X_t)] = \bar{\underline{\mathbb{E}}}_{\mathcal{Q}, \mathcal{M}}[g_{t_0}^-(X_{t_0})]$.*

So, in order to evaluate the function of interest, it remains to compute $\underline{\mathbb{E}}_{\mathcal{Q}, \mathcal{M}}[g_{t_0}^+(X_{t_0})]$. Since $g_{t_0}^+$ is a function on a single time point t_0 , this can again be done in polynomial time, using Equation (4).

6. Conclusions and Future Work

We considered the problem of performing inference with *imprecise continuous-time hidden Markov chains*; an extension of *imprecise continuous-time Markov chains* obtained by augmenting them with random *output* variables, which may be either discrete or continuous. Our main result is an efficient, polynomial runtime, algorithm to compute lower expectations of functions that depend on the state-space at any given time-point, given a collection of observations of the output variables.

In future work, we intend to further generalise this model, by also allowing for imprecise output variables. Furthermore, we also aim to develop algorithms for other inference problems, such as the problem of computing updated lower expectations of functions $f \in \mathcal{L}(\mathcal{X}_v)$ that depend on more than one time-point. Similarly, we aim to investigate predictive output inferences, i.e., the lower probability/density of observations, which has uses in classification problems. Another such problem is that of estimating state-sequences given observed output-sequences—as was previously done for (discrete-time) iHMM’s (De Bock and de Cooman, 2014).

Acknowledgments

The work in this paper was partially supported by the Research Foundation - Flanders (FWO) and the H2020-MSCA-ITN-2016 UTOPIAE, grant agreement 722734. The authors would also like to thank three anonymous referees for their helpful comments and suggestions.

References

- F. Cozman. *Alternatives to Lavine’s algorithm for calculation of posterior bounds given convex sets of distributions*. Carnegie Mellon University, The Robotics Institute, 1997.
- J. De Bock. *Credal networks under epistemic irrelevance: theory and algorithms*. PhD thesis, Ghent University, 2015.
- 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.
- G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052, 2010.
- T. Krak, J. De Bock, and A. Siebes. Imprecise continuous-time Markov chains. *Under Review*. Pre-print: <https://arxiv.org/abs/1611.05796>, 2016.
- T. Krak, J. De Bock, and A. Siebes. Efficient computation of updated lower expectations for imprecise continuous-time hidden Markov chains. *Extended ArXiv pre-print*: <https://arxiv.org/abs/1702.06791>, 2017.
- D. Škulj. Efficient computation of the bounds of continuous time imprecise Markov chains. *Applied Mathematics and Computation*, 250(C):165–180, Jan. 2015.
- P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.
- W. Wei, B. Wang, and D. Towsley. Continuous-time hidden Markov models for network performance evaluation. *Performance Evaluation*, 49(1):129–146, 2002.

Credal Sum-Product Networks

Denis Deratani Mauá

Institute of Mathematics and Statistics, Universidade de São Paulo (Brazil)

DENIS.MAU@USP.BR

Fabio Gagliardi Cozman

Escola Politécnica, Universidade de São Paulo (Brazil)

FGCOZMAN@USP.BR

Diarmuid Conaty

Cassio Polpo de Campos

Queen's University Belfast (United Kingdom)

DCONATY01@QUB.AC.UK

C.DECAMPOS@QUB.AC.UK

Abstract

Sum-product networks are a relatively new and increasingly popular class of (precise) probabilistic graphical models that allow for marginal inference with polynomial effort. As with other probabilistic models, sum-product networks are often learned from data and used to perform classification. Hence, their results are prone to be unreliable and overconfident. In this work, we develop credal sum-product networks, an imprecise extension of sum-product networks. We present algorithms and complexity results for common inference tasks. We apply our algorithms on realistic classification task using images of digits and show that credal sum-product networks obtained by a perturbation of the parameters of learned sum-product networks are able to distinguish between reliable and unreliable classifications with high accuracy.

Keywords: Sum-product networks; tractable probabilistic models; credal classification.

1. Introduction

Probabilistic models are usually built so that they can be used to produce inferences, that is, to draw quantitative (probabilistic) conclusions about the domain of interest. Probabilistic graphical models such as Bayesian networks and Markov Networks (Koller and Friedman, 2009; Darwiche, 2009) allow complex uncertain knowledge to be modeled succinctly; however, producing inferences with them is notoriously hard (Cooper, 1990; Roth, 1996; Darwiche, 2009).

Sum-Product Networks (SPNs) are a relatively new class of (precise) probabilistic graphical models that allow marginal inference in linear time in their size (Poon and Domingos, 2011). They have received increasing popularity in applications of machine learning due to their ability to represent complex and highly multidimensional distributions (Poon and Domingos, 2011; Cheng et al., 2014; Nath and Domingos, 2016; Amer and Todorovic, 2016). An SPN encodes an arithmetic circuit whose evaluation produces a marginal inference (Darwiche, 2003). The internal nodes of a SPN perform (weighted) sums and multiplications, while the leaves represent variable assignments. The sum nodes can be interpreted as latent variables, while the product nodes can be interpreted as encoding context-sensitive probabilistic independences. Thus, SPNs can be seen as a class of complex mixture distributions with tractable inference (Zhao et al., 2015; Peharz et al., 2016).

Imprecise probability models extend precise probabilistic models to accommodate the representation of incomplete and indeterminate knowledge (Walley, 1991; Augustin et al., 2014). For example, (separately specified) credal networks extend Bayesian networks by allowing sets of conditional probability measures to be associated with nodes in lieu of conditional probability measures (Cozman, 2000, 2005).

In this work, we develop the *Credal Sum-Product Networks* (CSPNs), a class of imprecise probability models which extend SPNs to the imprecise case. A CSPN is simply an SPN where the weights associated with sum nodes (i.e., the numerical parameters of the model) are allowed to vary inside a closed and convex set. Among other things, CSPNs can be used to analyze the robustness of conclusions supported by SPNs.

We begin by presenting some basic facts about SPNs in Section 2. Then in Section 3 we derive polynomial-time algorithms for computing upper and lower bounds on the marginal (unconditional) probability of an event; we also present a polynomial-time algorithm for computing upper and lower expectations when the structure is constrained so that every internal node has at most one parent. As many learning algorithms produce networks of this type (Gens and Domingos, 2013; Rooshenas and Lowd, 2014), this result is quite important and useful. We show that performing credal classification (i.e., verifying whether a class value dominates another value under maximality) is coNP-complete when the number of class values is unbounded. Since this task can be posed as the computation of a lower expectation, this result also shows hardness of computing expectation bounds on arbitrary (multivariate) functions. We show empirically in Section 4 that CSPNs are effective in assessing the reliability to classifications made with SPNs learned from data. Finally, we conclude the paper with a review of our contributions and some ideas for the future in Section 5.

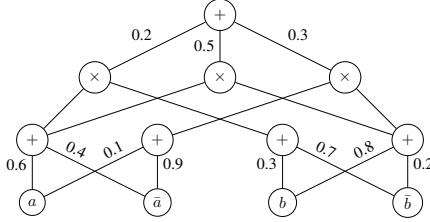
2. Sum-Product Networks

We use capital letters to denote both random variables and random vectors, with the former usually being indexed by a subscript: e.g., X_i . If X is a random vector, we call the set composed of the random variables in X its *scope*. The scope of a function that takes a random vector X as argument is the scope of X . In this work, we consider only finite-valued random variables, and leave the extension to random variables with infinite domains as future work.

We associate every random variable X_i taking values in $\{0, \dots, c_i - 1\}$ with a set of *indicator variables* $\{\lambda_{ij} : j = 0, \dots, c_i - 1\}$, each taking on values 0 and 1. If X_i is binary, we write x_i (resp., \bar{x}_i) to denote λ_{i1} (resp., $\lambda_{i0}\right)$. Any discrete multivariate distribution $P(X_V)$ can be written as a multilinear function of the corresponding indicator variables by $S(\lambda) = \sum_{x_V} \mathbb{P}(X_V = x_V) \prod_{i \in V} \lambda_{ix_i}$. For example, a Bernoulli distribution can be written as $S(x, \bar{x}) = \Pr(X = 1)x + \Pr(X = 0)\bar{x}$.

A SPN is a concise representation of the multilinear function representing a probability distribution. More formally, a SPN is a weighted rooted directed acyclic graph where internal nodes are associated to either sum or product operations and leaves are associated with indicator variables. Every arc from a sum node i to a child j is associated with a nonnegative weight w_{ij} , and every arc leaving a product node has weight one. The scope of a leaf node of the network is the respective random variable; the scope of an internal node is the union of the scopes of its children. If \mathbf{w} are the weights of a subnetwork $S_{\mathbf{w}}$, we denote by \mathbf{w}_i the weights in the subnetwork $S_{\mathbf{w}_i}^i$ rooted at node i , and by w_i the vector of weights w_{ij} associated with arcs from i to children j . Figure 1 shows an example of a SPN with scope $\{A, B\}$, where A and B are binary variables.

An SPN satisfies the following properties (Poon and Domingos, 2011; Peharz et al., 2015): (i) every indicator variable appears in at most one leaf node; (ii) the scope of any two children of a sum node are identical (completeness); (iii) the scopes of any two children of a product node are disjoint (decomposition); (iv) the sum of the weights associated with any sum node is one (normalization). Every discrete distribution can be represented by a SPN, and any SPN satisfying the those properties represents a valid distribution.

Figure 1: A sum-product network over binary random variables A and B .

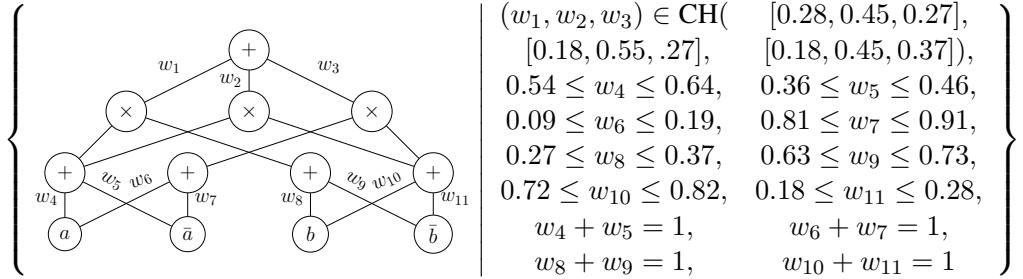
The evaluation of a SPN for a given configuration λ of the indicator variables is performed from the leaves toward the root. The leaves (indicator variables) propagate up their corresponding value (either 0 or 1) in the configuration λ . Sum (resp., product) nodes propagate the weighted sum (resp., product) of the values of their children multiplied by the corresponding arc weights. For example, the value of the SPN $S_w(a, \bar{a}, b, \bar{b})$ in Figure 1 at the point $\lambda = (1, 0, 0, 1)$ is 0.15 and corresponds to $\mathbb{P}(A = 1, B = 0)$, and for $\lambda = (1, 0, 1, 1)$ is 0.45 and corresponds to $\mathbb{P}(A = 1)$.

Let $\mathcal{E} \subseteq \{1, \dots, n\}$ be an index set, and $X_{\mathcal{E}}$ be a random vector of scope $\{X_i : i \in \mathcal{E}\}$. The marginal probability of some evidence $\{X_i = e_i : i \in \mathcal{E}\}$ induced by a SPN S can be obtained by evaluating the network at λ that is consistent with the evidence, and assigns one to all other indicator variables (Poon and Domingos, 2011). That is, $\lambda_{ij} = 0$ if $i \in \mathcal{E}$ and $e_i \neq j$, and $\lambda_{ij} = 1$ otherwise. Thus marginal probabilities can be computed in time linear in the network size (the number of nodes, arcs and weights). For example, the marginal probability $\mathbb{P}(B = 0) = 0.3$ induced by the SPN in Figure 1 can be obtained by evaluating $S(a, \bar{a}, b, \bar{b})$ at $\lambda = (1, 1, 0, 1)$. Conditional probabilities can either be obtained by evaluating the network at query and evidence (then dividing the result) or by applying Darwiche’s differential approach (Darwiche, 2003; Peharz et al., 2016).

A great deal of algorithms have been devised to “learn” SPNs from data (Dennis and Ventura, 2012; Gens and Domingos, 2013; Peharz et al., 2013, 2014; Lee et al., 2014; Rooshenas and Lowd, 2014; Dennis and Ventura, 2015; Adel et al., 2015; Rahman and Gogate, 2016). Most learning algorithms employ a greedy search on the space of SPNs augmenting an SPN in either a top-down or bottom-up fashion. For instance, Gens and Domingos (2013)’s algorithm starts with a single node representing the entire dataset, and recursively adds product and sum nodes that divide the dataset into smaller datasets until a stopping criterion is met. Product nodes are created using group-wise independence tests, while sum nodes are created performing clustering on the row instances. The weights associated with sum nodes are learned as the proportion of instances assigned to a cluster.

3. Credal Sum-Product Networks

Let S_w denote a SPN whose weights are w . We can obtain an imprecise sum-product network by allowing the weights w to vary in some space, subject to the constraint that they still define a SPN. More formally, a *Credal Sum-Product Network* (CSPN) is a set $\{S_w : w \in \mathcal{C}\}$, where \mathcal{C} is the Cartesian product of probability simplexes, and each probability simplex constrains only the weights associated with a single sum node. It is clear that a SPN is a CSPN where weights take values in a singleton \mathcal{C} , and that every choice of weights w inside \mathcal{C} specifies a SPN. Since each SPN induces a probability measure, the CSPN induces a *credal set*, that is, a (not necessarily convex) set

Figure 2: A credal sum-product network over variables A and B .

of probability measures (Levi, 1980). Figure 2 shows a CSPN obtained by ϵ -contamination of the SPN in Figure 1, with $\epsilon = 0.1$.

A SPN can be interpreted as a bilevel bipartite Bayesian network by identifying sum nodes with latent variables whose probability distributions are obtained from the corresponding weights (Zhao et al., 2015). The network has a layer of latent variables Y_1, \dots, Y_m corresponding to sum nodes of the network, and a layer of leaf variables X_1, \dots, X_n corresponding to (scopes of) indicator variables. There is an arc $Y_j \rightarrow X_i$ if and only if X_i is in the scope of the sum node (associated with) Y_j . Each variable Y_j has as many values as children, and its (unconditional) probabilities are specified as the associated weights. The (conditional) probabilities associated with a node X_i are specified as the weights entering the corresponding indicator variable (which depend on the value of the respective latent variables). Note that a variable X_i can have a large number of parents, so that obtaining this Bayesian network is often impracticable.

We can adapt a similar argument for CSPNs: sum nodes can be interpreted as latent variables in a credal network. This network is obtained exactly as the Bayesian network, except that conditional probability distributions are replaced by conditional credal sets.

3.1 Likelihood

The most trivial inference with CSPNs is to compute the minimum and maximum values obtained at a SPN for a given value λ of the indicator variables. This computation corresponds to computing the upper and lower likelihood of evidence, and can be performed in much the same way as the computation of marginal probabilities in SPNs, with the additional extra effort of solving a linear program at each node. To see this, consider a tree-shaped CSPN $\{S_w : w \in \mathcal{C}\}$ with root r . Since the structure is a tree, the subnetworks S^1, \dots, S^k rooted at the children of a node i do not share any weights. Hence, we have that $\min_w S_w(\lambda) = \min_{w_i} \sum_j w_{ij} \min_{w_j} S_{w_j}^j(\lambda)$. Thus, the problem of computing the minimum or maximum of a value λ decomposes into smaller similar problems. A much similar argument applies to CSPNs with cycles; simply break the cycles by duplicating nodes until the structure is a tree, and perform optimizations from the leaves toward the root. Every duplicated network receives the same values from the (duplicated) children; thus the optimizations are the same whether we “tie” the weights of identical parts or not. A more formal argument is given next.

Theorem 1 Consider a CSPN $\{S_{\mathbf{w}} : \mathbf{w} \in \mathcal{C}\}$, where \mathcal{C} is the Cartesian product of finitely-generated polytopes \mathcal{C}_i , one for each sum node i . Computing $\min_{\mathbf{w}} S_{\mathbf{w}}(\lambda)$ and $\max_{\mathbf{w}} S_{\mathbf{w}}(\lambda)$ takes $O(sL)$ time, where s is the number of nodes and arcs in the shared graphical structure and L is an upper bound on the cost of solving a linear program $\min_{\mathbf{w}_i} \sum_j c_{ij} w_{ij}$ subject to $w_i \in \mathcal{C}_i$.

Proof Consider the computation of $\min_{\mathbf{w}} S_{\mathbf{w}}(\lambda)$ (the case for max is analogous), and let $1, \dots, k$ denote the sum nodes of the network. By construction, the optimization is over weight vectors $\mathbf{w} = (w_1, \dots, w_k)$, where w_i denotes the weights associated with the sum node i , and vary in a finitely-generated polytope \mathcal{C}_i . Now start at the leaves. There are no weights associated, so these nodes simply propagate their values as in SPNs. Consider a sum node i , and assume that the weights of the subnetworks at its children have been optimized (and are hence fixed). The corresponding optimization is then $\min_{\mathbf{w}} \sum_p \mathbf{w}_p \sum_j w_{ij} S_{\mathbf{w}_j}^j(\lambda) + C_{\mathbf{w}}$, where the leftmost sum is over all paths from the root to i , the inner sum is over the children j of i , and $C_{\mathbf{w}}$ contains the subnetwork formed by nodes which are neither an ancestor nor a descendant of i (hence can be optimized independently of w_i); this expression defines a linear program with (finitely many) linear constraints $w_i \in \mathcal{C}_i$. Solving this linear program takes time $O(L)$. The result follows by induction on the height of subnetworks. ■

The algorithm to compute the minimum or maximum values at a configuration λ visits nodes from leaves toward the root: at product or indicator nodes, it evaluates the corresponding expression as in SPNs; at a sum node, it builds the corresponding linear program and calls a solver. Since linear programs can be solved in polynomial time, the overall time is also polynomial in the size of the input (which includes a description for the local polytopes). This leads to the following:

Corollary 2 Computing $\min_w S_w(\lambda)$ and $\max_w S_w(\lambda)$ takes at most polynomial time in CSPNs specified by finitely-generated polytopes.

3.2 Conditional Expectations

Each choice of the weights \mathbf{w} of a CSPN $\{S_{\mathbf{w}} : \mathbf{w} \in \mathcal{C}\}$ defines a SPN and hence induces a probability measure $\mathbb{P}_{\mathbf{w}}$. We can thus use the CSPN to compute upper and lower conditional expectations:

$$\max_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(f|X_{\mathcal{E}} = e) \quad \text{and} \quad \min_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(f|X_{\mathcal{E}} = e),$$

where $X_{\mathcal{Q}}$ are known as target variables, $f : X_{\mathcal{Q}} \rightarrow \mathbb{Q}$ is a function to rational numbers and $X_{\mathcal{E}} = e$ is the evidence. We will focus on the lower expectation, since the upper expectation can be obtained from $\max_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(f|e) = -\min_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(-f|e)$. This inference is however intractable (under the common assumptions in complexity theory):

Theorem 3 Assuming that f is encoded succinctly (e.g., sparsely by its non-zero terms only), computing lower/upper conditional expectations of f in CSPNs is NP-hard.

We defer the proof to Section 3.3, where we address the case of credal classification (that can be posed as the computation of a conditional expectation). The requirement of a succinct representation for f is necessary because an exponentially large input would give too much power to any algorithm (since polynomial time in the input would allow exponential time computations).

While the general case is NP-hard, there are useful subcases with tractable inference. We now present an algorithm for the computation of lower and upper conditional expectations when the network obtained by discarding leaves is a tree and $f : X_{\mathcal{Q}} \rightarrow \mathbb{Q}$ is a univariate function. The algorithm is based on the generalized Bayes rule, and uses the fact that, for any real μ :

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(f|X_{\mathcal{E}} = e) > \mu \iff \min_{\mathbf{w}} \sum_{q \in X_{\mathcal{Q}}} (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}}(X_{\mathcal{Q}} = q, X_{\mathcal{E}} = e) > 0, \quad (1)$$

provided that $\max_{\mathbf{w}} \mathbb{P}_{\mathbf{w}}(X_{\mathcal{E}} = e) > 0$ (this can be checked in polynomial time, see Section 3.1). We can also check efficiently whether $\min_{\mathbf{w}} \mathbb{P}_{\mathbf{w}}(X_{\mathcal{E}} = e) = 0$, and decide what to do in such extreme scenarios. If we can decide Inequality (1) for any μ , then we can perform a binary search to find the value of $\min_{\mathbf{w}} \mathbb{E}_{\mathbf{w}}(f|X_{\mathcal{E}} = e)$.

Theorem 4 *Computing lower/upper conditional expectations of a variable in CSPNs takes at most polynomial time when each internal node has at most one parent.*

Proof Let λ_e be the assignment of indicator variables that is consistent with $X_{\mathcal{E}} = e$ and assigns 1 to variables not in $X_{\mathcal{E}}$. As shown before, we can efficiently compute $\max_{\mathbf{w}} S_{\mathbf{w}}(\lambda_e) = \max_{\mathbf{w}} \mathbb{P}_{\mathbf{w}}(X_{\mathcal{E}} = e)$ and $\min_{\mathbf{w}} S_{\mathbf{w}}(\lambda_e) = \min_{\mathbf{w}} \mathbb{P}_{\mathbf{w}}(X_{\mathcal{E}} = e)$. To compute a lower conditional expectation we might do a binary search to find μ such that

$$\min_{\mathbf{w}} \sum_{q \in X_{\mathcal{Q}}} (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}}(X_{\mathcal{Q}} = q, X_{\mathcal{E}} = e).$$

To simplify notation we will write $\mathbb{P}_{\mathbf{w}}(q, e)$ to denote $\mathbb{P}_{\mathbf{w}}(X_{\mathcal{Q}} = q, X_{\mathcal{E}} = e)$. Now suppose that the CSPN has a product root node 0 with children $1, \dots, k$ and (without loss of generality) only node 1 has $X_{\mathcal{Q}}$ in its scope. Then, because the scopes of children of product nodes are fully disjoint and the internal graph of the CSPN forms a tree, we have that

$$\min_{\mathbf{w}_0} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_0}(q, e_0) = \left(\min_{\mathbf{w}_1} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_1}(q, e_1) \right) \cdot \prod_{j=2}^k \mathbb{P}_{\mathbf{w}_j}^*(e_j),$$

where e_j is the evidence for \mathcal{E}_j within the scope of child j (note that \mathcal{E}_1 might be empty and e_1 would disappear), $\mathbb{P}_{\mathbf{w}_j}^*(e_j) = \max_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_j)$ in the case that $\min_{\mathbf{w}_1} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_1}(q, e_1) < 0$ and $\mathbb{P}_{\mathbf{w}_j}^*(e_j) = \min_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_j)$ otherwise. Hence, if we assume that $S_{\mathbf{w}_1}^1(\lambda) = \min_{\mathbf{w}_1} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_1}(q, e_1)$ and that $S_{\mathbf{w}_j}^j(\lambda) = \mathbb{P}_{\mathbf{w}_j}^*(e_j)$, then from the computation scheme of the CSPN for a sum node, it is clear that $S_{\mathbf{w}_0}^0(\lambda) = \min_{\mathbf{w}_0} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_0}(q, e_0)$. The assumption $S_{\mathbf{w}_j}^j(\lambda) = \mathbb{P}_{\mathbf{w}_j}^*(e_j)$ is satisfied by definition for all children j that are leaf nodes and do not contain $X_{\mathcal{Q}}$. Moreover, if the node 0 is a product node and does not have $X_{\mathcal{Q}}$ in its scope, then

$$\min_{\mathbf{w}_0} \mathbb{P}_{\mathbf{w}_0}(e_0) = \prod_{j=1}^k \min_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_j) \text{ and } \max_{\mathbf{w}_0} \mathbb{P}_{\mathbf{w}_0}(e_0) = \prod_{j=1}^k \max_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_j),$$

and so it is immediate that $\min_{\mathbf{w}_0} S_{\mathbf{w}_0}^0(\lambda) = \min_{\mathbf{w}_0} \mathbb{P}_{\mathbf{w}_0}(e_0)$ if each $S_{\mathbf{w}_j}^j(\lambda) = \min_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_j)$ (analogous for the maximization).

If node 0 is a sum node with X_Q in its scope, then because the internal graph of the CSPN is a tree and expectations are linear, we have that

$$\min_{\mathbf{w}_0} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_0}(q, e_0) = \min_{w_0} \sum_{j=1}^k w_{0,j} \cdot \min_{\mathbf{w}_j} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_j}(q, e_0),$$

where $w_0 = (w_{0,1}, \dots, w_{0,k})$ varies in the corresponding polytope specifying the weights of the current node 0 (note that \mathcal{E}_0 might be empty and e_0 would disappear). Hence, if we assume that $S_{\mathbf{w}_j}^j(\lambda) = \min_{\mathbf{w}_j} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_j}(q, e_0)$, it is immediate from the local computation of the CSPN for a sum node that $S_{\mathbf{w}_0}^0(\lambda) = \min_{\mathbf{w}_0} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_0}(q, e_0)$. If node 0 is a sum node without X_Q , then

$$\min_{\mathbf{w}_0} \mathbb{P}_{\mathbf{w}_0}(e_0) = \min_{w_0} \sum_{j=1}^k w_{0,j} \cdot \min_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_0),$$

where $w_0 = (w_{0,1}, \dots, w_{0,k})$ varies in the polytope specifying the weights of the current node 0. Again, $\min_{\mathbf{w}_0} S_{\mathbf{w}_0}^0(\lambda) = \min_{\mathbf{w}_0} \mathbb{P}_{\mathbf{w}_0}(e_0)$ if each $S_{\mathbf{w}_j}^j(\lambda) = \min_{\mathbf{w}_j} \mathbb{P}_{\mathbf{w}_j}(e_0)$ (analogous for the maximization). Finally, if node 0 is a leaf node with scope X_Q , then

$$\min_{\mathbf{w}_0} \sum_q (f(q) - \mu) \cdot \mathbb{P}_{\mathbf{w}_0}(q) = f(q') - \mu,$$

where q' is the value of the variable X_Q associated to the $\lambda_{Q,q'}$ of the leaf node. Therefore, by using these expressions, we can perform the computation recursively and obtain the desired upper or lower conditional expectation in polynomial time. ■

3.3 Credal Classification

SPNs are most often constructed to perform probabilistic classification: to assign each object the assignment that maximizes the probability of a distinguished set of variables X_C given the realization of (a subset of) the remaining variables. Since CSPNs define more than a single SPN, there is more than one such possible maximizer. Many criteria have been devised for decision making with imprecise probability models. Here we adopt a very popular one, based on the principle of maximality, often called *credal classification* in the context of probabilistic classifiers.

Given distinguished variables X_C , evidence $e = \{X_i = e_i : i \in \mathcal{E}\}$ on some variables, and a credal set \mathcal{M} , we say that an assignment c' for X_C *credally dominates* another assignment c'' if ([Zaffalon, 2002](#))

$$\min_{\mathbb{P} \in \mathcal{M}} (\mathbb{P}(X_C = c', X_{\mathcal{E}} = e) - \mathbb{P}(X_C = c'', X_{\mathcal{E}} = e)) > 0.$$

There is a special case of $\mathbb{P}(X_{\mathcal{E}} = e) = 0$ to be treated—an advantage in CSPNs is that computing lower/upper marginals is efficient. According to the above definition, an assignment c' credally dominates class value c'' if $\mathbb{P}(X_C = c' | X_{\mathcal{E}} = e) > \mathbb{P}(X_C = c'' | X_{\mathcal{E}} = e)$ for all $\mathbb{P} \in \mathcal{M}$ where these conditional probabilities are defined. In the setting of CSPNs, credal dominance amounts to establishing whether $\min_{\mathbf{w}} (S_{\mathbf{w}}(\lambda_{c'}, \lambda_e) - S_{\mathbf{w}}(\lambda_{c''}, \lambda_e)) > 0$, where $\lambda_{c'}$ (resp., $\lambda_{c''}$) is the assignment of indicator variables associated with variables X_C consistent with c' (resp., c''), and λ_e is the

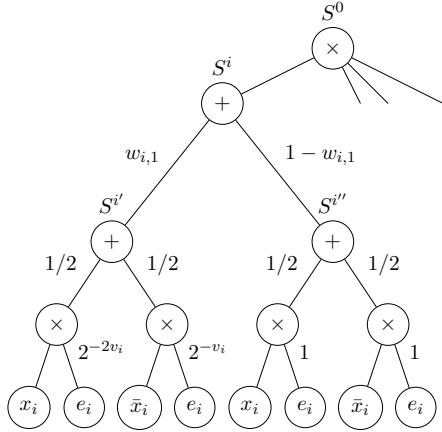


Figure 3: Fragment of the sum-product network used to solve PARTITION. We duplicate leaves for the sake of readability.

assignment of indicator variables consistent with evidence (if there are other variables, these have their indicator variables set to one to indicate their marginalization).

Lemma 5 *If we allow all weights \mathbf{w} of an SPN an additive variation $\varepsilon > 0$, then the result of $S(\lambda)$ will vary (additively) at most $O(s) \cdot \varepsilon$, where s is the number of nodes and arcs in the graphical structure. If we allow all \mathbf{w} a multiplicative variation $\varepsilon > 0$, then the result of $S(\lambda)$ will vary (multiplicatively) at most $\varepsilon^{O(s)}$.*

Proof Each sum node propagates an extra error of at most ε , while the product nodes propagate an extra error of at most $O(d) \cdot \varepsilon$, where d is its degree. So the result follows by induction. For the multiplicative error, we have that sum nodes contribute at most a factor ε to the error, while product nodes may contribute a factor $\varepsilon^{O(d)}$. Hence the overall result follows. ■

Theorem 6 *Credal classification is coNP-complete.*

Proof Membership in coNP is trivial: Given \mathbf{w} , computing $S_{\mathbf{w}}(\lambda_{c'}, \lambda_e) - S_{\mathbf{w}}(\lambda_{c''}, \lambda_e)$ is a polynomial time task. Hence, there is a polynomial certificate \mathbf{w} that confirms that $\min_{\mathbf{w}}(S_{\mathbf{w}}(\lambda_{c'}, \lambda_e) - S_{\mathbf{w}}(\lambda_{c''}, \lambda_e)) \leq 0$ if that is indeed the case, and since credal classification is the complement, membership follows.

Hardness follows by a reduction from the NP-hard problem PARTITION: Given a set of integers z_1, \dots, z_n , decide if there is a set $\mathcal{S} \subseteq \{1, \dots, n\}$ such that $\sum_{i \in \mathcal{S}} z_i = Z/2$, where $Z = \sum_i z_i$. First note that we can scale integers to become rationals in the unit interval without affecting complexity: Let $v_i = 2z_i/Z$; then set \mathcal{S} solves the original problem if and only if $\sum_{i \in \mathcal{S}} v_i = 1$.

Now, we build an CSPN over variables $X = (X_1, \dots, X_n, X_{n+1}, \dots, X_{2n})$ as in Figure 3, where the weights $w_{i,1}$ vary in $[0, 1]$, and let $\mathcal{C} = \{1, \dots, n\}$ and $\mathcal{E} = \{n+1, \dots, 2n\}$. Since the variables $X_i, i \in \mathcal{E}$, have their value fixed by the evidence e (say $X_i = 1$), we only show the corresponding value in the figure. The product node S^0 has children S^1, \dots, S^n . Note that for

$X_C = c'$, $S^{i'}$ computes 2^{-2v_i} while $S^{i''}$ computes 1; and for $X_C = c''$, $S^{i'}$ computes 2^{-v_i} and $S^{i''}$ computes 1. Because the weights are minimized at $\{0, 1\}$. Thus, we have that $S(\lambda_{c'}, \lambda_e) = 2^{-\sum_{i:w_{i,1}=1} 2v_i - n}$ and $S(\lambda_{c''}, \lambda_e) = 2^{-\sum_{i:w_{i,1}=1} v_i - n}$. Hence,

$$S_w(\lambda_{c'}, \lambda_e) - S_w(\lambda_{c''}, \lambda_e) = 2^{-n} \cdot (t^2 - t) = 2^{-n} \cdot t \cdot (t - 1), \text{ with } t = 2^{-\sum_{i:w_{i,1}=1} v_i}.$$

Now, deciding whether $\min_w(S_w(\lambda_{c'}, \lambda_e) - S_w(\lambda_{c''}, \lambda_e)) \leq -2^{-n-2}$ solves the partition problem since t would be 2^{-1} . With a small change in the model, we can move the threshold -2^{-n-2} to zero, as required in the classification problem. Therefore, credal classification is coNP-complete. However, we have to deal with the specification of 2^{-v_i} in polynomial time. To do so, we find rational numbers which approximate them, and in view of Lemma 5, we can find accurate enough results to separate between yes and no instances of PARTITION. ■

Since credal classification can be casted as the computation of the lower expectation of a univariate function, we have from Theorem 4 that:

Theorem 7 *Credal classification with a single class variable can be done in polynomial time in CSPNs when each internal node has at most one parent.*

4. Experiments

We evaluate the ability of CSPNs in distinguishing between robust and non-robust classifications in a handwritten digit recognition task. The dataset consists of 700 digitalized images of handwritten Arabic numerals ranging from 0 to 9 (70 images per digit). Each image consists of 20×30 pixels taking on values 0 and 1, and we associate every pixel with a binary variable. To assess the effect of dataset size, we consider two splits in training/test data: 50%/50% and 20%/80%. For each split, we learn a SPN from the training set using the approach discussed by Poon and Domingos (2011), and use it to classify each instance in the test set. Then, for each test instance, we find the maximum value of ϵ such that the CSPN obtained by imposing a local ϵ -contamination to each of the sum nodes produces a single classification under maximality (which is equivalent to *E-admissibility* in this case). Call this value the *classification robustness*. We repeat this procedure 10 times using different random partitions of the data into train and test parts. The curves show the accuracy (no. of correctly classified instances/no. of instances) of the SPN for instances with robustness at most a given ϵ (x-axis). The results are compiled into Figure 4. We see that the higher the robustness the greater the accuracy.

Examples of misclassified instances are given in Figure 5. For comparison, we also analyze a different approach to measure robustness: we compute the difference between the probability of the most probable class and the second most probable class. As we see in the figure, this measure correlates poorly with the accuracy.

In order to give a more quantitative perspective of the robustness value, we present in Table 1 some descriptive statistics for correctly and wrongly classified instances, using either robustness measure. We see a much clearer separation of the robustness values between correctly and incorrectly classified instances using the CSPN approach instead of the “best minus second best” probability approach.

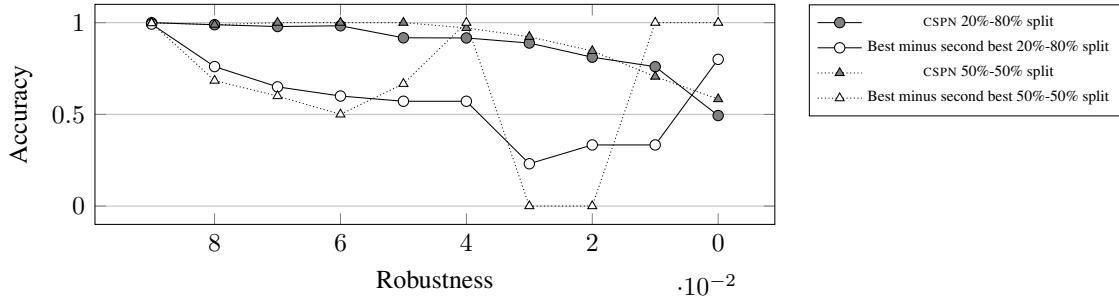


Figure 4: Average classification accuracy for instance below the given robustness (the x-axis shows the values times a constant 20 to be visually compatible with the probabilities), as explained in the text.

88885544433332

Figure 5: Examples of misclassified instances. Usually, number 8 is misclassified as 3, number 4 as number 1, and number 3 as 5 and 8. These classifications obtained low robustness values as given by the CSPN analysis (< 0.01), rightfully indicating the lack of statistical support.

5. Conclusion

Sum-product networks are tractable probabilistic graphical models that have shown competitive results in many machine learning tasks. In this work we developed the credal sum-product networks, a new class of imprecise probabilistic graphical models that extend sum-product networks to accommodate imprecision in the numerical parameters. We described algorithms and complexity for common inference tasks such as computing upper and lower bounds on the probability of evidence, computing conditional expectations and performing credal classification. We performed experiments that showed that credal sum-product networks can distinguish between reliable and unreliable classifications of sum-product networks, thus providing an important tool for the analysis of

Robustness Measure	CSPN		Best minus second best	
	Correct	Wrong	Correct	Wrong
1st quartile	0.0255	0.0012	0.0909	0.0627
median	0.0363	0.0029	0.0909	0.0880
3rd quartile	0.0461	0.0049	0.0909	0.0905
maximum	0.1524	0.0199	0.3333	0.3333
mean (std.dev.)	0.0369 ± 0.017	0.0043 ± 0.004	0.0976 ± 0.04	0.1042 ± 0.09

Table 1: Robustness values for split of 50% training and 50% testing, repeated 10 times. Overall classification accuracy of 99.31%.

such models. There are many open questions. We showed that verifying maximality is coNP-hard when the query involves multiple variables, but the problem admits an efficient solution if internal nodes have at most one parent and the test is over a single variable. In fact, we have showed a polynomial algorithm for computing conditional expectations in networks of that structure, which subsumes maximality. There remains to establish the complexity of verifying maximality and computing conditional expectations for single variables in general structures, and for multiple variables in tree-shaped networks. Our experiments here, however promising, are preliminary. In the future, we intend to perform a more thorough examination of the credal sum-product networks applied to robust analysis of sum-product networks.

Acknowledgments

This work was partially supported by CNPq (grants 308433/2014-9, 303920/2016-5) and FAPESP (grants 2016/01055-1). We greatly thank Renato Geh for making his source code and the handwritten digits dataset publicly available (at <http://github.com/RenatoGeh/gospn>).

References

- T. Adel, D. Balduzzi, and A. Ghodsi. Learning the structure of sum-product networks via an svd-based algorithm. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 32–41, 2015.
- M. R. Amer and S. Todorovic. Sum product networks for activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):800–813, 2016.
- T. Augustin, F. P. A. Coolen, G. De Cooman, and M. C. M. Troffaes. *Introduction to Imprecise Probabilities*. 2014.
- W.-C. Cheng, S. Kok, H. V. Pham, H. L. Chieu, and K. M. A. Chai. Language modeling with sum-product networks. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH’14)*, 2014.
- G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2–3):393–405, 1990.
- F. G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.
- F. G. Cozman. Graphical models for imprecise probabilities. In *International Journal of Approximate Reasoning*, volume 39, pages 167–184, 2005.
- A. Darwiche. A differential approach to inference in bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.
- A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.
- A. Dennis and D. Ventura. Learning the architecture of sum-product networks using clustering on variables. In *Advances in Neural Information Processing Systems 25*, pages 2042–2050, 2012.

- A. Dennis and D. Ventura. Greedy structure search for sum-product networks. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 932–938, 2015.
- R. Gens and P. Domingos. Learning the structure of sum-product networks. In *Proc. 30th Int. Conf. on Mach. Learning*, pages 873–880, 2013.
- D. Koller and N. Friedman. *Probabilistic Graphical Models*. The MIT press, 2009.
- S.-W. Lee, C. Watkins, and B.-T. Zhang. Non-Parametric Bayesian Sum-Product Networks, 2014.
- I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.
- A. Nath and P. Domingos. Learning tractable probabilistic models for fault localization. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1294–1301, 2016.
- R. Peharz, B. C. Geiger, and F. Pernkopf. Greedy part-wise learning of sum-product networks. In *Machine Learining and Knowledge Discovery in Databases*, volume 8189 LNAI, pages 612–627, 2013.
- R. Peharz, R. Gens, and P. Domingos. Learning selective sum-product networks. In *ICML Workshop on Learning Tractable Probabilistic Models*, volume 32, 2014.
- R. Peharz, S. Tschiatschek, F. Pernkopf, and P. Domingos. On theoretical properties of sum-product networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 744–752, 2015.
- R. Peharz, R. Gens, F. Pernkopf, and P. Domingos. On the latent variable interpretation in sum-product networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2016.
- H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *Proc. 27th Conf. on Uncertainty in Artif. Intell.*, pages 337–346, 2011.
- T. Rahman and V. Gogate. Merging strategies for sum-product networks: From trees to graphs. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 617–626, 2016.
- A. Rooshenas and D. Lowd. Learning sum-product networks with direct and indirect variable interactions. In *Proceedings of the 31st International Conference on Machine Learning*, pages 710–718, 2014.
- D. Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1):273–302, 1996.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.
- H. Zhao, M. Melibari, and P. Poupart. On the relationship between sum-product networks and bayesian networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 116–124, 2015.

Game Solutions, Probability Transformations and the Core

Enrique Miranda

Ignacio Montes

Dep. of Statistics and Operations Research

University of Oviedo (Spain)

MIRANDAENRIQUE@UNIOVI.ES

IMONTES@UNIOVI.ES

Abstract

We investigate the role of some game solutions, such the Shapley and the Banzhaf values, as probability transformations of lower probabilities. The first one coincides with the pignistic transformation proposed in the Transferable Belief Model; the second one is not efficient in general, leading us to propose a normalized version. We consider a number of particular cases of lower probabilities: minitive measures, coherent lower probabilities, as well as the lower probabilities induced by comparative or distortion models. For them, we provide some alternative expressions of the transformations and study when they belong to the core of the lower probability.

Keywords: game solutions; probability transformations; lower probabilities; belief functions; core; Shapley value; Banzhaf value; pignistic transformation.

1. Introduction

One important problem within imprecise probability theory is that of eliciting a (precise) probability measure from an imprecise model. This is usually referred to as a *probability transformation*, and has been approached in many different ways: we can consider for instance the probability measure that minimizes (some) distance to the lower probability (Baroni and Vicig, 2005) or that with the maximum entropy (Jaffray, 1995). The problem has been considered with particular attention by the belief function community, and a number of different transformations have been proposed (Smets, 2005; Voorbraak, 1989). Among these, one of the most widely used is the *pignistic* transformation, considered by Smets and proposed earlier by Dubois and Prade (1982) and Williams (1982). It turns out that this transformation coincides with what Shapley proposed in 1953 as a solution for a game. Under this formalism, the possibility space represents a set of players, and the non-additive measure of an event A is interpreted as the gain associated with a coalition from the players in A. The link allows us to obtain the pignistic transformation as the center of gravity (the average of the extreme points) of the set of probabilities associated with the non-additive measure, when the latter is 2-monotone.

Inspired by this result, in this paper we investigate game solutions as probability transformations. On the one hand, we deepen in the properties of the Shapley value, studying if it is also the center of gravity of the core under less restrictive conditions than 2-monotonicity. Moreover, we study for which imprecise probability models we can guarantee the consistency of the Shapley value with the lower probability it is induced from. In addition, we shall also study the role as a probability transformation of another popular solution proposed within game theory: the Banzhaf value.

After introducing some preliminary concepts in Section 2, in Sections 3–6 we investigate the properties of the Shapley and Banzhaf values for some particular types of lower probabilities: minitive measures, 2-monotone lower probabilities, coherent lower probabilities, or lower probabilities

induced by comparative or distortion models. We conclude the paper in Section 7 with some additional remarks. Due to the space limitations, proofs have been omitted.

2. Preliminary Concepts

2.1 Lower Probabilities

Consider a finite possibility space $\Omega = \{1, \dots, n\}$. A *lower probability* on $\Omega = \{1, 2, \dots, n\}$ is a function $\underline{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$ that is monotone ($A \subseteq B \Rightarrow \underline{P}(A) \leq \underline{P}(B)$) and normalized ($\underline{P}(\emptyset) = 0, \underline{P}(\Omega) = 1$). Its conjugate upper probability is given by $\overline{P}(A) = 1 - \underline{P}(A^c)$ for every $A \subseteq \Omega$, and its *core* is the set $\mathcal{M}(\underline{P})$ of additive models that are compatible with \underline{P} , in the sense that

$$\mathcal{M}(\underline{P}) = \{P : \mathcal{P}(\Omega) \rightarrow [0, 1] \text{ probability measure} : P(A) \geq \underline{P}(A) \forall A \subseteq \Omega\}.$$

We shall only consider in this paper lower probabilities \underline{P} whose core is non-empty. These are said to *avoid sure loss*. They are called *coherent* if they are moreover the lower envelope of their core, in the sense that $\underline{P}(A) = \min\{P(A) : P \in \mathcal{M}(\underline{P})\}$ for every $A \subseteq \Omega$. One particular family of coherent lower probabilities are the *2-monotone* ones, which are those satisfying $\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B)$ for any pair of subsets A, B of Ω .

This notion can be strengthened by considering *complete monotonicity*, which means that

$$\underline{P}(\bigcup_{i=1}^n A_i) \geq \sum_{i=1}^n \underline{P}(A_i) - \sum_{i,j \in \{1, \dots, n\}} \underline{P}(A_i \cap A_j) + \dots + (-1)^{n+1} \underline{P}(\bigcap_{i=1}^n A_i)$$

for every $n \in \mathbb{N}$ and every $A_1, \dots, A_n \subseteq \Omega$.

Completely monotone lower probabilities are also called *belief functions* in the theory of evidence (Shafer, 1976). One of their advantages is that they are uniquely determined by their *basic probability assignment* $m : \mathcal{P}(\Omega) \rightarrow [0, 1]$, by means of the formula

$$\underline{P}(A) = \sum_{B \subseteq A} m(B). \quad (1)$$

More generally, any lower probability is determined by its Möbius inverse, given by

$$m(B) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \underline{P}(A),$$

in the sense that this function m determines \underline{P} by means of Eq. (1); this Möbius inverse is non-negative if and only if \underline{P} is a belief function. In that case, the sets B with $m(B) > 0$ are called the *focal elements* of the belief function \underline{P} .

2.2 Game Solutions

Within game theory, the possibility space Ω is interpreted as a set of players, and $\underline{P}(A)$ is then regarded as the gain that is guaranteed by the coalition of the players in A . Under the assumption of transferable utility, the core of the game is the set distributions of the total payoff among the players that cannot be improved by a coalition. These distributions are referred to here as *solutions* of the game (they should not be mistaken with the alternative use of the term *solution* in game theory as a multifunction that assigns to each game a set of valid strategies).

Arguably the most important solution of a game is the so-called *Shapley value* (Shapley, 1953, 1971) that, for a player i , is given by

$$\Phi(\underline{P})(i) = \sum_{T \not\ni \{i\}} \frac{t!(n-t-1)!}{n!} (\underline{P}(T \cup \{i\}) - \underline{P}(T)), \quad (2)$$

where $t = |T|$. It is the only solution of the game that satisfies the properties of efficiency (in the sense defined below), symmetry, linearity and that is equal to zero on null players.

When the game \underline{P} is 2-monotone, $\Phi(\underline{P})$ corresponds to the center of gravity of the core, that is, the average of the extreme points of $\mathcal{M}(\underline{P})$ (Shapley, 1971). These are related to the permutations of Ω (Chateauneuf and Jaffray, 1989): any permutation σ defines an extreme point by means of the equation

$$P_\sigma(\{\sigma(1), \dots, \sigma(i)\}) := \underline{P}(\{\sigma(1), \dots, \sigma(i)\}) \text{ for } i = 1, \dots, n. \quad (3)$$

Thus, it holds that $\Phi(\underline{P})(i) = \frac{\sum_{\sigma \in S^\Omega} P_\sigma(\{i\})}{n!}$, where S^Ω denotes the set of permutations of Ω .

Interestingly, Shapley value of a belief function coincides with what Smets called its *pignistic transformation* within the Transferable Belief Model (Smets and Kennes, 1994), as shown in (Smets, 2005). This means that we can also compute the Shapley value as:

$$\Phi(\underline{P})(i) = \sum_{i \in B} \frac{m(B)}{|B|}. \quad (4)$$

The equivalence goes beyond belief functions, and as a consequence it can be used to justify the use of the pignistic transformation beyond this framework. See Aregui and Denoeux (2008); Monney et al. (2011) for some works making use of the pignistic transformation.

Another popular solution of a game is the so-called *Banzhaf value* (Banzhaf (1965); see also Webber (1988)), given by

$$B(\underline{P})(i) = \frac{1}{2^{n-1}} \sum_{T \not\ni \{i\}} \underline{P}(T \cup \{i\}) - \underline{P}(T). \quad (5)$$

However, and unlike the Shapley value, the equation above does not produce a probability mass function, because we may not have $\sum_{i \in \Omega} B(\underline{P})(i) = 1$ (in the language of game theory, if the sum of the values of the players does not agree with the total payoff $\underline{P}(\Omega)$ it means that the solution is not *efficient*). For this reason, it has been suggested to consider instead the *normalized Banzhaf value*, which is given by

$$\Psi(\underline{P})(i) = \frac{B(\underline{P})(i)}{\sum_{j \in \Omega} B(\underline{P})(j)}. \quad (6)$$

Although the normalized Banzhaf value does not share all the properties of the Banzhaf value (Dubey and Shapley, 1979), it has been axiomatized from the point of view of game theory by Van der Brink and Van der Laan (1998).

In this paper, we shall investigate the properties of the Shapley value and the normalized Banzhaf value as probability transformations of a lower probability. Specifically, we shall study for which types of lower probabilities they are guaranteed to belong to their core, as well as some simpler expressions for a number of particular cases.

3. Minitive Measures

We begin by considering a particular case of belief functions: minitive measures. They are also referred to as *consonant* belief functions, *necessity* measures or minitive lower probabilities.

Definition 1 A lower probability $\underline{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$ is called minitive when it satisfies

$$\underline{P}(A \cap B) = \min\{\underline{P}(A), \underline{P}(B)\} \quad \forall A, B \subseteq \Omega.$$

It was proven by [Nguyen et al. \(1997\)](#) that any minitive measure is in particular completely monotone, and therefore also 2-monotone. As a consequence, Shapley value can also be obtained in this case as the center of gravity of the elements of the core. On the other hand, the number of vertices of the core is smaller than $n!$ in this case, as it was shown by [Miranda et al. \(2003\)](#) to be equal to 2^{n-1} , at most. The reason for this is that minitive functions correspond to the particular case of completely monotone measures whose focal elements are *nested* ([Shafer, 1976](#)), in the sense that they are completely ordered by the inclusion relation, and this makes the extreme points associated with many different permutations of Ω to coincide.

Let \underline{P} be a minitive measure. In this subsection, we shall assume without loss of generality that its focal elements are the sets $\{1, \dots, j\}$ for $j = 1, \dots, n$; the results extend easily to the general case. Using the expression in Eq. (4), Dubois and Prade established the following formula:

Proposition 2 ([Dubois and Prade, 2002](#)) Let \underline{P} be a minitive measure, and denote by m its basic probability assignment. Then its Shapley value is given by:

$$\Phi(\underline{P})(i) = \sum_{j=i}^n \frac{m(\{1, \dots, j\})}{j} \quad \forall i = 1, \dots, n.$$

With respect to the normalized Banzhaf value, we have proven the following:

Proposition 3 Let \underline{P} be a minitive measure, and denote by m its basic probability assignment. Then its Banzhaf value is given by:

$$B(\underline{P})(i) = \frac{1}{2^{n-1}} \sum_{j=i}^n 2^{n-j} m(\{1, \dots, j\}) \quad \forall i = 1, \dots, n,$$

whence its normalized Banzhaf value is:

$$\Psi(\underline{P})(i) = \frac{\sum_{j=i}^n 2^{n-j} m(\{1, \dots, j\})}{\sum_{j=1}^n j \cdot 2^{n-j} m(\{1, \dots, j\})} \quad \forall i = 1, \dots, n.$$

Moreover, the probability measure $\Psi(\underline{P})$ belongs to the core $\mathcal{M}(\underline{P})$ of the minitive measure \underline{P} .

4. 2-Monotone Lower Probabilities

Next, we study in more detail the case of 2-monotone lower probabilities. As we mentioned before, for them the Shapley value always belongs to the core of $\mathcal{M}(\underline{P})$. Interestingly, the same property does not hold for the normalized Banzhaf value, not even in the particular case where \underline{P} is a belief function, as the following example shows:

Example 1 Let $\Omega = \{1, 2, 3, 4\}$, and consider the belief function associated with the basic probability assignment given by $m(\{1\}) = m(\{2, 3, 4\}) = 0.5$, and $m(A) = 0$ for any other A . Then it follows from Eq. (5) that $B(\underline{P}) = (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. As a consequence, the probability mass function of the normalized Banzhaf value is given by $\Psi(\underline{P}) = (\frac{4}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$. However, this does not belong to the core of \underline{P} : we have that $\Psi(\underline{P})(\{2, 3, 4\}) = \frac{3}{7} < \frac{1}{2} = \underline{P}(\{2, 3, 4\})$.

For comparison, in this case Eq. (2) tells us that $\Phi(\underline{P}) = (\frac{3}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. ♦

This means that the result we have established in Proposition 3 does not extend to arbitrary belief functions. It also illustrates the difference between the Shapley and the normalized Banzhaf values. On the other hand, it can be checked that in case of belief functions (and as a consequence also for their subclass of minitive measures) the sum $\sum_{i \in \Omega} B(\underline{P})(i)$ of the values given by Eq. (5) is always smaller than or equal to 1. This does not extend to arbitrary 2-monotone lower probabilities, as Example 2 will show.

Next, we shall investigate the properties of the Shapley and Banzhaf values for some other particular types of 2-monotone lower probabilities.

4.1 2-Monotone Lower Probabilities in a Three Element Space

Let us consider the particular case where the possibility space has three elements. In that case, it has been proven that a lower probability is 2-monotone if and only if it is coherent. Moreover, in the case of cardinality three 2-monotone lower probabilities are particular instances of *probability intervals* (de Campos et al., 1994), that is, they are uniquely determined by the constraints $[\underline{P}(\{i\}), \bar{P}(\{i\})]$ on singletons. In other words, it suffices to know in this case the lower and upper bounds on the gain of each player.

The following proposition gives an alternative expression for the Shapley and normalized Banzhaf values in this case:

Proposition 4 Given $\Omega = \{1, 2, 3\}$, it holds that, for every $i \in \Omega$,

$$\Phi(\underline{P})(i) = \frac{1}{3} + \frac{1}{2}[\underline{P}(\{i\}) + \bar{P}(\{i\})] - \frac{1}{6} \sum_{l=1}^3 [\underline{P}(\{l\}) + \bar{P}(\{l\})],$$

while the normalized Banzhaf value is

$$\Psi(\underline{P})(\{i\}) = \frac{4m(\{i\}) + m(\Omega) + 2 \sum_{j \neq i} m(\{i, j\})}{4 - m(\Omega)}.$$

Moreover, $\Psi(\underline{P})$ belongs to the core $\mathcal{M}(\underline{P})$.

As Example 1 shows, $\Psi(\underline{P})$ need not belong to the core for greater cardinalities of Ω .

4.2 Lower Probabilities Induced by a Distortion Model

Two particular cases of 2-monotone lower probabilities are those induced by a Pari-Mutuel Model (PMM for short) or a ε -contamination model; these two cases are usually referred to as *distortion* models. The PMM originated in horse racing. It considers a probability P_0 on $\mathcal{P}(\Omega)$ and a distortion

value $\delta > 0$. Using P_0 and δ , the PMM defines a lower probability \underline{P} by (Montes et al., 2017; Pelessoni et al., 2010; Walley, 1991):

$$\underline{P}(A) = \max\{(1 + \delta)P_0(A) - \delta, 0\}. \quad (7)$$

From (Montes et al., 2017), the lower probability \underline{P} induced by a PMM is in particular a probability interval, and as a consequence also 2-monotone. Thus, the Shapley value coincides with the center of gravity of the core.

The same applies to ε -contamination models, where we consider a probability P_0 and a contamination value $\varepsilon \in (0, 1)$, that represents the distortion made on P_0 . The ε -contamination model defines a lower probability by:

$$\underline{P}(A) = (1 - \varepsilon)P_0(A) + \varepsilon\underline{P}_\Omega(A), \quad (8)$$

where \underline{P}_Ω is the *vacuous* lower probability that assigns the value 1 to Ω and 0 otherwise. This lower probability is known to be, not only 2-monotone, but also completely monotone.

Although one may think that for a distortion based on the probability P_0 , the probability transformations associated with the Shapley and normalized Banzhaf values return P_0 , our next example shows that this is not the case:

Example 2 Consider the probability P_0 on $\{1, 2, 3\}$ given by $P_0(\{1\}) = 0.1$, $P_0(\{2\}) = 0.2$ and $P_0(\{3\}) = 0.7$. Take $\delta = \varepsilon = 0.3$, and denote by \underline{P}_δ and $\underline{P}_\varepsilon$ the PMM and ε -contamination they induce, respectively. Using Eqs. (7) and (8), these are given by:

A	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$\underline{P}_\delta(A)$	0	0	0.61	0.09	0.74	0.87	1
$\underline{P}_\varepsilon(A)$	0.07	0.14	0.49	0.21	0.56	0.63	1

We deduce from Eq. (3) that the extreme points of $\mathcal{M}(\underline{P}_\delta)$ and $\mathcal{M}(\underline{P}_\varepsilon)$ are given by:

σ	P_σ for $\mathcal{M}(\underline{P}_\delta)$	P_σ for $\mathcal{M}(\underline{P}_\varepsilon)$
(1, 2, 3)	(0, 0.09, 0.91)	(0.07, 0.14, 0.79)
(1, 3, 2)	(0, 0.26, 0.74)	(0.07, 0.44, 0.49)
(2, 1, 3)	(0.09, 0, 0.91)	(0.07, 0.14, 0.79)
(2, 3, 1)	(0.13, 0, 0.87)	(0.37, 0.14, 0.49)
(3, 1, 2)	(0.13, 0.26, 0.61)	(0.07, 0.44, 0.49)
(3, 2, 1)	(0.13, 0.26, 0.61)	(0.37, 0.14, 0.49)

Thus, the Shapley values are $\Phi(\underline{P}_\delta) = (0.08, 0.145, 0.775)$ and $\Phi(\underline{P}_\varepsilon) = (0.17, 0.24, 0.59)$, respectively, and none of them coincide with P_0 .

Similarly, the normalized Banzhaf values are given by $\Psi(\underline{P}_\delta) = (\frac{0.35}{4.09}, \frac{0.61}{4.09}, \frac{3.13}{4.09})$ and $\Psi(\underline{P}_\varepsilon) = (\frac{0.58}{3.7}, \frac{0.86}{3.7}, \frac{2.26}{3.7})$, which do not coincide with P_0 either. ♦

We now consider the PMM in the particular case where δ satisfies $\delta < \frac{P_0(\{i\})}{1 - P_0(\{i\})}$ for any $i = 1, \dots, n$. This can be shown (Walley, 1991) to correspond to the case where \underline{P} is strictly positive for any non-empty set. In that case we can give a simple expression for the Shapley and the normalized Banzhaf values.

Proposition 5 Let \underline{P} be the lower probability associated with the PMM determined by P_0, δ , and assume that $\delta < \frac{P_0(\{i\})}{1-P_0(\{i\})}$ for any $i = 1, \dots, n$. Then the Shapley value is given by $\Phi(\underline{P})(i) = (1 + \delta)P_0(\{i\}) - \frac{\delta}{n}$, while

$$\Psi(\underline{P})(i) = \frac{(1 + \delta)P_0(\{i\}) - \frac{\delta}{2^{n-1}}}{k}, \text{ where } k = (1 + \delta) - \frac{n\delta}{2^{n-1}}.$$

Moreover, both $\Phi(\underline{P}), \Psi(\underline{P})$ belong to the core $\mathcal{M}(\underline{P})$.

Next we establish a similar result for the ε -contamination models:

Proposition 6 Let \underline{P} be the lower probability associated with the ε -contamination determined by P_0, ε . Then the Shapley value is given by $\Phi(\underline{P})(i) = (1 - \varepsilon)P_0(\{i\}) + \frac{\varepsilon}{n}$, while the normalized Banzhaf value is

$$\Psi(\underline{P})(i) = \frac{(1 - \varepsilon)P_0(\{i\}) + \frac{\varepsilon}{2^{n-1}}}{k}, \text{ where } k = (1 - \varepsilon) + \frac{n\varepsilon}{2^{n-1}}.$$

Moreover, both $\Phi(\underline{P}), \Psi(\underline{P})$ belong to the core $\mathcal{M}(\underline{P})$.

A common choice for P_0 in a distortion model is the uniform distribution; see for example Utkin (2014) and Utkin and Wiencierz (2013). Our next result shows that for the ε -contamination model and for the PMM with small enough values of δ , the Shapley and normalized Banzhaf values coincide with P_0 if and only if P_0 is uniform.

Corollary 7 Let \underline{P} be the lower probability associated with either the PMM determined by P_0, δ , where δ satisfies $\delta < \frac{P_0(\{i\})}{1-P_0(\{i\})}$ for any $i = 1, \dots, n$ or a ε -contamination model. Then,

$$\Phi(\underline{P}) = P_0 \iff \Psi(\underline{P}) = P_0 \iff P_0(\{i\}) = \frac{1}{n} \forall i \in \Omega.$$

In fact, for the PMM we easily derive from the symmetry axioms satisfied by the Shapley and the Banzhaf values that, if P_0 is the uniform probability measure, then it coincides with the Shapley value of the PMM (P_0, δ) irrespective of the value of δ ; to see that the converse is not true in general, i.e., that Φ can be the uniform probability measure for other PMM (P_0, δ) , it suffices to consider that $\mathcal{M}(\underline{P})$ is the set of all probability measures for $\underline{P} = (P_0, \delta)$ provided δ is large enough (specifically, when $\delta \geq \frac{1}{P_0(A^c)}$ for every $A \neq \Omega$), and that in that case Φ becomes the uniform distribution. Similar comments apply to the normalized Banzhaf value.

5. Coherent Lower Probabilities

We consider next the case of coherent lower probabilities. It was established by Baroni and Vicig (2005, Proposition 5) in terms of the pignistic transformation that the Shapley value of a coherent lower probability need not be an element of the core, or, in other words, that the result for 2-monotone lower probabilities does not extend to arbitrary coherent lower probabilities. The very same example allows us to show that the normalized Banzhaf value need not belong to the core, either. In fact, as Example 1 shows, the normalized Banzhaf value is not guaranteed to be in the core even in the particular case of belief functions, and in the case of possibility spaces with four elements (the example by Baroni and Vicig considers a space of cardinality five).

In spite of this result, we can guarantee that the Shapley and Banzhaf values belong to the core in a number of particular cases. We begin by considering the case of coherent lower probabilities that are the lower envelope of two probability measures. They may arise for instance when we are aggregating the information from two different sources.

Proposition 8 *Consider two probability measures P_1, P_2 on $\mathcal{P}(\Omega)$ and let \underline{P} be the coherent lower probability they determine. Then $\Psi(\underline{P})(i) = \Phi(\underline{P})(i) = B(\underline{P})(i) = \frac{P_1(\{i\}) + P_2(\{i\})}{2}$ for every $i \in \Omega$.*

Interestingly, in the case considered in the proposition above the Banzhaf value is always normalized. On the other hand, the result does not extend to coherent lower probabilities that are the envelope of three probability measures, as Example 1 shows: note that the belief function in that example is the lower envelope of the family of probability measures with mass functions $\{(0.5, 0.5, 0, 0), (0.5, 0, 0.5, 0), (0.5, 0, 0, 0.5)\}$.

Another situation in which we can guarantee that the Shapley value of a coherent lower probability belongs to its core is when the possibility space has cardinality equal to four, as our next result shows:

Proposition 9 *Let $\Omega = \{1, 2, 3, 4\}$ and let $\underline{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$ be a coherent lower probability. Then, $\Phi(\underline{P})$ belongs to $\mathcal{M}(\underline{P})$.*

Example 1 shows that a similar result does not hold for the normalized Banzhaf value.

5.1 Comparative Lower Probabilities

Our attention shifts now to another useful model related to non-additive measures: comparative probabilities. These (de Finetti, 1931; Koopman, 1940) correspond to the case where the available information about the probability of the events is of qualitative nature, in the sense that we can only make statements of the type ‘the probability of A is at least as much as that of B’.

The mathematical study of comparative models can be involved, and for instance the existence of an additive model that is compatible with them (Kaplan and Fine, 1977; Kraft et al., 1959) is not guaranteed; we refer to Regoli (1996) for a survey of this topic. In (Miranda and Destercke, 2015), the particular case of *elementary* comparative probabilities was considered, where we only give qualitative assessments about the value of individual players.

With this in mind, given $\mathcal{I} \subseteq \Omega \times \Omega$, we call the (*elementary*) *comparative model* determined by \mathcal{I} the lower envelope \underline{P} of the set

$$\mathcal{M} := \{P \text{ probability measure} : P(\{i\}) \geq P(\{j\}) \forall (i, j) \in \mathcal{I}\}.$$

It was proven by Miranda and Destercke (2015) that the core of these models can be given quite a neat structure, and that we also have at most 2^{n-1} different extreme points. However, the lower probability induced by this core need not be 2-monotone in general (Miranda and Destercke, 2015, Section 4.3). Taking this into account, it is not surprising to see that the Shapley value need not coincide with the center of gravity of the core, as our next example shows:

Example 3 *Let us consider the comparative assessments*

$$P(\{1\}) \geq P(\{2\}), \quad P(\{1\}) \geq P(\{3\}), \quad P(\{2\}) \geq P(\{4\}), \quad P(\{3\}) \geq P(\{4\}).$$

If we consider the set of probability measures compatible with these assessments, it follows from ([Miranda and Destercke, 2015](#)) that the extreme points of this set are the probability measures

$$(1, 0, 0, 0), \left(\frac{1}{2}, \frac{1}{2}, 0, 0\right), \left(\frac{1}{2}, 0, \frac{1}{2}, 0\right), \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right), \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right).$$

From this we deduce that the lower probability \underline{P} associated with these assessments is given by $\underline{P}(A) = 0$ if $1 \notin A$, and

$$\begin{aligned} \underline{P}(\{1\}) &= \frac{1}{4}, & \underline{P}(\{1, 2\}) &= \underline{P}(\{1, 3\}) = \frac{1}{2}, & \underline{P}(\{1, 4\}) &= \frac{1}{3}, \\ \underline{P}(\{1, 2, 3\}) &= \frac{3}{4}, & \underline{P}(\{1, 2, 4\}) &= \underline{P}(\{1, 3, 4\}) = \frac{1}{2}, & \underline{P}(\Omega) &= 1. \end{aligned}$$

Now, from Eq. (2), the Shapley value is given by

$$\Phi(\underline{P}) = \left(\frac{41}{72}, \frac{13}{72}, \frac{13}{72}, \frac{5}{72}\right),$$

while the center of gravity of the core is given by $(\frac{31}{60}, \frac{13}{60}, \frac{13}{60}, \frac{3}{60})$. ♦

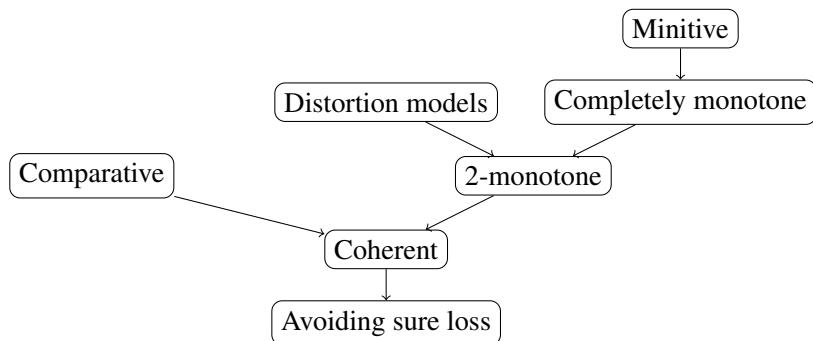
Nevertheless, it is possible to prove that both the Shapley and the normalized Banzhaf values belong to the core in this case:

Proposition 10 Let \underline{P} be a lower probability determined by elementary comparative probabilities. Then $\Phi(\underline{P})$ and $\Psi(\underline{P})$ belong to the core $\mathcal{M}(\underline{P})$.

6. Lower Probabilities Avoiding Sure Loss

The most general model of lower probabilities that we shall consider in this paper are those that avoid sure loss. They correspond to *balanced* games within game theory ([Shapley, 1967](#)). Recall that a lower probability $\underline{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$ is said to *avoid sure loss* ([Walley, 1991](#)) when its core $\mathcal{M}(\underline{P})$ is non-empty.

Note that a lower probability that avoids sure loss need not be coherent, because it may not be the lower envelope of its core. The relationship between the different models considered in this paper is summarised by the following figure, where the implication means an inclusion between the families:



Our next example provides an incoherent lower probability that avoids sure loss, and shows that in that case neither of the Shapley and the normalized Banzhaf values need belong to the core:

Example 4 Consider $\Omega = \{1, 2, 3\}$, and let us consider \underline{P} given by

$$\underline{P}(\{1\}) = \underline{P}(\{2\}) = \underline{P}(\{3\}) = 0, \quad \underline{P}(\{1, 2\}) = \frac{9}{12}, \quad \underline{P}(\{1, 3\}) = \frac{8}{12}, \quad \underline{P}(\{2, 3\}) = \frac{7}{12},$$

and of course with $\underline{P}(\Omega) = 1$. The core of \underline{P} is non-empty, as it includes for instance the probability measure P given by $(\frac{5}{12}, \frac{4}{12}, \frac{3}{12})$; in fact it can be checked that $\mathcal{M}(\underline{P})$ consists exactly of this probability measure. As a consequence, we see that the lower probability \underline{P} is not coherent, since for instance $\underline{P}(\{1\}) = 0 < \min\{P(\{1\}) : P \in \mathcal{M}(\underline{P})\}$.

Using Eq. (2), we obtain that the Shapley value of \underline{P} is given by

$$\Phi(\underline{P})(1) = \frac{9}{24}, \quad \Psi(\underline{P})(2) = \frac{8}{24}, \quad \Psi(\underline{P})(3) = \frac{7}{24}.$$

However, $\Phi(\underline{P})$ does not belong to the core of \underline{P} : we have that $\Phi(\underline{P})(\{1, 2\}) = \frac{17}{24} < \frac{18}{24} = \underline{P}(\{1, 2\})$.

By Eq. (6) the normalized Banzhaf value is given by:

$$\Psi(\underline{P})(1) = \frac{11}{30}, \quad \Psi(\underline{P})(2) = \frac{10}{30}, \quad \Psi(\underline{P})(3) = \frac{9}{30}.$$

Thus, it does not belong to the core, either. ♦

This example also shows that the result we have established for coherent lower probabilities in Proposition 4 does not extend to those avoiding sure loss, and also that the discussion about possibility spaces of cardinality three in Section 4.1 does not apply when coherence is not satisfied.

7. Conclusions

The results in this paper show that some of the nice properties of the Shapley value can be extended beyond the framework of 2-monotone lower probabilities and belief functions. With respect to the Banzhaf value, although the lack of efficiency leads to the definition of the normalized version, it is also possible to prove its consistency with the lower probability in a number of cases. Although in this paper we have focused on this consistency property, in the future we should deepen into the investigation of the mathematical properties of these models as probability transformations, in the vein of the work carried out by [Dezert et al. \(2012\)](#), so as to be able to compare them properly with the existing models.

More generally, we would like to continue this research by considering the probabilistic solutions of games considered in [\(Webber, 1988\)](#). In addition, we should also study the properties of other probability transformations, such as the maximum entropy one, for some of the imprecise probability models considered in this paper.

Acknowledgements

The research in this paper has been supported by project TIN2014-59543-P. We would also like to thank Paolo Vicig and the anonymous referees for some helpful comments.

References

- A. Aregui and T. Denoeux. Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, 49:575–594, 2008.
- J. Banzhaf. Weighted voting does not work: a mathematical analysis. *Rutgers Law Review*, 19: 317–343, 1965.
- P. Baroni and P. Vicig. An uncertainty interchange format with imprecise probabilities. *International Journal of Approximate Reasoning*, 40:147–180, 2005.
- 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.
- L. M. de Campos, J. F. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.
- B. de Finetti. Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329, 1931.
- J. Dezert, H. Han, Z.-G. Liu, and J.-M. Tacnet. Hierarchical DSmP transformation for decision-making under uncertainty. In *Proceedings of Fusion'2012*, pages 294–301, 2012.
- P. Dubey and L. S. Shapley. Mathematical properties of the Banzhaf power index. *Mathematics of Operations Research*, 4:99–131, 1979.
- D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North Holland, 1982.
- D. Dubois and H. Prade. Quantitative possibility theory and its probabilistic connections. In O. H. P. Grzegorzewski and M. Gil, editors, *Soft Methods in Probability, Statistics and Data Analysis*, pages 3–26. Physica-Verlag, 2002.
- J.-Y. Jaffray. On the maximum entropy probability which is consistent with a convex capacity. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 3:27–33, 1995.
- M. Kaplan and T. Fine. Joint orders in comparative probability. *Annals of Probability*, 5:161–179, 1977.
- B. Koopman. The bases of probability. *Bulletin of the American Mathematical Society*, 46:763–774, 1940.
- C. Kraft, J. Pratt, and A. Seidenberg. Intuitive probability on finite sets. *The Annals of Mathematical Statistics*, 30:408–419, 1959.
- E. Miranda and S. Destercke. Extreme points of the credal sets generated by comparative probabilities. *Journal of Mathematical Psychology*, 64/65:44–57, 2015.

- 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.
- P. Monney, M. Chan, and P. Romberg. A belief function classifier based on information provided by noisy and dependent features. *International Journal of Approximate Reasoning*, 52(3):335–352, 2011.
- I. Montes, E. Miranda, and S. Destercke. On the Pari-Mutuel Model seen as imprecise probabilities. 2017. Submitted for publication.
- H. T. Nguyen, N. T. Nguyen, and T. Wang. On capacity functionals in interval probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5:359–377, 1997.
- R. Pelessoni, P. Vicig, and M. Zaffalon. Inference and risk measurement with the pari-mutuel model. *International Journal of Approximate Reasoning*, 51(9):1145–1158, 2010.
- G. Regoli. Comparative probability and robustness. *Lecture Notes- Monograph Series*, 29:343–352, 1996.
- G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.
- L. S. Shapley. A value for n-person games. *Annals of Mathematical Studies*, 28:307–317, 1953.
- L. S. Shapley. On balanced sets and cores. *Naval Research Logistic Quarterly*, 14:453–460, 1967.
- L. S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1:11–26, 1971.
- P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133–147, 2005.
- P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2):191–234, 1994.
- L. Utkin. A framework for imprecise robust one-class classification models. *Journal of Machine Learning Research and Cybernetics*, 5(3):379–393, 2014.
- L. Utkin and A. Wiencierz. An imprecise boosting-like approach to regression. In *Proceedings of the 8th International Symposium on Imprecise Probability: Theories and Applications*, pages 345–354, 2013.
- R. Van der Brink and G. Van der Laan. Axiomatizations of the normalized Banzhaf value and the Shapley value. *Social Choice and Welfare*, 15:567–582, 1998.
- F. Voorbraak. A computationally efficient approximation of dempster-shafer theory. *International Journal of Man-Machine Studies*, 30:525–536, 1989.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- R. Webber. Probabilistic values for games. In A. Roth, editor, *The Shapley value. Essays in honour of L.S. Shapley*, pages 101–119. Cambridge University Press, 1988.
- P. Williams. Discussion of ‘Belief functions and parametric models’, by G. Shafer. *Journal of the Royal Statistical Society, Series B*, 44(3):341–343, 1982.

A Study of the Pari-Mutuel Model from the Point of View of Imprecise Probabilities

Ignacio Montes

Enrique Miranda

*Dep. of Statistics and Operations Research
University of Oviedo (Spain)*

IMONTES@UNIOVI.ES

MIRANDAENRIQUE@UNIOVI.ES

Sébastien Destercke

SEBASTIEN.DESTERCKE@HDS.UTC.FR

*UMR CNRS 7253 Heudiasyc, Sorbonne Université,
Université de Technologie de Compiègne CS 60319 - 60203 Compiègne cedex (France)*

Abstract

The Pari-Mutuel model is a distortion model that has its origin in horse racing. Since then, it has been applied in many different fields, such as finance or risk analysis. In this paper we investigate the properties of the Pari-Mutuel model within the framework of Imprecise Probabilities. Since a Pari-Mutuel model induces (2-monotone) coherent lower and upper probabilities, we investigate its connections with other relevant models within this theory, such as probability intervals and belief functions. We also determine the number of extreme points of the credal set induced by the Pari-Mutuel model and study how to combine the information given by multiple Pari-Mutuel models.

Keywords: pari-mutuel bets; credal sets; probability intervals; belief functions; information fusion.

1. Introduction

The Pari-Mutuel model (PMM, for short) is a betting scheme originated in horse racing, that has been used in other fields like economics, risk analysis or life insurance. It considers a probability P_0 which models the fair price for a bet fixed by an agent, usually called House. In order to ensure a positive gain, House transforms this fair gain into a slightly greater value given by $(1 + \delta)P_0$, where $\delta > 0$ is interpreted as the taxation from House. We refer to (Gerber, 1979; Peters et al., 2007; Terrell, 1994; Thaler and Ziemba, 1988) for some detailed studies on the PMM.

Using this interpretation, the PMM can be embedded into the Theory of Imprecise Probabilities: it determines lower and upper bounds for the probability of any event. These lower and upper probabilities satisfy the usual consistency requirement of coherence (Walley, 1991), and therefore they can be equivalently represented by means of the set of probability measures they bound. This set is a convex set of probabilities usually called credal set. Furthermore, the PMM satisfies the additional property of 2-monotonicity that offers computational advantages (Destercke, 2013).

To the best of our knowledge, there are few studies of the PMM from the point of view of imprecise probabilities. For example, (Pelessoni et al., 2010) studied the PMM as a risk measure and how to extend it from events to gambles, and (Utkin and Wiencierz, 2013) investigated how to use the PMM in classification problems.

In this paper, we further investigate the PMM from the point of view of Imprecise Probabilities. The rest of the paper is organized as follows: Section 2 recalls the definition and basic properties of the PMM. In Section 3 we investigate the connections between the PMM and other models from

Imprecise Probability Theory. In particular, we first prove that a PMM can be represented by means of a probability interval, and secondly we characterize the conditions a PMM must satisfy in order for its lower probability to be not only 2-monotone but also a belief function. Then, Section 4 studies some properties of the extreme points of the credal set induced by a PMM. On the one hand we investigate the form and the maximal number of extreme points of the credal set; on the other hand we give an upper bound of the number of extreme points. A number of procedures for merging different sources of information in the context of PMMs are investigated in Section 5. Due to space limitations, proofs as well as some less relevant explanations have been omitted.

2. Basic Notions About the Pari-Mutuel Model

Let $\mathcal{X} = \{x_1, \dots, x_n\}$ denote a finite universe and let P_0 be a probability measure defined on $\mathcal{P}(\mathcal{X})$. We shall assume throughout that $P_0(\{x_i\}) > 0$ for $i = 1, \dots, n$; the results generalize easily to the case where some elements have probability zero.

Given $\delta > 0$, the *pari-mutuel model* (PMM, for short) induced by P_0, δ , that we shall denote (P_0, δ) , is given by the following lower and upper probabilities:

$$\underline{P}(A) = \max\{(1 + \delta)P_0(A) - \delta, 0\} \text{ and } \overline{P}(A) = \min\{(1 + \delta)P_0(A), 1\} \quad \forall A \subseteq \mathcal{X}. \quad (1)$$

The functions $\underline{P}, \overline{P}$ are conjugate, meaning that $\overline{P}(A) = 1 - \underline{P}(A^c) \quad \forall A \subseteq \mathcal{X}$. Also, since $P_0(\{x_i\}) > 0 \quad \forall i = 1, \dots, n$, it holds that $\overline{P}(A) \geq P_0(A) > 0$ for every $A \subseteq \mathcal{X}$.

The interpretation of the parameter δ can be found in (Walley, 1991, Sec. 2.9.3). There, it is proven that $\overline{P}(A) - \underline{P}(A) \leq \delta$ for any A , and the equality is attained if and only if $\frac{1}{1+\delta} \leq P_0(A) \leq \frac{\delta}{1+\delta}$. In particular, this condition holds when $0 < \underline{P}(A) < \overline{P}(A) < 1$. Therefore, δ may be understood in terms of the imprecision allowed in the definition of $P_0(A)$.

Note also that, since the lower probability of a PMM can be obtained as a convex transformation of a probability measure, it follows (Denneberg, 1994) that \underline{P} is 2-monotone, meaning that

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B)$$

for any $A, B \subseteq \mathcal{X}$. As a consequence (Walley, 1991), $\underline{P}, \overline{P}$ are *coherent*, that is, they are respectively the lower and upper envelopes of the *credal set* associated with the PMM, given by

$$\mathcal{M}(P_0, \delta) = \{P \text{ probability} \mid \underline{P}(A) \leq P(A) \leq \overline{P}(A) \quad \forall A \subseteq \mathcal{X}\}. \quad (2)$$

3. PMM and Other Imprecise Probability Models

In this section, we study the connection between the PMM and other relevant imprecise probability models. In particular, we show that PMMs in a finite setting are particular instances of probability intervals, and study the conditions a PMM must satisfy in order to induce a belief function.

3.1 Connection Between PMM and Probability Intervals

Probability intervals on \mathcal{X} (de Campos et al., 1994; Tessem, 1992) are just lower probabilities defined on the singletons and their complementaries. Specifically, a probability interval is given by:

$$\mathcal{I} = \{[l_i, u_i] : i = 1, \dots, n\},$$

where it is assumed that $l_i \leq u_i$ and where the interpretation of $[l_i, u_i]$ is that the unknown or imprecisely specified probability of x_i belongs to the interval $[l_i, u_i]$. A probability interval determines a credal set by:

$$\mathcal{M}(\mathcal{I}) = \{P \text{ probability } | l_i \leq P(\{x_i\}) \leq u_i, i = 1, \dots, n\}, \quad (3)$$

and the lower and upper envelopes of $\mathcal{M}(\mathcal{I})$ determine coherent lower and upper probabilities by:

$$l(A) = \inf_{P \in \mathcal{M}(\mathcal{I})} P(A) \text{ and } u(A) = \sup_{P \in \mathcal{M}(\mathcal{I})} P(A) \forall A \subseteq \mathcal{X}. \quad (4)$$

A probability interval \mathcal{I} is called *reachable* (coherent in the terminology of [Walley \(1991\)](#)) whenever the functionals l, u determined by Eq. (4) satisfy $l(\{x_i\}) = l_i$ and $u(\{x_i\}) = u_i$ for all $i = 1, \dots, n$. This is equivalent to the following inequalities:

$$\sum_{j \neq i} l_j + u_i \leq 1 \text{ and } \sum_{j \neq i} u_j + l_i \geq 1 \quad \forall i = 1, \dots, n. \quad (5)$$

For a detailed study on probability intervals, we refer to ([de Campos et al., 1994](#)). See also ([Guo and Tanaka, 2010](#); [Skulj, 2009](#); [Tanaka et al., 2004](#)) for other relevant works on this topic.

By considering the restrictions to singletons of the lower and upper probabilities associated with a PMM, we can associate a reachable probability interval with any PMM. Interestingly, this probability interval keeps all the information about the PMM, in the sense that both determine the same credal set. In other words, PMMs are particular cases of reachable probability intervals, as our next result shows:

Theorem 1 *Let P_0 be a probability measure on $\mathcal{P}(\mathcal{X})$, $\delta > 0$ and (P_0, δ) the PMM they induce. Define the probability interval $\mathcal{I} = \{[l_i, u_i] : i = 1, \dots, n\}$ by $l_i = \underline{P}(\{x_i\})$ and $u_i = \overline{P}(\{x_i\})$, where $\underline{P}, \overline{P}$ are given by Eq. (1). Then, if $\mathcal{M}(\mathcal{I})$ denotes the credal set associated with \mathcal{I} by means of Eq. (3), it holds that:*

1. *The probability interval $\mathcal{I} = \{[l_i, u_i] : i = 1, \dots, n\}$ is reachable.*
2. *$\mathcal{M}(\mathcal{I}) = \mathcal{M}(P_0, \delta)$, or equivalently, $\underline{P}(A) = l(A)$ and $\overline{P}(A) = u(A)$ for any $A \subseteq \mathcal{X}$.*

Thus, the PMM is a particular case of probability interval. On the other hand, the latter model is more general, in the sense that not every reachable probability interval can be expressed in terms of a PMM.

Example 1 *Consider the four-element space $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ and the probability interval $\mathcal{I} = \{[l_i, u_i] : i = 1, \dots, 4\}$ given by:*

	x_1	x_2	x_3	x_4
l_i	0.2	0.1	0.3	0.2
u_i	0.4	0.2	0.5	0.4

which can be shown to be reachable using Eq. (5). To see that \mathcal{I} is not representable by a PMM (P_0, δ) , note that from the comments in Section 2, any set A such that $0 < \underline{P}(A) < \overline{P}(A) < 1$ should satisfy $\overline{P}(A) - \underline{P}(A) = \delta$. However, in this example it holds that:

$$\begin{aligned} 0 < l(\{x_1\}) = l_1 = 0.2 < 0.4 = u_1 = u(\{x_1\}) < 1 \text{ and} \\ 0 < l(\{x_2\}) = l_2 = 0.1 < 0.2 = u_2 = u(\{x_2\}) < 1, \end{aligned}$$

whence $u(\{x_1\}) - l(\{x_1\}) = 0.2$ and $u(\{x_2\}) - l(\{x_2\}) = 0.1$. Thus, the difference is not constant, and therefore l, u cannot be represented by means of a PMM. ♦

3.2 Connection Between PMM and Belief Functions

As we mentioned in Section 2, the lower probability of a PMM is 2-monotone. In this section we study under which conditions it is moreover *completely* monotone. Complete monotonicity means that for any $p \in \mathbb{N}$ and any sets $A_1, \dots, A_p \subseteq \mathcal{X}$, it holds that

$$\underline{P}(\cup_{i=1}^p A_i) \geq \sum_{J \subseteq \{1, \dots, n\}} (-1)^{|J|-1} \underline{P}(\cap_{i \in J} A_i).$$

A completely monotone lower probability is usually called a *belief function*. Belief functions (Shafer, 1976) are determined by their Möbius inverse $m : \mathcal{P}(\mathcal{X}) \rightarrow [0, 1]$, which is a mass function on the subsets of \mathcal{X} , by means of the formula $\underline{P}(A) = \sum_{B \subseteq A} m(B)$. The sets $A \subseteq \mathcal{X}$ such that $m(A) > 0$ are called the *focal elements* of \underline{P} . Conversely, the Möbius inverse m of a lower probability \underline{P} is determined by the formula

$$m(B) = \sum_{A \subseteq B} (-1)^{|B \setminus A|} \underline{P}(A), \quad (6)$$

and \underline{P} is a belief function if and only if the function m given by Eq. (6) satisfies $m(A) \geq 0$ for every $A \subseteq \mathcal{X}$.

We start with a simple result from which we deduce that in general the PMM is not 3-monotone, and therefore it is not completely monotone either.

Proposition 2 *Let \underline{P} be the lower probability associated with a PMM (P_0, δ) , with $|\mathcal{X}| \geq 3$. If there are different x_i, x_j, x_k such that $\underline{P}(\{x_i\}), \underline{P}(\{x_j\}), \underline{P}(\{x_k\}) > 0$, then \underline{P} is not 3-monotone.*

To see that the hypotheses of this proposition may be satisfied, let P_0 be the uniform distribution on $\{x_1, x_2, x_3\}$ and take $\delta = \frac{1}{3}$: it follows from Eq. (1) that $\underline{P}(\{x_1\}) = \underline{P}(\{x_2\}) = \underline{P}(\{x_3\}) = \frac{1}{9}$.

Next, we establish necessary and sufficient conditions for the PMM to induce a belief function. For this aim we define the *non-vacuity index* of a PMM as $k = \min\{|A| : \underline{P}(A) > 0\}$.

Theorem 3 *Let \underline{P} be the lower probability induced by a PMM (P_0, δ) by Eq. (1), and denote by k its non-vacuity index. \underline{P} is a belief function if and only if one of the following conditions is satisfied:*

(B1) $k = n$.

(B2) $k = n - 1$ and $\sum_{i=1}^n \underline{P}(\mathcal{X} \setminus \{x_i\}) \leq 1$.

(B3) $k < n - 1$, there exists a unique B with $|B| = k$ and $\underline{P}(B) > 0$, and $\underline{P}(A) > 0$ if and only if $B \subseteq A$.

(B4) $k < n - 1$, there exists a unique B with $|B| = k - 1$ and $\delta = \frac{P_0(B)}{1 - P_0(B)}$, and $\underline{P}(A) > 0$ if and only if $B \subset A$.

Proof (Sketch) For sufficiency, it suffices to take into account that cases (B1)–(B4) determine a belief function with focal elements:

(a) \mathcal{X} , in the case of (B1);

(b) $\{\mathcal{X}, \mathcal{X} \setminus \{x\} : \forall x \in \mathcal{X}\}$, in the case of (B2);

(c) $\{B, B \cup \{x\} : \forall x \notin B\}$, in the case of (B3); and

(d) $\{B \cup \{x\} : \forall x \notin B\}$, in the case of (B4).

For necessity, if \underline{P} is a belief function and we consider its associated non-vacuity index, we prove that the cases $k = n$, $k = n - 1$ and $k < n - 1$ determine the focal elements depicted in (a), (b) or (c)-(d) above, respectively, from which it follows that we are in cases (B1)–(B4) above. ■

[Boodgumarn et al. \(2013, Thm. 1\)](#) established that a sufficient condition for a probability interval to induce a belief function is that

$$\left| \left\{ i : u_i + \sum_{j \neq i} l_j < 1 \right\} \right| \leq 2. \quad (7)$$

Theorem 3 tells us that this condition is not necessary. Although it holds trivially under condition (B1) (i.e., for PMMs inducing a *vacuous* belief function), it is possible to find PMMs satisfying any of the conditions (B2)–(B4) and not the one in Eq. (7).

4. Extreme points induced by a PMM

Since the coherence of the PMM implies that it is uniquely determined by its (closed and convex) associated credal set, it becomes interesting to determine the extreme points of the set $\mathcal{M}(P_0, \delta)$ given by Eq. (2); this is particularly relevant if we want to use the PMM in some applied contexts, such as credal networks ([Antonucci and Cuzzolin, 2010](#); [Cozman, 2005](#)).

Recall that the extreme points of $\mathcal{M}(P_0, \delta)$ are the probability measures $P \in \mathcal{M}(P_0, \delta)$ such that if $P = \alpha P_1 + (1 - \alpha) P_2$ for some $\alpha \in (0, 1)$, $P_1, P_2 \in \mathcal{M}(P_0, \delta)$, then $P_1 = P_2$.

Since the lower probability of a PMM is 2-monotone, the extreme points of $\mathcal{M}(P_0, \delta)$ are associated with permutations of \mathcal{X} ([Chateauneuf and Jaffray, 1989](#)), in the following manner: if σ is a permutation of $\{1, \dots, n\}$, we consider the probability measure P_σ given by

$$\begin{aligned} P_\sigma(\{x_{\sigma(1)}\}) &= \overline{P}(\{x_{\sigma(1)}\}), \\ P_\sigma(\{x_{\sigma(k)}\}) &= \overline{P}(\{x_{\sigma(1)}, \dots, x_{\sigma(k)}\}) - \overline{P}(\{x_{\sigma(1)}, \dots, x_{\sigma(k-1)}\}) \quad \forall k = 2, \dots, n. \end{aligned} \quad (8)$$

Then, the extreme points of $\mathcal{M}(P_0, \delta)$ are $\{P_\sigma : \sigma \in S^n\}$, where S^n denotes the set of permutations of $\{1, \dots, n\}$. As a consequence, the number of extreme points of $\mathcal{M}(P_0, \delta)$ is bounded above by $n!$, the number of permutations of a n -element space. In this section, we study if this upper bound can be lowered in the particular case of the PMM.

4.1 Maximal Number of Extreme Points

We start our study by establishing two preliminary but helpful properties of the PMM. The first result shows that under some conditions, \overline{P} is not only sub-additive as a coherent upper probability, but also additive.

Lemma 4 *Let \overline{P} be the upper probability induced by a PMM (P_0, δ) by Eq. (1). If $\overline{P}(A) < 1$, then*

$$\overline{P}(A) = \sum_{x \in A} \overline{P}(\{x\}). \quad (9)$$

We deduce that if $\overline{P}(A \cup B) < 1$ and $A \cap B = \emptyset$, then $\overline{P}(A \cup B) = \overline{P}(A) + \overline{P}(B)$. Using Eq. (9), we can prove the second preliminary result, which gives the form of the extreme points in terms of \underline{P} and \overline{P} .

Lemma 5 Consider a PMM (P_0, δ) , and let $\underline{P}, \overline{P}$ be given by Eq. (1). The extreme point P_σ associated with the permutation σ by Eq. (8) is given by:

$$\begin{aligned} P(\{x_i\}) &= \overline{P}(x_i) \quad \forall i = \sigma(1), \dots, \sigma(j-1), \\ P(\{x_{\sigma(j)}\}) &= \underline{P}(\{x_{\sigma(j)}, \dots, x_{\sigma(n)}\}), \\ P(\{x_{\sigma(j+1)}\}) &= \dots = P(\{x_{\sigma(n)}\}) = 0, \end{aligned}$$

where $j \in \{1, \dots, n\}$ satisfies $\overline{P}(\{x_{\sigma(1)}, \dots, x_{\sigma(j-1)}\}) < \overline{P}(\{x_{\sigma(1)}, \dots, x_{\sigma(j)}\}) = 1$.

The above result is illustrated in the following example.

Example 2 Let $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$, P_0 the uniform probability distribution and $\delta = 0.5$. If we consider the permutation $\sigma = (1, 2, 3, 4)$, we obtain the extreme point P_σ given by:

$$\begin{aligned} P_\sigma(\{x_1\}) &= \overline{P}(\{x_1\}) = 1.5 \cdot 0.25 = 0.375, \\ P_\sigma(\{x_2\}) &= \overline{P}(\{x_2\}) = 1.5 \cdot 0.25 = 0.375, \\ P_\sigma(\{x_3\}) &= \underline{P}(\{x_3, x_4\}) = 1.5 \cdot 0.5 - 0.5 = 0.25, \\ P_\sigma(\{x_4\}) &= 0. \end{aligned}$$

In fact, it can be proven that the extreme points of $\mathcal{M}(P_0, \delta)$ are given by

$$\begin{aligned} P(\{x_i\}) &= \overline{P}(\{x_i\}) = 0.375, \\ P(\{x_j\}) &= \overline{P}(\{x_j\}) = 0.375, \\ P(\{x_k\}) &= \underline{P}(\{x_k, x_l\}) = 0.25, \\ P(\{x_l\}) &= 0, \end{aligned}$$

for any possible combination of i, j, k, l in $\{1, 2, 3, 4\}$. ♦

Next we use the results above to compute the maximal number of extreme points induced by a PMM. Note that from Theorem 1 we already know that any PMM is in particular a probability interval. This means that the number of extreme points induced by a PMM is upper bounded by the maximal number of extreme points induced by a probability interval. Next theorem shows that this upper bound can be attained.

Theorem 6 Given a PMM (P_0, δ) on \mathcal{X} , the maximal number of extreme points of $\mathcal{M}(P_0, \delta)$ is:

1. $\frac{n}{2} \binom{n}{2}$ if n is even;
2. $\frac{n+1}{2} \binom{n+1}{2}$ if n is odd.

Furthermore, these maxima are attainable, by considering P_0 a uniform distribution and $\delta \in \left(\frac{n-2}{n+2}, 1\right)$, if n is even, or $\delta \in \left(\frac{n-1}{n+1}, \frac{n+1}{n-1}\right)$ if n is odd.

The number of extreme points induced by a PMM (P_0, δ) where P_0 is the uniform probability measure has already been studied in (Utkin, 2014, Sect. 5.2.) and (Utkin and Wiencierz, 2013, Sect. 4.2). In this respect, note that, even if the definition of the PMM considered by Utkin and Wiencierz (2013) is slightly different from the one given in Section 2 (they consider instead $\underline{P}(A) = (1 + \delta)P_0(A) - \delta$ and $\bar{P}(A) = (1 + \delta)P_0(A) \forall A \subseteq \mathcal{X}$), both definitions determine the same credal set: the lower and upper probabilities in Eq. (1) correspond to the natural extensions of the ones considered by Utkin and Wiencierz (2013).

Remark also that the maximal number of extreme points for odd n can equivalently be expressed by $\binom{n+1}{\frac{n+1}{2}} \frac{n+1}{4}$. Therefore, the formula of the maximal number of extreme points of the credal set of a PMM coincides with that of probability intervals (Tessem, 1992).

4.2 Computing the Number of Extreme Points for an Arbitrary PMM

In this section, we establish a simple formula that provides an upper bound on the number of extreme points associated with a PMM. Let (P_0, δ) be a PMM, and define

$$\mathcal{L} = \{A \subseteq \mathcal{X} \mid \bar{P}(A) = 1\}. \quad (10)$$

This is a filter of subsets of \mathcal{X} , and as a consequence also a poset with respect to set inclusion. We can use it to bound the number of extreme points of a PMM.

Proposition 7 Consider a PMM (P_0, δ) , and let \mathcal{L} be given by Eq. (10). Then, the number of extreme points of $\mathcal{M}(P_0, \delta)$ is bounded above by:

$$\sum_{A \in \mathcal{L}} \left| \bigcap_{B \subseteq A, B \in \mathcal{L}} B \right|. \quad (11)$$

Furthermore, the number of extreme points coincides with this upper bound if and only if $P_0(A) > \frac{1}{1+\delta}$ for every $A \in \mathcal{L}$.

The following example illustrates the result.

Example 3 Consider a four-element space $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ with probabilities 0.1, 0.1, 0.3 and 0.5, respectively, and let $\delta = 0.3$. The poset (\mathcal{L}, \subseteq) is given by

$$\mathcal{L} = \{\mathcal{X}, \{x_2, x_3, x_4\}, \{x_1, x_3, x_4\}, \{x_3, x_4\}\}$$

Eq. (11) provides an upper bound for the number of extreme points of $\mathcal{M}(P_0, \delta)$. Specifically, it is easy to see that for any $A \in \mathcal{L}$, it holds that:

$$\left| \bigcap_{B \subseteq A, B \in \mathcal{L}} B \right| = |\{x_3, x_4\}| = 2;$$

therefore, the number of extreme points of $\mathcal{M}(P_0, \delta)$ is bounded by:

$$\sum_{A \in \mathcal{L}} \left| \bigcap_{B \subseteq A, B \in \mathcal{L}} B \right| = 2 + 2 + 2 + 2 = 8.$$

Moreover, this bound is tight, taking into account that $P_0(A) > \frac{1}{1+\delta} \forall A \in \mathcal{L}$, and applying Proposition 7. ♦

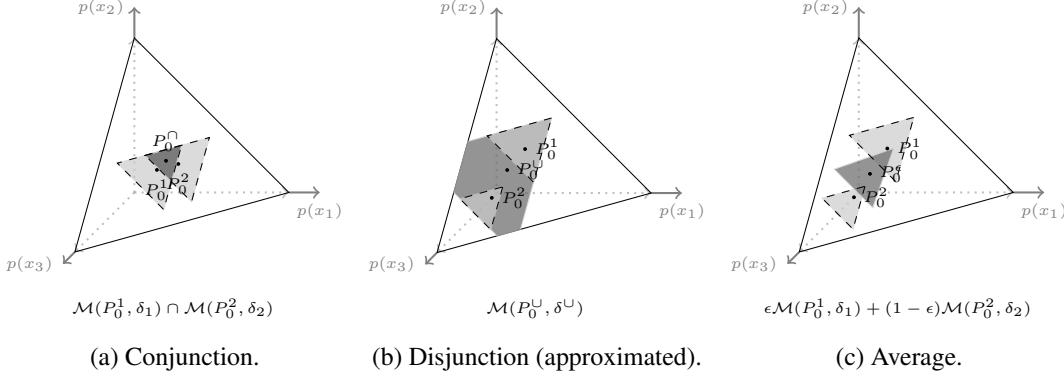


Figure 1: Illustration of combination rules. Initial credal sets are in light grey and delimited by dashed lines, combination results in dark grey.

However, when the additional condition given in Proposition 7 is not satisfied, the formula of Eq. (11) provides only an upper bound of the number of extreme points.

Example 4 Take $\mathcal{X} = \{x_1, x_2, x_3\}$, the uniform distribution P_0 on $\mathcal{P}(\mathcal{X})$ and $\delta = 0.5$. It holds that $\mathcal{L} = \{\{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \mathcal{X}\}$. By Eq. (11), the number of extreme points is bounded above by:

$$\sum_{A \in \mathcal{L}} \left| \bigcap_{B \subseteq A, B \in \mathcal{L}} B \right| = 2 + 2 + 2 + 0 = 6.$$

However, $\mathcal{M}(P_0, \delta)$ has only three extreme points: $(0.5, 0.5, 0)$, $(0.5, 0, 0.5)$ and $(0, 0.5, 0.5)$. Thus, the bound given by Eq. (11) is not tight. Note moreover that in this case $P_0(\{x_1, x_2\}) = \frac{2}{3} = \frac{1}{1+\delta}$. ♦

5. Information Fusion of PMMs

When two credal sets $\mathcal{M}(P_0^1, \delta_1)$ and $\mathcal{M}(P_0^2, \delta_2)$ are provided to describe our uncertainty over \mathcal{X} , one often needs to combine them into a single model. Three classical ways to achieve such a combination are to consider the conjunction (intersection), the disjunction (union) or the average (convex mixture) of the models. The results of these combinations is illustrated in Figure 1, where the specific used models are described in Examples 5, 6 and 7 for the conjunction, disjunction and average, respectively.

Before studying these three cases, we show a useful result which can be derived from Lemma 4.

Proposition 8 Let $\mathcal{M}(P_0, \delta)$ denote the credal set associated with a PMM (P_0, δ) by means of Eq. (2). Then, a probability measure P belongs to $\mathcal{M}(P_0, \delta)$ if and only if:

$$P(\{x\}) \leq (1 + \delta)P_0(\{x\}) \quad \forall x \in \mathcal{X}.$$

Thus, the credal set $\mathcal{M}(P_0, \delta)$ is not only determined by the restrictions of the lower and upper probabilities to singletons (as we know from the connection with probability intervals established in Theorem 1) but moreover that only the upper bounds on the singletons are necessary. This fact is instrumental in the derivation of the results of this section.

5.1 Conjunction

Let $\mathcal{M}(P_0^\cap, \delta^\cap) := \mathcal{M}(P_0^1, \delta_1) \cap \mathcal{M}(P_0^2, \delta_2)$ denote the credal set obtained by conjunctively combining $\mathcal{M}(P_0^1, \delta_1)$ and $\mathcal{M}(P_0^2, \delta_2)$. We then have the following result.

Proposition 9 *The set $\mathcal{M}(P_0^\cap, \delta^\cap)$ is non-empty if and only if*

$$\sum_{x \in \mathcal{X}} \min \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\}), 1\} \geq 1. \quad (12)$$

In that case, it is induced by the PMM (P_0^\cap, δ^\cap) such that

$$\delta^\cap = \left(\sum_{x \in \mathcal{X}} \min \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\})\} \right) - 1 \quad (13)$$

$$P_0^\cap(\{x\}) = \frac{\min \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\})\}}{1 + \delta^\cap}. \quad (14)$$

In the particular case where $P_0^1 = P_0^2$, Eq. (12) is always satisfied because:

$$\begin{aligned} \sum_{x \in \mathcal{X}} \min \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\}), 1\} \\ = \sum_{x \in \mathcal{X}} \min \{(1 + \min\{\delta_1, \delta_2\})P_0(\{x\}), 1\} \geq \sum_{x \in \mathcal{X}} P_0(\{x\}) = 1, \end{aligned}$$

and the values of δ^\cap and P_0^\cap given in Eqs. (13) and (14) become $\delta^\cap = \min\{\delta_1, \delta_2\}$ and $P_0^\cap = P_0$.

Example 5 Consider the space $\mathcal{X} = \{x_1, x_2, x_3\}$ and the two models given by $\delta_1 = \delta_2 = 0.3$ and:

$$P_0^1 = (0.3, 0.3, 0.4), \quad P_0^2 = (0.4, 0.3, 0.3),$$

that are such that $\mathcal{M}(P_0^1, \delta_1) \cap \mathcal{M}(P_0^2, \delta_2) \neq \emptyset$. Their conjunction is given by $P_0^\cap = (1/3, 1/3, 1/3)$ and $\delta^\cap = 0.17$. The result is illustrated on Figure 1a, where the initial two PMMs are in light gray, and the resulting conjunction is in dark gray. ♦

5.2 Disjunction

When the intersection of two credal sets is empty (they are conflicting), an alternative is to consider their union, that is to consider $\mathcal{M}(P_0^1, \delta_1) \cup \mathcal{M}(P_0^2, \delta_2)$ or its convex hull, since $\mathcal{M}(P_0^1, \delta_1) \cup \mathcal{M}(P_0^2, \delta_2)$ will not be convex in general.

The convex hull $\text{conv}(\mathcal{M}(P_0^1, \delta_1) \cup \mathcal{M}(P_0^2, \delta_2))$ will also not be induced by a PMM in general. However, we can easily provide a best outer-approximating PMM (P_0^\cup, δ^\cup) using the fact that any outer-approximation of $\mathcal{M}(P_0^1, \delta_1) \cup \mathcal{M}(P_0^2, \delta_2)$ must satisfy the constraint

$$\max \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\})\} \geq P(\{x\}) \quad \forall x \in \mathcal{X}.$$

Indeed, using the same arguments as in Proposition 9, we can define

$$\delta^\cup = \left(\sum_{x \in \mathcal{X}} \max \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\})\} \right) - 1$$

and

$$P_0^{\cup}(\{x\}) = \frac{\max \{(1 + \delta_1)P_0^1(\{x\}), (1 + \delta_2)P_0^2(\{x\})\}}{1 + \delta^{\cup}}$$

so that $\mathcal{M}(P_0^{\cup}, \delta^{\cup}) \supseteq \mathcal{M}(P_0^1, \delta_1) \cup \mathcal{M}(P_0^2, \delta_2)$. To see that this inclusion holds, simply note that for any event A , we have

$$\sum_{x \in A} \max \{\bar{P}^1(x), \bar{P}^2(x)\} \geq \max \left\{ \sum_{x \in A} \bar{P}^1(x), \sum_{x \in A} \bar{P}^2(x) \right\}$$

where \bar{P}^1, \bar{P}^2 are the upper probabilities induced by (P_0^1, δ_1) and (P_0^2, δ_2) , respectively.

Example 6 Consider the space $\mathcal{X} = \{x_1, x_2, x_3\}$ and the two models given by $\delta_1 = 0.2, \delta_2 = 0.3$ and:

$$P_0^1 = (0.3, 0.4, 0.3), \quad P_0^2 = (0.2, 0.2, 0.6),$$

for which $\mathcal{M}(P_0^1, \delta_1) \cap \mathcal{M}(P_0^2, \delta_2) = \emptyset$. Their outer-approximation is $P_0^{\cup} = (0.222, 0.297, 0.481)$ and $\delta^{\cup} = 0.62$. The result is illustrated on Figure 1b, where the initial two PMMs are in light gray, and the resulting outer-approximation of the disjunction is in dark gray. ♦

5.3 Mixture

The mixture of two PMMs, that is, the computation of

$$\mathcal{M}(P_0^{\epsilon}, \delta_{\epsilon}) := \epsilon \mathcal{M}(P_0^1, \delta_1) + (1 - \epsilon) \mathcal{M}(P_0^2, \delta_2)$$

for a given $\epsilon \in (0, 1)$ is straightforward when applying results established by [Moral and del Sagrado \(1998\)](#) for probability intervals. In particular, the model $\mathcal{M}(P_0^{\epsilon}, \delta_{\epsilon})$ is described by the constraints

$$\epsilon(1 + \delta_1)P_0^1(\{x\}) + (1 - \epsilon)(1 + \delta_2)P_0^2(\{x\}) \geq P(\{x\}) \quad \forall x \in \mathcal{X}$$

on a probability measure P . From this, we easily deduce that

$$\begin{aligned} 1 + \delta_{\epsilon} &= \sum_{x \in \mathcal{X}} \epsilon(1 + \delta_1)P_0^1(\{x\}) + (1 - \epsilon)(1 + \delta_2)P_0^2(\{x\}) \\ &= \epsilon(1 + \delta_1) \sum_{x \in \mathcal{X}} P_0^1(\{x\}) + (1 - \epsilon)(1 + \delta_2) \sum_{x \in \mathcal{X}} P_0^2(\{x\}) = \epsilon(1 + \delta_1) + (1 - \epsilon)(1 + \delta_2) \end{aligned}$$

$$\text{and } P_0^{\epsilon}(\{x\}) = \frac{\epsilon(1 + \delta_1)P_0^1(\{x\}) + (1 - \epsilon)(1 + \delta_2)P_0^2(\{x\})}{1 + \delta_{\epsilon}}.$$

Example 7 Consider the initial models of Example 6 with $\epsilon = 0.5$. We obtain the model $p_0^{\epsilon} = (0.248, 0.296, 0.456)$ and $\delta^{\epsilon} = 0.25$. The result is illustrated on Figure 1c, where the initial two PMMs are in light gray, and the resulting average is in dark gray. ♦

Other, more elaborate combinations can be derived from these basic ones; see for example ([Moral and del Sagrado, 1998](#); [Walley, 1982](#)).

6. Conclusion

This paper presents some advances on the study of the PMM as a model within Imprecise Probability Theory. Our results show that the PMM is a particular type of probability interval (Thm. 1). This means that any property satisfied by a probability interval is also satisfied by a PMM. In this paper, we have studied the extreme points of the credal set induced by a PMM, and proven that the maximal number of extreme points coincides with that of probability intervals (Thm. 6). In addition, we have established a formula that gives an upper bound for the number of extreme points and that is somewhat easier to apply.

With respect to the connection with other imprecise probability models, we have also given necessary and sufficient conditions for a PMM to induce a belief function, improving upon some results from the literature. Our results show that those belief functions that are attained as a PMM are quite specific, since the PMM imposes strong constraints on the focal elements. Although not reported here, from this it is easy to characterize in which cases the lower probability of a PMM is a minitive function. However, this only happens in even more particular scenarios.

Finally, we have also investigated the properties of the PMM when merging different sources of information, each providing a PMM. In particular, we have seen that the conjunction or the mixture of PMMs give rise to other PMM, while the disjunction of PMMs can be outer-approximated by a PMM. This gives simple tools to perform such combinations.

There are other practical aspects of uncertainty models that we did not study in the present paper, but that would deserve some attention, such as what happens when combine into a joint model PMM models issued from marginal variables. In particular, it would be worth checking whether such operations can be performed efficiently and preserve the form of the initial model, i.e., is the result still a PMM?

Acknowledgements

The research reported in this paper has been supported by project TIN2014-59543-P, and by the project Labex MS2T, financed 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

- A. Antonucci and F. Cuzzolin. Credal sets approximation by lower probabilities: application to credal networks. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 716–725. Springer, 2010.
- P. Boonpumarn, P. Thipwiwatpotjana, and W. Lodwick. When a probability interval is a random set. *Science Asia*, 39:319–327, 2013.
- 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.
- F. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

- L. M. de Campos, J. F. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.
- D. Denneberg. *Non-Additive Measure and Integral*. Kluwer Academic, Dordrecht, 1994.
- S. Destercke. Independence and 2-monotonicity: Nice to have, hard to keep. *International Journal of Approximate Reasoning*, 54(4):478–490, 2013.
- H. Gerber. *An introduction to mathematical risk theory*. Huebner Foundation, 1979.
- P. Guo and H. Tanaka. Decision making with interval probabilities. *European Journal of Operational Research*, 203(2):444–454, 2010.
- S. Moral and J. del Sagrado. Aggregation of imprecise probabilities. In *Aggregation and fusion of imperfect information*, pages 162–188. Springer, 1998.
- R. Pelessoni, P. Vicig, and M. Zaffalon. Inference and risk measurement with the pari-mutuel model. *International Journal of Approximate Reasoning*, 51(9):1145–1158, 2010.
- M. Peters, A. M-C. So, and Y. Ye. Pari-mutuel markets: mechanisms and performance. *Lecture Notes in Computer Science*, 4858:82–95, 2007.
- G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.
- D. Skulj. Discrete time Markov chains with interval probabilities. *International Journal of Approximate Reasoning*, 50(8):1314–1329, 2009.
- H. Tanaka, K. Sugihara, and Y. Maeda. Non-additive measures by interval probability functions. *Information Sciences*, 164:209–227, 2004.
- D. Terrell. A test on the gambler’s fallacy evidence from pari-mutuel gambles. *Journal of Risk and Uncertainty*, 8(3):309–317, 1994.
- B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7 (3–4):95–120, 1992.
- R. Thaler and W. Ziemba. Parimutuel betting markets: racetracks and lotteries. *Journal of Economic Perspectives*, 2:161–174, 1988.
- L. Utkin. A framework for imprecise robust one-class classification models. *Journal of Machine Learning Research and Cybernetics*, 5(3):379–393, 2014.
- L. Utkin and A. Wiencierz. An imprecise boosting-like approach to regression. In *Proceedings of the 8th International Symposium on Imprecise Probability: Theories and Applications*, pages 345–354, 2013.
- P. Walley. The elicitation and aggregation of beliefs. Statistics Research Report 23, University of Warwick, Coventry, 1982.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Efficient Algorithms for Checking Avoiding Sure Loss

Nawapon Nakharutai

NAWAPON.NAKHARUTAI@DURHAM.AC.UK

Matthias C. M. Troffaes

MATTHIAS.TROFFAES@DURHAM.AC.UK

Camila C. S. Caiado

C.C.D.S.CAIADO@DURHAM.AC.UK

Durham University

Durham (United Kingdom)

Abstract

Sets of desirable gambles provide a general representation of uncertainty which can handle partial information in a more robust way than precise probabilities. Here we study the effectiveness of linear programming algorithms for determining whether or not a given set of desirable gambles avoids sure loss (i.e. is consistent). We also suggest improvements to these algorithms specifically for checking avoiding sure loss. By exploiting the structure of the problem, (i) we slightly reduce its dimension, (ii) we propose an extra stopping criterion based on its degenerate structure, and (iii) we show that one can directly calculate feasible starting points in various cases, therefore reducing the effort required in the presolve phase of some of these algorithms. To assess our results, we compare the impact of these improvements on the simplex method and two interior point methods (affine scaling and primal-dual) on randomly generated sets of desirable gambles that either avoid or do not avoid sure loss. We find that the simplex method is outperformed by the primal-dual and affine scaling methods, except for very small problems. We also find that using our starting feasible point and extra stopping criterion considerably improves the performance of the primal-dual and affine scaling methods.

Keywords: avoiding sure loss; linear programming; benchmarking; simplex method; affine scaling method; primal-dual method; algorithm.

1. Introduction

Consider a subject modelling uncertainty about an experiment. One way of doing so, is by using gambles. A gamble is a (for instance, monetary) transaction that depends on the outcome of the experiment. To express her beliefs about the outcomes of the experiment, the subject can simply state which gambles she would accept, instead of directly specifying probabilities for the outcomes. A set of gambles that are considered acceptable to a subject is called a *set of desirable gambles*.

Williams (1975, 2007) was the first to give a full axiomatic treatment for sets of desirable gambles, and formalized a consistency principle called *avoiding sure loss*. This principle dictates that no combination of acceptable gambles should lead to a certain loss. Walley (1991, p.175) proposed a linear programming problem for checking avoiding sure loss, which was further studied and extended by various authors (Walley et al., 2004; Quaeghebeur, 2014). These linear programming problems can be solved by methods such as the primal-dual, simplex and affine scaling methods.

In the literature, to the best of our knowledge, there has been little to no discussion about which algorithm should be used to solve linear programming problems for avoiding sure loss. Walley (1991, p. 511) suggested that Karmarkar's method may be useful for solving large problems.

The main contribution of this paper is a comparative study and analysis of how we can solve linear programs for avoiding sure loss most effectively, by looking at the three methods mentioned above. We exploit the structure of this program and also the interactions between the structure

and the details of the algorithms. Specifically, we slightly reduce its dimension. We also propose an extra stopping criterion based on its degenerate structure. Finally, we show that one can directly calculate feasible starting points in various cases, therefore reducing the effort required in the presolve phase of some of these algorithms. Through a simulation study, we compare the impact of these improvements in the simplex method and two interior point methods (affine scaling and primal-dual) on randomly generated sets of desirable gambles that either avoid or do not avoid sure loss.

This paper is organised as follows. After reviewing linear programming problems and avoiding sure loss in Section 2, we give a linear programming problem and its dual for checking avoiding sure loss in Section 3. In Sections 4 to 6, we give a brief outline of primal-dual, simplex and affine scaling methods. We also study the different benefits of these three methods with respect to checking for avoiding sure loss. Algorithms for generating random sets of desirable gambles are outlined in Section 7 followed by a comparison of the efficiency of the methods. Section 8 concludes the paper.

2. Preliminaries

2.1 Linear programming problems

A linear programming problem is a problem of optimising a linear function (objective function) subject to constraints of linear equalities and linear inequalities. Because maximising a linear function is equivalent to minimising that function with a sign change, and because any linear inequality can be rewritten as a linear equality by adding non-negative slack variables, we have that every linear programming problem can be formulated as follows:

$$\min \mathbf{c}^T \mathbf{x} \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \text{ and } \mathbf{x} \geq \mathbf{0}, \quad (\text{P})$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank m and $m \leq n$. We call Eq. (P) the *primal* problem. The dual of Eq. (P) is:

$$\max \mathbf{b}^T \mathbf{y} \quad \text{subject to} \quad \mathbf{A}^T \mathbf{y} + \mathbf{t} = \mathbf{c} \text{ and } \mathbf{t} \geq \mathbf{0} \quad (\mathbf{y} \text{ free}). \quad (\text{D})$$

We can solve either the primal or the dual problem because they have the same solution.

A solution that satisfies all constraints is called a *feasible solution*. An *optimal solution* is a feasible solution that achieves the optimal value of the objective function. A *basic feasible solution* is a feasible solution with at most m non-zero variables. It can be shown that every basic feasible solution is an extreme point, and vice versa ([Fang and Puthenpura, 1993](#)).

A linear programming problem is *degenerate* when it has basic feasible solutions in which fewer than m variables are non-zero. As we shall see, one way to check avoiding sure loss is by solving a degenerate linear programming problem where $\mathbf{b} = \mathbf{0}$. In that case, the following lemma is helpful. It is a generalised version of an exercise in [Vanderbei \(2001, p. 42, exercise 3.4\)](#).

Lemma 1 *The linear programming problem $\min \mathbf{c}^T \mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \geq \mathbf{0}$ either has an optimal value that is zero, or is unbounded.*

This lemma is very useful since it tells us that as soon as we find a feasible solution with a negative objective function value, then the problem is unbounded.

2.2 Avoiding sure loss

Throughout this paper, let Ω be a finite set of uncertain outcomes. A gamble is a bounded real-valued function on Ω . We think of a gamble as an uncertain reward expressed in units of utility.

Let \mathcal{D} be a finite set of gambles that a subject is willing to accept; we refer to \mathcal{D} as the subject's set of desirable gambles. The desirability axioms state that a non-negative combination of desirable gambles should not produce a sure loss (Walley, 1991, p.151). In that case, we say that \mathcal{D} avoids sure loss. Formally:

Definition 2 (Walley, 1991, p.151) *The set $\mathcal{D} = \{f_1, \dots, f_n\}$ is said to avoid sure loss if for all $n \in \mathbb{N}$, $\lambda_1, \dots, \lambda_n \geq 0$ and $f_1, \dots, f_n \in \mathcal{D}$:*

$$\max_{\omega \in \Omega} \left(\sum_{i=1}^n \lambda_i f_i(\omega) \right) \geq 0. \quad (1)$$

3. Linear programs for checking avoiding sure loss

In this section, we study linear programming problems to check avoiding sure loss. The following linear programming problem is given by Walley (1991, p.175):

Theorem 3 (Walley, 1991, p.175) *The set $\mathcal{D} = \{f_1, \dots, f_n\}$ avoids sure loss if and only if the optimal value of the following linear programming problem is zero:*

$$(A1) \quad \min \alpha \quad (2)$$

$$\text{subject to } \forall \omega \in \Omega : \sum_{i=1}^n \lambda_i f_i(\omega) \leq \alpha \quad (3)$$

$$\text{where } \lambda_i \geq 0. \quad (4)$$

We propose an alternative linear programming problem, which is slightly smaller in size, and which has only non-negative variables:

Theorem 4 (Nakharutai, 2015, p.32) *Choose any $\omega^0 \in \Omega$. The set $\mathcal{D} = \{f_1, \dots, f_n\}$ avoids sure loss if and only if the optimal value of the following linear programming problem is zero:*

$$(A2) \quad \min \sum_{i=1}^n \lambda_i f_i(\omega^0) + \alpha \quad (5)$$

$$\text{subject to } \forall \omega \in \Omega \setminus \{\omega^0\} : \sum_{i=1}^n \lambda_i (f_i(\omega^0) - f_i(\omega)) + \alpha \geq 0 \quad (6)$$

$$\text{where } \lambda_i, \alpha \geq 0. \quad (7)$$

Note that (A1) and (A2) are fully degenerate because their right hand side is zero. Therefore, Theorem 1 applies to both of these problems.

When solving linear programs, typical algorithms such as primal-dual, simplex and affine scaling, require all free variables to be rewritten as a difference of two non-negative variables. So (A1) needs the introduction of an extra variable due to the presence of a free variable. Moreover, (A2) already has one fewer variable than (A1). Therefore, solving (A2) is easier than (A1).

Let's consider the dual of the problem (A2).

Theorem 5 ([Nakharutai, 2015](#), p.49) Choose any $\omega^0 \in \Omega$. The set $\mathcal{D} = \{f_1, \dots, f_n\}$ avoids sure loss if and only if the following linear programming problem has a feasible solution.

$$(B1) \quad \max \quad 0 \quad (8)$$

$$\text{subject to} \quad \forall f_j \in \mathcal{D} : \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega^0) - f_j(\omega))p(\omega) \leq f_j(\omega^0) \quad (9)$$

$$\sum_{\omega \in \Omega \setminus \{\omega^0\}} p(\omega) \leq 1 \quad (10)$$

$$\text{where} \quad p(\omega) \geq 0. \quad (11)$$

How should we choose ω^0 ? Looking at the primal problem (A2), it is not obvious which ω^0 we should choose. However, since optimality in the primal problem corresponds to feasibility in the dual problem ([Goh and X.Q.Yang, 2002](#), p.104), if we choose an ω^0 for which most values $f_j(\omega^0)$ are non-negative, then we can start (B1) closer to a feasible solution in the dual, and therefore closer to an optimal solution in the primal. For example, if there is an ω^0 for which $f_j(\omega^0) \geq 0$ for all j , then we immediately find a feasible solution by setting $p(\omega) = 0$ for all $\omega \neq \omega^0$.

4. Checking avoiding sure loss using primal-dual methods

4.1 Primal-dual methods

The primal-dual method solves the primal and dual problems simultaneously, see [Fang and Puthenpura \(1993\)](#); [Griva et al. \(2009\)](#) for more detail. It finds an optimal primal-dual solution $[\mathbf{x}^* \ \mathbf{y}^* \ \mathbf{t}^*]$ by repeatedly solving the following equalities:

$$\begin{bmatrix} \mathbf{Ax} - \mathbf{b} \\ \mathbf{A}^\top \mathbf{y} + \mathbf{t} - \mathbf{c} \\ \mathbf{x}^\top \mathbf{t} \end{bmatrix} = \mathbf{0} \quad \text{subject to} \quad \mathbf{x}, \mathbf{t} \geq \mathbf{0} \quad (12)$$

whilst keeping the variables \mathbf{x} and \mathbf{t} positive. The algorithm will stop when the primal residual $\mathbf{Ax} - \mathbf{b}$, dual residual $\mathbf{A}^\top \mathbf{y} + \mathbf{t} - \mathbf{c}$ and duality gap $\mathbf{x}^\top \mathbf{t}$ are small enough or when an unboundedness criterion is satisfied in either the primal or the dual problems ([Fang and Puthenpura, 1993](#)).

Theoretically, the primal-dual method generates iterates \mathbf{x} , \mathbf{y} , \mathbf{t} , where \mathbf{x} , $\mathbf{t} > 0$, that stay in the feasible region. However, in practical implementations, keeping \mathbf{x} , \mathbf{y} , \mathbf{t} in the feasible region is very difficult because of numerical problems. Therefore, practical implementations of primal-dual algorithms start with an arbitrary point $[\mathbf{x} \ \mathbf{y} \ \mathbf{t}]$ where \mathbf{x} , $\mathbf{t} > \mathbf{0}$ and generate iterate points that converge to an optimal solution. Even though there is no convergence proof, this approach seems to work well in practice ([Fang and Puthenpura, 1993](#)).

4.2 Corresponding linear programming problems

The primal problem (A2) can be easily written as (P) by adding slack variables:

$$(A3) \quad \min \quad \sum_{i=1}^n \lambda_i f_i(\omega^0) + \alpha \quad (13)$$

$$\text{subject to} \quad \forall \omega \in \Omega \setminus \{\omega^0\} : \sum_{i=1}^n \lambda_i (f_i(\omega^0) - f_i(\omega)) + \alpha - s(\omega) = 0 \quad (14)$$

$$\text{where} \quad \lambda_i, \alpha, s(\omega) \geq 0. \quad (15)$$

Note that these equalities are linearly independent due to the presence of distinct slack variables in each equality, so the corresponding matrix A has full rank. This will be the case for every system of equalities that we write down further as well; we will not repeat this point.

The dual problem (B1) can be written as (D) again by adding slack variables:

$$(B2) \quad \max \quad 0 \quad (16)$$

$$\text{subject to} \quad \forall f_j \in \mathcal{D} : \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega^0) - f_j(\omega)) p(\omega) + t_j = f_j(\omega^0) \quad (17)$$

$$\sum_{\omega \in \Omega \setminus \{\omega^0\}} p(\omega) + q = 1 \quad (18)$$

$$\text{where} \quad p(\omega), t_j, q \geq 0. \quad (19)$$

To apply the primal-dual method, (A3) and (B2) are easily written in the form of Eq. (12). Next, we will investigate the structure of the problems (A3) and (B2) to choose a suitable starting point.

4.3 Starting points and extra stopping criteria

For the dual problem (B2), a starting point can be $q^0 = p^0(\omega) = 1/|\Omega|$ for all $\omega \in \Omega \setminus \{\omega^0\}$, because these $p^0(\omega)$ and q^0 satisfy Eq. (18). For simplicity, we suggest choosing $t_j = 1$ for all j . Note that this solution may not be feasible (if it were, we would have solved the problem!).

For the primal problem (A3), we can find an initial interior feasible solution as follows:

Theorem 6 *An initial interior feasible solution of the specific linear programming problem*

$$\min \quad c_1 x_1 + c_2 x_2 + \dots + c_n x_n + \alpha \quad (20)$$

$$\text{subject to} \quad \forall j = 1, \dots, m : \sum_{k=1}^n a_{jk} x_k + \alpha - s_j = 0 \quad (21)$$

$$\text{where} \quad \alpha \geq 0, x_k \geq 0 \text{ and } s_j \geq 0. \quad (22)$$

is given by $x_k = 1$, $\alpha = 1 + \max\{0, -\delta\}$ with $\delta := \min_j \{\sum_{k=1}^n a_{jk}\}$ and $s_j = \alpha + \sum_{k=1}^n a_{jk}$.

Proof We must show that Eq. (21) is satisfied, and that all variables are strictly positive.

Clearly, Eq. (21) is satisfied by our choice of s_j , all $x_k = 1 > 0$, and $\alpha \geq 1 > 0$. Finally, note that also all $s_j > 0$ because

$$s_j = \alpha + \sum_{k=1}^n a_{jk} \geq \alpha + \delta \geq 1 - \delta + \delta > 0. \quad (23)$$

where we used the definitions of δ and α respectively. ■

Note that for problem (A3), even though we start with a feasible solution, the solution does not necessarily remain feasible due to numerical rounding errors. In particular, we can apply our extra stopping criterion from Theorem 1 only if the primal residual $\mathbf{Ax} - \mathbf{b}$ is small.

In the next two sections, we look at the simplex and affine scaling methods, and we also suggest suitable linear programming representations.

5. Checking avoiding sure loss using simplex methods

5.1 Simplex methods and pivoting

The simplex method is an iterative method. At every iteration, we move from a current extreme point to another extreme point that decreases the objective function value. If a linear programming problem is in the form

$$\min \mathbf{c}^\top \mathbf{x} \quad \text{subject to } \mathbf{Ax} + \mathbf{s} = \mathbf{b} \quad \text{where } \mathbf{x}, \mathbf{s}, \mathbf{b} \geq 0, \quad (\text{S})$$

then we immediately obtain a starting extreme point by setting $\mathbf{s} = \mathbf{b}$ and $\mathbf{x} = \mathbf{0}$.

A brief outline of the simplex method is given as follows (see [Fang and Puthenpura \(1993\)](#) for more detail). We first write (S) in the following format:

$$\begin{bmatrix} \mathbf{c}^\top & \mathbf{0}^\top & 0 \\ \mathbf{A} & \mathbf{I} & \mathbf{b} \end{bmatrix} \quad (24)$$

If there is a negative value in the top row, then we leave the current extreme point and move to an improved extreme point by performing a pivot via row operations. We repeat this until there is no negative value in the first row, or until we can no longer pivot.

5.2 Corresponding linear programming problems

Note that by multiplying Eq. (14) by -1 , problem (A3) can be rewritten as (S). Since all the right hand side constraints are zero, there is only one extreme point that is $\mathbf{0}$. Consequently, $\mathbf{0}$ is the only extreme point. In this case either the optimal value is zero or the problem is unbounded.

Since the value of the objective function is always zero, we cannot apply Theorem 1. It is also worth noting that problem (A3) is degenerate and that the simplex method may cycle in such cases ([Hoffman, 1953](#)). Cycling can be detected by checking Bland's rule or Lexicographic Rule ([Fang and Puthenpura, 1993](#), p.44) resulting in more calculations. Therefore, the simplex method may perform poorly when solving problem (A3).

We now look at the dual problem (B1). To convert (B1) into the standard form (S), if there exists $f_j(\omega^0) < 0$ for some j , then we multiply the corresponding constraint by -1 to make the right hand side non-negative and then add artificial variables. We obtain the following linear programming problem for which we can immediately give an initial extreme point.

Corollary 7 *The set $\mathcal{D} = \{f_1, \dots, f_n\}$ avoids sure loss if and only if the optimal value of the following linear programming problem is zero.*

$$(B3) \quad \min \sum_{j \in N} t_j \quad (25)$$

$$\text{subject to } \forall j \in N : \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega) - f_j(\omega^0)) p(\omega) - s_j + t_j = -f_j(\omega^0) \quad (26)$$

$$\forall j \in I \setminus N : \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega^0) - f_j(\omega)) p(\omega) + u_j = f_j(\omega^0) \quad (27)$$

$$\sum_{\omega \in \Omega \setminus \{\omega^0\}} p(\omega) + q = 1 \quad (28)$$

$$\text{where } p(\omega), q, s_j, t_j, u_j \geq 0 \quad (29)$$

with $I := \{1, \dots, n\}$ and $N := \{j \in N : f_j(\omega^0) < 0\}$.

We can then choose an initial extreme point to be $t_j = -f_j(\omega^0)$, $u_j = f_j(\omega^0)$, $q = 1$ and $p(\omega) = s_j = 0$.

Now, to check avoiding sure loss using the simplex method, we can solve either (A3) or (B3). If we want to avoid degeneracy and cycling, then we should solve (B3) rather than (A3).

6. Checking avoiding sure loss using affine scaling methods

6.1 Affine scaling methods

The idea of the affine scaling method is to generate a sequence of interior feasible solutions by repeatedly solving Eq. (P) such that the corresponding objective function values are decreasing. The assumption is that this sequence converges to the optimal solution if it exists. The method needs a starting interior feasible point, and we stop when the difference between objective function values is small enough or an unboundedness criterion is satisfied (see [Fang and Puthenpura \(1993\)](#); [Griva et al. \(2009\)](#); [Saigal \(1995\)](#) for more detail).

6.2 Extra stopping criteria

Recall that the right hand side constraints of (A3) are zero and the affine scaling method generates a sequence of interior feasible solutions. Therefore, we can apply Theorem 1. Specifically, the method can stop as soon as it finds a feasible solution with a negative objective function value. In this case, the problem is unbounded.

6.3 Initial feasible interior points

In practical implementations, there are several mechanisms for finding initial interior feasible points which require solving another linear programming problem. For the problem (A3), we do not need to use such mechanisms since we can apply Theorem 6 to find an initial interior feasible point. For the problem (B2), we can apply the following mechanism to obtain a starting interior feasible point ([Fang and Puthenpura, 1993](#)).

Given constraints $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$, we choose any $\mathbf{x}^0 > \mathbf{0}$ and calculate $\mathbf{y} = \mathbf{b} - \mathbf{Ax}^0$. If $\mathbf{y} = \mathbf{0}$, then \mathbf{x}^0 is an interior feasible solution. Otherwise, we solve

$$\min \gamma \quad \text{subject to} \quad \mathbf{Ax} + \mathbf{y}\gamma = \mathbf{b} \quad \text{where } \mathbf{x} \geq \mathbf{0}, \gamma \geq 0 \quad (30)$$

by the affine scaling method with an interior feasible solution $[\mathbf{x} \ \gamma] = [\mathbf{x}^0 \ 1]$. If the affine scaling method finds an optimal solution $[\mathbf{x}^* \ \gamma^*]$ such that $\gamma^* = 0$, then \mathbf{x}^* is an interior feasible solution of the original problem.

Let's write (B2) in the explicit form of Eq. (30). Let $\Omega \setminus \{\omega^0\} = \{\omega^1, \dots, \omega^m\}$, we choose $\mathbf{x}^0 = [p^0(\omega_1) \ \dots \ p^0(\omega_m) \ t_1^0 \ \dots \ t_n^0 \ q^0] > \mathbf{0}$ and calculate $\mathbf{y} := \mathbf{b} - \mathbf{Ax}^0$ as follows:

$$y_j := f_j(\omega^0) - \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega^0) - f_j(\omega))p^0(\omega) - t_j^0 \quad (31)$$

$$z := 1 - \left(\sum_{\omega \in \Omega \setminus \{\omega^0\}} p^0(\omega) + q^0 \right) \quad (32)$$

Note that we can choose $t_j^0 = 1$ for simplicity. We also choose $q^0 = p^0(\omega) = 1/|\Omega|$ for all $\omega \in \Omega \setminus \{\omega^0\}$ so that $z = 0$.

We then solve the following problem.

$$(B4) \quad \min \gamma \quad (33)$$

$$\text{subject to} \quad \forall f_j \in \mathcal{D} : \sum_{\omega \in \Omega \setminus \{\omega^0\}} (f_j(\omega^0) - f_j(\omega))p(\omega) + t_j - y_j\gamma = f_j(\omega^0) \quad (34)$$

$$\sum_{\omega \in \Omega \setminus \{\omega^0\}} p(\omega) + q = 1 \quad (35)$$

where \mathbf{x}^0 is an interior feasible point of (B4). After obtaining an optimal solution for (B4), if $\gamma^* = 0$, then \mathbf{x}^* is an interior feasible solution for (B2) (and therefore also an optimal solution for (B2)); otherwise, there is no feasible solution.

7. Algorithms and numerical results

7.1 Algorithms for generating random sets of gambles

We give two algorithms for generating random sets of desirable gambles \mathcal{D} that either avoid or do not avoid sure loss. In this section, Ω denotes the set of outcomes, $\Delta(\Omega)$ denotes the unit simplex over Ω and \mathcal{D} denotes a set of desirable gambles. Algorithm 1 generates a random set of desirable gambles that avoids sure loss. Starting from a \mathcal{D} that avoids sure loss, Algorithm 2 generates another gamble that can be added to violate consistency, thereby generating a set of desirable gambles that does not avoid sure loss.

7.2 Numerical results

We solve problem (A3) for checking avoiding sure loss by three methods. The simplex and primal-dual methods are available in MATLAB, while the affine scaling method is not. Unfortunately, the MATLAB implementation of the primal-dual method does not allow us to specify the initial

Algorithm 1: Generate a random set of desirable gambles \mathcal{D} that avoids sure loss

Input : Number of gambles $J := |\mathcal{D}|$

Number of outcomes $|\Omega|$

Number of probability mass functions k

Output: A set of desirable gambles \mathcal{D} that avoids sure loss

Stage 1. For each $i = 1 : k$, sample a single p_i uniformly from the unit simplex $\Delta(\Omega)$ as follows:

- (a) For each ω , sample $q_i(\omega)$ uniformly from $(0, 1)$.
- (b) For each ω , set $p_i(\omega) := (-\ln q_i(\omega)) / (-\sum_{\omega} \ln q_i(\omega))$.

Stage 2. Generate a set of desirable gambles \mathcal{D}

- (a) For each ω and j , sample $g_j(\omega)$ uniformly from $(0, 1)$.
 - (b) For each j , set $\underline{P}(g_j) := \min_{i=1}^k \sum_{\omega} p_i(\omega) g_j(\omega)$.
 - (c) Set $\mathcal{D} := \{g_j - \underline{P}(g_j); j \in J\}$.
-

Algorithm 2: Generate a random set of desirable gambles \mathcal{D} that does not avoid sure loss

Input : A set of desirable gambles $\mathcal{E} = \{f_1, \dots, f_J\}$ that avoids sure loss

$\delta > 0$

Output: A set of desirable gambles \mathcal{D} that does not avoid sure loss

- (a) For each ω , sample $g(\omega)$ uniformly from $(0, 1)$.
- (b) Solve the following linear program:

$$\begin{aligned}
 (\text{C}) \quad & \min \quad \beta \\
 \text{subject to} \quad & \forall \omega \in \Omega : \sum_{j=1}^J \lambda_j f_j(\omega) - \beta \leq -g(\omega) \\
 \text{where} \quad & \lambda_i \geq 0 \quad (\beta \text{ free}).
 \end{aligned}$$

- (c) Set $\underline{P}(g) := \beta + \delta$.
 - (d) Set $\mathcal{D} := \mathcal{E} \cup \{g - \underline{P}(g)\}$.
-

starting point and to add an extra stopping criterion (Theorem 1). To compare these three methods, we wrote our own implementation of the improved affine scaling and the improved primal-dual methods. Specifically, the extra stopping criterion and our method for the initial interior feasible point were included in our implementation of these two algorithms.

We generate two types of random sets of desirable gambles. For each type, we consider the scenarios $|\mathcal{D}| = J = 2^i$ for $i \in \{1, 2, \dots, 8\}$ and $|\Omega| = 2^j$ for $j \in \{1, 2, \dots, 8\}$. We also fixed $k = 2^4$; varying k had little impact on the results. We first generate a set that avoids sure loss using Algorithm 1. Next, we generate a set that does not avoid sure loss using Algorithm 2 with $\delta = 0.05$ and with \mathcal{E} provided by Algorithm 1. We then benchmark the primal-dual, simplex and affine scaling methods by measuring their computational times applied to each generated set.

For each set of desirable gambles, we assume that we do not know whether it avoids sure loss or not, and pose the linear programming problem in the format of the primal problem (A3). We then solve each case using these three methods. For each method, we run the algorithm twice to remove any warm-up effects that can happen in the first run, and we only measure the corresponding computational time taken in the second run. We repeat the process 1000 times and present a summary of the results in Fig. 1.

Figure 1 shows the average computational time spent during each method when checking avoiding sure loss. In the left column, the sets of desirable gambles avoid sure loss and in the right column, they do not avoid sure loss. Each row represents a different number of desirable gambles. The vertical axis represents the computational time. The horizontal axis shows the number of outcomes. The computational time is averaged over 1000 random sets of desirable gambles. The error bars on the figures represent approximate 95% confidence intervals on the mean computation time. These are barely visible due to the sufficiently large sample size.

Overall, in the first two rows, where we compare the three methods, the simplex method is always outperformed by the improved primal-dual and the improved affine scaling methods. Regardless of whether we avoid sure loss or not, the improved primal dual method is faster than the improved affine scaling method except when we do not avoid sure loss and the number of desirable gambles is small.

In the last two rows, we compare our suggested improvements on the primal-dual method. When we avoid sure loss, using our feasible starting point shows a very slight improvement, although it is barely noticeable. The extra stopping criterion does not help at all in this case, quite logically so, because it will never be invoked in this case. What is important here is that it also does not hinder performance; the overhead of the extra check is thus negligible. When we do not avoid sure loss, both the extra stopping criterion and the feasible starting point considerably improve performance. Using both improvements gives the best results.

8. Conclusion

We studied and improved methods to solve linear programming problems efficiently for checking avoiding sure loss. By exploiting the structure of the linear programming problem, we first slightly reduced its dimension. Secondly, we proposed an extra stopping criterion based on its degenerate structure. We also showed that one can directly calculate feasible starting points in various cases, therefore reducing the effort required in the presolve phase of some of these algorithms.

We compared the impact of these improvements on linear programming methods (simplex, affine scaling, and primal-dual) on randomly generated sets of desirable gambles that either avoid

EFFICIENT ALGORITHMS FOR CHECKING AVOIDING SURE LOSS

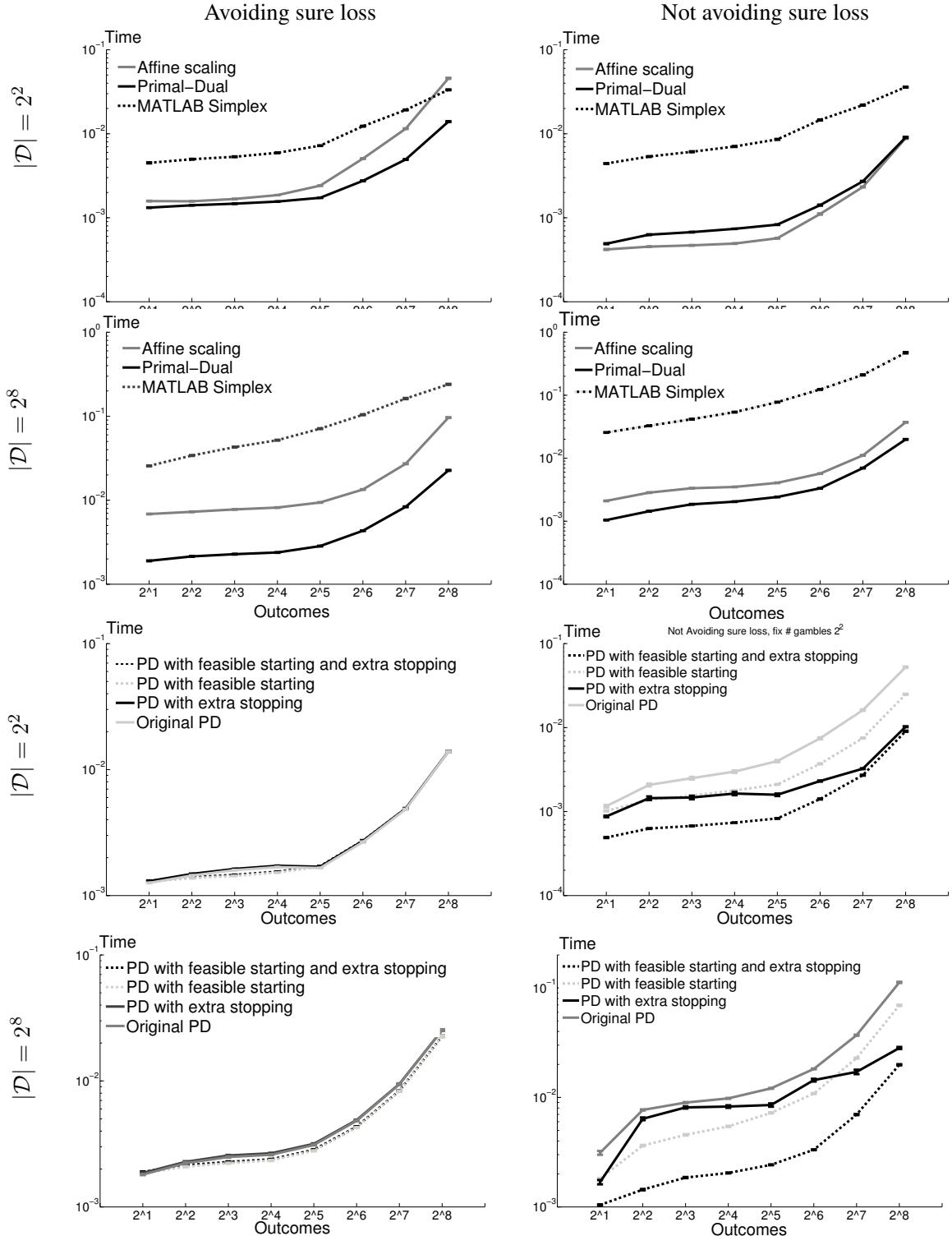


Figure 1: Comparison plots of the average computational time for three methods and for different improved primal-dual methods (PD).

or do not avoid sure loss. In our simulation, we found that the improved primal-dual and improved affine scaling methods outperform the simplex method. We found that both affine scaling and primal-dual methods benefit from the extra stopping criterion and feasible starting points. Overall, the improved primal-dual method is faster than the improved affine scaling method except when we do not avoid sure loss and the number of desirable gambles is small.

In future work, we will explore suitable starting points for the dual problems, algorithms for choosing ω^0 for large problems, and different structures for the credal set.

Acknowledgments

We would like to acknowledge support for this project from Development and Promotion of Science and Technology Talents Project (Royal Government of Thailand scholarship).

References

- S.-C. Fang and S. Puthenpura. *Linear Optimization and Extensions: Theory and Algorithms*. Springer Science+Business Media New York, 1993.
- C. Goh and X.Q.Yang. *Duality in optimization and variational inequalities*. Taylor and Francis, London, 2002.
- I. Griva, S. G. Nash, and A. Sofer. *Linear and Nonlinear Optimization Second edition*. SIAM, Philadelphia, 2009.
- A. J. Hoffman. Cycling in the simplex algorithm. *National Bureau of Standards, Washington, D.C*, 1953.
- N. Nakharutai. Computing with lower previsions using linear programming. Master's thesis, Durham University, 2015. Unpublished thesis.
- E. Quaeghebeur. *A Propositional CONEstrip Algorithm*, pages 466–475. Springer International Publishing, 2014. ISBN 978-3-319-08852-5. doi:[10.1007/978-3-319-08852-5_48](https://doi.org/10.1007/978-3-319-08852-5_48).
- R. Saigal. *Linear programming : a modern integrated analysis*. Springer Science+Business Media New York, 1995.
- R. J. Vanderbei. *Linear Programming: Foundations and Extensions, Second edition*. Springer, 2001.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- P. Walley, R. Pelessoni, and P. Vicig. Direct algorithms for checking consistency and making inferences from conditional probability assessments. *Journal of Statistical Planning and Inference*, 126:119–151, 2004.
- P. M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex, 1975.
- P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007. doi:[10.1016/j.ijar.2006.07.019](https://doi.org/10.1016/j.ijar.2006.07.019).

Towards a Cautious Modelling of Missing Data in Small Area Estimation

Julia Plass

Aziz Omar

Thomas Augustin

Department of Statistics, LMU Munich,

Germany (Plass, Omar, Augustin)

JULIA.PLASS@STAT.UNI-MUENCHEN.DE

AZIZ.OMAR@STAT.UNI-MUENCHEN.DE

AUGUSTIN@STAT.UNI-MUENCHEN.DE

Department of Mathematics, Insurance and Applied Statistics, Helwan University

Egypt (Omar)

Abstract

In official statistics, the problem of sampling error is rushed to extremes when not only results on sub-population level are required, which is the focus of Small Area Estimation (SAE), but also missing data arise. When the nonresponse is wrongly assumed to occur at random, the situation becomes even more dramatic, since this potentially leads to a substantial bias. Even though there are some treatments jointly considering both problems, they are all reliant upon the guarantee of strong assumptions on the missingness. For that reason, we aim at developing cautious versions of well known estimators from SAE by exploiting the results from a recently suggested likelihood approach, capable of including tenable partial knowledge about the nonresponse behaviour in an adequate way. We generalize the synthetic estimator and propose a cautious version of the so-called LREG-synthetic estimator in the context of design-based estimators. Then, we elaborate why the approach above does not directly extend to model-based estimators and proceed with some first studies investigating different missingness scenarios. All results are illustrated through the German General Social Survey 2014, also including area-specific auxiliary information from the German Federal Statistical Office's data report.

Keywords: small area estimation; LREG-synthetic estimator; missing data; partial identification; sensitivity analysis; likelihood; logistic regression; logistic mixed model; German General Social Survey.

1. Introduction

Survey methodology distinguishes between sampling and non-sampling error (cf., e.g., [Biemer, 2010](#)). Sampling error occurs when only a subset, but not the whole population can be included in a survey, yet the aim is to generalize the results beyond the units that have been sampled. Sampling error is especially severe if the population is composed of several sub-populations and the samples drawn from these sub-populations are not large enough to permit a satisfying precision on sub-population level. A set of methods has been introduced to tackle such situations and is referred to as *Small Area Estimation* (SAE). The main approach of SAE is to use additional data sources, such as administrative records and census data, as auxiliary data in an attempt to increase the effective sample size (cf., e.g., [Münich et al., 2013](#); [Rao and Molina, 2015](#)).

A common non-sampling error encountered in inference is item-nonresponse. Applying the EM-algorithm and Multiple Imputations are the recent practices (cf., e.g., [Little and Rubin, 2014](#)). Both techniques force point-identifiability, i.e. uniqueness of parameters, by requiring the assump-

tion that the missingness is occurring randomly (MAR), i.e. independently of the true underlying value of the variable of interest given covariates. Since the MAR assumption is generally not testable and wrongly imposing it may cause a substantial bias, results have to be treated with caution.

According to the methodology of partial identification in the spirit of [Manski \(2003\)](#), one does not have to insist on strong assumptions to obtain a result at all. Allowing for partially identified parameters enables to incorporate tenable knowledge only. In this way, one receives imprecise – but credible – results, which are refined if additional knowledge about the missingness is available. In this context, there are already several approaches refraining from strong assumptions on the missingness process (cf., e.g., [Couso and Dubois, 2014](#); [Denœux, 2014](#)). These cautious procedures also represent a popular field of research of the ISIPTA symposia (cf.,e.g., [Cattaneo and Wiencierz, 2012](#); [Schollmeyer and Augustin, 2015](#); [Utkin and Coolen, 2011](#)). Since we may not conjure information about the missingness process or make other strong modelling assumptions (cf., e.g., [Couso and Sánchez, 2016](#); [Hüllermeier, 2014](#)), uncertainty due to nonresponse has to be interpreted as lack of knowledge. Thus, approaches, explicitly communicating the associated uncertainty, are indispensable. In the context of official statistics this point was recently stressed by [Manski \(2015\)](#).

Since nonresponse may seriously reduce the already small sample size in SAE jointly considering both issues is especially challenging. As far as we know, already existing approaches dealing with nonresponse in SAE are based on strong assumptions on the missingness process, as MAR or the missing not at random (NMAR) assumption plus strict distributional assumptions. Thus, considering a cautious approach for dealing with nonresponse in SAE represents the core of this paper. To pursue this goal, in Section 2 we start by introducing the notation for the setting considered here followed by an introduction to our application using the German General Social Survey. Afterwards, we give a basic overview about prominent design-based estimators applicable in our situation in Section 3. Two design-based estimators, the classical synthetic estimator and the LREG-synthetic estimator, are generalized in Section 4. While cautious versions are given for the case of including no missingness assumptions at all, the case of including weak assumptions is considered for both estimators by relying on the cautious likelihood approach developed in [Plass et al. \(2015\)](#). In Section 5 the results are illustrated by means of the application example. In Section 6 we discuss why our approach cannot be directly extended to prominent model-based estimators and then perform a first sensitivity analysis. Section 7 concludes by summarizing the major points and giving some remarks on further research.

2. Setting

Technically, our setting is as follows: Let the population U under study have a total size of N units, and be divided into M non-overlapping domains (areas) U_i , each containing units j , $j = 1, \dots, N_i$ with N_i as the size of U_i , $i = 1, \dots, M$. Let Y be a binary variable of interest that is assumed to have a relation with a set of k precisely observed categorical covariates X_1, \dots, X_k through a certain model. Cross classifying the categorical covariates forms a k -dimension table with a total number of cells v , where the g -th cell – representing the g -th subgroup of the population – contains known joint absolute frequency $X_i^{[g]}$, $g = 1, \dots, v$, $i = 1, \dots, M$. To infer about π_i , the probability of a certain category of Y in area i , a sample s of size n is selected, such that a sample s_i of size n_i is selected from area i with $\sum_{i=1}^M n_i = n$. Within s_i , sample units j , $j = 1, \dots, n_i$ ($j \in s_i$) are selected with inclusion probability $1/w_{ij}$, where w_{ij} are the usual sample weights. Sample values of the covariates, denoted by x_{1ij}, \dots, x_{kij} , are assumed to be completely observed, while

of sample values of Y , denoted by y_{ij} , some are missing. Accordingly, s_i is partitioned into $s_{i,obs}$ and $s_{i,mis}$ that refer to sample units with observed and unobserved values of Y , respectively. If we additionally split by g , the samples are denoted by $s_i^{[g]}$, $s_{i,obs}^{[g]}$ and $s_{i,mis}^{[g]}$.

Application example: To illustrate the setting (and later on the results), we rely on the German General Social Survey (GGSS) ([GESIS Leibniz Institute for the Social Sciences, 2016](#)). We are interested in the area-specific ratio of people at risk of poverty, where German federal states are the areas completely partitioning the overall domain “Germany” (i.e. $M = 17$)¹. We construct a binary response variable with values “poor” and “rich” by comparing the collected equivalent income measured on the OECD modified scale with the poverty risk threshold given by 60% of the median net equivalent income, i.e. 986.65€ for year 2014 ([DESTATIS, Statistisches Bundesamt, 2016b](#)). The poverty variable shows 454 missing values. As covariates, we use the highest school leaving certificate, which – for ease of presentation – is dichotomized, distinguishing between categories “no Abitur”² and “Abitur” only, as well as sex.³ We base the analysis on the sample with $|s| = 3466$, $|s_{obs}| = 3012$, $|s_{mis}| = 454$. The German Federal Statistical Office’ data report ([DESTATIS, Statistisches Bundesamt, 2016a](#)) provides area-specific totals $X_i^{[g]}$, $i = 1, \dots, M$, $g = 1, \dots, v$, split by the values of the covariates, i.e. the absolute frequencies of the four subgroups “male-no Abitur”, “male-Abitur”, “female - no Abitur ” and “female - Abitur” in area i .

3. Theoretical Background of Design-Based Estimators

SAE techniques result in producing estimators $\hat{\pi}_i$ for area of interest i , $i = 1, \dots, M$, that are either design-based or model-based.⁴ In this paper, we mainly refer to design-based estimators, while we consider model-based ones in Section 6 only. Design-based estimators are either direct estimators that only use data from the targeted area, or indirect estimators that rely on data from other areas as well. This is justified under the assumption of similarity between the areas made to *borrow strength* from other areas.

The Horvitz-Thompson (HT) estimator ([Horvitz and Thompson, 1952](#)) $\hat{\pi}_{i,HT} = \frac{1}{N_i} \sum_{j=1}^{n_i} w_{ij} y_{ij}$ for an area i , well known in sampling theory, provides a method to estimate the mean of subpopulation (area) i , thereby accounting for the different sampling probabilities of respondents by sampling weights. The so-called *synthetic estimator* from SAE is a design-based indirect estimator, which is built upon the HT estimator, incorporating not only information from the area of interest, but averaging over all M areas. Thus, the area specific probability π_i is estimated as

$$\hat{\pi}_{i,SYN} \equiv \hat{\pi}_{SYN} = \frac{1}{N} \sum_{i=1}^M \sum_{j \in s_i} w_{ij} y_{ij} = \frac{1}{N} \sum_{i=1}^M N_i \cdot \hat{\pi}_{i,HT}, \quad \forall i = 1, \dots, M. \quad (1)$$

Since there is no distinction between areas and sample information is included about the response variable only, it merely serves as a basis for further estimators.

-
1. Although Germany is divided into 16 federal states, the GGSS differentiates between 17 ones, additionally distinguishing between “former East-Berlin” and “former West-Berlin”.
 2. The “Abitur” is the general qualification for university entrance in Germany.
 3. Since there should not be any regional differences with regard to covariate sex, the reason for the inclusion of this covariate rather lies in the interest of illustrating the subgroup specific analysis in a proper way than in an increase of explanatory power in the subject matter context.
 4. While properties of design-based estimators (e.g. bias and variance) are evaluated under sampling distribution over all samples with population parameters held fixed, model-based estimators usually condition on the selected sample, and inference regarding them is carried out with respect to the underlying model (cf., e.g., [Rao and Molina, 2015](#)).

An estimator that employs sample data as well as area specific auxiliary information on the joint totals X_{1i}, \dots, X_{ki} is the GREG-synthetic estimator (cf. Särdnal et al., 1992), where we here use its logistic version, the *LGreg-synthetic estimator* (cf. Lehtonen and Veijanen, 1998). Applying the LGreg-synthetic estimator is split into two steps:

First, the regression coefficients $\beta_0, \beta_1, \dots, \beta_k$ are estimated by means of a standard logistic regression model linking π_{ij} , i.e. the probability for individual $j, j = 1, \dots, n_i$ in $s_i, i = 1, \dots, M$, to have the value $y_{ij} = 1$, to the linear predictor containing the individual auxiliary information, here always assuming that all interactions are incorporated.⁵ Referring to the application example, we consider two covariates, hence the model includes $\beta_0, \beta_1, \beta_2$ and an interaction $\beta_{1:2}$, expressing the joint effect of both covariates. According to the aim of borrowing strength, one obtains global regression coefficients. From the estimated global regression coefficients, by applying the response function of a standard logistic regression model, we receive global predictions that only depend on the values of the covariate, but are independent of the area. To stress this, we write $\hat{\pi}^{[g]}, g = 1, \dots, v$, instead of $\hat{\pi}_{ij}$ in our case of categorical covariates. The calculation of these predictions becomes simpler here: Due to the strict monotonicity of the response function, the categorical nature of the covariates and the inclusion of all interactions, a unique relation between the regression coefficients and the predictions can be shown (as, e.g., addressed in Plass et al., 2017). Consequently, we can directly calculate the subgroup specific predictions by

$$\hat{\pi}^{[g]} = \sum_{i=1}^M \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}}, \quad (2)$$

with $n^{[g]}$ denoting the cell-count in subgroup $g, g = 1, \dots, v$.

Second, area-specific information is used: In our setting, the original LGreg-estimator (cf., e.g., Lehtonen and Veijanen, 1998, p.52) for a certain area of interest i can be expressed as

$$\hat{\pi}_{i,LGreg} = \sum_{g=1}^v \left(\underbrace{\sum_{j \in s_{i,g}} w_{ij} y_{ij}}_{\text{HT-part}} + \underbrace{\hat{\pi}^{[g]} (X_i^{[g]} - \sum_{j \in s_{i,g}} w_{ij})}_{\text{correction term}} \right) / N_i. \quad (3)$$

It can be understood as the HT estimator corrected by a term accounting for under- and overrepresentation of certain constellations of covariates in the sample, present in case of $X_i^{[g]} > \sum_{j \in s_{i,g}} w_{ij}$ and $X_i^{[g]} < \sum_{j \in s_{i,g}} w_{ij}$, respectively. The subgroup specific representation in (3) will turn out to be beneficial in context of developing a cautious version (cf. Section 4.2 and 4.3).

4. Cautious Versions of Design-based Estimators under Nonresponse

Since the already established ways of dealing with nonresponse in SAE require strong assumptions, we aim at improving the presented prominent estimators by striving for a proper reflection of the available information on the missingness process. For this purpose, we use the framework of the cautious approach developed for the more general case of coarse⁶ categorical data in Plass et al.

-
- 5. This is quite natural in this context, since only then the full information about the subgroup specific information, also provided by the auxiliary information in terms of totals, is used.
 - 6. The data problem only distinguishes between fully observed and completely unobserved values, while coarse data additionally include partial observations, e.g. in the sense of grouped data (cf. Heitjan and Rubin, 1991).

(2015) and further extended in Plass et al. (2017) to practically frame the inclusion of auxiliary information. We start by recalling the basic elements of this approach in the following section.

4.1 A Cautious Approach for Dealing with Nonresponse

An observation model \mathcal{Q} is used as a medium to frame the procedure of incorporating auxiliary information on the incompleteness. Restricting to the missing data problem and a binary response variable and considering the problem for subgroup g , $g = 1, \dots, v$, the model $\mathcal{Q}^{[g]}$ is determined by the set of missingness parameters $q_{na|y}^{[g]}$, i.e. the probability associated with refusing the answer (“na”), given a certain subgroup g and the true value $y \in \{0, 1\}$ of the response variable.⁷ In the spirit of partial identification, one can start by incorporating “no” assumptions⁸ on $q_{na|y}^{[g]}$, then restricting these missingness parameters successively by certain conceivable conditions. The cautious approach includes this observation model into a classical categorical likelihood problem. For this purpose, a connection between the parameters $\pi^{[g]}$ and $p_{\mathbf{y}}^{[g]}$ is established via the observation model, where $p_{\mathbf{y}}^{[g]}$ refers to the observed value $\mathbf{y} \in \{0, 1, na\}$, thus treating the missing values as a category of its own. The invariance of the likelihood allows to rewrite the log-likelihood in terms of $p_{\mathbf{y}}^{[g]}$, which can be uniquely maximized in terms of the parameters of interest by relying on the theorem of total probability, receiving

$$\ell(\pi^{[g]}, q_{na|0}^{[g]}, q_{na|1}^{[g]}) = n_1^{[g]} \left(\ln(\pi^{[g]}) + \ln(1 - q_{na|1}^{[g]}) \right) + n_0^{[g]} \left(\ln(1 - \pi^{[g]}) + \ln(1 - q_{na|0}^{[g]}) \right) + n_{na}^{[g]} \left(\ln(\pi^{[g]} q_{na|1}^{[g]} + (1 - \pi^{[g]}) q_{na|0}^{[g]}) \right), \quad (4)$$

where $n_1^{[g]}$, $n_0^{[g]}$ and $n_{na}^{[g]}$ refer to the respective observed cell counts within subgroup g , which later on have to be replaced by appropriate sample weights. By maximizing the log-likelihood in (4), we determine the generally set-valued⁹ estimators, whose one-dimensional projections can be represented by the lower and upper bounds of intervals, namely $\underline{\hat{\pi}}^{[g]}$, $\bar{\hat{\pi}}^{[g]}$, $\underline{\hat{q}}_{na|0}^{[g]}$, $\bar{\hat{q}}_{na|0}^{[g]}$, $\underline{\hat{q}}_{na|1}^{[g]}$ and $\bar{\hat{q}}_{na|1}^{[g]}$. Thereby, $\underline{\hat{\pi}}^{[g]}$ is attained under $\bar{\hat{q}}_{na|0}^{[g]}$ and $\underline{\hat{q}}_{na|1}^{[g]}$, while $\bar{\hat{\pi}}^{[g]}$ is associated with $\underline{\hat{q}}_{na|0}^{[g]}$ and $\bar{\hat{q}}_{na|1}^{[g]}$.

By considering $q_{na|1}^{[g]} = R \cdot q_{na|0}^{[g]}$, with missing ratio $R \in \mathcal{R} \subseteq \mathbb{R}_0^+$ (also cf. Nordheim (1984)),¹⁰ and \mathcal{R} as the set of missing ratios, assumptions about the missingness can be incorporated. Specific values of R are associated with a particular missingness scenario, thus point-identifying $\pi^{[g]}$. For instance, $R = 1$ represents the missingness scenario under gMAR¹¹, requiring $q_{na|1}^{[g]} = q_{na|0}^{[g]}$. Partial (weak) assumptions, like incorporating $R \in \mathcal{R}$ into (4), thus refine the result obtained from the log-likelihood optimization without the inclusion of any missingness assumptions. Since it can be shown that $\underline{\hat{\pi}}^{[g], \mathcal{R}}$, $\bar{\hat{\pi}}^{[g], \mathcal{R}}$ and $\underline{\hat{q}}_{na|1}^{[g], \mathcal{R}}$ as well as $\bar{\hat{\pi}}^{[g], \mathcal{R}}$, $\underline{\hat{q}}_{na|0}^{[g], \mathcal{R}}$ and $\bar{\hat{q}}_{na|1}^{[g], \mathcal{R}}$, i.e. the bounds under the partial assumptions expressed by $\mathcal{R} = [\underline{R}, \bar{R}]$, are achieved under missingness ratios \underline{R} and \bar{R} , respectively, one does not have to optimize the log-likelihood for all values in $[\underline{R}, \bar{R}]$, but optimizing under \underline{R} and \bar{R} is sufficient. While $\mathcal{R} = [0, 1]$ corresponds to $q_{na|1}^{[g]} \leq q_{na|0}^{[g]}$, a cautious version of

7. Referring to the framework of analyzing contingency tables, it is natural to drop the reference to individual j .

8. In fact, we confine ourselves to very general assumptions detailed in Plass et al. (2017).

9. The mapping relating $\hat{\pi}^{[g]}$ to $\hat{p}_{\mathbf{y}}^{[g]}$ is generally not injective.

10. Here we consider a different R than in Plass et al. (2015).

11. Conditioning on subgroup g generalizes the typical MAR assumption.

gMAR is given by $\mathcal{R} = [\max(0, 1 - \tau), 1 + \tau]$, $\tau \geq 0$, where the degree of cautiousness is given by the definition of the neighborhood τ (cf. Plass et al., 2017).

4.2 Cautious SAE: Including no Missingness Assumptions

In case of considering $\mathcal{R} = \mathbb{R}_0^+$, i.e. incorporating no assumption on the missingness, the result of the cautious likelihood approach (Plass et al., 2015, p. 251) can be shown to correspond to the one obtained from cautious data completion, plugging in all potential precise sample outcomes compatible with the observations (cf. Augustin et al., 2014, §7.8). Thus, here the lower and upper bound of the synthetic estimator in (1) can be calculated in this case by considering the extreme cases of regarding all missing values as $y_{ij} = 0$, $\forall j \in s_{i,mis}$, $i = 1, \dots, M$, or all as $y_{ij} = 1$, $\forall j \in s_{i,mis}$, $i = 1, \dots, M$:

$$\hat{\pi}_{i,SYN} = \frac{1}{N} \sum_{i=1}^M \sum_{j \in s_{i,obs}} w_{ij} y_{ij}, \quad \bar{\pi}_{i,SYN} = \frac{1}{N} \sum_{i=1}^M \left(\sum_{j \in s_{i,obs}} w_{ij} y_{ij} + \sum_{j \in s_{i,mis}} w_{ij} \right). \quad (5)$$

In order to study the bounds $\hat{\pi}_{i,LGREG}$ and $\bar{\pi}_{i,LGREG}$, it turns out to be beneficial to break the summation over all areas into a term for area i^* ¹² of interest and a summation over all other areas $i \neq i^*$. With the regularity condition that sampling weights within area i are equal such that $w_{ij} = w_i$, $\forall j = 1, \dots, n_i$, and defining $n^{[g]}$ and $n_i^{[g]}$ to be respectively the number of units in s and s_i existing in subgroup g , $g = 1, \dots, v$, $i = 1, \dots, M$, we can rewrite $\hat{\pi}_{i^*,LGREG}$ in (3) as

$$\sum_{g=1}^v \left(\left(\sum_{\substack{i=1 \\ i \neq i^*}}^M \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n_i^{[g]}} \right) \left(X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*} \right) + \sum_{j \in s_{i^*}^{[g]}} \frac{y_{i^*j}}{n^{[g]}} \left(X_{i^*}^{[g]} - w_{i^*} (n_{i^*}^{[g]} + n^{[g]}) \right) \right) / N_{i^*}, \quad (6)$$

$$\text{with } \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}} = \sum_{j \in s_{i,obs}^{[g]}} \frac{y_{ij}}{n^{[g]}} + \sum_{j \in s_{i,mis}^{[g]}} \frac{y_{ij}}{n^{[g]}} \quad \text{and} \quad \sum_{j \in s_{i^*}^{[g]}} \frac{y_{i^*j}}{n^{[g]}} = \sum_{j \in s_{i^*,obs}^{[g]}} \frac{y_{i^*j}}{n^{[g]}} + \sum_{j \in s_{i^*,mis}^{[g]}} \frac{y_{i^*j}}{n^{[g]}},$$

when missing data are included. The problem consists of finding the values of y_{ij} for the nonrespondents that minimize (maximize) Equation (6). Since Equation (6) is a sum of subgroup specific quantities, optimization for each subgroup g , $g = 1, \dots, v$, separately is sufficient. Provided that $X_{i^*}^{[g]} \geq n_{i^*}^{[g]} w_{i^*}$, we can directly infer that the term referring to the areas $i \neq i^*$ is minimized (maximized) if all the y_{ij} 's, $j \in s_{i,mis}$ are equal to zero (one). Otherwise, the other extreme allocation of zeros and ones is chosen to obtain the minimum (maximum). Analogous considerations can be accomplished in the term associated with area i^* , now based on the condition $X_{i^*}^{[g]} \geq w_{i^*} (n_{i^*}^{[g]} + n^{[g]})$.

4.3 Cautious SAE: First Attempts to Include (Partial) Missingness Assumptions

When partial assumptions in the sense of $R \in [\underline{R}, \bar{R}]$ are tenable, it is useful to express the cautious synthetic estimator and the LGREG-synthetic estimator in terms of $\hat{\pi}^{\mathcal{R}}$, $\hat{q}_{na|0}^{\mathcal{R}}$ and $\hat{q}_{na|1}^{\mathcal{R}}$ obtained by optimizing a log-likelihood as given in (4) under the constraints expressed by R . By again splitting

12. Whenever a differentiation between quantities summing up over all regions and quantities referring to a specific region is needed, we explicitly write i^* for the region under consideration.

$j \in s_i$ into $j \in s_{i,obs}$ and $j \in s_{i,mis}$, the lower bound for the synthetic estimator is received as¹³

$$\hat{\pi}_{SYN}^{\mathcal{R}} = \frac{1}{N} \sum_{i=1}^M \left(\sum_{j \in s_{i,obs}} w_{ij} y_{ij} + \hat{q}_{na|i1}^{\mathcal{R}} \cdot \hat{\pi}_i^{\mathcal{R}} \cdot \sum_{j \in s_i} w_{ij} \right), \quad (7)$$

where $\hat{q}_{na|i1}^{\mathcal{R}} \cdot \hat{\pi}_i^{\mathcal{R}} \cdot \sum_{j \in s_i} w_{ij}$ is the – here smallest – estimated weighted number of nonrespondents with $y_{ij} = 1$, $j \in s_{i,mis}$, under the missingness assumption in focus. Thereby, the included estimators are received by refraining from a subgroup specific consideration, thus regarding $\ell(\pi^{\mathcal{R}}, q_{na|0}^{\mathcal{R}}, q_{na|1}^{\mathcal{R}})$ instead of $\ell(\pi^{[g],\mathcal{R}}, q_{na|0}^{[g],\mathcal{R}}, q_{na|1}^{[g],\mathcal{R}})$ (cf. (4)). Analogously, $\hat{\pi}_{SYN}^{\mathcal{R}}$ is achieved by using $\hat{q}_{na|i1}^{\mathcal{R}}$ and $\hat{\pi}_i^{\mathcal{R}}$ within (7).

To derive the cautious LGREG-synthetic estimator described by $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$ and $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$, we base our presentation on the lower bound, while the upper bound is obtained analogously vice versa. Basically, there are two ways to generalize the LGREG-synthetic estimator to a cautious version: One could either consider one overall likelihood or make consistent use of the fact that the LGREG-synthetic estimator is a combination of two estimators, a global one motivated by the idea of “borrowing strength” and another one referring to area i^* . Here, we address the second possibility, while the first one should be studied in further research. For this purpose, we start by maximizing two log-likelihoods, namely $\ell(\pi^{[g],\mathcal{R}}, q_{na|0}^{[g],\mathcal{R}}, q_{na|1}^{[g],\mathcal{R}})$ and $\ell(\pi_{i^*}^{[g],\mathcal{R}}, q_{na|i^*0}^{[g],\mathcal{R}}, q_{na|i^*1}^{[g],\mathcal{R}})$, under \underline{R} and \overline{R} to derive the respective projections of the generally set-valued estimators. In a next step, we then approach the calculation of $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$ by including those estimators that minimize

$$\sum_{g=1}^v \left(\underbrace{\sum_{j \in s_{i^*,obs}^{[g]}} w_{i^*} y_{i^*j} + \hat{q}_{na|i^*1}^{[g],\mathcal{R}} \hat{\pi}_{i^*}^{[g],\mathcal{R}} \cdot \sum_{j \in s_{i^*}^{[g]}} w_{i^*j}}_{\text{HT-part}} + \underbrace{\hat{\pi}^{[g],\mathcal{R}} (X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*})}_{\text{correction term}} \right) / N_{i^*}, \quad (8)$$

which is a version of the classical LGREG-synthetic estimator in Equation (3), where the HT-part is represented in terms of $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ and $\hat{q}_{na|i^*1}^{[g],\mathcal{R}}$, guaranteeing for the partial assumptions under consideration. Due to the distinct estimation of $\pi^{[g]}$ and $\pi_{i^*}^{[g]}$, we now try to take the associated dependence into account: The interrelation between both estimators may be clearly inferred by considering the representations

$$\hat{\pi}_{i^*}^{[g]} = \left(\sum_{j \in s_{i^*}^{[g]}} y_{ij} \right) / n_{i^*}^{[g]} \quad \text{and} \quad \hat{\pi}^{[g]} = \left(\sum_{\substack{i=1 \\ i \neq i^*}}^M \sum_{j \in s_i^{[g]}} y_{ij} + \sum_{j \in s_{i^*}^{[g]}} y_{ij} \right) / n^{[g]} \quad (9)$$

(here for ease of representation given without splitting into $s_{i,obs}$ and $s_{i,mis}$), both including respondents from area i^* .¹⁴ Whenever $X_{i^*}^{[g]} > n_{i^*}^{[g]}$, we achieve $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$ if $\hat{\pi}_{i^*}^{[g],\mathcal{R}}, \hat{q}_{na|i^*1}^{[g],\mathcal{R}}, \hat{\pi}^{[g],\mathcal{R}}$ are taken in (8). This choice is possible in this case, since individuals $j \in s_{i^*}^{[g]}$ are assumed to have the

13. For more details see the preliminary version of a technical report available at <http://jpllass.userweb.mwn.de/forschung.html>.

14. While in (6) a splitting into terms for area i^* and areas $i \neq i^*$ was achieved, this cannot be accomplished here. Note that $\sum_{i=1}^M \sum_{j \in s_i^{[g]}} \frac{y_{ij}}{n^{[g]}}$ and $\sum_{j \in s_{i^*}^{[g]}} \frac{y_{i^*j}}{n^{[g]}}$, appearing in Equation (6), are different from (9) and cannot be regarded as estimated probabilities due to the different reference in numerator and denominator.

same values within both estimated probabilities in (9). Considering the situation of $X_{i^*}^{[g]} < n_{i^*}^{[g]}$, this is not the case. While $\hat{\pi}^{[g],\mathcal{R}}$ is supposed to be maximal, $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ and $\hat{q}_{na|i^{*1}}^{[g],\mathcal{R}}$ should be minimal to minimize (8). To proceed, we give a reasonable way out of this situation. Thereby, we distinguish between the case (i), where the correction term in (8) is of greater importance compared to the HT-part and case (ii), considering the opposite situation.

Case (i): The lower bound of the LGREG-synthetic estimator should be obtained by selecting $\bar{\hat{\pi}}^{[g],\mathcal{R}}$. In this way, for all individuals $j \in s_{i^*}^{[g]}$ the lowest possible scenario compatible with the partial knowledge is assumed, such that the inclusion of $\bar{\hat{\pi}}^{[g],\mathcal{R}}$ and $\bar{\hat{q}}_{na|i^{*1}}^{[g],\mathcal{R}}$ directly follows. This is supported by Equation (6), indicating that bounds of $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ are included instead of estimators referring to a scenario between.¹⁵

Case (ii): $\underline{\hat{\pi}}^{[g],\mathcal{R}}$, $\bar{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$ and $\hat{q}_{na|i^{*1}}^{[g],\mathcal{R}}$ are incorporated for $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$, while $\hat{\pi}^{[g],\mathcal{R}}$ is improvable by assuming the upper missingness scenario for individuals from $i \neq i^*$. A practical compromise is the inclusion of a pooled estimator

$$\hat{\pi}_{\text{pooled}}^{[g]} = \left(\bar{\hat{\pi}}_{i \neq i^*}^{[g]} \cdot n_{i \neq i^*}^{[g]} + \hat{\pi}_{i^*}^{[g]} \cdot n_{i^*}^{[g]} \right) / n^{[g]}, \quad (10)$$

to receive $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$, where $\hat{\pi}_{i \neq i^*}^{[g],\mathcal{R}}$ can also be obtained from the cautious log-likelihood calculated based on all data except from area i^* . Analogously, a pooled version can be determined for the calculation of $\bar{\hat{\pi}}_{i^*,LGREG}^{\mathcal{R}}$.

Because of the under-/overweighting of certain subgroups in the sample, automatically some $(X_{i^*}^{[g]} - n_{i^*}^{[g]} w_{i^*})$ will be positive and others negative, such that the distinction of different cases can not be avoided. The development of a criterion evaluating the “importance” of the HT-term and the correction term used in our argument should be part of further research. Thereby, also the results and conditions from Section 4.2 should be taken into consideration. Up to then, we choose the minimum of the results from case (i) and (ii) to obtain a suggestion for $\hat{\pi}_{i^*,LGREG}^{\mathcal{R}}$.

5. Results from the Application Example

The area-specific poverty rate is the focus of our illustration explained in Section 2. Yet, we explicitly avoid making conclusions on the poverty in a substance matter sense, considering this application as a first illustration of technical aspects of the elaborated cautious estimators only. Here, additionally to the case without assuming anything about the missingness process, we studied the weak assumption that rich respondents tend to refuse the income question more often compared to poor ones, i.e. $R \in [0, 1]$ (assum. 1), as well as a cautious version of MAR, here incorporating $R \in [0.3, 1.7]$ (assum. 2). Although subgroup specific assumptions were feasible in the context of the LGREG-synthetic estimator, we here impose the same missingness assumption on all subgroups.

By applying Equations (7) and (8) to the (weighted) marginal sample data,¹⁶ we can calculate the cautious synthetic estimator and the LGREG-synthetic estimator for the different situations of

15. From Equation (6) we could conclude that either all or no virtual values $y_{ij}, j \in s_{i^*,mis}$, have to be equal to 1 to obtain $\hat{\pi}_{i^*,LGREG}$ and $\bar{\hat{\pi}}_{i^*,LGREG}$ in the case of no assumptions. If partial assumptions are included, this applies in the sense that this does not have to be satisfied for all, but for the minimum/maximum number of virtual values that is consistent with the partial missing assumption ending up with $\hat{\pi}_{i^*}^{[g],\mathcal{R}}$ or $\bar{\hat{\pi}}_{i^*}^{[g],\mathcal{R}}$.

16. In the GGSS, respondents from East-Germany are oversampled, such that weights are required in the analysis (0.564 (East Germany), 1.205 (West Germany), cf. Koch et al. (1994)).

	no assum.	assum. 1	assum. 2
$[\hat{\pi}_{SYN}, \bar{\pi}_{SYN}]$	[0.167, 0.300]	[0.167, 0.193]	[0.175, 0.208]

Table 1: Bounds for the synthetic estimator under various missingness assumptions

Federal state	no assum.		assum. 1		assum. 2	
	$\hat{\pi}_{i,LGREG}$	$\bar{\pi}_{i,LGREG}$	$\hat{\pi}_{i,LGREG}$	$\bar{\pi}_{i,LGREG}$	$\hat{\pi}_{i,LGREG}$	$\bar{\pi}_{i,LGREG}$
BW	0.129	0.366	0.129	0.210	0.141	0.224
BY	0.088	0.233	0.088	0.133	0.091	0.141
HB	0.077	0.405	0.115	0.193	0.125	0.206
HH	0.009	0.196	0.014	0.075	0.019	0.083

Table 2: Bounds for the LREG-synthetic estimator under various missingness assumptions

partial knowledge (cf. Table 1 and Table 2, respectively). The practically weak assumptions already induce a remarkable refinement of the intervals obtained under no assumptions.¹⁷. Due to the separate likelihood optimization that in some cases led us to the pooled version, including different bounds for i^* and $i \neq i^*$, the lower bound from “no assum.” and “assum. 1” do not necessarily have to coincide here. This gives rise to an overall likelihood approach that admittedly refrains from “borrowing strength” within the missingness process, but implicitly accounts for interrelations.

6. First Studies Towards a Cautious Model-based Estimator under Nonresponse

Until now, we focused on models dealing with the small sample size by incorporating observations from other areas on the one hand and area-specific auxiliary information on the other hand. To account for between-area variation beyond that explained by auxiliary variables, model-based estimators relying on mixed models establish a basis. Model-based estimators incorporate data from different areas through a model that depends on the level of aggregation of the auxiliary variables. The well known Fay-Herriot (FH) area-level model, introduced by [Fay III and Herriot \(1979\)](#) for linear regression, has been further developed for categorical regression by [MacGibbon and Tomberlin \(1989\)](#). By relying on the logistic mixed model, they include area specific random effects $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$ into the linear predictor of a standard logistic regression model. Based on this model, we can make predictions contributing to the final model-based estimators.

Since we aim at applying the cautious likelihood approach, we consider the likelihood in the mixed model context first. Generally, the marginal likelihood of the i -th area is received by averaging over the probability distribution of the random effects u_i (cf., e.g., [Booth and Hobert, 1999](#)). Since thereby almost always analytically intractable integrals are involved, numerical methods are required for the maximization. Consequently, the cautious likelihood approach is stretched to the limits of its direct applicability if model-based estimators are of interest.

Nevertheless, we proceed with some studies to get a first impression about the predictions obtained from a mixed model if refrained from strong assumptions on the missingness process. Since

17. We use the official abbreviations of the federal states, here BW and BY for Baden-Wuerttemberg and Bavaria, and HB and HH for the federal city states (H) Bremen and Hamburg.

the random effects u_i and the regression coefficients are estimated simultaneously with the aid of approximation methods, we can no longer establish a direct connection between the subgroup specific probabilities and the regression coefficients, as we did in Section 4. Hence, we here start with a first sensitivity analysis, estimating β_0, \dots, β_k and u_i under different types of missingness mechanisms. Since for a part of our research question, i.e. getting a first impression about the bounds of the estimated random effects, an area-specific missingness behaviour is of high interest, we simplify the databases classifying the federal states into four regions (“northeast”, . . . , “southwest”), thus substantially reducing the scenarios that have to be considered within a corresponding missing type. Moreover restricting to the covariate “Abitur” (yes/no), we investigate the impact of two different missing types over a grid of values: The first missing type requires independence of the covariates, whereas the second type depends on the covariate and the area.

While the estimated random effects tend to show no systematic reaction to different missingness scenarios, the regression estimates¹⁸ attain the bounds in the extreme missingness situations. Consequently, by focusing on the scenarios that either regard all or no missing values as $y_{ij} = 1$, we apparently can at least give an estimator based on the best-worst-case estimation of the regression coefficients, here denoted by $\hat{\pi}^\beta \in [\hat{\pi}^\beta, \bar{\hat{\pi}}^\beta]$. For this purpose, we use $\hat{\beta}_0, \dots, \hat{\beta}_k, \hat{u}_i$ obtained for the extreme cases to determine the individual prediction bounds. Again, in our categorical case it turns out to be sufficient to calculate the bounds of $\hat{\pi}^{[g],\beta}$, now not only split by the values of the covariate, but also the region. Using $\hat{\pi}^{[g],\beta}$ and the area-specific totals $X_i^{[g]}$, the bounds of a model-based estimator, relying on the best-worst estimation of β , can be calculated.

7. Conclusion

By exploiting the cautious likelihood approach (cf. [Plass et al., 2015](#)), we considered an opportunity to adapt the LGREG-synthetic estimator for nonresponse, without the need of strict and often practically untenable assumptions about the missingness process. The included observation model is a powerful medium to make use of frequently available, partial assumptions about the missingness, where results from the application example corroborated that very weak assumptions may already suffice to substantially refine the results obtained without the inclusion of any missingness assumptions. Further research should be devoted to a more extensive consideration of the here proposed method characterized by separate likelihood optimizations. Although some first investigations of cautious model-based estimators were accomplished, due to the technically different situation, a more detailed study should be part of future research. In addition, comparing the magnitude of both principally differing sources of uncertainty induced by the problems in focus (i.e. sampling uncertainty as well as lack of knowledge associated to SAE and nonresponse, respectively) is notably worthwhile. For this purpose, uncertainty regions (cf. [Vansteelandt et al., 2006](#)), covering both types of uncertainties, should be investigated.

Acknowledgments

The first author thanks the LMUMentoring program, providing financial support for young, female researchers. The second author thanks the government of Egypt and the German Academic

¹⁸ cf. figure in the prelim. version of a technical report mentioned in footnote 9.

Exchange Service (DAAD) for their joint financial support. We are very grateful for the helpful remarks of all three anonymous reviewers and especially appreciate the constructive suggestions of one rather critical reviewer, improving the presentation.

References

- 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, Chichester, 2014.
- P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opin. Q.*, 74:817–848, 2010.
- J. Booth and J. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, 61:265–285, 1999.
- M. Cattaneo and A. Wiencierz. Likelihood-based imprecise regression. *Int. J. Approx. Reason.*, 53: 1137–1154, 2012. [based on an ISIPTA ’11 paper].
- I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reason.*, 55:1502–1518, 2014.
- I. Couso and L. Sánchez. Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach. *Inf. Sci. (Ny)*, 358:129–150, 2016.
- T. Denœux. Likelihood-based belief function: Justification and some extensions to low-quality data. *Int. J. Approx. Reason.*, 55:1535–1547, 2014.
- Destatis, Statistisches Bundesamt. Micro-census 2014 – Destatis: Results: Federal states, year, sex, general school education, 2016a. <https://www.genesis.destatis.de> [accessed: 04.02.2017].
- Destatis, Statistisches Bundesamt. EU-SILC 2014 – Destatis: Living conditions, risk of poverty, 2016b. <https://www.destatis.de> [accessed: 04.02.2017].
- R. Fay III and R. Herriot. Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.*, 74:269–277, 1979.
- GESIS Leibniz Institute for the Social Sciences. German General Social Survey – ALLBUS 2014. GESIS Data Archive, Cologne, 2016. ZA5242 Data file Version 1.0.0.
- D. Heitjan and D. Rubin. Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253, 1991.
- D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47(260):663–685, 1952.
- E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.*, 55:1519–1534, 2014.
- A. Koch, S. Gabler, and M. Braun. Konzeption und Durchführung der “Allgemeinen Bevölkerungs-umfrage der Sozialwissenschaften” (ALLBUS) 1994. *ZUMA-Arbeitsbericht*, 94, 1994.

- R. Lehtonen and A. Veijanen. Logistic generalized regression estimators. *Surv. Methodol.*, 24: 51–56, 1998.
- R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd edition, 2014.
- B. MacGibbon and T. Tomberlin. Small area estimation of proportions via empirical Bayes techniques. *Surv. Methodol.*, 15:237–252, 1989.
- C. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.
- C. Manski. Credible interval estimates for official statistics with survey nonresponse. *J. Econometrics*, 191:293–301, 2015.
- R. Münnich, J. Burgard, and M. Vogt. Small Area-Statistik: Methoden und Anwendungen. *ASzA Wirtschafts- und Sozialstatistisches Archiv*, 6:149–191, 2013.
- 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.
- J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *Proc ISIPTA ’15*, pages 247–256. SIPTA, 2015.
- J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, and C. Heumann. Reliable categorical regression analysis for non-randomly coarsened data. Preliminary version of a technical report available at <http://jplass.userweb.mwn.de/forschung.html>, 2017.
- J. Rao and I. Molina. *Small Area Estimation*. Wiley, 2nd edition, 2015.
- C. Särdnal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 1992.
- G. Schollmeyer and T. Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reason.*, 56:224–248, 2015. [based on an ISIPTA ’13 paper].
- L. Utkin and F. Coolen. Interval-valued regression and classification models in the framework of machine learning. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *Proc ISIPTA ’11*, pages 371–380. SIPTA, 2011.
- S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, 16:953–979, 2006.

Efficient Computation of Belief Theoretic Conditionals

Lalintha G. Polpitiya

Kamal Premaratne

Manohar N. Murthi

Dilip Sarkar

University of Miami

Coral Gables, Florida (USA)

LALINTHA@UMIAMI.EDU

KAMAL@MIAMI.EDU

MMURTHI@MIAMI.EDU

SARKAR@MIAMI.EDU

Abstract

Dempster-Shafer (DS) belief theory is a powerful general framework for dealing with a wider variety of uncertainties in data. As in Bayesian probability theory, the conditional operation plays a critical role in DS theoretic strategies for evidence updating and fusion. A major limitation associated with the application of DS theoretic techniques for reasoning under uncertainty is the absence of a feasible computational framework to overcome the prohibitive computational burden this conditional operation entails. This paper addresses this critical challenge via a novel generalized conditional computational model — *DS-Conditional-One* — which allows the conditional to be computed in significantly less computational and space complexity. This computational model also provides valuable insight into the DS theoretic conditional itself and can be utilized as a tool for visualizing the conditional computation. We provide a thorough analysis and experimental validation of the utility, efficiency, and implementation of the proposed data structures and algorithms for carrying out both the *Dempster's conditional* and *Fagin-Halpern conditional*, the two most widely utilized DS theoretic conditional strategies.

Keywords: Dempster-Shafer belief theory; Dempster's conditional; Fagin-Halpern conditional; data structures; algorithms; computational complexity.

1. Introduction

The Dempster-Shafer (DS) belief theory (Dempster, 1967, 1968; Shafer, 1976), also referred to as evidence theory, is a powerful and convenient framework that can handle a wide variety of data imperfections (Shafer, 1990; Smets, 1999). With the greater expressiveness and flexibility in evidential reasoning and decision-making that they offer, DS theoretic (DST) methods are finding increased utilization in numerous application scenarios and have generated an active research field (Yager and Liu, 2008; Denœux, 2016).

Motivation. As in the Bayesian methods, the conditional operation plays a pivotal role in DST strategies for evidence updating and fusion, and in general, for reasoning under uncertainty. Among these various notions that have been proposed over the years, perhaps the most widely used DST conditional notion is the *Dempster's conditional* (Shafer, 1976; Klawonn and Smets, 1992; Nguyen and Smets, 1993; Xu and Smets, 1996; Smets, 2002). On the other hand, the *Fagin-Halpern (FH) conditional* can be considered as the most natural generalization of the probabilistic conditional notion because of its close connection with the inner and outer conditional probability measures (Fagin and Halpern, 1990). The recent work on the DST conditional approach (Premaratne et al., 2009; Wickramarathne et al., 2011) is based on this FH conditional.

Challenges. In spite of the advantages they offer, DST implementations in current use are restricted to smaller frames of discernment (FoDs) because of the prohibitive computational burden

that larger FoDs impose on existing methods. While this difficulty has been addressed via several approximation methods (Yager and Liu, 2008; Denœux, 2016), such approaches usually require one to compromise the quality of the generated results for computational efficiency, and some approaches cannot be extended for computing the DST *conditionals*. Exact (or sufficiently precise) computation of conditionals is of paramount importance because the quality of results generated from DST strategies depend directly on the precision of the conditional. A review of current implementations (Yager and Liu, 2008; Augustin et al., 2014; Denœux, 2016; SIPTA, 2017) confirms that work is needed to overcome these computational limitations associated with the DST conditionals. A fast Möbius transform (FMT), which is analogous to the fast Fourier transform (FFT), has been developed and employed for efficient precise computation of DST notions (Thoma, 1989; Kennes, 1992). Polpitiya et al. (2016) proposes several data structures which enable highly efficient exact computation of the DST notions of belief and plausibility, but it does not address the computation of DST conditionals.

As for the Dempster’s conditional, perhaps the most thorough discussion for carrying out its precise computation appears in Klauw and Smets (1992) and Smets (2002). It provides a matrix calculus based algorithm to compute Dempster’s conditional masses. However, this approach is feasible only on smaller frames because of the matrix operations it requires. It is not applicable for FH conditional computation. As for the FH conditional, the work in Wickramarathne et al. (2013) provides a method to identify the propositions that retain non-zero support after FH conditioning, but it does not address conditional computation of these propositions.

Contributions. The main contribution of this paper is a completely new generalized model for computing DST *conditionals*. This conditional computational model — *DS-Conditional-One* — offers significantly greater flexibility and computational capability for implementation of DST conditional strategies. We provide the DS-Conditional-One computational model along with its complexity analysis, experimental validation of the utility, efficiency, implementation of the associated data structures and algorithms. This model can be employed to compute both the FH and Dempster’s conditional beliefs of an *arbitrary* proposition. This is exactly the challenge that Shafer refers to in Shafer (1990, p.348), viz., “*It remains to be seen how useful the fast Möbius transform will be in practice. It is clear, however, that it is not enough to make arbitrary belief function computations feasible.*”

By reducing the number of operations being executed, the proposed approach takes significantly less computational and space complexity when compared with other approaches for conditional computation. As an example, our experiment results demonstrate that the average computational time taken to compute the conditional belief of an arbitrary proposition by the proposed approach is less than 2 (μ s) for a FoD of size 10 and 0.7 (ms) for a FoD of size 20 (\sim 1 million focal elements). This new model can also be utilized as a visualization tool for conditional computations and in analyzing characteristics of conditioning and updating operations. All software routines are available at ProFuSELab (2017). We believe that this computational model and the associated data structures constitute a significant step toward filling the void between what the DST framework can offer for reasoning under uncertainty and the practical implementation of DST strategies.

This paper is organized as follows: Section 2 provides a review of essential DST notions and computational tools. Our DS-Conditional-One computational model and our algorithms for efficient computation of DST conditionals appear next in Sections 3 and 4, respectively. The experimental results are provided next in Section 5. Finally, Section 6 offers some concluding remarks.

2. Preliminaries: DS Belief Theory

2.1 DST Basic Notions

In DS theory, the *frame of discernment (FoD)* refers to the set of all possible mutually exclusive and exhaustive propositions (Shafer, 1976). We consider the case where the FoD is finite and we denote it as $\Theta = \{\theta_0, \theta_1, \dots, \theta_{n-1}\}$. Proposition $\{\theta_i\}$, which is referred to as a *singleton*, represents the lowest level of discernible information. The power set of Θ , denoted by 2^Θ , form all the propositions of interest in DS theory. A proposition that is not a singleton is referred to as a *composite*. The set $A \setminus B$ denotes all singletons in $A \subseteq \Theta$ that are not included in $B \subseteq \Theta$, i.e., $A \setminus B = \{\theta_i \in \Theta \mid \theta_i \in A, \theta_i \notin B\}$. We use \bar{A} to denote $\Theta \setminus A$ and $|A|$ to denote the cardinality of A . Note that $|\Theta| = n$.

In DS theory, the ‘support’ that is being strictly allocated to a proposition is captured via

Definition 1 (Basic Belief Assignment (BBA) or Masses) *The mapping $m : 2^\Theta \mapsto [0, 1]$ is said to be a basic belief assignment (BBA) or a mass assignment if*

$$m(\emptyset) = 0 \text{ and } \sum_{A \subseteq \Theta} m(A) = 1.$$

■

The mass of a composite proposition is free to move into its individual singletons, which allows one to model the notion of *ignorance*. Complete ignorance can be modeled via the *vacuous BBA*, viz., $m(\Theta) = 1$ and $m(A) = 0, \forall A \neq \Theta$. Propositions that possess nonzero mass are referred to as *focal elements*; the set of all focal elements in a FoD is referred to as its *core* \mathfrak{F} , i.e., $\mathfrak{F} = \{A \subseteq \Theta \mid m(A) > 0\}$. Note that $|\mathfrak{F}|$ is the number of focal elements. $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ is referred to as the *body of evidence (BoE)*.

Definition 2 (Belief) *Given a BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$, the belief and plausibility functions are the mappings $Bl : 2^\Theta \mapsto [0, 1]$ and $Pl : 2^\Theta \mapsto [0, 1]$, respectively, where*

$$Bl(A) = \sum_{B \subseteq A} m(B); \quad Pl(A) = \sum_{\substack{B \subseteq \Theta \\ B \cap A \neq \emptyset}} m(B).$$

■

The *belief* assigned to a proposition takes into account the support for all of its subsets. It is easy to see that, $Pl(A) = 1 - Bl(\bar{A}) \geq Bl(A), \forall A \subseteq \Theta$. So, the *plausibility* measures the extent to which a proposition is plausible, i.e., the amount of belief not strictly supporting the complement of the proposition. Propositions that possess nonzero belief are denoted by $\widehat{\mathfrak{F}}$, i.e., $\widehat{\mathfrak{F}} = \{A \subseteq \Theta \mid Bl(A) > 0\}$.

Given a valid belief function $Bl : 2^\Theta \mapsto [0, 1]$, one may generate the corresponding BBA $m : 2^\Theta \mapsto [0, 1]$ via the Möbius transform (Shafer, 1976)

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} Bl(B), \quad \forall A \subseteq \Theta. \quad (1)$$

The following notation will be useful for our work:

$$\mathcal{S}(A; B) = \sum_{\substack{0 \neq C \subseteq A; \\ 0 \neq D \subseteq B}} m(C \cup D). \quad (2)$$

So, $\mathcal{S}(A; B)$ denotes the sum of all masses of propositions that ‘straddle’ both $A \subseteq \Theta$ and $B \subseteq \Theta$.

The following result is of critical importance for our work.

Proposition 3 Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \subseteq \Theta$. For $B \subseteq \Theta$, consider the mappings $\Gamma_A : 2^\Theta \mapsto [0, 1]$ and $\Pi_A : 2^\Theta \mapsto [0, 1]$, where

$$\Gamma_A(B) = \sum_{\emptyset \neq X \subseteq \bar{A}} m((A \cap B) \cup X); \quad \Pi_A(B) = \sum_{Y \subseteq (A \cap B)} \Gamma_A(Y).$$

Then the following are true:

- (i) $\Gamma_A(A \cap B) = \Gamma_A(B)$ and $\Pi_A(A \cap B) = \Pi_A(B)$. So, w.l.o.g., we assume that $B \subseteq A$.
- (ii) $\Gamma_A(\emptyset) = Bl(\bar{A})$.

Proof These follow by direct substitution. ■

2.2 Fagin-Halpern (FH) Conditional

FH conditional can be considered the most natural generalization of the probabilistic conditional notion because of its close connection with the inner and outer conditional probability measures in probability theory (Fagin and Halpern, 1990).

Definition 4 (Fagin-Halpern (FH) Conditional) (Fagin and Halpern, 1990) Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \in \widehat{\mathfrak{F}}$. The conditional belief $Bl(B|A)$ of B given the conditioning event A is

$$Bl(B|A) = \frac{Bl(A \cap B)}{Bl(A \cap B) + Pl(A \cap \bar{B})}. \quad ■$$

The conditional plausibility $Pl(B|A)$ of B given A is computed as $Pl(B|A) = 1 - Bl(\bar{B}|A)$. Of course, once the conditional beliefs of all the propositions are computed, one may obtain the corresponding conditional BBA via a Möbius transform of the type in (1).

Suppose the BoE $\{\Theta, \mathfrak{F}, m(\cdot)\}$ is being conditioned w.r.t. the proposition $A \in \widehat{\mathfrak{F}}$. The propositions that retain a nonzero mass after conditioning are referred to as the *conditional focal elements*; the set of all such conditional focal elements is referred to as the *conditional core* \mathfrak{F}_A , i.e., $\mathfrak{F}_A = \{B \subseteq A \in \widehat{\mathfrak{F}} \mid m(B|A) > 0\}$.

In our work, we will exploit several previous results related to the conditional core (Kulasekere et al., 2004; Wickramarathne et al., 2013). Of particular importance is the following result:

Lemma 5 (Kulasekere et al., 2004) Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \in \widehat{\mathfrak{F}}$. Then,

- (i) $m(B|A) = 0$ whenever $\bar{A} \cap B \neq \emptyset$, and
- (ii) $Bl(B|A)$ can be expressed as

$$Bl(B|A) = \frac{Bl(A \cap B)}{Pl(A) - \mathcal{S}(\bar{A}; A \cap B)}, \quad B \subseteq A. \quad ■$$

Note that, (i) states that FH conditioning annuls those propositions that ‘straddle’ the conditioning proposition A and its complement \bar{A} . So, w.l.o.g., for FH conditioning, one may consider only those propositions $B \subseteq A$.

For our work, we will need the following alternate expression for the FH conditional:

Proposition 6 Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \in \widehat{\mathfrak{F}}$. Then, we may express $Bl(B|A)$ as

$$Bl(B|A) = \frac{Bl(A \cap B)}{1 - Bl(\bar{A}) - \mathcal{S}(\bar{A}; A \cap B)}, \quad B \subseteq \Theta. \quad \square$$

Proof This follows directly from Lemma 5(ii) by using the fact that $Bl(A) = 1 - Pl(\bar{A})$. ■

2.3 Dempster's Conditional

Dempster's conditional is perhaps the most widely employed DST conditional notion.

Definition 7 (Dempster's Conditional) (*Shafer, 1976*) Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \subseteq \Theta$ s.t. $Bl(\bar{A}) \neq 1$, or equivalently, $Pl(A) \neq 0$. The conditional belief $Bl(B|A)$ of B given the conditioning event A is

$$Bl(B|A) = \frac{Bl(\bar{A} \cup B) - Bl(\bar{A})}{1 - Bl(\bar{A})}. \quad \blacksquare$$

One may compute the corresponding conditional mass $m(B|A)$ and $Pl(B|A)$ from $Bl(B|A)$. Similarly to FH conditioning, Dempster's conditioning also annuls masses of all those propositions that 'straddle' the conditioning proposition A and its complement \bar{A} . So, w.l.o.g., for Dempster's conditioning, one may consider only those propositions $B \subseteq A$.

For our work, we will need the following alternate expression for the Dempster's conditional:

Proposition 8 Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m(\cdot)\}$ and $A \subseteq \Theta$ s.t. $Bl(\bar{A}) \neq 1$. Then, $Bl(B|A)$ can be expressed as

$$Bl(B|A) = \frac{Bl(A \cap B) + \mathcal{S}(\bar{A}; A \cap B)}{1 - Bl(\bar{A})}, \quad B \subseteq \Theta. \quad \square$$

Proof This follows directly from Definition 7 by using the fact that $Bl(\bar{A} \cup B) = Bl(\bar{A} \cup (A \cap B)) = Bl(\bar{A}) + Bl(A \cap B) + \mathcal{S}(\bar{A}; A \cap B)$. \blacksquare

Propositions 6 and 8 highlight an important fact: the three quantities $Bl(\bar{A})$, $Bl(A \cap B)$, and $\mathcal{S}(\bar{A}; A \cap B)$ fully determine both FH and Dempster's conditionals $Bl(B|A)$ and $Bl(B|A)$, respectively. It is this fact that we exploit for computing the conditionals of an arbitrary proposition.

2.4 The REGAP Property

The work in [Polpitiya et al. \(2016\)](#) proposes new data structures — *DS-Vector*, *DS-Matrix* and *DS-Tree* — and computationally efficient algorithms for computing the basic DST operations of belief and plausibility. For this purpose, the authors utilize what is referred to as the *REGAP* (*REcursive Generation of and Access to Propositions*) property.

To be more specific, consider the FoD $\Theta = \{\theta_0, \theta_1, \dots, \theta_{n-1}\}$. Suppose we desire to determine the belief potential $Bl(A)$ associated with $A = \{\theta_{k_0}, \theta_{k_1}, \dots, \theta_{k_{|A|-1}}\} \subseteq \Theta$. Then, *REGAP*(A) recursively generates all the $2^{|A|} - 1$ propositions whose masses are required to compute $Bl(A)$, viz., all subsets of A (including A itself). It is implemented in the following manner: Start with $\{\emptyset\}$. First insert the singleton $\{\theta_{k_0}\} \in A$. Only one proposition is associated with this singleton, viz., $\{\emptyset\} \cup \{\theta_{k_0}\} = \{\theta_{k_0}\}$ itself. Next insert another singleton $\{\theta_{k_1}\} \in A$. The new propositions that are associated with this singleton are $\{\emptyset\} \cup \{\theta_{k_1}\} = \{\theta_{k_1}\}$ and $\{\theta_{k_0}\} \cup \{\theta_{k_1}\} = \{\theta_{k_0}, \theta_{k_1}\}$. Inserting the next singleton $\{\theta_{k_2}\} \in A$ brings the new propositions $\{\emptyset\} \cup \{\theta_{k_2}\} = \{\theta_{k_2}\}$, $\{\theta_{k_0}\} \cup \{\theta_{k_2}\} = \{\theta_{k_0}, \theta_{k_2}\}$, $\{\theta_{k_1}\} \cup \{\theta_{k_2}\} = \{\theta_{k_1}, \theta_{k_2}\}$, and $\{\theta_{k_0}, \theta_{k_1}\} \cup \{\theta_{k_2}\} = \{\theta_{k_0}, \theta_{k_1}, \theta_{k_2}\}$. In essence, when a new singleton is added, new propositions associated with it can be recursively generated by adding the new singleton to each existing proposition. Of course, all propositions of interest within the FoD Θ can be generated by *REGAP*(Θ), i.e., when $A = \Theta$.

The propositions recursively generated via the REGAP property can be represented as a vector, DS-Vector, a matrix, DS-Matrix, or a tree, DS-Tree, and utilized to capture a BoE. We will utilize this REGAP property and the DS-Matrix structure in this work too.

3. DS-Conditional-One Computational Model

DS-Conditional-One is a computational model that enables one to compute the FH and Dempster's conditional beliefs of an *arbitrary* proposition. DS-Conditional-One model facilitates the representation, access, and efficient computation of the quantities that are needed to compute these conditionals (see Propositions 6 and 8).

Henceforth, we will denote the conditioning proposition A , its complement \bar{A} , and the conditioned proposition B as $\{a_0, a_1, \dots, a_{|A|-1}\}$, $\{\alpha_0, \alpha_1, \dots, \alpha_{|\bar{A}|-1}\}$, and $\{b_0, b_1, \dots, b_{|B|-1}\}$, respectively. Here, $\Theta = \{\theta_0, \theta_1, \dots, \theta_{n-1}\}$ denotes the FoD and $a_i, \alpha_j, b_k \in \Theta$. When dealing with FH and Dempster's conditioning, it is implicitly assumed that $A \in \mathfrak{F}$ and $Bl(\bar{A}) \neq 1$, respectively.

Furthermore, we will represent singletons of the conditioning event $A = \{a_0, a_1, \dots, a_{|A|-1}\}$ as *column singletons* and singletons of the complement of conditioning event $\bar{A} = \{\alpha_0, \alpha_1, \dots, \alpha_{|\bar{A}|-1}\}$ as *row singletons* in a DS-Matrix. See Fig. 1.

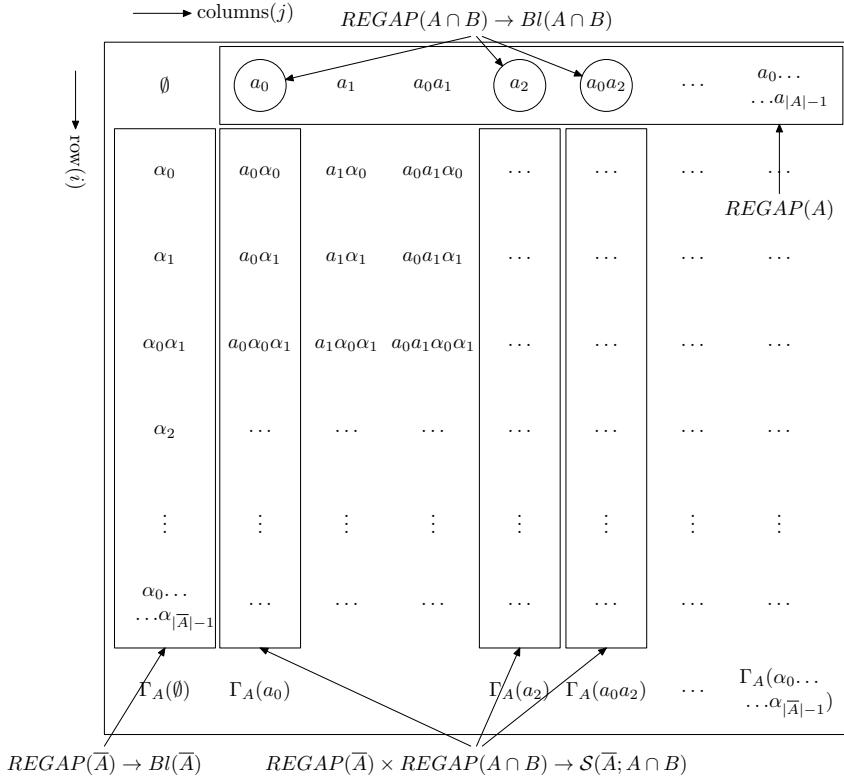


Figure 1: **DS-Conditional-One model.** Quantities related to $Bl(B|A)$ computation when $A = \{a_0, a_1, \dots, a_{|A|-1}\}$ and $\bar{A} = \{\alpha_0, \alpha_1, \dots, \alpha_{|\bar{A}|-1}\}$, and $B = \{a_0, a_2\} \subseteq A$.

The proposed DS-Conditional-One computational model allows direct identification of $REGA P(A)$, $REGAP(\bar{A})$, $REGAP(A \cap B)$, $(REGAP(\bar{A}) \times REGAP(A \cap B))$, $(REGAP(A) \times REGAP(\bar{A}))$, and $\Gamma_A(C)$, $\forall C \subseteq B$. Among these, the following three quantities are required to compute both FH and Dempster's conditional beliefs (see Propositions 6 and 8): **(a)** $REGAP(A \cap B)$: Use this to compute $Bl(A \cap B)$ (see Algorithm 1). **(b)** $REGAP(\bar{A})$: Use this to compute $Bl(\bar{A})$ (see Algorithm 2). **(c)** $(REGAP(\bar{A}) \times REGAP(A \cap B))$, the Cartesian product of $REGAP(\bar{A})$ and $REGAP(A \cap B)$: Use this to compute $\mathcal{S}(\bar{A}; A \cap B)$ (see Algorithm 3).

Fig. 1 depicts these quantities for $A = \{a_0, a_1, \dots, a_{|A|-1}\}$ and $B = \{a_0, a_2\} \subseteq A$

In the algorithms to follow, we use a lookup table named *power* to enhance the computational efficiency. It contains 2 to the power of singleton indexes in increasing order and it is implemented using a dynamic array that replaces run-time computation of 2 to the power values with a simpler array indexing operation. $power[i]$, the i -th entry of the *power* table, refers to 2^i . *index[]* is a dynamic array which keeps the indexes of subset propositions of $A \cap B$.

Algorithm 1 Compute $Bl(A \cap B)$ (with complexity $\mathcal{O}(2^{|A \cap B|})$)

```

1: procedure BLB(Singletons  $A$ , Singletons  $B$ , DS-Matrix  $BBA$ )
2:    $belief \leftarrow 0$ 
3:    $count \leftarrow 0$ 
4:   for each  $a_i$  in  $A \cap B$  do
5:      $index[count] \leftarrow power[i]$ 
6:      $temp \leftarrow count$ 
7:      $count \leftarrow count + 1$ 
8:     for  $j \leftarrow 0, temp - 1$  do
9:        $index[count] \leftarrow index[j] + power[i]$ 
10:       $count \leftarrow count + 1$ 
11:    end for
12:  end for
13:  for  $i \leftarrow 0, power[|A \cap B|] - 2$  do
14:     $belief \leftarrow belief + BBA[0][index[i]]$ 
15:  end for
16:  Return  $belief$ 
17: end procedure

```

Time Complexity of Algorithm 1. This computes $Bl(A \cap B)$ in $\mathcal{O}(2^{|A \cap B|})$ complexity. *Line #1:* The algorithm inputs are the conditioning event A , conditioned event B , and the DS-Matrix BBA . *Lines #4-12:* The outer loop is executed $|A \cap B|$ times. *Lines #8-11:* The inner loop is executed $temp - 1$ times. It can be shown that for $\ell = 0, 1, 2, \dots, |A \cap B| - 1$, $temp = (2^\ell - 1)$. *Lines #5 and #9* are constant time operations. Thus, the computational complexity of *lines #4-12* is given by

$$\sum_{\ell=0}^{|A \cap B| - 1} (1 + temp) = \sum_{\ell=0}^{|A \cap B| - 1} 2^\ell = 2^{|A \cap B|} - 1 = \mathcal{O}(2^{|A \cap B|}). \quad (3)$$

Lines #13-15: The required number of iterations is $2^{|A \cap B|} - 1$ and the complexity of this segment is $\mathcal{O}(2^{|A \cap B|})$. *Line #16:* The algorithm output is $Bl(A \cap B)$.

Algorithm 2 Compute $Bl(\bar{A})$ (with complexity $\mathcal{O}(2^{|\bar{A}|})$)

```

1: procedure BLCOMP(Singletons  $\bar{A}$ , DS-Matrix  $BBA$ )
2:    $belief \leftarrow 0$ 
3:   for  $i \leftarrow 1, power[|\bar{A}|] - 1$  do
4:      $belief \leftarrow belief + BBA[i][0]$ 
5:   end for
6:   Return  $belief$ 
7: end procedure

```

Time Complexity of Algorithm 2. This computes $Bl(\bar{A})$ in $\mathcal{O}(2^{|\bar{A}|})$ complexity. *Line #1:* The algorithm inputs are the complement of conditioning event \bar{A} and the DS-Matrix BBA . *Lines #3-5:* The required number of iterations is $2^{|\bar{A}|} - 1$ and the computational complexity of this segment is $\mathcal{O}(2^{|\bar{A}|})$. *Line #6:* The algorithm output is the belief potential $Bl(\bar{A})$.

Algorithm 3 Compute $S(\bar{A}; A \cap B)$ (with complexity $\mathcal{O}(2^{|\bar{A}|+|A \cap B|})$)

```

1: procedure STRAD(Singletons  $\bar{A}$ , Singletons  $A$ , Singletons  $B$ , DS-Matrix  $BBA$ )
2:    $belief \leftarrow 0$ 
3:    $count \leftarrow 0$ 
4:   for each  $a_i$  in  $A \cap B$  do
5:      $index[count] \leftarrow power[i]$ 
6:      $temp \leftarrow count$ 
7:      $count \leftarrow count + 1$ 
8:     for  $j \leftarrow 0, temp - 1$  do
9:        $index[count] \leftarrow index[j] + power[i]$ 
10:       $count \leftarrow count + 1$ 
11:    end for
12:   end for
13:   for  $i \leftarrow 1, power[|\bar{A}|] - 1$  do
14:     for  $j \leftarrow 0, power[|A \cap B|] - 2$  do
15:        $belief \leftarrow belief + BBA[i][index[j]]$ 
16:     end for
17:   end for
18:   Return  $belief$ 
19: end procedure

```

Time Complexity of Algorithm 3. This computes $S(\bar{A}; A \cap B)$ in $\mathcal{O}(2^{|\bar{A}|+|A \cap B|})$ complexity. *Line #1:* The algorithm inputs are the complement of conditioning event \bar{A} , the conditioning and conditioned propositions A and B , respectively, and the DS-Matrix BBA . *Lines #4-12:* Subset propositions of $A \cap B$ are generated via $REGAP(A \cap B)$. Computational complexity of this segment is $\mathcal{O}(2^{|A \cap B|})$, which can be obtained from equation 3. *Lines #13-17:* The outer loop is executed $(2^{|\bar{A}|} - 1)$ times. *Lines #14-16:* The inner loop is executed $(2^{|A \cap B|} - 1)$ times. Complexity of an access operation is $\mathcal{O}(1)$. Thus, the computational complexity of lines #13-17 is $(2^{|\bar{A}|} - 1)(2^{|A \cap B|} - 1) = \mathcal{O}(2^{|\bar{A}|+|A \cap B|})$. *Line #18:* The algorithm output is $S(\bar{A}; A \cap B)$.

Space Complexity of Algorithms 1, 2, and 3. The matrix in Fig. 1 is of size $2^{|A|} \times 2^{|\bar{A}|}$. Hence, the space complexity associated with each algorithm above is $\mathcal{O}(2^{|\Theta|})$.

Note that, in the DS-Conditional-One model, $REGAP(A)$ captures all propositions that may contribute to the conditional core \mathfrak{F}_A , and $REGAP(\bar{A})$ and $(REGAP(A) \times REGAP(\bar{A}))$, the Cartesian product of $REGAP(A)$ and $REGAP(\bar{A})$, capture all propositions whose masses are annulled (as identified by Lemma 5 (Kulasekere et al., 2004)). See Fig 1.

4. Efficient Computation of DST Conditionals

4.1 Computation of the FH Conditional Belief of an Arbitrary Proposition

To compute the FH conditional belief of an arbitrary proposition B , one can now use the expression in Proposition 6, where $Bl(A \cap B)$, $Bl(\bar{A})$ and $\mathcal{S}(\bar{A}; A \cap B)$ are obtained via Algorithms 1, 2, and 3, respectively. Thus the computational complexity of this computation remains as $\mathcal{O}(2^{|\bar{A}|+|A \cap B|})$.

As an example, to compute $Bl(B|A)$, where $B = \{a_0, a_2\}$, we may proceed as follows: **(a)** $REGAP(A \cap B)$ captures the propositions that contribute to $Bl(A \cap B)$. Use Algorithm 1 to compute this. **(b)** $REGAP(\bar{A})$ captures the propositions that contribute to $Bl(\bar{A})$. Use Algorithm 2 to compute this. Note that $Bl(\bar{A})$ is represented by $\Gamma_A(\emptyset)$ in Fig. 1. **(c)** The Cartesian product $(REGAP(\bar{A}) \times REGAP(A \cap B))$ captures the propositions that contribute to $\mathcal{S}(\bar{A}; A \cap B)$. Use Algorithm 3 to compute this. $\mathcal{S}(\bar{A}; A \cap B) = \Gamma_A(\{a_0\}) + \Gamma_A(\{a_2\}) + \Gamma_A(\{a_0, a_2\})$.

Then, $Bl(B|A)$ for $B = \{a_0, a_2\}$ is computed as

$$Bl(B|A) = \frac{Bl(A \cap B)}{1 - \Gamma_A(\{\emptyset\}) - \Gamma_A(\{a_0\}) - \Gamma_A(\{a_2\}) - \Gamma_A(\{a_0, a_2\})}. \quad (4)$$

4.2 Computation of the Dempster's Conditional Belief of an Arbitrary Proposition

To compute the Dempster's conditional belief of an arbitrary proposition B , one can use the expression in Proposition 8, where $Bl(A \cap B)$, $Bl(\bar{A})$ and $\mathcal{S}(\bar{A}; A \cap B)$ are obtained via Algorithms 1, 2, and 3, respectively. Thus the computational complexity is $\mathcal{O}(2^{|\bar{A}|+|A \cap B|})$.

Consider the same example as before, viz., $B = \{a_0, a_2\}$. Then, we may compute $Bl(B\|A)$ as

$$Bl(B\|A) = \frac{Bl(A \cap B) + \Gamma_A(\{a_0\}) + \Gamma_A(\{a_2\}) + \Gamma_A(\{a_0, a_2\})}{1 - \Gamma_A(\{\emptyset\})}. \quad (5)$$

Computation of the Dempster's Conditional Mass Using Specialization Matrix. It is noteworthy that Klawonn and Smets (1992) and Smets (2002) have proposed a matrix calculus based algorithm for *direct* computation of Dempster's conditional *masses*. It employs a $2^{|\Theta|} \times 2^{|\Theta|}$ -sized stochastic matrix \mathfrak{S}_A (with each entry '0' or '1') referred to as the conditioning specialization matrix and a $2^{|\Theta|} \times 1$ -sized vector $m(\cdot)$ containing the BoE's focal elements. Then $m(\cdot\|A) = \mathfrak{S}_A \cdot m(\cdot)$ yields Dempster's conditioning masses *without normalization*. The computational and space complexity of the specialization matrix multiplication is $\mathcal{O}(2^{|\Theta|} \times 2^{|\Theta|})$, a prohibitive burden even for modest FoD sizes.

5. Experiments

Recall that Algorithms 1, 2, and 3 yield all the parameters (viz., $Bl(A \cap B)$, $Bl(\bar{A})$, and $\mathcal{S}(\bar{A}; A \cap B)$) required for both FH and Dempster's conditional belief computations. Once these quantities are

computed, computational times for both conditional belief computations are similar because they require constant time (see Propositions 6 and 8).

For a given FoD size, we selected a random set of focal elements, with randomly selected mass values, and conducted 10,000 conditional computations for randomly chosen propositions A and $B \subseteq A$. Table 1 lists the average computational times taken by the DS-Conditional-One model and the specialization matrix based method in Klawonn and Smets (1992) and Smets (2002).

With the DS-Conditional-One model (which applies to *both* FH and Dempster's conditionals), we use a 'brute force' approach to compute all the conditional beliefs (i.e., compute the conditional belief of every proposition); we then use the FMT to get the conditional masses for all the propositions (Shafer, 1976; Fagin and Halpern, 1990). The specialization matrix based method (which applies to the Dempster's conditional *only*) yields the conditional masses of *all* propositions, but the time taken already far exceeds what the DS-Conditional-One model takes (even including the FMT). So we did not compute the conditional beliefs with the specialization matrix based method (which would have required the FMT).

All conditional computations for an *arbitrary* proposition were done on an iMac running Mac OS X 10.12.3 (with 2.9GHz Intel Core i5 processor and 8GB of 1600MHz DDR3 RAM). Conditional computations for *all* propositions were done on the same iMac for smaller FoDs and on a supercomputer (<http://ccs.miami.edu/pegasus>) for larger FoDs (underlined in Table 1). The complete C++ library is available at ProFuSELab (2017).

Method →		DS-Conditional-One Model			Specialization Matrix
Conditional →		FH or Dempster's			Dempster's
FoD		$Bl(B A)$ or $Bl(B\ A)$	$Bl(B A)$ or $Bl(B\ A)$	$m(B A)$ or $m(B\ A)$	$m(B\ A)$
$ \Theta $	Max. $ \mathfrak{F} $	(Arbitrary)	(All)	(All)	(All)
2	3	0.0005	0.0011	0.0016	0.0011
4	15	0.0005	0.0038	0.0050	0.0063
6	63	0.0006	0.0128	0.0170	0.0696
8	255	0.0009	0.0517	0.0679	1.0154
10	1,023	0.0017	0.2428	0.3090	<u>93.1590</u>
12	4,095	0.0040	1.3528	1.6186	<u>1485.6300</u>
14	16,383	0.0120	<u>18.4885</u>	<u>22.4995</u>	<u>25051.8200</u>
16	65,535	0.0405	<u>146.1480</u>	<u>151.9600</u>	***
18	262,143	0.1516	<u>1,087.2800</u>	<u>1,113.5300</u>	***
20	1,048,575	0.6011	<u>8,485.4500</u>	<u>8,862.9800</u>	***

Table 1: DS-Conditional-One model versus specialization matrix based method. Average computational times (ms). (*** denotes computations not completed within a feasible time).

The significant speed advantage offered by the proposed computational model over the specialization matrix based approach is evident from Table 1. For larger FoDs, the computational burden associated with the specialization matrix based approach becomes prohibitive because of its space complexity of $\mathcal{O}(2^{|\Theta|} \times 2^{|\Theta|})$. For example, an FoD of size 20 would need 128 ($= 2^{20} \times 2^{20}/8$) GB of memory to represent the specialization matrix, if each matrix entry occupies only 1 bit.

With increasing FoD size, the computational time requirement of the DS-Conditional-One model is significantly less compared to what the specialization matrix based approach requires.

6. Concluding Remarks

This paper provides a general framework for computation of DST conditionals. The DS-Conditional-One model that we propose can also serve as a tool for visualization and further analysis of the conditional computation process. We believe that the algorithms we have developed constitute a significant step forward in harnessing the strengths of DST methods in practical applications.

The efficiency of these algorithms is mainly because of the significantly reduced number of operations that are executed. Computational complexity associated with conditional belief computation of an arbitrary proposition is $\mathcal{O}(2^{|A|+|A \cap B|})$. This is a significant improvement over the $\mathcal{O}(2^{|\Theta|} \times 2^{|\Theta|})$ complexity associated with the specialization matrix based approach. The DS-Conditional-One model also provides a significant advantage in terms of memory usage: it requires a $\mathcal{O}(2^{|\Theta|})$ space complexity versus $\mathcal{O}(2^{|\Theta|} \times 2^{|\Theta|})$ for the specialization matrix based approach.

Another advantage of the proposed approach is that it can be utilized for either the FH conditional or Dempster's conditional belief computations. An outcome of this research is a conditional computation library (in C++) which is available at [ProFuSELab \(2017\)](#). We expect that this library will be useful for practical application of DST methods.

Our current research work is focused on conditional computations on potentially dynamic FoDs (where the singletons may have to be removed or new singletons may have to be appended as operations are carried out). This would be of immense value for enhanced resource utilization. It also appears possible to further enhance the algorithms that we have developed via parallel computing optimizations because of the underlying matrix structure.

Acknowledgments

This work is based on research supported by the U.S. Office of Naval Research (ONR) via grant #N00014-10-1-0140 and the U.S. National Science Foundation (NSF) via grant #1343430.

References

- T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, Chichester, UK, May 2014.
- A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.*, 38(2):325–339, Apr 1967.
- A. P. Dempster. A generalization of Bayesian inference. *J. R. Stat. Soc. Ser. B*, 30(2):205–247, 1968.
- T. Denœux. 40 years of Dempster–Shafer theory. *Int. J. Approx. Reason.*, 79(C):1–6, Dec 2016.
- R. Fagin and J. Y. Halpern. A new approach to updating beliefs. In *Proc. 6th Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 347–374, Cambridge, MA, Jul 1990.
- R. Kennes. Computational aspects of the Möbius transformation of graphs. *IEEE Trans. Syst. Man. Cybern.*, 22(2):201–223, Mar./Apr. 1992.

- F. Klawonn and P. Smets. The dynamic of belief in the transferable belief model and specialization-generalization matrices. In *Proc. 8th Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 130–137, Stanford, CA, Jul 1992.
- E. C. Kulasekere, K. Premaratne, D. A. Dewasurendra, M.-L. Shyu, and P. H. Bauer. Conditioning and updating evidence. *Int. J. Approx. Reason.*, 36(1):75–108, 2004.
- H. T. Nguyen and P. Smets. On dynamics of cautious belief and conditional objects. *Int. J. Approx. Reason.*, 8(2):89–104, Feb 1993.
- L. G. Polpitiya, K. Premaratne, M. N. Murthi, and D. Sarkar. A framework for efficient computation of belief theoretic operations. In *Proc. 19th Int. Conf. Information Fusion (FUSION)*, pages 1570–1577, Heidelberg, Germany, Jul 2016.
- K. Premaratne, M. N. Murthi, J. Zhang, M. Scheutz, and P. H. Bauer. A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. In *Proc. 12th Int. Conf. Information Fusion (FUSION)*, pages 2122–2129, Seattle, WA, Jul 2009.
- ProFuSELab. Conditional Computation Library, 2017. URL <http://profuselab.github.io/Conditional-Computation-Library>.
- G. Shafer. *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ, 1976.
- G. Shafer. Perspectives on the theory and practice of belief functions. *Int. J. Approx. Reason.*, 4 (5-6):323–362, Oct 1990.
- SIPTA. Software tools for imprecise probabilities, 2017. URL <http://www.sipta.org/index.php?id=sfw>.
- P. Smets. Practical uses of belief functions. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 612–621, Stockholm, Sweden, Jul 1999.
- P. Smets. The application of the matrix calculus to belief functions. *Int. J. Approx. Reason.*, 31(1):1–30, 2002.
- H. M. Thoma. *Factorization of belief functions*. PhD thesis, Dept. Stat., Harvard Univ., Cambridge, MA, 1989.
- T. L. Wickramarathne, K. Premaratne, M. N. Murthi, M. Scheutz, S. Kübler, and M. Pravia. Belief theoretic methods for soft and hard data fusion. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2388–2391, Prague, Czech Republic, May 2011.
- T. L. Wickramarathne, K. Premaratne, and M. N. Murthi. Toward efficient computation of the Dempster–Shafer belief theoretic conditionals. *IEEE Trans. Cybern.*, 43(2):712–724, Apr 2013.
- H. Xu and P. Smets. Reasoning in evidential networks with conditional belief functions. *Int. J. Approx. Reason.*, 14(2):155–185, 1996.
- R. R. Yager and L. Liu, editors. *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer-Verlag, Berlin Heidelberg, 2008.

The CWI World Cup Competition: Eliciting Sets of Acceptable Gambles

Erik Quaeghebeur

*Delft University of Technology (TU Delft)
Delft (The Netherlands)*

E.R.G.QUAEGHEBEUR@TUDELFT.NL

Chris Wesseling

Emma Beauxis-Aussalet

Teresa Piovesan

Tom Sterkenburg

*Centrum Wiskunde & Informatica (CWI)
Amsterdam (The Netherlands)*

CHRIS.WESSELING@CWI.NL

EMMANUELLE.BEAUXIS-AUSSALET@CWI.NL

TPIOVESAN@CWI.NL

TOM@CWI.NL

Abstract

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. 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. The experiment consisted of a betting competition for the 2014 FIFA World Cup: For each match bets were assigned based on the sets of acceptable gambles elicited from the participants. A new algorithm was designed for generating fair bets for assignment. Participant feedback indicated that improving the usability and transparency of the interface would ease the elicitation procedure. The experiment's results underlined that imprecision is an essential aspect of real-life uncertainty modeling.

Keywords: elicitation; gamble; acceptability; desirability; user interface; experiment; fair bet.

1. Introduction

In practical applications of uncertainty models, e.g., in expert systems, we need concrete values for their parameters. For example, the conditional probability tables of a Bayesian network need to be filled in. Such values can be obtained by learning them from data or by eliciting them from domain experts, or a combination thereof ([Druzdzel and van der Gaag, 2000](#)). This paper introduces a procedure for eliciting quantities describing the uncertainty about some phenomenon or experiment.

Typically, the uncertainty is modeled in terms of probabilities, such as in a Bayesian network. Elicitation of probabilities is commonplace and well-studied ([Spetzler and Staël von Holstein, 1975](#); [Cooke, 1991](#); [Renooij, 2001](#); [O'Hagan et al., 2006](#)). Apart from other issues such as various biases, there is agreement that eliciting probabilities directly and as precise numbers is often problematic due to a lack of familiarity with probability theory and the absence of a concrete context. Therefore, (i) targeted graphical interfaces such as scales and lottery wheels are designed, (ii) verbal descriptions of probability values are used, or (iii) the elicitation problem is reformulated as a betting problem. Another recurring idea is to use qualitative information such as comparative probability (see, e.g., [Druzdzel and van der Gaag, 1995](#)).

Uncertainty can also be modeled in alternative ways. One approach is to use generalizations of probabilities, such as imprecise-probabilistic models ([Walley, 1991](#); [Augustin et al., 2014](#)). Given that

the theory of imprecise probabilities encompasses probability intervals and comparative probability, imprecise-probabilistic techniques are better suited to deal with the results of an elicitation procedure, as even some of the most ardent ‘precise’ probabilists admit ([O’Hagan and Oakley, 2004](#), Section 3.3).

In this paper, assuming the elicitation problem can be formulated in betting terms, we discuss an interface to elicit coherent sets of acceptable gambles, also called desirable gambles ([Walley 1991](#), Appendix F; [Quaeghebeur 2014](#)). Roughly speaking, the gambles (random variables) in such a set are those for which the elicitee’s expectation is at least zero. Our interface can then, e.g., be used for eliciting the parameters of a credal network that is defined in terms of sets of acceptable gambles ([De Bock and de Cooman, 2015](#)). Moreover, sets of acceptable gambles can equivalently be transformed into the more classical imprecise-probabilistic models, credal sets (convex sets of probabilities) and lower expectations (previsions); so the procedure can also be used for eliciting, e.g., the parameters of classical—credal set based—credal networks. One can always obtain a single probability measure by selecting it from an elicited credal set in a principled way (see, e.g., [Druzdzel and van der Gaag, 1995](#)), opening up the option for also eliciting, e.g., Bayesian networks.

The interface for eliciting sets of acceptable gambles we present is designed for three-outcome possibility spaces, i.e., involving three mutually exclusive and exhaustive events. The design ideas could be adapted for the much simpler case of a two-element possibility space. For larger possibility spaces, the interface can be used in combination with an appropriate decomposition thereof. For example, a marginal extension theorem ([Quaeghebeur, 2014](#), Theorem 1.2) guarantees that we can coherently combine a (marginal) coherent set of desirable gambles on some partition and (conditional) coherent sets of desirable gambles on the partition elements. So working with a hierarchical partitioning of the possibility space with partition elements of cardinality three or less is an option; using two-outcome spaces only would result in reduced expressiveness.

Next to the interface itself, we discuss a real-life experiment conducted as an exploratory test of our elicitation interface. It was organized around the 2014 FIFA World Cup. But first, we start by giving a brief primer on the theory of coherent sets of acceptable gambles.

2. Sets of Acceptable Gambles

Coherent sets of acceptable gambles are an imprecise-probabilistic model originally introduced by [Williams \(1976](#), Section IV). The idea essentially lay dormant until this model was advocated by [Walley \(1991](#), Appendix F; [2000](#), Section 6) using the term ‘desirable gambles’.

2.1 Essential Concepts

The *possibility space* Ω describes the events about which there is uncertainty. For this paper we may assume it is a finite set. Formally, a *gamble* is a real-valued function on the possibility space. It represents a positive or negative payoff that depends on the unknown actual realization $\omega \in \Omega$.

An elicitee finds a gamble g on Ω *acceptable* if she is in some sense committed to the following transaction: Once the realization $\omega \in \Omega$ is determined, she gets the payoff $g(\omega)$. The set of gambles the elicitee assesses to be acceptable is denoted by \mathcal{A} . We assume it to be finite, which is reasonable in an elicitation context. A consequence is that all nontrivial checks and computations can be done using linear programming ([Quaeghebeur, 2013](#)).

We consider accepting a gamble that is everywhere negative to be irrational. Based on the assumption that the gamble payoffs are expressed in a linear utility—e.g., small amounts of money—we also consider positive linear combinations of acceptable gambles to be acceptable. Consequently,

an assessment \mathcal{A} , even if it does not contain negative gambles, may nevertheless be irrational; to wit, the elicitree can be forced to *incur a sure loss* by combining some of the gambles she accepts. Formally, this happens if there are real coefficients $\lambda_g \geq 0$ such that $\sum_{g \in \mathcal{A}} \lambda_g g < 0$, where the sum and inequality are taken pointwise, i.e., hold for all ω in Ω .

To the above assumptions, we add that it is irrational to not accept nonnegative gambles. So we arrive at the following set of *coherence* axioms, which describes the essential properties a *deductively closed* set of acceptable gambles \mathcal{D} should satisfy:

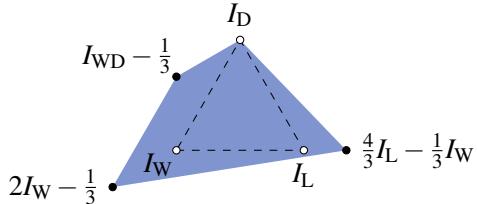
$$\begin{array}{ll} \text{Avoiding Sure Loss: } g < 0 \Rightarrow g \notin \mathcal{D}, & \text{Addition: } g, h \in \mathcal{D} \Rightarrow g + h \in \mathcal{D}, \\ \text{Accepting Partial Gains: } g \geq 0 \Rightarrow g \in \mathcal{D}, & \text{Positive Homogeneity: } g \in \mathcal{D}, \lambda_g > 0 \Rightarrow \lambda_g g \in \mathcal{D}. \end{array}$$

This set of axioms forces \mathcal{D} to be a convex cone in the linear space of gambles on Ω that includes the positive orthant and does not intersect the negative orthant. Note that this set of axioms allows gambles that are strictly negative on some nontrivial event $B \subset \Omega$ and zero on its complement to be acceptable: this is interpreted as the elicitree considering the event B to be (practically) impossible.

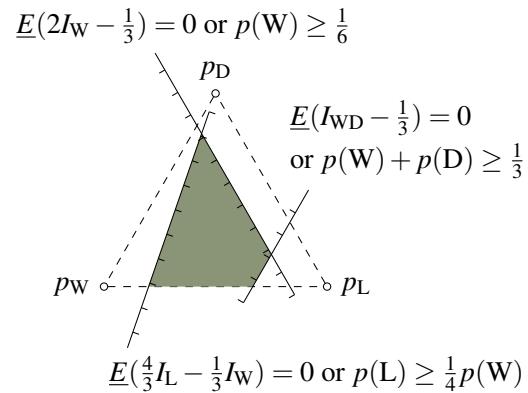
Given an elicited assessment \mathcal{A} that avoids sure loss, we can use the last three axioms in a generative way, to extend the assessment to a coherent set of acceptable gambles

$$\mathcal{D} := \{f + \sum_{g \in \mathcal{A}} \lambda_g g : \text{gamble } f \geq 0, \text{ coefficients } \lambda_g \geq 0\}.$$

This deductive closure is called the *natural extension* of \mathcal{A} . An illustration of an assessment that avoids sure loss and its natural extension is given in Figure 1a.



- (a) The dashed triangle delimits the positive octant. It is spanned by the space's unit vectors I_ω —white dots—which we will often look at as singleton indicator functions: $I_{\omega'}(\omega)$ with $\omega, \omega' \in \Omega$ is equal to 1 if $\omega = \omega'$ and 0 otherwise. Also, $I_{\omega\omega'} := I_\omega + I_{\omega'}$. An interpretation example: accepting the gamble $2I_W - \frac{1}{3} = (\frac{5}{3}, -\frac{1}{3}, -\frac{1}{3})$ means being prepared to lose $\frac{1}{3}$ for the opportunity of winning $\frac{5}{3}$ when W occurs. The drawing depicts an assessment \mathcal{A} of acceptable gambles—black dots—which avoids sure loss, and its natural extension \mathcal{D} —shaded—which is the convex conic hull of \mathcal{A} and $\{I_W, I_D, I_L\}$.



- (b) The dashed triangle delimits the probability simplex, which is spanned by the degenerate probability mass functions—white dots—for which $p_{\omega'}(\omega)$ is equal to 1 if $\omega = \omega'$ and 0 otherwise. This drawing depicts the credal set associated to the assessment presented in Figure 1a—shaded—and the lower expectations that define it—stubbled lines.

Figure 1: We consider the possibility space $\{\text{W}, \text{D}, \text{L}\}$ (for ‘Win’, ‘Draw’, and ‘Loss’). Of the resulting 3-dimensional space of gambles, Figure 1a shows the plane with gambles whose payoffs sum to one. Figure 1b shows the plane containing the resulting probability simplex.

2.2 Relationship with Other Models

Although modeling the uncertainty of an elicitee with the set of gambles she accepts is quite natural and direct, it does differ from the usual, probability-based approach. Let us therefore have a look at how a coherent set of acceptable gambles is related to more common models: expectation operators and sets of probability mass functions.

A coherent set of acceptable gambles \mathcal{D} determines, for any gamble h on Ω , the supremum acceptable buying price $\underline{E}(h) := \sup\{\alpha \in \mathbb{R} : h - \alpha \in \mathcal{D}\}$ and infimum acceptable selling price $\overline{E}(h) := \inf\{\beta \in \mathbb{R} : \beta - h \in \mathcal{D}\} = -\underline{E}(-h)$. The nonlinear operators \underline{E} and \overline{E} are called *lower* and *upper expectations* or *previsions* (Walley, 1991). They satisfy $\underline{E} \leq \overline{E}$ pointwise, i.e., for all gambles. So the gambles the elicitee finds acceptable are essentially those with lower expected payoff greater than or equal to zero.

With a lower expectation \underline{E} , we can associate a *credal set* $\mathcal{M} := \{p : \underline{E} \leq E_p\}$, consisting of all the probability mass functions p whose expectation E_p dominates \underline{E} . In this definition, the inequalities are again pointwise. Our set of axioms forces \mathcal{M} to be a convex subset of the probability simplex. We give an illustration in Figure 1b.

3. The Elicitation Interface

We first briefly discuss elicitation of probability mass functions and credal sets to provide some context and contrast. Then we move on to acceptable gambles.

3.1 Probability Elicitation

Looking at Figure 1b, an interface to elicit a probability mass function on a possibility space of three elements presents itself naturally: Allow the elicitee to indicate a point of the probability simplex.

To elicit a credal set, the above idea should be extended in a way that allows the elicitee to delimit a convex subset of the probability simplex. The three most obvious general approaches are:

- A *direct approach* is to allow multiple points to be selected and take their convex hull. The main advantage is the point-and-click nature, but the elicitee will have difficulty interpreting her actions.
- *Bounding the expectation of gambles* or, more specifically, providing probability intervals is easier to interpret. This can be achieved by ‘placing’ stubbled lines (see Figure 1b), but will result in a more involved interface.
- An *interpretation-agnostic approach* is to partition the simplex into a limited number of points and convex sets which can be selected and combined, e.g., based on comparative probabilities. The main advantage is the point-and-click nature, but there will be non-expressible elicitee attitudes.

The literature focuses mostly on interfaces for eliciting precise probabilities and continuous distributions, nowadays often interactive and on-line (Bastin et al., 2013; Morris et al., 2014). In the imprecise probabilities literature we can find thoughtful consideration of the issue of elicitation (see, e.g., Piatti et al., 2010), but mostly only elicitation interfaces for binary variables—e.g., ‘Win’ vs. ‘No Win’—are considered.

3.2 Gamble Space Representation

Walley (1991, Section 4.1) already considered the direct elicitation of acceptable gambles. But, since the credal set representation of imprecise probabilities has received most attention, it seems that these ideas never led to the concrete design of elicitation interfaces until now.

The space of gambles we have to consider is three-dimensional, because—as with the probability simplex interface—we require our interface to be two-dimensional due to practical display technology limitations. But doing this for sets of acceptable gambles is not so straightforward as for credal sets (cf. Section 3.1): The representation of Figure 1a was based on the restriction to the plane of gambles whose values sum to one. However, not all coherent sets of acceptable gambles can be compactly depicted in such a representation; as an extreme example, the above-mentioned plane is strictly contained in the coherent set consisting of those gambles with components that sum to zero or more.

Nevertheless, because of the Positive Homogeneity axiom, we know that we can represent a coherent set of acceptable gambles—a convex cone—on a two-dimensional surface. For example, we could take its intersection with a sphere centered at the origin or some other suitable two-dimensional surface and then do a projection.

The nature of the projection is influenced by the following considerations: (i) because of the Accepting Partial Gains and Avoiding Sure Loss axioms, the positive octant and the negative octant do not need to be represented prominently or faithfully; (ii) the representation should be essentially invariant under a permutation of the elementary events to avoid introducing biases between them; and (iii) to allow for intuitive exploration by the elicitee, the representation should be a continuous deformation of the points in all but the positive and negative octants.

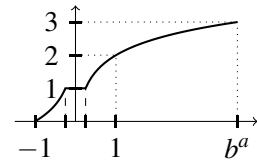
These considerations lead us to a polar projection, where the poles are defined by the line corresponding to constant gambles, i.e., those for which the payoff is equal for all possible outcomes. On the right, we show the example of a spherical such projection found in the United Nations emblem.



To decide on the exact surface to project and the projection center—our ‘North Pole’—we refocus on the interpretation of the projection points as gambles. To provide the elicitee with a reference value when selecting acceptable gambles, we should anchor them by fixing either their maximum or minimum value. We fix the minimum value, as this bounds potential losses and so may—hypothetically—mitigate effects of risk-aversion. We here take -1 as the normalized minimum value. Consequently, the surface we consider is the set of gambles $\{f : \min f = -1\}$, namely, the convex cone with apex $(-1, -1, -1)$ and extreme rays $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. The apex is also the projection center. The projection is illustrated in Figure 2.

The illustration of our representation provided by Figure 3a allows us to pinpoint an important deficiency: the limited range due to the linear scale used. For example, the difference between gamble values one and two times the stake is practically speaking much more important than between five and six times the stake. (The same argument can be used for probability values.) In such a context where relative magnitude takes precedence over absolute magnitude, using a logarithmic scale is a better choice. Because nonpositive values are used in our representation and the constant gamble -1 corresponds to the center of our representation, we use a custom scaling that is based on a ‘saturating’ logarithm:

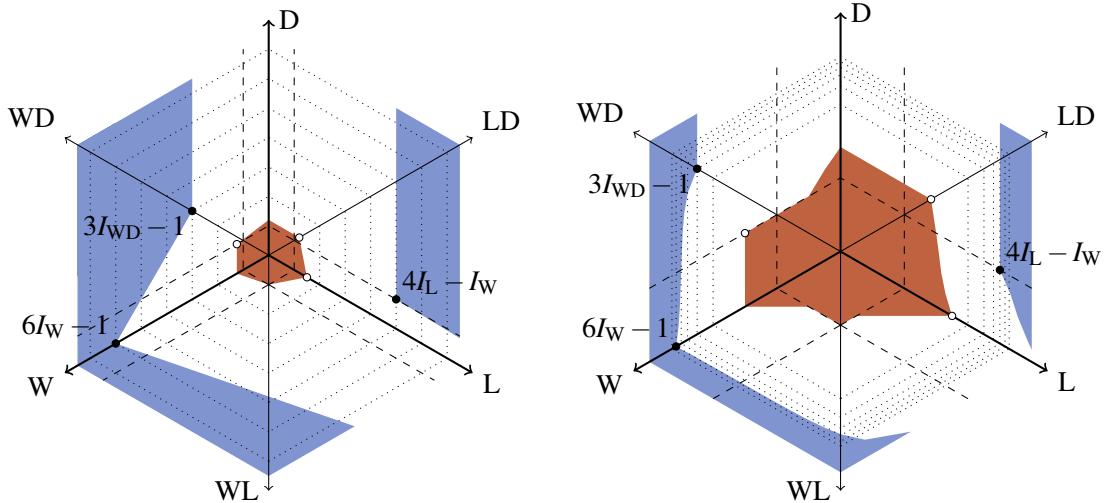
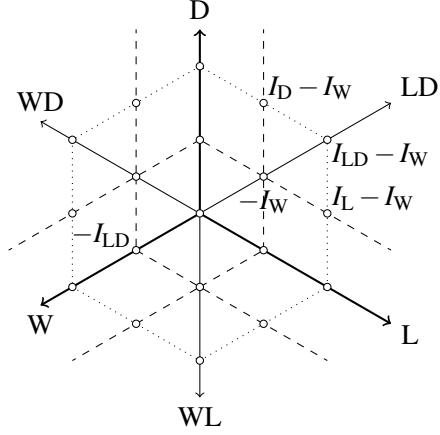
$$f_b^a(x) = \begin{cases} 0 - \frac{1}{a} \log_b(-x), & x \in [-1, -b^{-a}], \quad (\text{values in } [0, 1]) \\ 1, & x \in [-b^{-a}, b^{-a}], \quad (\text{saturating value } 1) \\ 2 + \frac{1}{a} \log_b(x), & x \in (b^{-a}, b^a], \quad (\text{values in } (1, 3]) \end{cases} \quad (1)$$



where $b > 1$ and $a > 0$ determine the smallest representable absolute value b^{-a} and a practical upper bound b^a . The impact on our representation of this scale is illustrated in Figure 3b.

The considerations underlying our representation and its technical details are quite involved. However, the elicitee need not be aware of these to use our concrete implementation, described next.

Figure 2: This is a polar projection of the gamble-space subset $\{f : \min f = -1\}$. The white dots correspond to gambles of interest: The central one represents the apex, the gamble with constant value -1 . The others—some labeled—represent differences of event indicator functions; for example $-I_{LD} = (0, -1, -1)$ and $I_L - I_W = (-1, 0, 1)$. The thick and thin axes point towards gambles with higher payoffs for the corresponding events. The dashed lines form the set of *contingent gambles*, i.e., those that are zero on some event. The dotted line indicates the set of ‘even’ gambles, with a maximum payoff equal to the stake, i.e., one.



(a) This drawing uses our proposed projection of the gamble space. In comparison to Figure 2, the scale we use here is smaller and we now show dotted lines for the loci of gambles with maximum payoff one to six.

(b) This is Figure 3a rescaled using the function of Equation (1) with parameters $b = 2$ and $a = 4$. The effect of the logarithm can, e.g., be seen in the nonlinearity of lines connecting the assessment gambles—black dots. A consequence of saturation is the disappearance of lines between white dots ‘into’ the negative octant (central hexagon).

Figure 3: Translations of Figure 1a to alternate gamble space representations (renormalization to satisfy $\min f = -1$). The gambles in the set \mathcal{A} assessed to be acceptable are represented by black dots. Their natural extension \mathcal{D} , represented by a closed convex polytope before, is now (partially) represented by the disconnected shaded area on the outside. We have added the pointwise additive inverse of the gambles assessed to be acceptable—white-filled dots. These determine the open convex polytope of ‘*rejected*’ gambles—shaded area in the center—that would cause a sure loss if one or more of them were to be assessed acceptable.

3.3 Eliciting Acceptable Gambles

With deployment convenience in mind, our representation was turned into a concrete elicitation interface by implementing it in SVG and Javascript (ECMAScript) so that it can be used in current web browsers. We relied on the library d3js ([Bostock et al., 2011](#)) for visualization and the library NumericJS ([Loisel, 2012](#)) for linear programming functionality.

The biggest and only substantive change we had to make was a discretization of the gamble space. There are two reasons for this:

- We have not found a way to calculate and represent the sets of acceptable and rejected gambles (cf. Figure 3a) fast enough to obtain a responsive interface. This is mainly due to the nonlinear character of their borders, which is a result of the logarithmic rescaling. We can work around this issue by discretizing the representation.
- We wish to show the values of the gamble over which the elicitee is hovering with her pointer. We do not want to show a large number of significant digits of these values, because it is unrealistic to expect the elicitee’s uncertainty attitudes to be so fine-grained; this would therefore be distracting. But now, if we show only a few significant digits and wish to make sure that the numbers shown correspond to the gamble under the pointer, we must discretize the representation.

The result is shown in Figure 4a; it can be used without detailed knowledge of the representation.

The biggest computational challenge we faced when implementing the interface was finding the natural extension \mathcal{D} efficiently enough to make it responsive. To tackle this, we split up the problem into different subroutines. The ones that provide the most important efficiency gain are the *propagation* routines:

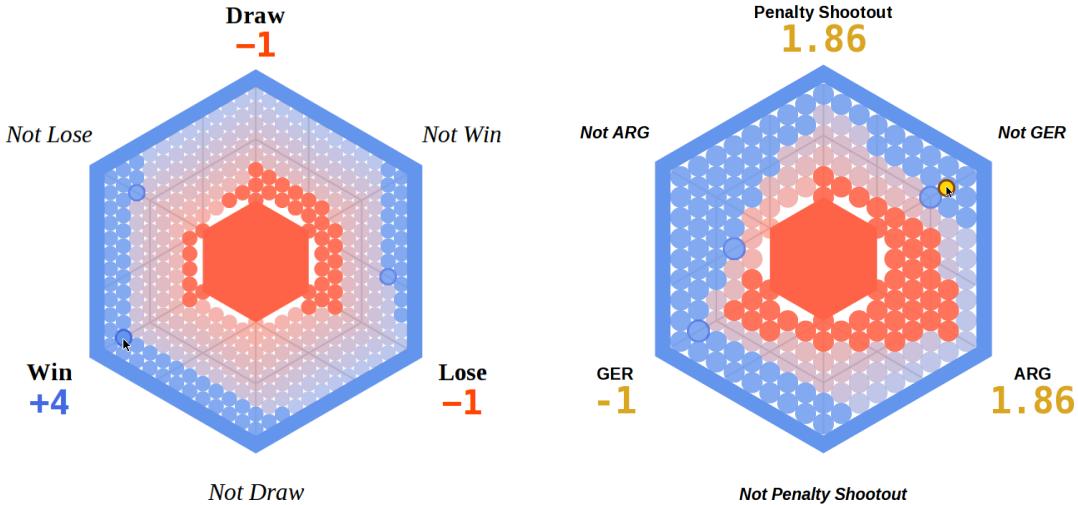
- In our interface, for each dot—i.e., gamble f in the discretization—, there is a unique dot in its negation’s neighborhood whose negation—up to scaling—strictly dominates the gamble f ; e.g., $(-\frac{1}{2}, 1, -1)$ negation-dominates $(\frac{1}{4}, -1, \frac{1}{2})$. We pre-calculate this *negation structure*. So when a dot is marked as accepted, we can mark the dot it is negation-dominated by as rejected due to the Avoiding Sure Loss and Addition axioms.
- In our interface, for each dot, the neighboring dots either pointwise dominate it or are dominated by it; e.g., $(\frac{1}{4}, -1, \frac{1}{2})$ dominates $(\frac{1}{4}, -1, \frac{1}{4})$. We pre-calculate this *dominance structure*. So when a dot is marked as accepted (or rejected), we can recursively propagate this status to all its dominating (dominated) dots due to the Accepting Partial Gains and Addition axioms. All unmarked dots that neighbor an accepted (rejected) dot are gathered in a list of accept (reject) *candidates*, which is kept up to date while propagating.

In our procedure, whenever a dot is marked as accepted or rejected, this change is fully propagated before continuing with the outer *search* routine:

- We iterate over the accept (or reject) candidates and check whether they should be marked acceptable (rejected). The iteration order is determined by the heuristic that dots ‘low’ (‘high’) in the dominance structure should come first, as they are most promising from the propagation perspective. The candidate lists are pre-populated by applying propagation to the assessment.

The subroutine that checks the status of a dot effectively calculates the natural extension by determining whether the dot’s lower (or upper) expectation is nonnegative (strictly negative). Calculating a lower (upper) prevision requires linear programming, a relatively computationally demanding task in a web browser. Propagation sufficiently reduces the number of prevision calculations in practice.

Once the interface was ready for action, we moved to test it in a practical experiment with the goal of getting usage data and general feedback. This is the topic of the next section.



- (a) This shows the interface we implemented with $\{-1, -\frac{1}{2}, -\frac{1}{4}, -\frac{1}{8}, 0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$ as the discretization values. We have chosen the assessment gambles—e.g., $(4, -1, -1)$ —to mimic Figure 3b.
- (b) This shows the experiment’s interface variant, with $\{-1, -\frac{1}{2}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ as the set of discretization values. The assigned gamble appears as the light (yellow) dot hovered over by the pointer.

Figure 4: Elicitation interface screenshots. Each dot represents a gamble; when hovering over it, the payoff for each outcome is shown; assessment dots are a bit darker and have a border.

4. The Experiment

In 1982, [Walley \(1991, Appendix I\)](#) ran an experiment for eliciting lower and upper probabilities concerning the outcome of matches in that year’s FIFA World Cup. There were 17 academic participants. Their assessments were evaluated using the possible pairwise bets between them (cf. Section 4.2). This experiment has later been repeated in the imprecise probabilities community, but more as a diversion than in search of data. Others, such as [Winkler \(1971\)](#), ran earlier, precise probability elicitation experiments organized around sports competitions.

We organized our experiment around the 2014 FIFA World Cup. Whereas in 1982 pen and paper was used, we can now deliver a point-and-click interface accessible over the internet. Moreover, we can calculate the natural extension on-line and thus make sure the elicitree avoids sure loss.

4.1 A Betting Competition Website

We set up the competition as a betting website. We provided the following instructions:

The [...] Competition is a game in which you gamble against other participants. Each of the possibly many matches for which you enter the ‘gambling pool’, you stake €1 (or [...], e.g., \$1), so you can only lose this amount or less per match. The exact amounts you can win or lose depend on the other players’ choices.

You play by expressing your opinion about the outcome of the match in terms of gambles that are acceptable to you. [...] An algorithm will look for a fair bet between you and the other participants. [A bet is a set of gambles assigned to the participants.] If a bet is found, and you are included in it, a single acceptable gamble will be assigned to you and appear in the gamble selection interface for that match. This assigned gamble determines your potential winnings [...]

and losses. You may change your set of acceptable gambles up until an hour before the start of the match; [...]. The match's outcome determines your actual winnings or losses.

Actually, any winnings will not go to you, but, by participating, you commit yourself to pay your losses to the Red Cross/Crescent [...]

The website and its backend were developed using the [Django framework](#). Figure 4b shows a match screen with an assigned gamble.

Participants were recruited by word of mouth initially, then through academic mailing lists, and via social networks near the end of the competition. Participation was anonymous; only a hash of the sign-up email address was stored. A total of 80 people participated, providing assessments for 32 of the 64 matches (due to time constraints, the website was launched and tweaked while the World Cup was already ongoing), for a total of 488 gambles assessed to be acceptable. For 20 of those 32 matches bets were possible, for a total of 100 assigned gambles (cf. Section 4.2). The sum of the lower common expected winnings for those gambles was 37.86 currency units, and in the end the total amount won—and lost—was 47.19 currency units. Even though we could not enforce participants paying to the Red Cross/Crescent, adding the element of having real money at stake, even if not much, was important because it incentivizes them to take the elicitation task more seriously.

This experiment is not just meant as a one-off setup to test our gamble selection interface. Many of its elements can be used as inspiration for practical elicitation experiments. The competitive aspect can be used when eliciting from multiple experts (cf. [Lichtendahl and Winkler, 2007](#)). The repetition aspect is relevant when considering time series, e.g., in a context of weather forecasting.

4.2 Generating Fair Bets

For his experiment, [Walley \(1991\)](#), Appendices I and H6) scored the participants by arranging specific pairwise bets between them for each match, whenever possible. These bets assigned opposite gambles to each that were acceptable given their assessments. The gambles were moreover fair in the sense that their lower expectations—supremum acceptable buying prices (see Section 2.2)—coincided.

We used the same acceptability requirement but moved away from the pairwise approach to a global one, in which for each match a single bet was generated, i.e., the assigned gambles sum to zero. One reason is that we wished the stake per match to be at most one. An advantage of such a setup is that the set of potential bets is in general much larger—e.g., it includes all convex combinations of pairwise bets. Therefore we needed a criterion to choose a single one: we decided on a utilitarian one by maximizing the sum of *identical* acceptable buying prices for the assigned gambles. Our notion of fairness is this maximal common price instead of [Walley's](#) common maximal price.

A bet satisfying the constraints described above can be computed using mixed-integer linear programming. Its formulation is independent of the size of the possibility space Ω . Participant j in J has specified an assessment \mathcal{A}_j . He may be included in the bet or not, encoded by the binary variable b_j . If included, he will be assigned a nontrivial gamble h_j . His acceptable buying price for this gamble is α_j and must—by fairness—be identical to the common buying price α_* . So we have the following program (gamble constraints must be read pointwise):

$$\text{maximize } \sum_{j \in J} \alpha_j = \alpha_* \sum_{j \in J} b_j \quad (2)$$

$$\text{subject to } \sum_{j \in J} h_j = 0, \quad (3)$$

$$\text{and for all } j \text{ in } J: h_j - \alpha_j \geq \sum_{g_j \in \mathcal{A}_j} \lambda_{j,g_j} g_j \text{ with } \lambda_{j,g_j} \geq 0, \quad (4)$$

$$0 \leq \alpha_j = b_j \alpha_* \text{ and } -1 \leq h_j = b_j h_j \text{ with } b_j \in \{0, 1\}. \quad (5)$$

Constraint (4) expresses that $h_j - \alpha_j$ must lie in the cone \mathcal{D}_j spanned by \mathcal{A}_j and the first orthant (cf. Section 2.1). Objective (2) then forces α_j to be an acceptable buying price of h_j (cf. Section 2.2). Constraint (3) guarantees that the h_j gambles form a bet between the participants. Constraints (5) give us the freedom to exclude participants from the bet, force the acceptable prices for those included to coincide, and force the stakes to be one or less.

However, the products $b_j\alpha_*$ and b_jh_j make the Constraints (5) nonlinear. Luckily, we can replace them by an equivalent set of linear constraints: Notice first that $h_j \geq -1$ together with Constraint (3) implies the bound $h_j \leq |J| - 1$. Then the Constraint (4) with all $\lambda_{j,g_j} = 0$ further implies that also $\alpha_j \leq |J| - 1$. The existence of these bounds allows us—given $b_j \in \{0, 1\}$ —to replace (5) by

$$\alpha_* - (|J| - 1)(1 - b_j) \leq \alpha_j \leq \alpha_*, \quad 0 \leq \alpha_j \leq (|J| - 1)b_j, \quad -b_j \leq h_j \leq (|J| - 1)b_j. \quad (6)$$

We used the linear programming library GLPK (Makhorin, 2014) without practical efficiency issues. An instance of an assigned gamble calculated using this program can be seen in Figure 4b.

4.3 Experimental Results

The aim of our exploratory experiment was to obtain feedback about the gamble selection interface, get a view of the types of assessments people provide, test the fair bet generation procedure on real-life data, and have some fun doing it.

We provided a form where participants could optionally enter feedback and information about themselves, such as gender, age, and experience relevant to the competition. However, almost no one made use of it. We did get quite a bit of feedback through personal communications with participants we knew, both laymen and people experienced in uncertainty modeling. As was anticipated by the human-computer interaction (HCI) expert in our team, the interface was found to be too complex: it needs to be simplified, explained more extensively, or a combination thereof.

There were 194 match assessments in total, of which a good 20% was complete in the sense that all dots were marked—after natural extension—, so corresponding to some probability mass function. (A nice anecdote: The few participants who used complete models almost exclusively all had greater losses than winnings.) For the others, with strictly imprecise-probabilistic assessments, the degree of completeness varied over the whole range between just a few and all but a few marked dots.

Something generally orthogonal to completeness is the number of selected dots per assessment:

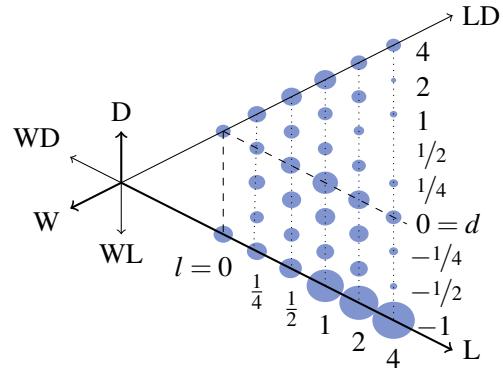
#dots:	1	2	3	4	5	6	7	8
#assessments:	54	52	47	26	8	5	1	1

We see that the number of selected dots is mostly concentrated in relatively small numbers. In fact, whenever four or more dots were registered, there usually were some that were actually redundant after natural extension, i.e., were implied by other selected dots. So participants kept things simple.

Regarding the distribution of the dot selection, Figure 5 shows that mainly dots on the axes and corresponding to contingent gambles were chosen, but not overwhelmingly so. It nevertheless indicates that restricting attention to these loci may be a way to simplify the interface.

The bet generation procedure worked as intended, but we noticed that on average a third and sometimes up to more than 60% of the participants that gave an assessment for a match—typically relatively more imprecise (incomplete)—were not included in the bet. Once the match outcome is determined, the assigned gamble results in a payoff; such feedback gives the participant an idea of the quality of his assessments (cf. scoring rules; see, e.g., Winkler, 1971). Therefore it would be useful to modify the bet generation procedure to include more participants.

Figure 5: This drawing depicts the observed relative dot selection frequency. Because of symmetry—outcome identity irrelevance—all dots were mapped to the subregion of dots $(-1, d, l)$ (cfr. Figure 3b). The possible values for d and l are respectively shown on the right and at the bottom of the drawing. The *area* of the circles is proportional to the relative number of selections of that dot; the largest circle, at $(-1, -1, 4)$, corresponds to 12.5% of selections.



5. Conclusions

We designed a gamble selection interface that is based on a representation of the space of gambles tailored to elicitation. We coded an efficient—responsive—implementation and used it in an experiment. In support of this experiment, we developed a novel procedure for generating fair bets.

From the experiment, we learned that the interface can be effectively used, but also that it needs to be made more usable and transparent. Furthermore, given that the majority of assessments made by the participants were imprecise, a more generally important conclusion is that *imprecision is a non-negligible aspect of uncertainty*: models that do not allow for it to be expressed may lead to gambles—i.e., any decision under uncertainty—that its users are actually not willing to commit to.

Follow-up work should focus on improvements to the interface, user guidelines, and bet generator, and experimental comparison to alternative interfaces (see, e.g., Section 3.1).

Acknowledgments

This work is part of the *Safe Statistics* project at the CWI financed by the Netherlands Organisation for Scientific Research (NWO). Erik Quaeghebeur was an ERCIM “Alain Bensoussan” Fellow, receiving funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 246016. Tom Sterkenburg is also affiliated with the Faculty of Philosophy of the University of Groningen. Teresa Piovesan is partially funded by the European Project SIQS.

References

- T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, 2014. doi:[10.1002/9781118763117](https://doi.org/10.1002/9781118763117).
- L. Bastin et al. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environ. Modell. Software*, 39:116–134, 2013. doi:[10.1016/j.envsoft.2012.02.008](https://doi.org/10.1016/j.envsoft.2012.02.008).
- M. Bostock, V. Ogievetsky, and J. Heer. D³: Data-driven documents. *IEEE Trans. Visual Comput. Graphics*, 17(12):2301–2309, 2011. doi:[10.1109/TVCG.2011.185](https://doi.org/10.1109/TVCG.2011.185). URL <http://d3js.org/>.
- R. M. Cooke. *Experts in Uncertainty*. Oxford University Press, 1991.

- J. De Bock and G. de Cooman. Credal networks under epistemic irrelevance: The sets of desirable gambles approach. *Int. J. Approx. Reason.*, 56:178–207, 2015. doi:[10.1016/j.ijar.2014.07.002](https://doi.org/10.1016/j.ijar.2014.07.002).
- M. J. Druzdzel and L. C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, page 141–148. Morgan Kaufmann, 1995.
- M. J. Druzdzel and L. C. van der Gaag. Building probabilistic networks: Where do the numbers come from? *IEEE Trans. Knowl. Data Eng.*, 12(4):481–486, 2000. doi:[10.1109/TKDE.2000.868901](https://doi.org/10.1109/TKDE.2000.868901).
- K. C. Lichtendahl, Jr. and R. L. Winkler. Probability elicitation, scoring rules, and competition among forecasters. *Manage. Sci.*, 53(11):1745–1755, 2007. doi:[10.1287/mnsc.1070.0729](https://doi.org/10.1287/mnsc.1070.0729).
- S. Loisel. Numeric javascript, 2012. URL <http://numericjs.com/>.
- A. Makhorin. GNU Linear Programming Kit, 2014. URL <http://www.gnu.org/software/glpk/>.
- D. E. Morris, J. E. Oakley, and J. A. Crowe. A web-based tool for eliciting probability distributions from experts. *Environ. Modell. Software*, 52:1–4, 2014. doi:[10.1016/j.envsoft.2013.10.010](https://doi.org/10.1016/j.envsoft.2013.10.010).
- A. O’Hagan and J. E. Oakley. Probability is perfect, but we can’t elicit it perfectly. *Reliab. Eng. Syst. Saf.*, 85(1–3):239–248, 2004. doi:[10.1016/j.ress.2004.03.014](https://doi.org/10.1016/j.ress.2004.03.014).
- A. O’Hagan et al. *Uncertain Judgements: Eliciting Experts’ Probabilities*. Wiley, 2006.
- A. Piatti, A. Antonucci, and M. Zaffalon. Building knowledge-based systems by credal networks: a tutorial. In A. R. Baswell, editor, *Advances in Mathematics Research*, volume 11, page 227–279. Nova Science Publishers, 2010.
- E. Quaeghebeur. The CONESTrip algorithm. In R. Kruse et al., editors, *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, page 45–54. Springer, 2013. doi:[10.1007/978-3-642-33042-1_6](https://doi.org/10.1007/978-3-642-33042-1_6).
- E. Quaeghebeur. Desirability. In [Augustin et al. \(2014\)](#), chapter 1, page 1–27.
- S. Renooij. Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review*, 16(3):255–269, 2001. doi:[10.1017/S0269888901000145](https://doi.org/10.1017/S0269888901000145).
- C. S. Spetzler and C.-A. S. Staël von Holstein. Probability encoding in decision analysis. *Manage. Sci.*, 22(3):340–358, 1975. doi:[10.1287/mnsc.22.3.340](https://doi.org/10.1287/mnsc.22.3.340).
- P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman & Hall, 1991.
- P. Walley. Towards a unified theory of imprecise probability. *Int. J. Approx. Reason.*, 24(2–3):125–148, 2000. doi:[10.1016/S0888-613X\(00\)00031-1](https://doi.org/10.1016/S0888-613X(00)00031-1).
- P. M. Williams. Indeterminate probabilities. In M. Przelecki et al., editors, *Formal Methods in the Methodology of Empirical Sciences*, page 229–246. D. Reidel Publishing Company, 1976. doi:[10.1007/978-94-010-1135-8_16](https://doi.org/10.1007/978-94-010-1135-8_16).
- R. L. Winkler. Probabilistic prediction: Some experimental results. *J. Am. Stat. Assoc.*, 66(336):675–685, 1971.

Errors Bounds for Finite Approximations of Coherent Lower Previsions

Damjan Škulj

*University of Ljubljana, Faculty of Social Sciences
Ljubljana (Slovenia)*

DAMJAN.SKULJ@FDV.UNI-LJ.SI

Abstract

Coherent lower previsions are general probabilistic models allowing incompletely specified probability distributions. However, for complete description of a coherent lower revision – even on finite underlying sample spaces – an infinite number of assessments is needed in general. Therefore, they are often only described approximately by some less general models, such as coherent lower probabilities or in terms of some other finite set of constraints. The magnitude of error induced by the approximations has often been neglected in the literature, despite the fact that it can be significant with substantial impact on consequent decisions. An apparent reason is that no widely used general method for estimating the error seems to be available at the moment. The goal of this paper is to provide such a method. The proposed method allows calculating an upper bound for the error of a finite approximation of coherent lower revision on a finite underlying sample space. An estimate of the maximal error is especially useful in the cases where calculating assessments is computationally demanding. Our method is based on convex analysis applied to credal sets, which in the case of finite sample spaces correspond to convex polyhedra.

Keywords: lower revision; partially specified lower revision; credal set; convex polyhedron; quadratic programming.

1. Introduction

One of the most popular and also most general models of imprecise probabilities are *coherent lower revisions* (see, e.g., [Miranda, 2008](#); [Troffaes and De Cooman, 2014](#)). A coherent lower revision \underline{P} is an imprecise probability model based on judgements about the lower or upper expectations on a set of bounded maps \mathcal{K} from a sample space \mathcal{X} to real numbers, also called *gambles*. The set of all gambles on a given underlying sample space will be denoted by \mathcal{L} . In this paper, all sample spaces are finite, therefore, we do not address any measurability or countable additivity conditions. The *judgement* or *assessment* $\underline{P}(f) = a$ states that every precise probability distribution P compatible with \underline{P} must satisfy $E_P(f) \geq a$, that is $\underline{P}(f)$ means that the expectation of f is at least a . *Coherence* in this context means that the judgements on the set of gambles allow, for every gamble f , the existence of at least one precise probability distribution P compatible with \underline{P} for which $E_P(f) = \underline{P}(f)$. The expectation functionals with respect to precise (finitely additive) probability distributions are often called *linear revisions*.

A coherent lower revision \underline{P} specified on a set of gambles \mathcal{K} can have multiple possible extensions to a larger set, say $\mathcal{H} \supset \mathcal{K}$. In other words, there can be multiple coherent lower revisions that coincide on a set of gambles. In particular, a coherent lower revision may be approximated by a more specific model, such as *coherent lower probability* (see, e.g., [Antonucci and Cuzzolin, 2010](#)), in which case its restriction to *indicator gambles* is only known, i.e. an *indicator gamble* 1_A

is a map $\mathcal{X} \rightarrow \mathbb{R}$ such that $1_A(x)$ equals 1 if $x \in A$ and 0 otherwise. We will write 1_x instead of $1_{\{x\}}$ for elements $x \in \mathcal{X}$.

In this paper we investigate the following problem. Let \underline{P} be a coherent lower prevision on the set \mathcal{L} of all gambles on a finite sample space \mathcal{X} . Its full description would in general require detailed information on the set of compatible precise models, called *credal set*, which often is unavailable. Suppose that instead we know the values of \underline{P} on a set of gambles \mathcal{K} . The restriction $\underline{P}_{\mathcal{K}}$ approximates \underline{P} and the natural question arises, how accurate is this approximation. Given the restriction, \underline{P} is an extension of $\underline{P}_{\mathcal{K}}$, which in general is not unique. Therefore, we would like to know by how much can another extension deviate from \underline{P} . That is, we want to find the maximal distance between two arbitrary extensions of a coherent lower prevision on a finite set \mathcal{K} to the set of all gambles.

In our analysis we first show that the maximal possible distance is always reached when one of the extensions is the *natural extension*. Consequently, much of the analysis is done on the credal set of the natural extension with the special emphasis on its extreme points. Our main result gives an upper bound for the maximal distance in terms of distances between the extreme points.

The paper is structured as follows. In Section 2 we review basic concepts of imprecise probabilities with the emphasis on coherent lower previsions. In Section 3 we analyze basic properties of credal sets as convex polyhedra and apply some general concepts of convex analysis to the case of credal sets. Our main results are stated in Section 4.

2. Notation and basic results

In this section we introduce the notation and review the concepts used in the paper. When possible we will stick with the standard terminology used in the theory of imprecise probabilities, which will sometimes be supplemented by the standard terminology of convex analysis, linear algebra and optimization.

GAMBLES.

Throughout this paper let \mathcal{X} represent a finite set, a *sample space*, and \mathcal{L} the set of all real-valued maps on \mathcal{X} , also called *gambles*. Equivalently, \mathcal{L} may be viewed as the set of vectors in $\mathbb{R}^{|\mathcal{X}|}$. The set of gambles will be endowed by the standard inner product $f \cdot g = \sum_{x \in \mathcal{X}} f(x)g(x)$, which generates the l^2 norm: $\|f\| = \sqrt{f \cdot f} = \sqrt{\sum_{x \in \mathcal{X}} f(x)^2}$, and the Euclidean distance between vectors: $d(f, g) = \|f - g\|$, which will be used by default throughout the paper.

LINEAR PREVISIONS.

A *linear prevision* P is an expectation functional with respect to some probability mass vector p on \mathcal{X} . It maps a gamble f into a real number $P(f)$. Usually, we will write $P(f) = \sum_{x \in \mathcal{X}} p(x)f(x) =: P \cdot f$. The set of linear previsions is therefore a subset of the dual space of \mathcal{L} . The inner product notation is introduced because we will often use linear functionals of the form $f \mapsto p \cdot f$ where the vector p will not necessarily be a probability mass vector. We will then use the inner product notation to avoid misinterpretations. Without danger of confusion we will therefore interpret a linear prevision P as a vector with the same length as gambles in \mathcal{L} .

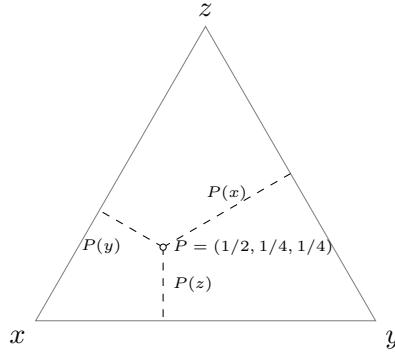


Figure 1: Probability simplex: the distance from a side denotes the probability of the element at the opposite vertex.

PROBABILITY SIMPLEX.

If the sample space \mathcal{X} contains exactly three elements, say $\mathcal{X} = \{x, y, z\}$, the probability mass vectors can be represented as points of the form $(p(x), p(y), p(z))$ in \mathbb{R}^3 . However, since the restriction $p(x) + p(y) + p(z) = 1$ applies, they in fact form a two dimensional space, which can be depicted as an equilateral triangle with vertices x , y and z . Given any point in this triangle, the sum of distances to its sides is constantly equal to its altitude, which equals $\frac{\sqrt{3}}{2}a$, where a is the common length of the sides. Taking $a = \frac{2}{\sqrt{3}}$ makes the altitude equal to 1. The distance of a point from each side now denotes the probability of the point in the opposite vertex. (See Figure 1.) Probability simplex diagrams are very useful to illustrate concepts of imprecise probabilities; however, one needs to be cautious not to be misled by specifics of low dimensional probability spaces.

COHERENT LOWER PREVISIONS.

A *coherent lower prevision* on an arbitrary set of gambles \mathcal{K} is a mapping $\underline{P}: \mathcal{K} \rightarrow \mathbb{R}$ that allows the representation

$$\underline{P}(f) = \min_{\mathcal{M}} P(f) \quad (1)$$

for every $f \in \mathcal{K}$, where \mathcal{M} is a closed and convex set of linear previsions. Note that unless \mathcal{K} is the set of all gambles, there may be multiple sets \mathcal{M} that fit into equation (1) (this is also one of the motivations for this paper); however, there is a unique maximal such set. We will denote the maximal such set with $\mathcal{M}(\underline{P})$ and call it the *credal set* of \underline{P} .

THE NATURAL EXTENSION.

Given a coherent lower prevision \underline{P} on \mathcal{K} , it is possible to extend it to the set of all gambles \mathcal{L} in possibly several different ways, and again, there is unique minimal extension, called the *natural extension*:

$$E(f) = \min_{P \in \mathcal{M}(\underline{P})} P(f). \quad (2)$$

Note that replacing $\mathcal{M}(\underline{P})$ with another set \mathcal{M} of linear previsions satisfying the equation (1) would result in some other extension of \underline{P} .

A mapping $\underline{P}: \mathcal{K} \rightarrow \mathbb{R}$, where \mathcal{K} is a linear (vector) space, is a coherent lower prevision if and only if it satisfies the following axioms (Miranda, 2008) for all $f, g \in \mathcal{K}$ and $\lambda \geq 0$:

- (P1) $\underline{P}(f) \geq \inf_{x \in \mathcal{X}} f(x)$ [accepting sure gains];
- (P2) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ [non-negative homogeneity];
- (P3) $\underline{P}(f + g) \geq \underline{P}(f) + \underline{P}(g)$ [superlinearity].

An easy consequence of the definitions is :

- (P4) $\underline{P}(f + \lambda 1_{\mathcal{X}}) = \underline{P}(f) + \lambda$ for any $\lambda \in \mathbb{R}$ and $f \in \mathcal{L}$ [constant additivity].

3. Credal set as a convex polyhedron

A credal set is a closed and convex set of linear previsions. Since every linear prevision can be uniquely represented as a probability mass vector, a credal set can be represented as a convex set of probability mass vectors. The set \mathcal{M} is therefore the maximal set of $|\mathcal{X}|$ -dimensional vectors p satisfying:

$$p \cdot f \geq \underline{P}(f) \quad \text{for every } f \in \mathcal{K}, \quad (3)$$

$$p \cdot 1_x \geq 0 \quad \text{for every } x \in \mathcal{X} \text{ and} \quad (4)$$

$$p \cdot 1_{\mathcal{X}} = 1. \quad (5)$$

In the sequel we will assume that the set \mathcal{K} is finite. When needed, we will index its elements as f_i for $i \in \{1, \dots, n\}$.

According to the above, it would be suitable to extend the domain of \underline{P} with the gambles of the form 1_x for every $x \in \mathcal{X}$. Doing so, though, may result in a non-coherent lower prevision, because other constraints may already imply that $\underline{P}(1_x) \geq 0$, where the inequality may even be strict. Therefore we adopt the following convention:

Convention 1 *The domain \mathcal{K} of all lower previsions used will contain all gambles of the form 1_x together with the value $\underline{P}(1_x) = 0$, unless $\underline{P}(1_x) \geq 0$ is already implied by other values of \underline{P} on \mathcal{K} .*

Assuming the above convention, the credal set of coherent lower prevision \underline{P} is the set of vectors p satisfying constraints (3) and (5).

In the case where \mathcal{K} is finite, the corresponding credal set is a *convex polyhedron*. Strictly speaking, it is an \mathcal{H} -polyhedron, which means that it is bounded and an intersection of a finite number of half spaces. According to Theorem 14.3 in Gruber (2007) every \mathcal{H} -polyhedron in an \mathbb{R}^m is also a \mathcal{V} -polyhedron, which means that it is a convex combination of a finite number of extreme points.

Example 1 Let \underline{P} be a lower prevision on $\mathcal{K} = \{f_1, \dots, f_5\}$ where

$$\begin{aligned} f_1 &= (0, 1, 0.5) & f_2 &= (0, 0.5, 1) & f_3 &= (0.15, 0, 1) \\ f_4 &= (1, 0, 0.6) & f_5 &= (0.2, 1, 0) \end{aligned}$$

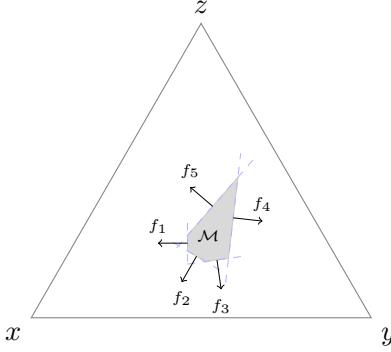


Figure 2: Credal set from Example 1 as an intersection of half planes: their support lines are dashed, gambles $f_i \in \mathcal{K}^+$ are depicted as normal vectors to faces.

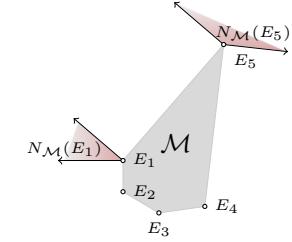


Figure 3: Normal cones at extreme points are the non-negative hulls of the normal vectors of adjacent faces.

and

$$\begin{array}{lll} \underline{P}(f_1) = 0.46 & \underline{P}(f_2) = 0.4 & \underline{P}(f_3) = 0.25 \\ \underline{P}(f_4) = 0.44 & \underline{P}(f_5) = 0.4 & \end{array}$$

The credal set corresponding to \underline{P} is depicted in Figure 2 as an intersection of half-planes.

FACES AND EXTREME POINTS OF A FINITELY GENERATED CREDAL SET.

The *faces* of a credal set \mathcal{M} are the sets of the form $\mathcal{M}_f = \{P \in \mathcal{M}: P(f) = \underline{E}(f)\}$, where f is an arbitrary gamble. The smallest faces are exactly the extreme points and the faces of codimension 1 are called *facets*¹. The set of all extreme points of \mathcal{M} will be denoted by $\mathcal{E}(\mathcal{M})$ or simply \mathcal{E} . The set of extreme points of a face \mathcal{M}_f will be denoted by \mathcal{E}_f , and $\mathcal{E}_f \subseteq \mathcal{E}$ holds.

Example 2 The extreme points of the credal set from Example 1 are

$$\begin{array}{lll} E_1 = (0.4, 0.32, 0.28) & E_2 = (0.43, 0.35, 0.23) & E_3 = (0.39, 0.42, 0.19) \\ E_4 = (0.32, 0.48, 0.20) & E_5 = (0.15, 0.37, 0.48) & \end{array}$$

(See Figure 3.)

Let $f \in \mathcal{K}$ be a gamble and $\underline{P}(f)$ its lower prevision. Then the lower prevision of the gamble $f - \underline{P}(f)1_{\mathcal{X}}$ equals 0. Moreover, setting $\underline{P}(f - \underline{P}(f)1_{\mathcal{X}}) = 0$ is equivalent to setting the lower prevision of f to $\underline{P}(f)$, by constant additivity. Following this idea, we extend a credal set \mathcal{M} to the set of vectors

$$\hat{\mathcal{M}} = \{p: p \cdot (f - \underline{P}(f)1_{\mathcal{X}}) \geq 0, \text{ for every } f \in \mathcal{K}\}, \quad (6)$$

1. The codimension 1 is meant relative to the dimension of \mathcal{M} . That is $\dim \mathcal{M}_f = \dim \mathcal{M} - 1$. Note also that a credal set is at most of dimension $|\mathcal{X}| - 1$ because of the constraint $P(1_{\mathcal{X}}) = 1$.

which is a convex cone, with the *basis* \mathcal{M} . This means that every $p \in \hat{\mathcal{M}}$ is of the form $p = \lambda P$ for some $\lambda \geq 0$ and $P \in \mathcal{M}$. This is easily seen by noticing that every $p \in \hat{\mathcal{M}}$ has non-negative components, which is guaranteed by Convention 1. Dividing $\mathbf{0} \neq p \in \hat{\mathcal{M}}$ by the sum of its components then results in a vector P whose components are non-negative, sum to one and clearly satisfy the same linear constraints as p , except (5).

Given a credal set \mathcal{M} , the *cone of (almost) desirable gambles* contains exactly those gambles in \mathcal{L} whose lower revision is non-negative:

$$\mathcal{D} = \{f \in \mathcal{L}: P(f) \geq 0 \text{ for every } P \in \mathcal{M}\} = \{f \in \mathcal{L}: \underline{P}(f) \geq 0\}. \quad (7)$$

The gambles f with $\underline{P}(f) = 0$ are sometimes called *marginally desirable*.

3.1 Normal cones of credal sets

THE NORMAL CONE.

Let

$$\mathcal{C} = \{x \in \mathbb{R}^n: Ax \leq b\}, \quad (8)$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$ a vector, be a convex polyhedron and x a point on its boundary. According to Gruber (2007), the *normal cone* at x is the set

$$N_{\mathcal{C}}(x) = \{u: u \cdot y \leq u \cdot x \text{ for all } y \in \mathcal{C}\} = \{u: u \cdot (y - x) \leq 0 \text{ for all } y \in \mathcal{C}\}. \quad (9)$$

In our case, let \mathcal{M} be a credal set defined by constraints of the form (3) and (5) and E its boundary point. The normal cone of \mathcal{M} at E is the set

$$N_{\mathcal{M}}(E) = \{f: E(f) \leq P(f) \text{ for every } P \in \mathcal{M}\}. \quad (10)$$

The normal cone is thus the set of gambles f that satisfy $E(f) = \underline{P}(f)$.

Proposition 2 (Gruber (2007) Proposition 14.1.) *Let \mathcal{C} be a convex polyhedron defined as in (8) and x its boundary point. Let $a_i \cdot x = b_i$ hold for exactly $i \in I \subseteq \{1, \dots, m\}$, where a_i denotes i -th row of the matrix A . Then $N_{\mathcal{C}}(x) = \text{pos}\{a_i: i \in I\}$, where pos denotes the non-negative hull.*

Corollary 3 *Let \mathcal{M} be a credal set defined by constraints (3) and (5). Then the set of (almost) desirable gambles \mathcal{D} corresponding to \mathcal{M} is the normal cone of $\hat{\mathcal{M}}$ at $\mathbf{0}$ and we have that $\mathcal{D} = \text{pos}\{f - \underline{P}(f)\mathbf{1}_X: f \in \mathcal{K}\}$.*

Proof The set $\hat{\mathcal{M}}$ is a convex cone whose support hyperplanes are exactly the sets of the form $H_f = \{p: p \cdot (f - \underline{P}(f)\mathbf{1}_X) = 0\}$ for $f \in \mathcal{K}$, and the origin is exactly the intersection of all support hyperplanes: $\mathbf{0} \cdot (f - \underline{P}(f)\mathbf{1}_X) = 0$ for every $f \in \mathcal{K}$. We can therefore apply Proposition 2. ■

Remark 4 In Augustin et al. (2014) Chapter 1, the set constructed as \mathcal{D} in Corollary 3 is called the *natural extension* of the assessment \mathcal{K} . The fact that the set of desirable gambles is the non-negative hull of marginally desirable assessments in \mathcal{K} with included strictly positive gambles can also be found in Chapter 2 of the mentioned book. In our case, strictly positive gambles are included because of Convention 1.

Corollary 5 Let \mathcal{M} be a credal set defined by constraints of the form (3) and (5), $E \in \mathcal{M}$ a linear prevision and h a gamble such that $E(h) = \underline{P}(h)$. For every gamble $f \in \mathcal{K}$ let $\tilde{f} = f - \underline{P}(f)$, and thus $\underline{P}(\tilde{f}) = 0$ for all $f \in \mathcal{K}$.

Suppose that $E(\tilde{f}_i) = 0$ for exactly $i \in I \subseteq \{1, \dots, n\}$. Then there exist $\alpha_i \geq 0$ for every $i \in I$ and $\beta \in \mathbb{R}$ so that

$$h = \sum_{i \in I} \alpha_i \tilde{f}_i + \beta 1_{\mathcal{X}}. \quad (11)$$

Proof Let $h \in \mathcal{L}$ be a gamble such that $E(h) = \underline{P}(h)$. Set $g = h - \underline{P}(h)$. Then, for every $p \in \hat{\mathcal{M}}$ (see (6)), $p = \alpha P$ for some $P \in \mathcal{M}$ and $\alpha \geq 0$. Therefore $p \cdot g = \alpha P \cdot g \geq 0 = E \cdot g$, whence $g \in N_{\hat{\mathcal{M}}}(E)$. By Proposition 2, $g = \sum_{i \in I} \alpha_i \tilde{f}_i$ for some non-negative constants α_i . Hence $h = \sum_{i \in I} \alpha_i \tilde{f}_i + \underline{P}(h)1_{\mathcal{X}}$, which proves the proposition. ■

Note that Equation (11) still holds if \tilde{f}_i are replaced by f_i .

4. The distance between coherent lower previsions

4.1 The definition of the distance

Let \underline{P} and \underline{P}' be two coherent lower previsions on the set of all gambles \mathcal{L} on a finite set \mathcal{X} . We define the distance² between \underline{P} and \underline{P}' as

$$d(\underline{P}, \underline{P}') = \max_{f \in \mathcal{L}} \frac{|\underline{P}(f) - \underline{P}'(f)|}{\|f\|}, \quad (12)$$

where the norm $\|f\| = \sqrt{f \cdot f}$ is the Euclidean norm in $\mathbb{R}^{|\mathcal{X}|}$. Clearly, the following alternative definition is equivalent: $d(\underline{P}, \underline{P}') = \max_{\substack{f \in \mathcal{L} \\ \|f\|=1}} |\underline{P}(f) - \underline{P}'(f)|$.

It is readily verified that the above distance function induces a metric in the set of all lower previsions on \mathcal{L} . In this section we will analyze the maximal possible distance between two coherent lower previsions that coincide on a finite set of gambles.

Suppose that \underline{P} is a lower prevision on \mathcal{L} , and the only information about it are the values on a finite set of gambles $\mathcal{K} \subset \mathcal{L}$. That is, $\underline{P}(f)$ are given for every $f \in \mathcal{K}$. We denote the restriction of \underline{P} to \mathcal{K} by $\underline{P}_{\mathcal{K}}$. We also adopt Convention 1. The natural extension \underline{E} is the minimal (or the least committal) extension of $\underline{P}_{\mathcal{K}}$. This implies that $\underline{P}(f) \geq \underline{E}(f)$ for every $f \in \mathcal{L}$. Therefore, given another extension \underline{P}' of $\underline{P}_{\mathcal{K}}$, we have that

$$|\underline{P}(f) - \underline{P}'(f)| \leq \max\{\underline{P}(f) - \underline{E}(f), \underline{P}'(f) - \underline{E}(f)\}, \quad (13)$$

which implies that $d(\underline{P}, \underline{P}') \leq \max\{d(\underline{P}, \underline{E}), d(\underline{P}', \underline{E})\}$. As we are interested in the maximal possible distance between coherent lower previsions coinciding on \mathcal{K} , it will therefore be enough to focus to the case where one of them is the natural extension of $\underline{P}_{\mathcal{K}}$.

4.2 Maximal distance to the natural extension

Let \underline{E} and \underline{P} be respectively the natural extension of $\underline{P}_{\mathcal{K}}$ and another extension, and \mathcal{M} and \mathcal{C} respectively their credal sets. As described in previous sections, both are convex sets and the natural extension is a convex polyhedron with extreme points $\mathcal{E}(\mathcal{M})$.

2. For another distance function between coherent lower previsions, see, e.g., Škulj and Hable (2013).

Assuming the above notations, we start with the following proposition.

Proposition 6 *Take some $f \in \mathcal{K}$ and let \mathcal{M}_f be the corresponding face of \mathcal{M} . Then $\mathcal{C} \cap \mathcal{M}_f \neq \emptyset$.*

Proof Clearly, \mathcal{M}_f contains exactly all linear previsions P in \mathcal{M} such that $P(f) = \underline{P}(f)$. If no $P \in \mathcal{C}$ belongs to \mathcal{M}_f , this then implies that $P(f) > \underline{P}(f)$ for every $P \in \mathcal{C}$, and since \mathcal{C} is compact, this would imply that $\min_{P \in \mathcal{C}} P(f) > \underline{P}(f)$, which contradicts the assumptions. ■

Corollary 7 *Let $h \in \mathcal{L}$ be an arbitrary gamble. Then:*

- (i) $\underline{P}(h) \leq \max_{P \in \mathcal{M}_f} P(h)$ for every $f \in \mathcal{K}$;
- (ii) $\underline{P}(h) \leq \min_{f \in \mathcal{K}} \max_{P \in \mathcal{M}_f} P(h)$; the inequality is tight in the sense that for every $h \in \mathcal{L}$ an extension of $\underline{P}_{\mathcal{K}}$ exists that gives equality in the equation.
- (iii) $\underline{P}(h) \leq \min_{f \in \mathcal{K}} \max_{E \in \mathcal{E}_f} E(h)$ where \mathcal{E}_f is the set of extreme points of the face \mathcal{M}_f ; and the inequality is again tight.

Proof (i) is an immediate consequence of Proposition 6.

The inequality in (ii) is a direct consequence of (i). It remains to prove that there is an extension of $\underline{P}_{\mathcal{K}}$ where the equality is reached.

Let \mathcal{M}_f be a face of \mathcal{M} and let $P_f \in \arg \max_{P \in \mathcal{M}_f} P(h)$. Let \mathcal{M}' be the convex hull of $\{P_f : f \in \mathcal{K}\}$ and \underline{P}' the corresponding coherent lower prevision, which coincides with \underline{P} on \mathcal{K} by construction, and thus must satisfy the inequality (ii). For every $P \in \mathcal{M}'$, on the other hand, we have that $P = \sum_{f \in \mathcal{K}} \alpha_f P_f$, for some collection of values $\alpha_f \geq 0$ for every $f \in \mathcal{K}$ and $\sum_{f \in \mathcal{K}} \alpha_f = 1$. Thus,

$$P(h) = \sum_{f \in \mathcal{K}} \alpha_f P_f(h) \geq \min_{f \in \mathcal{K}} P_f(h) = \min_{f \in \mathcal{K}} \max_{P \in \mathcal{M}_f} P(h) \quad (14)$$

Hence, $\underline{P}'(h) = \min_{P \in \mathcal{M}'} P(h) \geq \min_{f \in \mathcal{K}} \max_{P \in \mathcal{M}_f} P(h)$, which combined with the above reverse inequality gives the required equality.

The fact that extremal values are reached in extreme points easily implies (iii). ■

Now we can express the maximal possible distance between two arbitrary extensions of $\underline{P}_{\mathcal{K}}$ in terms of its natural extension alone.

Corollary 8 *Let \underline{E} be the natural extension and \underline{P} and \underline{P}' two other extensions of $\underline{P}_{\mathcal{K}}$, and $h \in \mathcal{L}$ a gamble. Then $|\underline{P}(h) - \underline{P}'(h)| \leq \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} P(h) - \underline{E}(h)$ and*

$$d(\underline{P}, \underline{P}') \leq \max_{\|h\|=1} \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} P(h) - \underline{E}(h). \quad (15)$$

Proof The first inequality is a direct consequence of Corollary 7(iii) and Eq. (13). The second inequality is an immediate consequence of the first one, definition of the distance between two coherent lower previsions and the fact that $\underline{E}(h)$ is less than $P(h)$ for every feasible P . ■

Equation (15) gives the maximal possible distance between two unknown extensions of $\underline{P}_{\mathcal{K}}$ entirely in terms of its natural extension. However, as an optimization problem it is not solvable in any

apparently applicable way. We will therefore apply it to derive a practically computable upper bounds.

By the definition of \underline{E} we have:

$$d(\underline{P}, \underline{P}') \leq \max_{\|h\|=1} \max_{E \in \mathcal{E}} \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} P(h) - E(h) \quad (16)$$

$$= \max_{E \in \mathcal{E}} \max_{\|h\|=1} \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} P(h) - E(h) \quad (17)$$

by interchanging $\max_{\|h\|=1}$ and $\min_{f \in \mathcal{K}}$:

$$\leq \max_{E \in \mathcal{E}} \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} \max_{\|h\|=1} P(h) - E(h) \quad (18)$$

$$= \max_{E \in \mathcal{E}} \min_{f \in \mathcal{K}} \max_{P \in \mathcal{E}_f} d(P, E), \quad (19)$$

where $d(P, E)$ is the Euclidean distance between extreme points P and E .

Now denote

$$\bar{d}(E, f) = \max_{P \in \mathcal{E}_f} d(P, E), \quad (20)$$

which is the maximal Euclidean distance between an extreme point E and a face \mathcal{M}_f . Thus we obtain the following formula:

$$d(\underline{P}, \underline{P}') \leq \max_{E \in \mathcal{E}} \min_{f \in \mathcal{K}} \bar{d}(E, f). \quad (21)$$

Since E and P in the above expressions are (extreme) points in $\mathbb{R}^{|\mathcal{X}|}$, their Euclidean distances can be found easily by calculating the Euclidean norms $\|P - E\|$. Particularly, calculating $\bar{d}(E, f)$ requires calculating the Euclidean distances between E and all extreme points of the face \mathcal{M}_f . Finally, the RHS expression in (21) is calculated by finding $\bar{d}(E, f)$ for all pairs of extreme points and gambles in \mathcal{K} .

4.3 Improved bounds

Equation (21) gives an upper bound for the difference between coherent lower previsions coinciding on a set of gambles, however, the estimate is systematically too conservative. This is caused by the fact that extreme points E can only maximize expression (16) for some h if $E(h) = \underline{E}(h)$. This means that the domain for h in (18) should be restricted to those gambles h that reach the lowest value $\underline{E}(h)$ in E . In other words, h should belong to the normal cone $N_{\mathcal{M}}(E)$.

Therefore, instead of taking the Euclidean distance between E and P in (20), we should take the following distance:

$$d_E(E, P) = \max_{h \in N_{\mathcal{M}}(E)} \frac{|P(h) - E(h)|}{\|h\|}, \quad (22)$$

which we call the *normed distance* between E and P .

The geometrical intuition behind replacing Euclidean distance with the above distance function is the following. Given a gamble h , the difference $P(h) - E(h)$ can be viewed as the inner product $(P - E) \cdot h$, which depends on the angle between $(P - E)$ and h . As the normal cone contains elements that are orthogonal to $P - E$ for adjacent extreme points P , we may expect that the other

elements are nearly orthogonal too, especially in the case of narrow normal cones. In Figure 3 such situation can be observed in the case of the normal cone of E_1 , in contrast to the case of E_5 , where the normal cone is wide. Therefore, we would, for instance, expect that the normed distances between E_1 and its adjacent extreme points would be significantly smaller than the Euclidean distance, in contrast the case of E_5 . Analytically we demonstrate this in Example 3.

In the sequel we represent the calculation of the normed distance in the form of a quadratic programming problem.

MINIMUM NORM ELEMENTS OF THE NORMAL CONE.

Consider an element h of the form (11). Given a pair of expectation functionals E and P , the distance $P(h) - E(h)$ does not depend on β . In order to maximize the normed distance (22), we must consider the representative with the minimum norm, as the norm appears in the denominator of the expression. The characterization of the minimal norm element of the form (11) follows.

Proposition 9 *Let h be a gamble. Then $\|h + \beta 1_{\mathcal{X}}\| \geq \|h\|$ for every $\beta \in \mathbb{R}$ if and only if $h \cdot 1_{\mathcal{X}} = 0$.*

Proof We have that $\|h + \beta 1_{\mathcal{X}}\|^2 = \|h\|^2 + \beta^2 + 2\beta h \cdot 1_{\mathcal{X}}$, which has minimum in $\beta = -h \cdot 1_{\mathcal{X}}$. Hence the minimizing β equals 0 exactly if $h \cdot 1_{\mathcal{X}}$ does. ■

Corollary 10 *Let E, h and I be as in Corollary 5 and let f'_i be the unique vectors such that $f_i - f'_i = c1_{\mathcal{X}}$ and $f'_i \cdot 1_{\mathcal{X}} = 0$ for every $i \in I$. Then, as follows from Corollary 5, there exist some $\alpha'_i \geq 0$ for every $i \in I$ and $\beta' \in \mathbb{R}$ so that*

$$h = \sum_{i \in I} \alpha'_i f'_i + \beta' 1_{\mathcal{X}}. \quad (23)$$

Moreover,

$$\left\| \sum_{i \in I} \alpha'_i f'_i \right\| \leq \left\| \sum_{i \in I} \alpha'_i f'_i + \beta' 1_{\mathcal{X}} \right\| \text{ for every } \beta \in \mathbb{R}. \quad (24)$$

Proof Since $f'_i \cdot 1_{\mathcal{X}} = 0$, we have that $(\sum_{i \in I} \alpha'_i f'_i) \cdot 1_{\mathcal{X}} = 0$, whence by Proposition 9 it follows that this is the minimal-norm gamble of the form (23). ■

Let I and f'_i , for $i \in I$, be as in Corollary 10 and let $\underline{\alpha}: I \rightarrow [0, \infty)$ be a map and $\beta \in \mathbb{R}$ a constant (we will write α_i instead of $\alpha(i)$). Then we define $h(\underline{\alpha}, \beta) = \sum_{i \in I} \alpha_i f'_i + \beta 1_{\mathcal{X}}$. Clearly, $h(\underline{\alpha}, \beta) \in N_{\mathcal{M}}(E)$ and every element of $N_{\mathcal{M}}(E)$ is of the form $h(\underline{\alpha}, \beta)$, by Corollary 5.

Corollary 11 *The following equality holds:*

$$\max_{(\underline{\alpha}, \beta)} \frac{|E(h(\underline{\alpha}, \beta)) - P(h(\underline{\alpha}, \beta))|}{\|h(\underline{\alpha}, \beta)\|} = \max_{\underline{\alpha}} \frac{|E(h(\underline{\alpha}, 0)) - P(h(\underline{\alpha}, 0))|}{\|h(\underline{\alpha}, 0)\|} \quad (25)$$

Proof Since $|E(h + \beta 1_{\mathcal{X}}) - P(h + \beta 1_{\mathcal{X}})| = |E(h) - P(h)|$, the maximum of the expression is achieved at h with the minimum norm, which is the one with $\beta = 0$. ■

THE CALCULATION OF THE NORMED DISTANCE BETWEEN EXPECTATION FUNCTIONALS.

Take two linear expectation functionals P and $E \in \mathcal{M}$ and let I and f'_i for $i \in I$ be as in Corollary 10. Our goal is to find the normed distance (22). The absolute value in the numerator of (22) can be omitted because $E(h) = \min_{P \in \mathcal{M}} P(h)$ for every $h \in N_{\mathcal{M}}(E)$. By Corollary 11, every $h \in N_{\mathcal{M}}(E)$ that can minimize the above expression is of the form $h(\underline{\alpha}, 0)$. Since E and P are themselves vectors too, we can denote $D = P - E$, and write $P(h) - E(h) = (P - E) \cdot h = D \cdot h$.

Now we can decompose every f'_i for $i \in I$ as $f'_i = \lambda_i D + u_i$, so that $D \cdot u_i = 0$. Given that $h = \sum_{i \in I} \alpha_i f'_i$, we obtain $h = (\underline{\alpha} \cdot \underline{\lambda}) D + \underline{\alpha} \cdot U$, where U is the matrix whose rows are u_i , $\underline{\lambda}$ is the column vector with components λ_i and the vectors f'_i are also written as row vectors. We also assume $\underline{\alpha}$ to be a column vector.

Further we have that $\|h\|^2 = h \cdot h = \|D\|^2 \underline{\alpha} \cdot \underline{\lambda} \underline{\lambda}^t \underline{\alpha}^t + \underline{\alpha} \cdot U U^t \underline{\alpha}^t$. Now denote $\Pi = \|D\|^2 \underline{\lambda} \underline{\lambda}^t + U U^t$ and write $\|h\|^2 = \underline{\alpha} \Pi \underline{\alpha}^t$. Clearly, Π is a symmetric and positive semi-definite matrix.

Moreover, we have that $P(h) - E(h) = D \cdot (\underline{\alpha} \cdot \underline{\lambda}) D = (\underline{\alpha} \cdot \underline{\lambda}) \|D\|^2$. Our goal is the maximization of expression (22). Thus we need to maximize

$$\varphi(\underline{\alpha}) = \frac{(\underline{\alpha} \cdot \underline{\lambda}) \|D\|^2}{\sqrt{\underline{\alpha} \Pi \underline{\alpha}^t}} \quad (26)$$

over the set of all I -vectors $\underline{\alpha}$ with non-negative components. Clearly, for every non negative constant k we have that $\varphi(k\underline{\alpha}) = \varphi(\underline{\alpha})$. Moreover, only those $\underline{\alpha}$ for which the numerator in $\varphi(\underline{\alpha})$ is positive are of interest, and then multiplying $\underline{\alpha}$ by a suitable positive constant can ensure that the numerator is 1. Maximizing $\varphi(\underline{\alpha})$ is then equivalent to minimizing the nominator, which yields the following quadratic programming problem:

Minimize:

$$\underline{\alpha} \Pi \underline{\alpha}^t \quad (27)$$

subject to

$$(\underline{\alpha} \cdot \underline{\lambda}) \|D\|^2 = 1 \quad (28)$$

$$\underline{\alpha} \geq 0 \quad (29)$$

Example 3 Consider the lower prevision \underline{P} from Example 1. We will calculate the distance $d_{E_1}(E_1, E_5)$, where $E_1 = (0.4, 0.32, 0.28)$ and $E_5 = (0.15, 0.37, 0.48)$. First we have:

$$D = E_5 - E_1 = (-0.2462, 0.0492, 0.1969),$$

and its norm, which is the Euclidean distance between the two extreme points is $\|D\| = 0.3191$. The positive basis of $N_{\mathcal{M}}(E_1)$ consists of the transformed gambles

$$\begin{aligned} f'_1 &= f_1 - f_1 \cdot 1_{\mathcal{X}} / 3 = (-0.5, 0.5, 0) \\ f'_5 &= f_5 - f_5 \cdot 1_{\mathcal{X}} / 3 = (-0.2, 0.6, -0.4). \end{aligned}$$

(see Corollary 10).

We have $f'_1 = 1.451D + (-0.1429, 0.4286, -0.2857)$, and since f'_5 is orthogonal to D , it follows that $u_5 = f'_5$ and $\lambda_2 = 0$. Thus $\underline{\lambda} = \begin{bmatrix} 1.451 \\ 0 \end{bmatrix}$ and $U = \begin{bmatrix} -0.14 & 0.43 & -0.29 \\ -0.20 & 0.60 & -0.40 \end{bmatrix}$ which gives $\Pi = \|D\|^2 \underline{\lambda} \underline{\lambda}^t + UU^t = \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.56 \end{bmatrix}$. Taking $\underline{\alpha} = (\alpha_1, \alpha_2)^t$, we obtain the objective function to be minimized: $\underline{\alpha} \Pi \underline{\alpha}^t = 0.5\alpha_1^2 + 0.8\alpha_1\alpha_2 + 0.56\alpha_2^2$ subject to $\|D\|^2 \underline{\alpha} \cdot \underline{\lambda} = \|D\|^2 \lambda_1 \alpha_1 = 1$ whence $\alpha_1 = 6.7708$. Substituting α_1 in the objective function we obtain $\underline{\alpha} \Pi \underline{\alpha}^t = 22.9219 + 5.41664\alpha_2 + 0.56\alpha_2^2$, which has to be minimized subject to $\alpha_2 \geq 0$. The minimum is obtained for $\alpha_2 = 0$, with the minimal value of objective function $\underline{\alpha} \Pi \underline{\alpha}^t$ equal to 22.9219. Now $d_{E_1}(E_1, E_5) = \varphi(\underline{\alpha}) = 1/\sqrt{22.9219} = 0.2089$. Note that this is significantly less than the Euclidean distance between the points, which is equal to $\|D\| = 0.3191$.

5. Conclusions and further work

This paper provides as its main contribution a practically computable upper bound for the difference between any two extensions of a coherent lower prevision given on an arbitrary finite set of gambles. The problem is relevant for many applications of the theory of imprecise probabilities, where complete description of lower previsions or their credal sets is often infeasible.

A drawback of the proposed method is that it requires finding all extreme points of credal sets in question. The number of the extreme points in general grows exponentially with the number of constraints, which makes the method computationally demanding.

In future, faster and perhaps less accurate methods could be developed to quickly asses maximal possible error of finite approximation of coherent lower previsions could be developed based on the results proposed in this paper. The method might also be simplified for special cases of approximations, such as coherent lower probabilities.

References

- A. Antonucci and F. Cuzzolin. Credal sets approximation by lower probabilities: application to credal networks. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 716–725. Springer, 2010.
- T. Augustin, F. P. Coolen, G. de Cooman, and M. C. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.
- P. Gruber. *Convex and Discrete Geometry*. Springer-Verlag Berlin Heidelberg, 2007. doi:[10.1007/978-3-540-71133-9](https://doi.org/10.1007/978-3-540-71133-9).
- E. Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628 – 658, 2008. ISSN 0888-613X. doi:[10.1016/j.ijar.2007.12.001](https://doi.org/10.1016/j.ijar.2007.12.001). In Memory of Philippe Smets (1938-2005).
- M. C. Troffaes and G. De Cooman. *Lower previsions*. John Wiley & Sons, 2014.
- D. Škulj and R. Hable. Coefficients of ergodicity for Markov chains with uncertain parameters. *Metrika*, 76(1):107–133, dec 2013. ISSN 0026-1335. doi:[10.1007/s00184-011-0378-0](https://doi.org/10.1007/s00184-011-0378-0).

New Distributions for Modeling Subjective Lower and Upper Probabilities

Michael Smithson

The Australian National University, Canberra (Australia)

MICHAEL.SMITHSON@ANU.EDU.AU

Abstract

This paper presents an investigation of approaches to modeling lower and upper subjective probabilities. A relatively unexplored approach is introduced, based on the fact that every cumulative distribution function (CDF) with support $(0,1)$ has a “dual” CDF that obeys the conjugacy relation between coherent lower and upper probabilities. A new 2-parameter family of “CDF-Quantile” distributions with support $(0,1)$ is extended via a third parameter for the purpose of modeling lower-upper probabilities. The extension exploits certain properties of the CDF-Quantile family, and the fact that continuous CDFs on $(0,1)$ random variables form an algebraic group that is closed under composition. This extension also yields models for testing specific models of lower-upper probability assignments. Finally, the new models are applied to a real data-set, and compared with the alternative approaches for their relative advantages and drawbacks.

Keywords: probability judgment; distribution; quantile regression; generalized linear model.

1. Introduction

This paper presents an investigation of approaches to modeling lower and upper subjective probabilities. This investigation springs from two motivational sources. First, it is motivated by the many applications in which interval-valued probability assignments play a role in human probability judgments, whether as input into decision making and forecasting or as risk communication (e.g., Budescu et al., 2014). Second, it is motivated by recent developments for modeling random variables on the $(0,1)$ interval, which have resulted in a new family of probability distributions with $(0,1)$ support, described by Smithson and Merkle (2014) and elaborated in Smithson and Shou (2017).

We begin with a brief description of conventional methods for modeling lower-upper probabilities, followed by the introduction of a heretofore unexplored modeling approach. Then the new family of distributions is introduced, and extended for the purpose of modeling lower-upper probabilities via the methods described previously. Finally, the models are applied to real data-sets, and compared for their relative advantages and drawbacks.

Conventional statistical approaches to modeling lower-upper probability assignments treat them as a pair of dependent random variables. One type of method ignores the ordering and simply models the dependency either via a “subject-effect” parameter or a covariance. A somewhat more sophisticated regression-style approach uses a binary dummy predictor that takes a value of 0 for the lower probabilities and 1 for the upper probabilities and respects the ordering by restricting the coefficient to being non-negative by exponentiating it (e.g., Smithson et al., 2012).

This paper introduces another approach to modeling lower-upper probabilities, in which the probability distributions modeling the lower and upper probability assignments share parameters but take two different forms. This pair of distributions is determined by the so-called “conjugacy” relation between coherent lower and upper probabilities. Let $p_L(A) = W(p(A), \theta)$, be a lower

probability with respect to probability $p(A)$ so that $0 \leq W(p(A), \theta) \leq p(A)$, for real-valued θ . The conjugate upper probability is $p_U(A) = 1 - p_L(\sim A)$, so that $p_U(A) = 1 - W(1 - p(A), \theta)$.

A version of this relationship may be identified in cumulative distribution functions (CDFs) for random variables on the (0,1) interval. Consider a CDF, $G(x, \theta)$, for $0 \leq x \leq 1$, with a location parameter, θ , so that $G(0, \theta) = 0$, $G(1, \theta) = 1$, and G is monotonically increasing in x . Define $G_D(x, \theta) = 1 - G(1 - x, \theta)$, which clearly also is a CDF. G_D is the *conjugate dual* of G , which follows by observing that

$$1 - G_D(1 - x, \theta) = 1 - [1 - G(1 - (1 - x), \theta)] = G(x, \theta) \quad (1)$$

As a simple example, consider $G(x, \theta) = x^\theta$, for $\theta > 0$. Then $G_D(x, \theta) = 1 - (1 - x)^\theta$. When $\theta < 1$ G is the upper CDF, when $\theta = 1$ we have the uniform distribution so that $G = G_D$, and when $\theta > 1$ G is the lower CDF.

A second example is the beta distribution. It is easy to show that if X is distributed $\text{beta}(\omega, \tau)$ then G_D is the CDF of a random variable, X_D , say, that is distributed $\text{beta}(\tau, \omega)$, i.e., the PDF of X flipped around 1/2. The absolute difference between their means, $|(\omega - \tau)/(\omega + \tau)|$, gives a convenient index of the distance between the lower and upper distributions. Reparameterizing the beta distribution so that the parameters are the mean, $\mu = \omega/(\omega + \tau)$, and precision, $\phi = \omega + \tau$, it is clear that the mean and precision of X jointly determine the magnitude of the difference between its distribution and that of and its conjugate dual X_D .

One- and two-parameter distributions of the kinds illustrated here have very limited flexibility regarding the location of G and G_D ; typically the corresponding PDFs are mirror-images of one another centred on 1/2. Nevertheless, while these pairs of distributions may not be very useful for modeling real data, the concepts involved turn out to have such applications when applied to the family of distributions introduced in the next section.

2. CDF-Quantile Distributions

The family of distributions presented here is elaborated in [Smithson and Shou \(2017\)](#) and [Shou and Smithson \(2016\)](#) implement them in the R package `cdfquantile` for generalized linear modeling. This family is a special case of the T-X family presented by [Aljarrah, et al. \(2014\)](#), although it was independently described in [Smithson and Merkle \(2014\)](#). Let $G(x, \mu, \sigma)$ denote a CDF for random variable X with support $(0, 1)$, a real-valued location parameter μ and positive scale parameter σ . We define G as follows:

$$G(x, \mu, \sigma) = F[U(H^{-1}(x), \mu, \sigma)] \quad (2)$$

where F is a CDF with support denoted by D_1 , H is an invertible CDF with support denoted by D_2 , and $U : D_2 \rightarrow D_1$ is an appropriate transform for incorporating parameters μ and σ . We limit the domains D_1 and D_2 to pairs taken from $(-\infty, \infty)$ and/or $(0, \infty)$, and the following cases of U .

For $D_1 = (-\infty, \infty)$ and $D_2 = (-\infty, \infty)$ we put

$$U(y, \mu, \sigma) = (y - \mu)/\sigma. \quad (3)$$

For $D_1 = (-\infty, \infty)$ and $D_2 = (0, \infty)$ we put

$$U(y, \mu, \sigma) = (\log(y) - \mu)/\sigma. \quad (4)$$

For $D_1 = (0, \infty)$ and $D_2 = (-\infty, \infty)$ we put

$$U(y, \mu, \sigma) = \exp(-\mu/\sigma) \exp(y/\sigma). \quad (5)$$

Finally, for $D_1 = (0, \infty)$ and $D_2 = (0, \infty)$ we put

$$U(y, \mu, \sigma) = \exp(-\mu/\sigma) y^{1/\sigma}. \quad (6)$$

If all the functions are differentiable then the PDF $g(x, \mu, \sigma)$ has an explicit expression. If F is invertible, then for every γ such that $G(x, \mu, \sigma) = \gamma$, the quantile functions corresponding to the cases described in equations (3) to (6) are as follows. For $D_1 = (-\infty, \infty)$ and $D_2 = (-\infty, \infty)$ we put

$$G^{-1}(\gamma, \mu, \sigma) = H[\sigma F^{-1}(\gamma) + \mu]. \quad (7)$$

For $D_1 = (-\infty, \infty)$ and $D_2 = (0, \infty)$ we put

$$G^{-1}(\gamma, \mu, \sigma) = H[\exp(\sigma F^{-1}(\gamma) + \mu)]. \quad (8)$$

For $D_1 = (0, \infty)$ and $D_2 = (-\infty, \infty)$ we put

$$G^{-1}(\gamma, \mu, \sigma) = H[\mu + \sigma \log(F^{-1}(\gamma))]. \quad (9)$$

Finally, for $D_1 = (0, \infty)$ and $D_2 = (0, \infty)$ we put

$$G^{-1}(\gamma, \mu, \sigma) = H\left[\exp(\mu)(F^{-1}(\gamma))^\sigma\right]. \quad (10)$$

Smithson and Shou (2017) present 36 members of the *CDF-Quantile* family by employing six standard distributions for F and H : The logistic, Cauchy, t with df = 2, arc-sinh, Burr VII, and Burr VIII distributions. All of these have explicit PDF, CDF, and quantile functions. Smithson and Shou observe that F and H may exchange roles. The resulting pairs of distributions are "quantile-duals" of one another in the sense that one's CDF is the other's quantile, with the appropriate parameterization. This duality is due to the fact that (0, 1) is both the domain and range of these functions. Smithson and Shou denote these distributions with the nomenclature F - H (e.g., Cauchit-Logistic and Logit-Cauchy).

Smithson and Shou (2017) show that the CDF-Quantile family members share the following properties:

1. The family can model a wide variety of distribution shapes, with different skew and kurtosis coverage from the beta or the Kumaraswamy.
2. (Proposition 1, from Smithson and Shou (2017)) Members are self-dual in the sense that $g(x, \mu, \sigma) = g(1 - x, -\mu, \sigma)$. Moreover, $G = G_D$, so the conjugate-CDF duals in this family consists of identical distributions.
3. (Proposition 2) The median is solely a function of μ , so that μ is genuinely a location parameter.
4. (Proposition 3) The parameter σ is a dispersion parameter.

5. (Proposition 4) Members of this family fall into four subfamilies distinguished by behavior at the boundaries of the $[0, 1]$ interval, including a subfamily whose density is finite in the limits at 0 and at 1.

Thus, the CDF-Quantile family enables a wide variety of quantile regression models for random variables on the $(0, 1)$ interval with predictors for both location and dispersion parameters, and simple interpretations of those parameters. Smithson and Shou demonstrate that members of the family can out-perform the beta and other two-parameter distributions in fitting real data. Because they have explicit CDFs and quantile functions, the CDF-Quantile family is well-suited for multivariate models using copulas, and an example of this application will be presented later in this paper. [Shou and Smithson \(2017\)](#) fit a trivariate copula model to real data as a demonstration of how this may be done using their `cdfquantreg` package in conjunction with the R package `copula`.

3. Introducing a Third Parameter to the CDF-Quantile Family

The fact that $G = G_D$ for the entire CDF-Quantile family implies that they may be well-suited to testing the conjugate-CDF model of lower and upper probabilities via the introduction of a third parameter. Unlike two-parameter distributions such as the beta distribution, for a three-parameter distribution the third parameter can determine the difference between a CDF and its conjugate dual CDF.

There are several ways to introduce a third parameter, but we will focus on doing so through a composition operator. Marshall and Olkin (2007, pp. 494-495) state that the class **G** of CDFs G whose support is $(0,1)$ form an algebraic group. This is true of continuous CDFs. The class of continuous CDFs is closed under the composition operation $G_1 \bullet G_2 = G_1(G_2)$, and this operation also is associative. The uniform distribution is the identity. Likewise, for any G in **G**, the quantile function G^{-1} also is in **G**. The quantile-dual relation described in the preceding section is a special case of this type of closure.

A straightforward way to introduce a third parameter is via an invertible monotonic function applied either at the outermost or innermost level of the CDF or the quantile function. Applying an invertible $(0, 1) \rightarrow (0, 1)$ transformation W to the innermost level of the CDF, for instance, we have

$$G(x, \mu, \sigma, \theta) = F[U(H^{-1}(W(x, \theta)), \mu, \sigma)] \quad (11)$$

and

$$G^{-1}(\gamma, \mu, \sigma, \theta) = W^{-1}[H(U^{-1}(F^{-1}(\gamma), \mu, \sigma)), \theta] \quad (12)$$

If we additionally require that $W(0, \theta) = 0$, $W(1, \theta) = 1$ and W monotonically increasing in x , then W behaves as a CDF. The conjugate dual CDF therefore is

$$G_D(x, \mu, \sigma, \theta) = F[U(H^{-1}(1 - W(1 - x, \theta)), \mu, \sigma)]. \quad (13)$$

Several kinds of CDFs for W and application of the CDF-composition operator are available from the literature on lifetime distributions. A power (resilience) parameter or a frailty parameter can be introduced in this way, by applying the CDF-composition operator. The relevant CDF is x^θ , for some $\theta > 0$. Slightly less obviously, introducing a tilt parameter also involves a CDF-composition, because, for $\theta > 0$, it is a composition of the CDF $x/(x + \theta(1 - x))$ with $G(x, \mu, \sigma)$. Likewise, a hazard parameter can be introduced via composition using the CDF

$1 - \exp \left[-(-\log(1-x))^{\theta} \right]$, for $\theta > 0$; and a Laplace transform parameter with the CDF $(1 - e^{-\theta x}) / (1 - e^{-\theta})$, for real θ .

In the cases where the composition is $G \bullet W$, the introduction of the third parameter yields a three-parameter CDF-Quantile family with distinct CDFs and conjugate dual CDFs (i.e., $G \neq G_D$) and possessing certain properties paralleling those derived by [Smithson and Shou \(2017\)](#) for the two-parameter family. The following Proposition is an extension of Proposition 1 (the self-dual property) from [Smithson and Shou \(2017\)](#).

Proposition 5.1: Let $W(x, \theta)$ be defined as earlier, so that it behaves as a CDF. Let

$$G(W(x, \theta), \mu, \sigma) = F[U(H^{-1}(W(x, \theta)), \mu, \sigma)].$$

Then if the CDFs F and H satisfy certain symmetry conditions (in the 4 cases detailed below),

$$1 - G(W(1-x, \theta), -\mu, \sigma) = G(1 - W(1-x, \theta), \mu, \sigma). \quad (14)$$

Now define

$$G^{-1}(Z_1(\gamma, \mu, \sigma), \theta) = W^{-1}[H(U^{-1}(F^{-1}(\gamma), \mu, \sigma)), \theta],$$

and

$$G^{-1}(Z_2(\gamma, \mu, \sigma), \theta) = 1 - W^{-1}[1 - H(U^{-1}(F^{-1}(\gamma), \mu, \sigma)), \theta].$$

These are the quantile functions corresponding to the conjugate dual CDFs $G(W(x, \theta), \mu, \sigma)$ and $G(1 - W(1-x, \theta), \mu, \sigma)$, respectively. Then $G^{-1}(Z_1(\gamma, \mu, \sigma), \theta)$ and $G^{-1}(Z_2(\gamma, \mu, \sigma), \theta)$ behave as conjugate lower-upper probabilities.

Proof: The identity in equation (14) has four cases, corresponding to the four combinations of domains in the CDF-Quantile family.

Case 1: For $D_1 = (-\infty, \infty)$ and $D_2 = (-\infty, \infty)$ when $-H^{-1}(x) = H^{-1}(1-x)$ and $f(x) = f(-x)$, $1 - G(W(1-x, \theta), -\mu, \sigma, \theta) = 1 - F[(H^{-1}(W(1-x, \theta)) + \mu)/\sigma]$
 $= 1 - F[(-H^{-1}(1 - W(1-x, \theta)) + \mu)/\sigma] = F[(H^{-1}(1 - W(1-x, \theta)) - \mu)/\sigma]$
 $= G(1 - W(1-x, \theta), \mu, \sigma, \theta).$

Case 2: For $D_1 = (-\infty, \infty)$ and $D_2 = (0, \infty)$ when $H^{-1}(x) = 1/H^{-1}(1-x)$ and $f(x) = f(-x)$, $1 - G(W^{-1}(1-x, \theta), -\mu, \sigma, \theta) = 1 - F[(\log(H^{-1}(W(1-x, \theta))) + \mu)/\sigma]$
 $= 1 - F[(-\log(H^{-1}(1 - W(1-x, \theta))) + \mu)/\sigma] = F[(\log(H^{-1}(1 - W(1-x, \theta))) - \mu)/\sigma]$
 $= G(1 - W(1-x, \theta), \mu, \sigma, \theta).$

Case 3: For $D_1 = (0, \infty)$ and $D_2 = (-\infty, \infty)$ when $H^{-1}(x) = 1/H^{-1}(1-x)$ and $F(x) = 1 - F(1/x)$, $1 - G(1-x, -\mu, \sigma) = 1 - F[(H^{-1}(W(1-x, \theta)) \exp(\mu))^{1/\sigma}]$
 $= 1 - F[(H^{-1}(1 - W(1-x, \theta)))^\sigma (\exp(\mu))^{1/\sigma}] = F[(H^{-1}(1 - W(1-x, \theta)) \exp(-\mu))^{1/\sigma}]$
 $= G(1 - W(1-x, \theta), \mu, \sigma, \theta).$

Case 4. For $D_1 = (0, \infty)$ and $D_2 = (0, \infty)$ when $-H^{-1}(x) = H^{-1}(1-x)$ and $F(x) = 1 - F(1/x)$, $1 - G(1-x, -\mu, \sigma) = 1 - F[\exp((-H^{-1}(W(1-x, \theta)) + \mu)/\sigma)]$
 $= 1 - F[\exp((-H^{-1}(1 - W(1-x, \theta)) + \mu)/\sigma)] = F[\exp((H^{-1}(1 - W(1-x, \theta)) - \mu)/\sigma)]$
 $= G(1 - W(1-x, \theta), \mu, \sigma, \theta).$

The conjugacy relationship immediately follows immediately by observing that, in the definition of

the quantile functions, $H(U^{-1}(F^{-1}(\gamma), \mu, \sigma))$ fulfills the role of x in the function W^{-1} . *End of proof.*

The conjugate dual CDFs straddle the CDF $G(x, \mu, \sigma)$ and the resultant lower and upper quantile functions straddle the quantile function $G^{-1}(\gamma, \mu, \sigma)$. That is, the location of the conjugate-dual pair is determined by μ , which makes them flexible enough to be worthy candidates for modeling real data. Propositions 2-4 in [Smithson and Shou \(2017\)](#) also hold for these three-parameter CDF-Quantile distributions because W is monotonically increasing in x and we can write the quantile function as $W^{-1}[H(U^{-1}(F^{-1}(\gamma), \mu, \sigma)), \theta]$. Thus, the median is solely a function of μ and θ , and σ still is a dispersion parameter. Moreover, the θ parameter has an interpretation as a risk-attitude parameter, because it determines the difference between the lower and upper CDFs (and likewise the difference between the corresponding quantile functions). This three-parameter family therefore is suited to ascertaining whether samples of lower and upper probability assignments behave as though they come from populations with conjugate dual distributions.

4. Examples and Applications

4.1 $G \bullet W$ Conjugate Duals

In this subsection we will survey two examples of three-parameter CDF-Quantile distributions of the $G \bullet W$ type, each one corresponding to a well-known kind of parameterization borrowed from the life distributions literature. These include the power parameter (which in this case corresponds to a frailty parameter) and the tilt parameter. The Cauchit-Cauchy distribution will be used throughout this subsection for illustrative purposes (it also is employed in the data-fitting example in the next subsection).

Starting with the power parameter, $W(x, \theta) = x^\theta$ and so $1 - W(1 - x, \theta) = 1 - (1 - x)^\theta$. Applied to the Cauchit-Cauchy distribution, we have the conjugate CDF duals As its name suggests, both F and H are Cauchy CDFs, the power parameter (exponentiated) model simply replaces x with x^θ , and the conjugate-dual CDF pair is

$$G(x, \mu, \sigma) = \frac{1}{2} + \frac{\arctan((\tan((2\pi x^\theta - \pi)/2) - \mu)/\sigma)}{\pi} \quad (15)$$

and

$$G_D(x, \mu, \sigma) = \frac{1}{2} + \frac{\arctan((\tan((2\pi(1 - (1 - x)^\theta) - \pi)/2) - \mu)/\sigma)}{\pi} \quad (16)$$

When $\theta < 1$ then $G > G_D$, and when $\theta > 1$ then $G < G_D$.

The tilt parameter, as mentioned earlier, uses the CDF $W(x, \theta) = x/(x + \theta(1 - x))$. Applying it to the Cauchit-Cauchy distribution yields the conjugate CDF duals

$$G(x, \mu, \sigma) = \frac{1}{2} + \frac{\arctan((\tan((2\pi x/(x + \theta(1 - x)) - \pi)/2) - \mu)/\sigma)}{\pi} \quad (17)$$

and

$$G_D(x, \mu, \sigma) = \frac{1}{2} + \frac{\arctan((\tan((2\pi\theta x/(1 + x(\theta - 1)) - \pi)/2) - \mu)/\sigma)}{\pi} \quad (18)$$

This model behaves as a rescaled version of the constant-odds-ratio imprecise probability model described in [Walley \(1991\)](#) and elsewhere. When $\theta < 1$ then $G > G_D$, and when $\theta > 1$ then

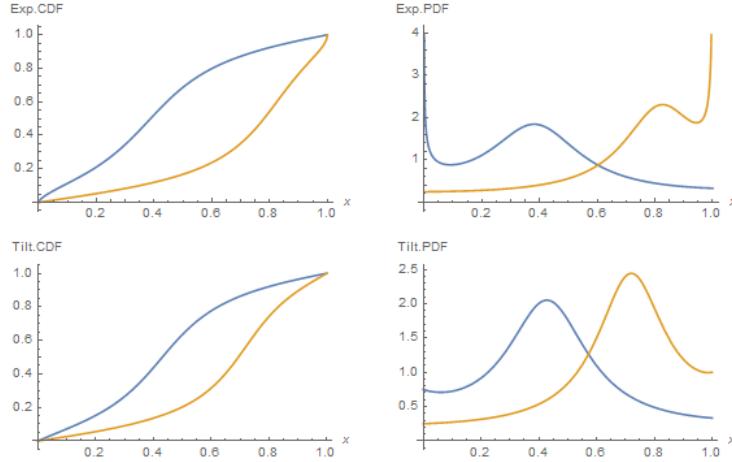


Figure 1: Power- and Tilt-Parameter Conjugate Dual Distributions

$G < G_D$. Figure 1 displays the pairs of CDFs and PDFs for the exponentiated and tilt parameter models when $\mu = 0.1$, $\sigma = 0.5$, and $\theta = 1.5$.

Finally, it is worth mentioning that because any CDF whose support is $(0,1)$ can play the role of W , a one-parameter version of any member of the CDF-Quantile family may be used in that capacity, with θ as the location parameter. These alternatives would seem to present a forbiddingly large variety of models for analysts to consider. However, it turns out that under some conditions all of them can be very similar to one another with appropriate choices of θ . For many practical modeling purposes we may restrict attention to a subset of such models, such as the power and tilt parameter (constant odds-ratio) models, but at this stage of research on these models the best procedure for selecting among them remains an open topic for further investigation. The next section presents examples of model-fitting with a real data-set, demonstrating that conjugate dual lower-upper CDF models can fit lower-upper probability assignments quite well.

4.2 Fitting Models to Data

We now present an example of model-fitting that compares the conjugate lower-upper distributions with appropriate alternatives for modeling lower-upper probability assignments. The fourth Intergovernmental Panel on Climate Change (IPCC) report utilizes verbal phrases such as “likely” and “unlikely” to describe the uncertainties in climate science. [Budescu et al. \(2009\)](#) conducted an experimental study of lay interpretations of these phrases, using 13 sentences from the IPCC report, in which they asked 223 participants to provide lower, “best”, and upper numerical estimates of the probabilities to which they believed each sentence referred. For example, participants were presented with the sentence “The Greenland ice sheet and other Arctic ice fields likely contributed no more than 4 m of the observed sea level rise.”, and asked to consider the probability they thought the report authors may have had in mind for the term “likely” in this sentence. Participants were required to provide their lowest, highest, and their best numerical estimates of this probability. Budescu et al. found that participants’ “best” estimates were more regressive (toward the middle of the $[0, 1]$ interval) than the IPCC stipulations, but they did not report systematic analyses of the lower and upper estimates.

I present 11 models fitted to the lower and upper probability estimates in the Budescu et al. data. The first three models are based on the two-parameter CDF-Quantile distribution. Model 1 is just the two-parameter distribution, as defined in equation (3), with intercept-only submodels $\hat{\mu} = \beta_0$ and $\hat{\sigma} = \exp(\delta_0)$. Model 2 has conditional parameter estimates, with submodels $\hat{\mu} = \beta_0 + \beta_1 x$ and $\hat{\sigma} = \exp(\delta_0 + \delta_1 x)$, where $x = 0$ for lower probabilities and $x = 1$ for upper probabilities. Model 3, in addition to the submodels from Model 2, also estimates the dependency between the lower and upper estimates via a t-copula with CDF-Quantile margins. This model therefore also includes estimates of the t-copula dependency parameter, ρ , and degrees of freedom parameter, ϕ .

Models 4-7 are based on the 3-parameter power (exponentiated) CDF-Quantile distribution, as in the CDF defined in equation (11) with $W(x, \theta) = x^\theta$. Model 4 has intercept-only submodels $\hat{\mu} = \beta_0$, $\hat{\sigma} = \exp(\delta_0)$, and $\hat{\theta} = \exp(\gamma_0)$. Model 5 is the conjugate-dual model, as defined in equations (11) and (13). This has the same intercept-only submodels as Model 4 but is a two-component distribution mixture model with a fixed mixture parameter, so that the first CDF, G , is weighted 1 and the second, G_D , is weighted 0 for the upper probabilities and the reverse weighting is applied to the lower probabilities. Technically, it is a four-parameter model although the mixture parameter is not being estimated. Model 6 has conditional parameter estimates, $\hat{\mu} = \beta_0 + \beta_1 x$ and $\hat{\sigma} = \exp(\delta_0 + \delta_1 x)$ with $x = 0$ and 1 for lower and upper probabilities, but an intercept-only submodel $\hat{\theta} = \exp(\gamma_0)$. Model 7 has the conditional μ and σ submodels in Model 6 plus $\hat{\theta} = \exp(\gamma_0 + \gamma_1 x)$. Finally, models 8-11 are based on the tilt-parameter CDF-Quantile distribution, as in the CDF defined in equation (11) with $W(x, \theta) = x/(x + \theta(1 - x))$. These models have the same variants as Models 4-7.

The best-fitting models from the CDF-Quantile family are from the “finite-tailed” subfamily, whose members have defined, finite densities at 0 and 1 (Smithson and Shou, 2017). The best-fitting distribution from this subfamily is the Cauchit-Cauchy, so the models considered here are mainly limited to that distribution. Table 1 displays goodness-of-fit statistics for the 11 models. The top section of the table presents these results for the three models using the two-parameter Cauchit-Cauchy. The middle section contains the power-parameter (exponentiated) models, and the lower section contains the tilted-parameter models. The “Params” column displays the number of parameters in each model, the “2LL” column shows twice the log-likelihood of the fitted models, and the “AIC” column is the Akaike Information Criterion, $AIC = -2LL + 2p$, where p is the number of parameters in the Params column.

Remarkably, the 4-parameter conjugate-dual models fit the data better than most of the 5- and 6-parameter conditional models and better than the 6-parameter copula model. The conjugate-dual power-parameter model is superior to the conjugate-dual tilted-parameter model, and is outperformed only by the 6-parameter conditional tilted-parameter model. Likewise, the conjugate-dual tilted-parameter model is out-performed only by the 5- and 6-parameter conditional tilted-parameter models and the 6-parameter conditional power-parameter model.

These results are not due to some kind of fluke in the Cauchit-Cauchy distribution. Other members of the finite-tailed subfamily have similar fits for their conjugate-dual models. For instance, the T2-T2 and the Cauchit-ArcSinh conjugate-dual power-parameter models have AIC’s of -2159 and -2062, respectively, and both of these out-perform their respective 5- and 6-parameter conditional power-parameter counterparts.

Figure 2 shows the fitted distributions from the conjugate-dual model (top half of the figure) and the 6-parameter conditional exponentiated model. The two pairs of fitted distributions are strikingly similar and the conjugate-dual AIC is the better of the two. The facts that the 4-parameter

Table 1: Cauchit-Cauchy Models and Fits

Model	Description	Params.	2LL	AIC
1	2-parameter	2	595	-591
2	2-parameter condit. μ, σ	4	1378	-1370
3	2-parameter condit. t-copula	6	1584	-1572
4	exponentiated 3-param.	3	616	-609
5	conjugate-dual exponentiated	4	2378	-2372
6	exponentiated condit. μ, σ	5	1392	-1382
7	exponentiated condit. μ, σ, θ	6	1967	-1955
8	tilted 3-param.	3	880	-874
9	conjugate-dual tilted	4	1736	-1730
10	tilted condit. μ, σ	5	2152	-2142
11	tilted condit. μ, σ, θ	6	3118	-3106

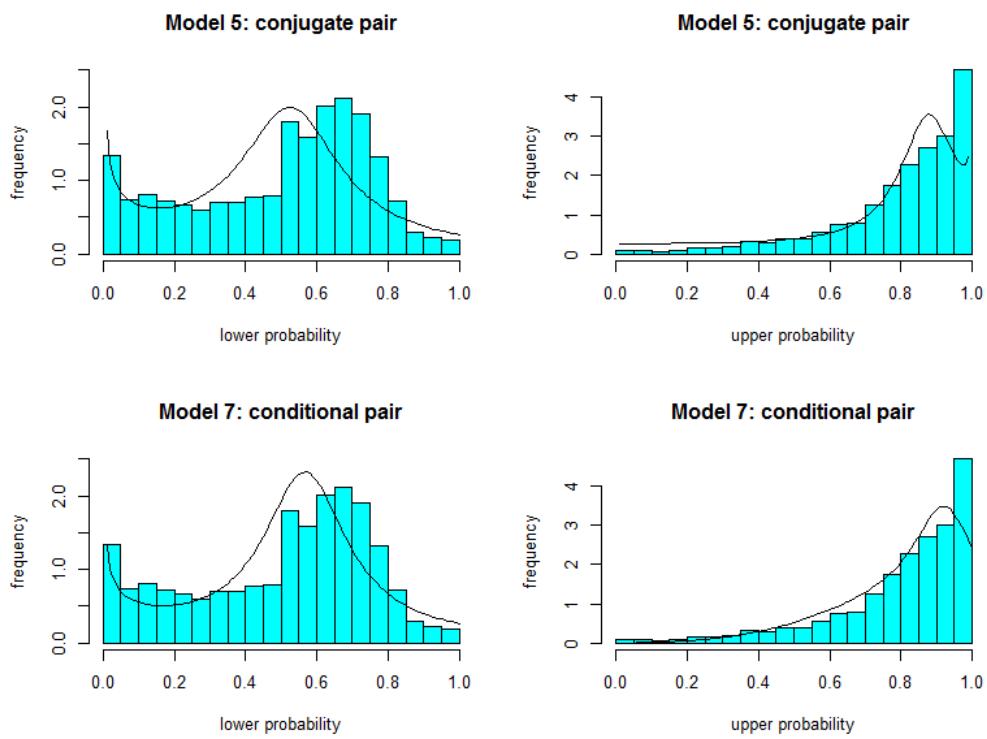


Figure 2: IPCC Data and Fitted Distributions

conjugate-dual model fits the data better than a regression model with 6 parameters and that the fitted distribution shapes are reasonably similar to the empirical distributions lend plausibility to the seemingly unlikely hypothesis that human lower-upper probability judgments are distributed approximately as conjugate-dual distributions.

The exponentiated Cauchit-Cauchy 4-parameter conjugate-dual and 6-parameter conditional regression models may be compared further via the 5-number summaries in Table 2. When compared with their empirical counterparts (rows 1 and 4 in the table), the conditional model is more accurate than the conjugate-dual model at the 10th quantile, but the reverse is the case for most of the other quantiles. Both models appear to be fairly accurate in the middle 50% of the distributions. Again, this is an intriguing outcome for the conjugate-dual model, given that only three of its four parameters are being estimated from the data.

Table 2: Quantiles and Exponentiated Model Quantile Estimates

Model	Estimate	.1	.25	.5	.75	.9
	empirical lower	0.092	0.301	0.570	0.699	0.779
5	conjugate-dual lower	0.059	0.303	0.535	0.688	0.825
7	conditional lower	0.091	0.378	0.584	0.713	0.834
	empirical upper	0.540	0.729	0.858	0.948	0.998
5	conjugate-dual upper	0.298	0.684	0.863	0.935	0.977
7	conditional upper	0.495	0.672	0.846	0.935	0.975

That said, there are practical and technical issues in estimating both conjugate-dual and regression models for the 3-parameter CDF-Quantile distributions. For several of these distributions, maximum-likelihood estimations of conjugate-dual models of the IPCC data failed to converge, and regression models yielded high correlations between the parameter estimates for μ and θ (although the latter problem did not occur for any of the successful conjugate-dual models). Moreover, as [Smithson and Shou \(2017\)](#) observe, model diagnostics and related aspects of model evaluation for the 2-parameter CDF-Quantile family have yet to be completely thought through. Thus, the questions of effective estimation procedures and diagnostics for these models are active topics of research. Nonetheless, the evidence from the example in this section suggests that a sufficiently well-specified conjugate-dual model using 3-parameter CDF-Quantile distributions can be used to test a specific type of coherent lower-upper probability relationship.

5. Conclusions and Future Directions

A new family of probability distributions, the CDF-Quantile family, shows promise in modeling probability judgments. The two-parameter version of the family has been sufficiently well-explored by [Smithson and Shou \(2017\)](#) to have been made available for generalized linear modeling via the `cdfquantreg` package in R and a SAS macro, as presented by Shou and Smithson ([2016, 2017](#)), and those authors also have demonstrated that these distributions can model probabilities better than other two-parameter distributions such as the beta. This paper has presented an investigation of the application of the CDF-Quantile family to modeling imprecise distributions of probabilities, by extending it to incorporate a third parameter.

Because CDFs whose support is the (0,1) interval are closed under composition, and due to the properties of the CDF-Quantile distributions, three-parameter extensions via the composition of CDF functions yield conjugate dual pairs of CDFs. This result may hold some theoretical interest. A future line of research may elaborate the connections between these conjugate duals and imprecise probability frameworks. There is a natural link with probability boxes (p-boxes, as coined by [Ferson](#)

et al. (2003)), given that the conjugate-dual CDFs form a p-box. Conjugate duals are noteworthy cases of p-boxes because the “width” of the gap between them is determined in a different way from the data-driven methods to which Ferson et al. (2003) refer. To my awareness, p-boxes have not been systematically studied regarding methods of fitting them to lower-upper probability data.

Some conjugate-dual models, in turn, have been found to fit a data-set reasonably well, raising the possibility that human lower-upper probability assignments may approximate a conjugacy relationship in their CDFs. Further research will determine whether these findings generalize to other such data-sets, if elicitation methods influence the results, and what judgment mechanisms or heuristics account for the phenomenon. However, perhaps the first priority is to ascertain the connections between the θ parameter, measurement error, and sampling error.

Finally, the three-parameter CDF-Quantile distributions also beg for further investigation. The overview in this paper only skims their characteristics, and little is known about the advantages and drawbacks of alternative parameterization methods for θ (e.g., power versus tilt parameters). Preliminary investigations suggest that the high correlations between parameter estimates may be a pervasive problem for three-parameter distributions on the unit interval (including three-parameter generalizations of the beta distribution). Likewise, as mentioned earlier, much remains to be developed and explored regarding parameter estimation methods and model diagnostics, even for the two-parameter CDF-Quantile family. The primary goals here have been to introduce this extension of the CDF-Quantile family and to make a case that it holds some promise for modeling distributions of lower-upper probability assignments. Accordingly, this paper may be regarded as a preliminary exploration of three-parameter CDF-Quantile distributions, with the unexpected finding that conjugate-dual distributions may be useful for modeling lower-upper probability assignments.

References

- M. A. Aljarrah, C. Lee, and F. Famoye. On generating t-x family of distributions using quantile functions. *Journal of Statistical Distributions and Applications*, 1(1):2, 2014.
- D. V. Budescu, S. Broomell, and H.-H. Por. Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, 20(3):299–308, 2009.
- D. V. Budescu, H.-H. Por, S. B. Broomell, and M. Smithson. The interpretation of ipcc probabilistic statements around the world. *Nature Climate Change*, 4(6):508–512, 2014.
- S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz. Constructing probability boxes and dempster-shafer structures. *Sandia National Laboratories*, pages 143–180, 2003.
- A. W. Marshall and I. Olkin. *Life Distributions*. Springer, 2007.
- Y. Shou and M. Smithson. *cdfquantreg: Quantile Regression for Random Variables on the Unit Interval*, 2016. R package version 1.1.0.
- Y. Shou and M. Smithson. *cdfquantreg: An R package for CDF-Quantile Regression*, 2017. The Australian National University, Canberra, Australia.
- M. Smithson and E. C. Merkle. *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. Chapman and Hall/CRC Press, 2014.

M. Smithson and Y. Shou. Cdf-quantile distributions for modeling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology*, 2017. doi:[10.1111/bmsp.12091](https://doi.org/10.1111/bmsp.12091).

M. Smithson, D. V. Budescu, S. B. Broomell, and H.-H. Por. Never say not: Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning*, 53(8):1262–1270, 2012.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Linear Core-Based Criterion for Testing Extreme Exact Games

Milan Studený

Václav Kratochvíl

*Institute of Information Theory and Automation of the CAS
CZ-18208, Prague, Pod Vodárenskou věží 4 (Czech Republic)*

STUDENY@UTIA.CAS.CZ

VELOREX@UTIA.CAS.CZ

Abstract

The notion of a (discrete) coherent lower probability corresponds to a game-theoretical concept of an exact (cooperative) game. The collection of (standardized) exact games forms a pointed polyhedral cone and the paper is devoted to the extreme rays of that cone, known as extreme exact games. A criterion is introduced for testing whether an exact game is extreme. The criterion leads to solving simple linear equation systems determined by (the vertices of) the core polytope (of the game), which concept corresponds to the notion of an induced credal set in the context of imprecise probabilities. The criterion extends and modifies a former necessary and sufficient condition for the extremity of a supermodular game, which concept corresponds to the notion of a 2-monotone lower probability. The linear condition we give in this paper is shown to be necessary for an exact game to be extreme. We also know that the condition is sufficient for the extremity of an exact game in an important special case. The criterion has been implemented on a computer and we have made a few observations on basis of our computational experiments.

Keywords: extreme exact game; coherent lower probability; core; credal set; supermodular game; 2-monotone lower probability; min-representation; oxytrophic game.

1. Introduction

The notion of a *coherent lower probability* and that of an induced *credal set* (of discrete probability distributions) are traditional topics of interest in the theory of imprecise probabilities. These notions correspond to game-theoretical concepts of an *exact game* and its *core* (polytope), widely used in the context of cooperative coalition games. The analogy is even broader: a lower probability avoiding sure loss corresponds to a weaker concept of a balanced game while a *2-monotone lower probability* (= capacity) corresponds to a stronger concept of a *supermodular game*, named also a convex game.

The discrete case is considered here: the sample space (= frame of discernment) for distributions is a fixed finite set N having at least two elements. The elements of N correspond to players in the context of cooperative game theory and to random variables in yet another context of probabilistic conditional independence structures. The collection of coherent lower probabilities on N , where $n = |N|$, is a polytope in a 2^n -dimensional real vector space, while the set of non-negative exact games is a pointed polyhedral cone whose extreme rays are generated just by extreme points of that polytope. This paper offers a method to test whether a ray is extreme in the cone of exact games, which implicitly gives a method to test extreme coherent lower probabilities.

Some effort to develop criteria to recognize the extremity of an exact game was exerted earlier by Rosenmüller (2000, § 4 of chapter 5) in his book on game theory. He offered one necessary and one sufficient condition for the extremity based on a *min-representation* of the exact game; however, these conditions have a limited scope because they are applicable only in quite special situations. Nevertheless, in this paper we follow the idea of min-representation and propose a more

general criterion based on the list of vertices of the respective *core*, which provides a *standard min-representation* of any exact game. Our condition is always necessary for the extremity of an exact game and we conjecture it is also sufficient, which is the case in a certain special case.

Being motivated by questions raised by [Maass \(2003\)](#), [Quaeghebeur and de Cooman \(2008\)](#) became interested in *extreme lower probabilities* and computed these in the case of small $n = |N|$. [Antonucci and Cuzzolin \(2010\)](#) considered an enlarging transformation of a credal set with a finite number of extreme points, when the respective (coherent) lower probability is computed and then a larger credal set is induced by the lower probability. Note that their second step, namely representing a coherent lower probability by the vertices of the induced credal set, corresponds to our standard min-representation of an exact game.

It is always useful to be aware of the correspondence between concepts from different areas. For example, [Wallner \(2005\)](#) confirmed a conjecture by Weichselberger that the credal set induced by a (coherent) lower probability has at most $n!$ vertices. However, the same result was achieved already by [Derks and Kuipers \(2002\)](#) in the context of cooperative game theory. They also made an interesting observation that whenever a core of an exact game has $n!$ vertices then it has the maximal number of $2^n - 2$ facets and gave an example of a game in the relative interior of the exact cone whose respective core does not have the maximal number of $n!$ vertices.

The criterion we offer here is a modification of the criterion from [\(Studený and Kroupa, 2016\)](#), where a necessary and sufficient condition was provided for a supermodular game being extreme in the cone of (standardized) supermodular games. That result was motivated by the research on conditional independence structures ([Studený, 2005](#)), in which context extreme supermodular games encode submaximal structural conditional independence models. The supermodular criterion leads to solving a simple linear equation system determined by certain combinatorial structure (of the core), which concept was pinpointed earlier by [Kuipers et al. \(2010\)](#). The difference here is that testing the extremity in the supermodular cone leads to one linear equation system, while testing the extremity in the exact cone may require solving several such equation systems.

What is an added value of this contribution is that we have also implemented both criteria and provide a web platform for testing the extremity of a supermodular/exact game in the respective cone for reasonably limited number of players. Of course, this can also be used to test the extremity of coherent lower probabilities. However, we have intentionally chosen to deal with games because this approach allows one to utilize the profits of integer arithmetics implementation.

In our paper we assume that the reader is familiar with basic concepts in polyhedral geometry, namely a polytope (= bounded polyhedron) and its faces/facets/vertices. The structure of the paper is as follows. In the next section ([§ 2](#)) we recall basic concepts and facts. In [§ 3](#) the concept of a min-representation of an exact game and the question of its uniqueness are discussed. After that our criterion is formulated ([§ 4](#)). In Conclusions ([§ 5](#)) we give a few remarks based on our computational experiments. The Appendix contains some proofs.

2. Notation, basic definitions and facts

Let N be a finite non-empty set of *variables*, $|N| \geq 2$, and $\mathcal{P}(N) := \{S : S \subseteq N\}$ its power set. The symbol \mathbb{R}^N will denote the set of real vectors whose components are indexed by elements of N . Analogously, $\mathbb{R}^{\mathcal{P}(N)}$ is the collection of real functions on $\mathcal{P}(N)$ (= vectors with components indexed by subsets of N). Given $S \subseteq N$, the vector $\chi_S \in \mathbb{R}^N$ will denote the zero-one indicator of S . Given $v, x \in \mathbb{R}^N$, their scalar product will be $\langle v, x \rangle := \sum_{i \in N} v_i \cdot x_i$.

2.1 Game-theoretic concepts

By a (cooperative) *game* we will understand a set function $m \in \mathbb{R}^{\mathcal{P}(N)}$ with $m(\emptyset) = 0$.

Definition 1 (core, exact game, supermodular game)

Let $m : \mathcal{P}(N) \rightarrow \mathbb{R}$, $m(\emptyset) = 0$, be a game. Its *core* is a polytope in \mathbb{R}^N defined by

$$C(m) := \{x \in \mathbb{R}^N : \sum_{i \in N} x_i = m(N) \text{ & } \forall S \subseteq N \quad \sum_{i \in S} x_i \geq m(S)\}.$$

The symbol $\text{ext } C(m)$ will be used to denote the set of extreme points (= vertices) of $C(m)$. A game m is *balanced* if $C(m) \neq \emptyset$. A balanced game is called *exact* if

$$\forall S \subseteq N \quad \exists x \in C(m) \quad \sum_{i \in S} x_i = m(S).$$

A game m is *supermodular* if it satisfies the supermodularity inequalities

$$\forall C, D \subseteq N \quad m(C) + m(D) \leq m(C \cup D) + m(C \cap D).$$

A game m is called ℓ -standardized (ℓ stands for “lower”; in game theory = zero-normalized) if $m(S) = 0$ for any $S \subseteq N$, $|S| \leq 1$. Denote the class of exact ℓ -standardized games by $E_\ell(N)$.

A well-known fact is that any supermodular game, named traditionally *convex* in game theory, is exact (Csóka et al., 2011, §4). A non-negative exact game m normalized by $m(N) = 1$ is nothing but a coherent lower probability; see (Walley, 1991, Corollary 3.3.4).

The fact that, for any $S \subseteq N$, $\{x \in C(m) : \sum_{i \in S} x_i = m(S)\}$ is a face of $C(m)$ allows one to observe that any exact game m satisfies a formally stronger condition

$$\forall S \subseteq N \quad \exists x \in \text{ext } C(m) \quad \sum_{i \in S} x_i = m(S). \quad (1)$$

Indeed, every face of a polytope is the convex hull of extreme points of the whole polytope contained in the face. A necessary condition for the exactness of a game m is that it is *superadditive*:

$$\forall A, B \subseteq N \quad A \cap B = \emptyset \quad m(A) + m(B) \leq m(A \cup B).$$

Indeed, given disjoint $A, B \subseteq N$ there exists $x \in C(m)$ with $m(A \cup B) = \sum_{i \in A} x_i + \sum_{i \in B} x_i$ and one has both $m(A) \leq \sum_{i \in A} x_i$ and $m(B) \leq \sum_{i \in B} x_i$. In particular, any ℓ -standardized exact game is non-decreasing with respect to inclusion and non-negative.

It can be derived from results in (Csóka et al., 2011, §3) that the collection of exact games is a rational polyhedral cone. Thus, non-negative exact games on $\mathcal{P}(N)$ form a pointed rational cone and the same holds for $E_\ell(N)$. Degenerate non-negative exact games are superset indicators for singletons in N , which correspond to crisp degenerate probabilities in the context of imprecise probabilities. Since any non-negative exact game can be written as the sum of an ℓ -standardized exact game and of a conic combination of these degenerate exact games the question of testing the extremity in the cone of non-negative exact games reduces to testing the extremity in $E_\ell(N)$.

Definition 2 (extreme exact game)

An ℓ -standardized exact game $m : \mathcal{P}(N) \rightarrow \mathbb{R}$ is *extreme* if it generates an extreme ray of $E_\ell(N)$.

It can be derived from the fact that $E_\ell(N)$ is a rational cone that any *extreme* ℓ -standardized exact game is a multiple of an integer-valued function $m : \mathcal{P}(N) \rightarrow \mathbb{Z}$. In particular, when testing the extremity of an exact game one can limit oneself to integer-valued functions.

3. The concept of a min-representation

A useful property of an exact game is that it can be represented as the minimum of a finite collection of additive games. Specifically, every $x \in \mathbb{R}^N$ defines an additive game

$$x \in \mathbb{R}^{\mathcal{P}(N)} \text{ by the formula } x(S) := \sum_{i \in S} x_i \quad \text{for any } S \subseteq N,$$

and every exact game can be obtained as the set-wise minimum of a finite collection of such additive games. This leads to the following concept.

Definition 3 (regular min-representation)

We say that $m \in \mathbb{R}^{\mathcal{P}(N)}$ has a *min-representation* (by additive functions) if a non-empty finite set $\mathcal{R} \subseteq \mathbb{R}^N$ exists such that

$$\forall S \subseteq N \quad m(S) = \min_{x \in \mathcal{R}} \sum_{i \in S} x_i. \quad (2)$$

Every $x \in \mathcal{R}$ is then assigned the corresponding *tightness class* of sets

$$\mathcal{T}_x^m := \{S \subseteq N : m(S) = \sum_{i \in S} x_i\}. \quad (3)$$

We say that a min-representation $\mathcal{R} \subseteq \mathbb{R}^N$ of a game m is *regular* if, for any $x \in \mathcal{R}$,

- (i) $\sum_{i \in N} x_i = m(N)$, and
- (ii) the linear hull of $\{\chi_S : S \in \mathcal{T}_x^m\} \subseteq \mathbb{R}^N$ is whole \mathbb{R}^N .

Note that an equivalent formulation of the regularity condition (ii) is that the only vector in \mathbb{R}^N which is orthogonal to all vectors from $\{\chi_S : S \in \mathcal{T}_x^m\}$ is the zero vector. There exists at least one regular min-representation for every exact game.

Proposition 4 (min-representations of exact games)

A game $m \in \mathbb{R}^{\mathcal{P}(N)}$ is exact iff it admits a min-representation \mathcal{R} satisfying (i) for any $x \in \mathcal{R}$. Every exact game has a regular min-representation given by the list of vertices of its core: $\overline{\mathcal{R}} = \text{ext } C(m)$. A min-representation $\mathcal{R} \subseteq \mathbb{R}^N$ of an exact game m is regular iff $\mathcal{R} \subseteq \text{ext } C(m)$.

The proof of Proposition 4 is shifted to Appendix, § A.1. In particular, any exact game m has the largest regular min-representation which we consider to be a kind of *standard min-representation* of m . Note that a simple example of a non-exact game exists which has a min-representation.

3.1 On uniqueness of regular min-representations

In general, one can have several regular min-representations of an exact game. On the other hand, sometimes only one regular min-representation exists, which happens iff the next condition holds.

Definition 5 (oxytrophic game)

We say that an exact game $m : \mathcal{P}(N) \rightarrow \mathbb{R}$ is *oxytrophic* if $\forall x \in \text{ext } C(m)$

$$\exists S \subseteq N \text{ with } \sum_{i \in S} x_i = m(S) \text{ such that } \forall y \in \text{ext } C(m), y \neq x \quad m(S) < \sum_{i \in S} y_i. \quad (4)$$

This relevant mathematical concept has already appeared in the literature and we have simply taken over the terminology by Rosenmüller (2000, § 3 of chapter 5). The following gives an example of an oxytrophic game, which is extreme in $E_\ell(N)$.

Example 1 Put $N = \{a, b, c, d\}$ and consider $\mathcal{R} \subseteq \mathbb{R}^N$ consisting of 4 vectors (x_a, x_b, x_c, x_d) , namely $(1, 1, 1, 1), (2, 2, 0, 0), (2, 0, 2, 0), (0, 2, 2, 0)$. Then the formula (2) gives

$$m(abcd) = 4, m(abc) = 3, m(abd) = m(acd) = m(bcd) = m(ab) = m(ac) = m(bc) = 2,$$

and $m(S) = 0$ for other $S \subseteq N$. One can verify by computation that $\mathcal{R} = \text{ext } C(m)$, which allows one to check the condition (4) for any $x \in \text{ext } C(m)$: $(1, 1, 1, 1)$ has one respective set $S = abc$, while $(2, 2, 0, 0)$ has even two respective sets $S = c$ and $S = cd$, etc. In particular, m is oxytrophic. Moreover, m is also an example of an (extreme) exact game which is not supermodular: $m(ac) + m(bc) = 4 > 3 = m(abc) + m(c)$.

An interesting observation is that in case $|N| = 3$ the ℓ -standardized oxytrophic games are just the zero game and extreme exact games. However, in case $|N| = 4$ an extreme exact game exists which is not oxytrophic. The next example is even a supermodular game.

Example 2 Put $N = \{a, b, c, d\}$ and introduce $m(abcd) = 2, m(abc) = m(abd) = m(acd) = 1$, and $m(S) = 0$ for other $S \subseteq N$. Then the core $C(m)$ has seven vertices (x_a, x_b, x_c, x_d) , namely four substantial ones denoted by

$$\mathcal{R} : \quad (2, 0, 0, 0), \quad (0, 1, 1, 0), \quad (0, 1, 0, 1), \quad (0, 0, 1, 1),$$

and three additional ones, namely

$$(1, 1, 0, 0), \quad (1, 0, 1, 0), \quad (1, 0, 0, 1).$$

The vectors in \mathcal{R} satisfy (4): $S = bcd$ for $(2, 0, 0, 0)$, $S = ad$ for $(0, 1, 1, 0)$, $S = ac$ for $(0, 1, 0, 1)$ and $S = ab$ for $(0, 0, 1, 1)$. However, the remaining 3 vertices of $C(m)$ do not satisfy (4) and m is not oxytrophic. On the other hand, every regular min-representation involves \mathcal{R} and vectors in \mathcal{R} provide a min-representation of m . Thus, \mathcal{R} is the least regular min-representation of m .

On the other hand, an exact game can have several inclusion-minimal regular min-representations (see later Example 5). The following example shows that an oxytrophic game need not be extreme.

Example 3 Put $N = \{a, b, c, d\}$ and $m(abcd) = 2, m(S) = 1$ for $S \subseteq N, |S| = 3$, while $m(ab) = m(cd) = 1$, and $m(S) = 0$ for other $S \subseteq N$. Then $\mathcal{R} = \text{ext } C(m)$ has four vectors (x_a, x_b, x_c, x_d) , namely $(0, 1, 0, 1), (0, 1, 1, 0), (1, 0, 0, 1)$ and $(1, 0, 1, 0)$. The vectors in \mathcal{R} satisfy (4): $S = ac$ for $(0, 1, 0, 1)$, $S = ad$ for $(0, 1, 1, 0)$, $S = bc$ for $(1, 0, 0, 1)$ and $S = bd$ for $(1, 0, 1, 0)$. Thus, m is oxytrophic. On the other hand, m is the sum of two other supermodular games m^1 and m^2 , where m^1 is the indicator of supersets of ab and m^2 is the indicator of supersets of cd .

4. The criterion: a conjecture and results

Assume now that $m \in E_\ell(N)$ is an ℓ -standardized exact game. Then the core $C(m)$ consists of non-negative vectors and the same holds for its vertices: $\text{ext } C(m) \subseteq [0, \infty)^N$. In this section we formulate our linear core-based criterion.

4.1 Some arrangement

To formalize our conjecture let us choose and fix an auxiliary index set Υ for the vertices of the core of m and imagine (= have) the vertex set $\text{ext } C(m)$ arranged in the form of a real array $x \in \mathbb{R}^{\Upsilon \times N}$ whose rows are indexed by Υ and columns by N :

$$x := [x(\tau, i)]_{\tau \in \Upsilon, i \in N} \in \mathbb{R}^{\Upsilon \times N} \quad \text{where } \text{ext } C(m) = \{[x(\tau, i)]_{i \in N} : \tau \in \Upsilon\}.$$

Recall that the minimization formula (2) for $\mathcal{R} = \text{ext } C(m)$ means that m is obtained by set-wise minimization in the array x over its rows:

$$\forall S \subseteq N \quad m(S) = \min_{\tau \in \Upsilon} \sum_{i \in S} x(\tau, i).$$

In this context, the tightness classes (3) correspond to elements of Υ :

$$\mathcal{T}_\tau := \{S \subseteq N : m(S) = \sum_{i \in S} x(\tau, i)\} \quad \text{for any } \tau \in \Upsilon.$$

For computational and implementation reasons, it is advisable to consider a special big zero-one *tightness array* encoding all tightness classes. This indicator array ι has rows indexed by Υ and columns by subsets of N :

$$\iota := [\iota(\tau, S)]_{\tau \in \Upsilon, S \subseteq N} \in \{0, 1\}^{\Upsilon \times \mathcal{P}(N)} \quad \text{where } \iota(\tau, S) = \begin{cases} 1 & \text{if } m(S) = \sum_{i \in S} x(\tau, i), \\ 0 & \text{otherwise.} \end{cases}$$

Note that ι serves as computer encoding of the concept of a combinatorial *core structure* mentioned in (Studený and Kroupa, 2016). By Proposition 4, the concept of a *regular min-representation* of m corresponds in this context to a special subset of the set of rows, namely $\Gamma \subseteq \Upsilon$ satisfying

$$\forall S \subseteq N \quad m(S) = \min_{\tau \in \Gamma} \sum_{i \in S} x(\tau, i). \quad (5)$$

To test whether $\Gamma \subseteq \Upsilon$ satisfies (5) one can consider the restricted tightness array $\iota_\Gamma \in \{0, 1\}^{\Gamma \times \mathcal{P}(N)}$ to rows in Γ and check whether each column in ι_Γ contains least one 1. Thus, a computer can be used to find *all inclusion-minimal* regular min-representations of m on basis of ι .

4.2 The linear equation systems

Every regular min-representation $\Gamma \subseteq \Upsilon$ satisfying (5) can be ascribed a system of linear constraints on the respective sub-array specified by rows in Γ :

$$y_\Gamma = [y(\tau, i)]_{\tau \in \Gamma, i \in N} \in \mathbb{R}^{\Gamma \times N}.$$

Specifically, the constraints are as follows:

- (a) $\forall \tau \in \Gamma \ \forall i \in N$ with $\{i\} \in \mathcal{T}_\tau \quad y(\tau, i) = 0,$
- (b) $\forall S \subseteq N, |S| \geq 2, \ \forall \tau, \rho \in \Gamma \text{ with } S \in \mathcal{T}_\tau \cap \mathcal{T}_\rho \quad \sum_{i \in S} y(\tau, i) = \sum_{i \in S} y(\rho, i).$

It is not difficult to observe that the starting restricted array $x_\Gamma \in \mathbb{R}^{\Gamma \times N}$ satisfies the constraints (a)-(b); this is because these constraints are determined by x_Γ through ι_Γ . Informally, the conjecture is that the extremity means that, for any min-representation Γ , the equation system (a)-(b) has unique solution up to a real multiple.

Conjecture 6 An ℓ -standardized exact game $m \in E_\ell(N)$ is *extreme* in $E_\ell(N)$ iff, for every $\Gamma \subseteq \Upsilon$ satisfying (5), every real solution $y_\Gamma \in \mathbb{R}^{\Gamma \times N}$ to (a)-(b) is a multiple of $x_\Gamma \in \mathbb{R}^{\Gamma \times N}$, that is,

$$\exists \beta \in \mathbb{R} \quad \forall \tau \in \Gamma \quad \forall i \in N \quad y(\tau, i) = \beta \cdot x(\tau, i).$$

The constraints (a)-(b) for fixed $\Gamma \subseteq \Upsilon$ can be written in the form of a matrix equality

$$\mathbf{C}_\Gamma \cdot y_\Gamma = \mathbf{0}, \text{ where } \mathbf{C}_\Gamma \text{ is an appropriate } \textit{constraint matrix} \text{ with entries in } \{-1, 0, +1\}.$$

The rows of \mathbf{C}_Γ encode the constraints and its columns correspond to the elements of $\Gamma \times N$. The matrix is sparse: every constraint of type (a) is encoded by a row with one non-zero component while any constraint of type (b) for $S \subseteq N$, $|S| \geq 2$, is encoded by a row containing $|S|$ -times a component $+1$, $|S|$ -times a component -1 and 0 otherwise.

The number of rows can be economized because some of the constraints of type (b) follow from the others. For example, whenever $S \subseteq N$, $|S| \geq 2$, belongs to $\mathcal{T}_\tau \cap \mathcal{T}_\rho \cap \mathcal{T}_\sigma$ for different $\tau, \rho, \sigma \in \Gamma$ then only two constraints

$$\sum_{i \in S} y(\tau, i) - \sum_{i \in S} y(\rho, i) = 0 \quad \text{and} \quad \sum_{i \in S} y(\tau, i) - \sum_{i \in S} y(\sigma, i) = 0$$

are enough. Therefore, if $\lambda(S)$, for $S \subseteq N$, denotes the number of 1's in the respective column of the tightness array ι_Γ , then the economized number of rows in \mathbf{C}_Γ is

$$\sum_{S \subseteq N: |S|=1} \lambda(S) + \sum_{S \subseteq N: |S| \geq 2} [\lambda(S) - 1].$$

Testing of the condition from Conjecture 6 for fixed $\Gamma \subseteq \Upsilon$ can be realized by computing the *nullity* of the matrix \mathbf{C}_Γ , which is the dimension of the space of solutions y_Γ to $\mathbf{C}_\Gamma \cdot y_\Gamma = \mathbf{0}$. Any solution to (a)-(b) is a multiple of x_Γ iff the nullity is 1; otherwise the nullity exceeds 1.

The following observation is useful to avoid testing all regular min-representations.

Proposition 7 Given an ℓ -standardized exact game $m \in E_\ell(N)$ assume the situation from § 4.1 and take $\Gamma \subseteq \Upsilon$ satisfying (5). If every real solution $y_\Gamma \in \mathbb{R}^{\Gamma \times N}$ to (a)-(b) is a multiple of $x_\Gamma \in \mathbb{R}^{\Gamma \times N}$ then the same holds in case of any Σ such that $\Gamma \subseteq \Sigma \subseteq \Upsilon$.

The proof of Proposition 7 is shifted to Appendix, § A.2. The consequence of this observation is that to test the condition from Conjecture 6 it is enough to consider only the inclusion-minimal regular min-representations; this simplification may spare the computational time. What we have actually shown in the proof of Proposition 7 is that

$$\text{whenever } \Gamma \subseteq \Upsilon \text{ satisfies (5) and } \Gamma \subseteq \Sigma \subseteq \Upsilon \text{ then } \text{null}(\mathbf{C}_\Gamma) \geq \text{null}(\mathbf{C}_\Sigma) \geq 1,$$

meaning that the nullity (of constraint matrices) achieves its maximal value at one of the inclusion-minimal regular min-representations; see later Example 4 for illustration.

4.3 Our theoretical results

Proposition 8 Given a non-zero game $m \in E_\ell(N)$, the condition from Conjecture 6, that is,

$$\forall \Gamma \subseteq \Upsilon \text{ satisfying (5), every real solution } y_\Gamma \in \mathbb{R}^{\Gamma \times N} \text{ to (a)-(b) is a multiple of } x_\Gamma \in \mathbb{R}^{\Gamma \times N},$$

is necessary for m being extreme in $E_\ell(N)$.

The proof of Proposition 8 is shifted to Appendix, § A.3. Another comment is that, in case m is a supermodular function, a necessary and sufficient condition for its extremity in the supermodular cone is that the condition from Conjecture 6 holds for the largest set $\Gamma = \Upsilon$ (Studený and Kroupa, 2016). The relation is illustrated by the next example.

Example 4 Put $N = \{a, b, c, d\}$, $m(abcd) = 4$, $m(S) = 2$ for $S \subseteq N$ with $|S| = 3$, $m(S) = 1$ for any $S \subseteq N$ with $|S| = 2$ except $m(cd) = 0$ and $m(S) = 0$ for remaining $S \subseteq N$. One can easily verify that m is a supermodular game. The core $\bar{\mathcal{R}} = \text{ext } C(m)$ consists of 13 vertices. To confirm that m is extreme in the supermodular cone one can use our method: the nullity of the respective constraint matrix \mathbf{C}_Υ is 1. However, this is not the only $\Gamma \subseteq \Upsilon$ with $\text{null } (\mathbf{C}_\Gamma) = 1$; there exists $\Sigma \subset \Upsilon$ with $|\Sigma| = 9$ such that $\text{null } (\mathbf{C}_\Gamma) = 1$ iff $\Sigma \subseteq \Gamma \subseteq \Upsilon$.

On the other hand, m is not an extreme exact game for it can be written as the sum of the game

$$m^1(abcd) = 2, \quad m^1(S) = 1 \text{ for } S \subseteq N, |S| = 3, \text{ and } \quad m^1(ac) = m^1(bc) = m^1(bd) = 1$$

(vanishing otherwise) and the game

$$m^2(abcd) = 2, \quad m^2(S) = 1 \text{ for } S \subseteq N, |S| = 3, \text{ and } \quad m^2(ab) = m^2(ad) = 1$$

(vanishing otherwise). Both m^1 and m^2 appear to be extreme in $E_\ell(N)$. The core $C(m^1)$ has three vertices (x_a, x_b, x_c, x_d) , namely $(1, 1, 0, 0)$, $(0, 1, 1, 0)$ and $(0, 0, 1, 1)$, while $C(m^2)$ has four of them: $(1, 1, 0, 0)$, $(1, 0, 1, 0)$, $(1, 0, 0, 1)$ and $(0, 1, 0, 1)$. The set $\mathcal{R}^* := \text{ext } [C(m^1) \oplus C(m^2)]$, where $X \oplus Y := \{x + y : x \in X \& y \in Y\}$ denoted the Minkowski sum of sets $X, Y \subseteq \mathbb{R}^N$, defines a regular min-representation of m . The respective index set $\Gamma^* \subseteq \Upsilon$ has 10 elements and one can construct two different solutions $y_{\Gamma^*} \in \mathbb{R}^{\Gamma^* \times N}$ to (a)-(b) on the basis of the standard min-representations of m^1 and m^2 . Nevertheless, one even has $\text{null } (\mathbf{C}_{\Gamma^*}) = 4$ in this case.

However, Γ^* does not provide the least regular min-representation of m . We found using a computer 27 inclusion-minimal regular min-representations of 10 permutation types; 8 of them have only four vectors (3 types) and 19 of them have five vectors (7 types). The nullities for the above mentioned inclusion-minimal min-representations are 6 and 7, which is the maximal nullity.

We also achieved the following partial converse result, whose proof we skip due to lack of space.

Proposition 9 Given non-zero $m \in E_\ell(N)$ such that the least $\mathcal{R} \subseteq \text{ext } C(m)$ satisfying (2) exists, the condition from Conjecture 6 is sufficient for m being extreme in $E_\ell(N)$.

The idea of the proof of Proposition 9 is that different solutions to (a)-(b) are constructed on the basis of standard min-representations of potential summands of m . Note that the condition from Proposition 9 involves the special case of oxytrophic $m \in E_\ell(N)$. On the other hand, an extreme exact game exists not having the least regular min-representation as the following example shows.

Example 5 Put $N = \{a, b, c, d\}$ and define $m(abcd) = 3$, $m(abc) = m(abd) = m(ab) = 2$, $m(acd) = m(bcd) = m(ac) = m(bc) = 1$ with $m(S) = 0$ for remaining for $S \subseteq N$. Then the set $\overline{\mathcal{R}} = \text{ext } C(m)$ has 5 vectors (x_a, x_b, x_c, x_d) . Three of them satisfy the oxytropy condition (4):

$$\mathcal{R} : (2, 0, 1, 0), \quad (0, 2, 1, 0), \quad (1, 1, 0, 1),$$

and two of them not: $(2, 1, 0, 0)$ and $(1, 2, 0, 0)$. Adding of any of two other vectors to \mathcal{R} turns it into an inclusion-minimal regular min-representation.

5. Conclusions

We have prepared a web platform for testing the extremity of an ℓ -standardized integer-valued exact game, available at

<http://gogo.utia.cas.cz:3838/exact-and-supermodular/>.

It also allows one to test the extremity of supermodular games in the supermodular cone.

We have tested our criterion on 41 permutation types of 398 extreme ℓ -standardized exact games over 4 variables; these were also earlier listed by Quaeghebeur and de Cooman (2008). What we have found out is that 20 of these types are oxytrophic; one of them is mentioned in Example 1. The remaining types are not, but for 19 of these the least min-representation exists; one of them is mentioned in Example 2. We also found 2 types of extreme exact games for which two inclusion-minimal regular min-representations exist; one of these 2 types is given in Example 5.

In all 41 cases the necessary condition from Proposition 8 is valid: the nullities of the respective constraint matrices are 1. Proposition 9 is applicable in great majority of 39 cases, when the least regular min-representation exists. Thus, in these 39 cases our linear criterion allows one to confirm the extremity. However, in the remaining 2 cases one cannot apply Proposition 9 to confirm the extremity and an open question is whether our condition from Conjecture 6 is sufficient for the extremity of an exact game in general.

Acknowledgments

We thank our colleague Tomáš Kroupa for re-computing extreme exact games over 4 variables. This research has been supported by the grant project of GAČR, number 16-12010S.

Appendix A. Proofs

A.1 Proof of Proposition 4

If m has a min-representation \mathcal{R} satisfying (i) for any $x \in \mathcal{R}$ then (2) implies $\emptyset \neq \mathcal{R} \subseteq C(m)$ and the condition of exactness for m is evident. Conversely, given an exact game m we put $\overline{\mathcal{R}} = \text{ext } C(m)$ and use (1) to observe that (2) holds with $\overline{\mathcal{R}}$ in place of \mathcal{R} . The regularity condition (i) for $\overline{\mathcal{R}}$ is evident. To verify (ii) consider a fixed $x \in \overline{\mathcal{R}} = \text{ext } C(m)$ and realize that the vectors in $\mathcal{V} := \{\chi_S \in \mathbb{R}^N : S \in \mathcal{T}_x^m\} \cup \{-\chi_N\}$ belong to the (inner) normal cone of (the least face of $C(m)$ containing) the vector x defined by

$$N_x := \{v \in \mathbb{R}^N : \forall y \in C(m) \langle v, y \rangle \geq \langle v, x \rangle\} \equiv \{v \in \mathbb{R}^N : \langle v, x \rangle = \min_{y \in C(m)} \langle v, y \rangle\};$$

indeed, for $v = \chi_S$, $S \in \mathcal{T}_x^m$, one has $\langle v, y \rangle = \sum_{i \in S} y_i \geq m(S) = \sum_{i \in S} x_i = \langle v, x \rangle$ for any $y \in C(m)$. The cone N_x is the conic hull of \mathcal{V} , which observation can be derived from Farkas's lemma: if $t \in \mathbb{R}^N$ is not in the conic hull of \mathcal{V} then $w \in \mathbb{R}^N$ exists such that $\langle v, w \rangle \geq 0$ for any $v \in \mathcal{V}$ while $\langle t, w \rangle < 0$. The former condition allows one to show that $y^\varepsilon := x + \varepsilon \cdot w$ belongs to $C(m)$ for some small $0 < \varepsilon$. The latter one implies that $\langle t, y^\varepsilon \rangle - \langle t, x \rangle = \langle t, y^\varepsilon - x \rangle = \varepsilon \cdot \langle t, w \rangle < 0$, implying that $t \notin N_x$.

The next observation is that, for any $x \in C(m)$, x is a vertex of $C(m)$ iff its normal cone N_x is full-dimensional. This result holds for any polytope $P \subseteq \mathbb{R}^N$ in place of $C(m)$. To see why this is the case the reader is advised to consult (Ziegler, 1995, § 7.1) for basic facts about the collection of normal cones for a polytope P , named the *normal fan* of the polytope. It is easy to realize that the lattice of normal cones is anti-isomorphic to the face-lattice of P . Specifically, the latter means, for $x, y \in P$, that one has $N_y \subseteq N_x$ iff $F[y] \supseteq F[x]$, where $F[x]$ denotes the least face of P containing $x \in P$. To this end realize that, for any $v \in N_x$ and $z \in P$, $\langle v, x \rangle = \langle v, z \rangle$ iff $v \in N_z$, which allows one to observe $F[x] := \bigcap_{v \in N_x} \{z \in P : \langle v, x \rangle = \langle v, z \rangle\} = \{z \in P : N_x \subseteq N_z\}$. Hence,

$$x \text{ is a vertex of } P \Leftrightarrow F[x] = \{x\} \Leftrightarrow N_x \text{ is a maximal cone} \Leftrightarrow N_x \text{ has the dimension } |N|.$$

By the former observation, the linear hull of N_x is the linear hull of $\{\chi_S : S \in \mathcal{T}_x^m\}$, which implies the condition (ii) for $x \in \overline{\mathcal{R}}$.

Thus, it follows from above arguments that any min-representation $\mathcal{R} \subseteq \text{ext } C(m)$ is regular. Conversely, given a regular min-representation \mathcal{R} of m , its elements belong to the core of m and the second regularity condition (ii) for $x \in \mathcal{R}$ implies that the respective normal cone N_x is full-dimensional, which happens only in case x is a vertex of $C(m)$.

A.2 Proof of Proposition 7

In case $\Gamma \subseteq \Sigma \subseteq \Upsilon$, it is evident that whenever $y_\Sigma \in \mathbb{R}^{\Sigma \times N}$ satisfies (a)-(b) with Σ then its restriction $y_\Gamma \in \mathbb{R}^{\Gamma \times N}$ to $\Gamma \times N$ satisfies (a)-(b) with Γ . The restriction mapping $y_\Sigma \mapsto y_\Gamma$ is linear and we show that it is one-to-one (under the assumptions from § 4.1). Thus, we assume that $y_\Sigma^1, y_\Sigma^2 \in \mathbb{R}^{\Sigma \times N}$ are two solutions to (a)-(b) with Σ such that their restrictions to $\mathbb{R}^{\Gamma \times N}$ coincide, that is $y_\Gamma^1 = y_\Gamma^2$, and we are going to show $y_\Sigma^1 = y_\Sigma^2$.

Consider a fixed $\tau \in \Sigma \setminus \Gamma$ and our goal is to verify that $y^1(\tau, i) = y^2(\tau, i)$ for any $i \in N$. To this end, we show that, for any $S \in \mathcal{T}_\tau$ one has $\sum_{i \in S} y^1(\tau, i) = \sum_{i \in S} y^2(\tau, i)$ and then apply the fact that the vectors $\{\chi_S : S \in \mathcal{T}_\tau\}$ linearly generate \mathbb{R}^N (see Definition 3 and Proposition 4). In case $S \in \mathcal{T}_\tau$, $S = \{i\}$, use (a) for Σ to observe $y^1(\tau, i) = 0 = y^2(\tau, i)$. In case $S \in \mathcal{T}_\tau$, $|S| \geq 2$, use the assumption that $\Gamma \subseteq \Upsilon$ satisfies (5) and find $\rho \in \Gamma$ such that $S \in \mathcal{T}_\rho$. The constraints (b) with Σ then imply

$$\sum_{i \in S} y^1(\tau, i) \stackrel{(b)}{=} \sum_{i \in S} y^1(\rho, i) = \sum_{i \in S} y^2(\rho, i) \stackrel{(b)}{=} \sum_{i \in S} y^2(\tau, i),$$

which gives what is desired. Thus, if every solution to (a)-(b) with Γ is a multiple of x_Γ then every solution to (a)-(b) with Σ must be a multiple of x_Σ .

A.3 Proof of Proposition 8

To verify the necessity of the condition it is enough to show that its negation implies that m is a convex combination of $m^1, m^2 \in \mathsf{E}_\ell(N)$ none of which is a multiple of m .

For this purpose assume, under the situation described in § 4.1, that there exists $\Gamma \subseteq \Upsilon$ satisfying (5) such that the equation system (a)-(b) has a solution $y \in \mathbb{R}^{\Gamma \times N}$ which is not a multiple of $x_\Gamma \in \mathbb{R}^{\Gamma \times N}$. Note that the facts $m \neq 0$ and (5) imply that $x_\Gamma \neq \mathbf{0}$.

The first observation is that, for any $y \in \mathbb{R}^{\Gamma \times N}$ satisfying (a)-(b), an ℓ -standardized game $t \in \mathbb{R}^{\mathcal{P}(N)}$ exists such that

$$\forall \gamma \in \Gamma \quad \forall S \in \mathcal{T}_\gamma \quad t(S) = \sum_{i \in S} y(\gamma, i). \quad (6)$$

To this end realize that (a)-(b) for y together imply the next *consistency condition*

$$\forall S \subseteq N \quad \forall \tau, \rho \in \Gamma \text{ with } S \in \mathcal{T}_\tau \cap \mathcal{T}_\rho \quad \sum_{i \in S} y(\tau, i) = \sum_{i \in S} y(\rho, i).$$

Since (5) implies $\mathcal{P}(N) = \bigcup_{\gamma \in \Gamma} \mathcal{T}_\gamma$, one can correctly define t using (6). This game t is uniquely determined through (6); moreover, the function $y \in \mathbb{R}^{\Gamma \times N} \mapsto t \in \mathbb{R}^{\mathcal{P}(N)}$ is linear by definition. Finally, the fact that m is ℓ -standardized together with the condition (a) for y imply that t must be ℓ -standardized, too.

Consider the line L in $\mathbb{R}^{\Gamma \times N}$ passing through y and x_Γ , namely the vectors

$$z_\varepsilon := (1 - \varepsilon) \cdot x_\Gamma + \varepsilon \cdot y \quad \text{for any } \varepsilon \in \mathbb{R}.$$

Observe that L does not contain the zero vector in $\mathbb{R}^{\Gamma \times N}$ as otherwise y is a multiple of x_Γ . As vectors in L satisfy (a)-(b), ℓ -standardized games q_ε , $\varepsilon \in \mathbb{R}$, exist such that

$$\forall \varepsilon \in \mathbb{R} \quad \forall \gamma \in \Gamma \text{ with } S \in \mathcal{T}_\gamma \quad \sum_{i \in S} z_\varepsilon(\gamma, i) = q_\varepsilon(S).$$

The next step is to show that, for sufficiently small $|\varepsilon|$, one has

$$\forall \gamma \in \Gamma \text{ with } S \notin \mathcal{T}_\gamma \quad \sum_{i \in S} z_\varepsilon(\gamma, i) > q_\varepsilon(S), \quad (7)$$

which implies, for those small $|\varepsilon|$, that

$$q_\varepsilon(S) = \min_{\gamma \in \Gamma} \sum_{i \in S} z_\varepsilon(\gamma, i) \quad \text{for any } S \subseteq N.$$

This implies that $z_\varepsilon(\gamma, *) \in \mathbb{R}^N$, $\gamma \in \Gamma$, belong to the core $C(q_\varepsilon)$; in particular, $q_\varepsilon \in E_\ell(N)$.

To ensure (7) for small $|\varepsilon|$, consider a fixed $\gamma \in \Gamma$ and $S \subseteq N$, $S \notin \mathcal{T}_\gamma$, and choose $\pi \in \Gamma$ such that $S \in \mathcal{T}_\pi$, by (5). The definitions of \mathcal{T}_γ and \mathcal{T}_π then imply

$$0 < \sum_{i \in S} x_\Gamma(\gamma, i) - m(S) = \sum_{i \in S} x_\Gamma(\gamma, i) - \sum_{i \in S} x_\Gamma(\pi, i).$$

This allows one to write

$$\begin{aligned} \sum_{i \in S} z_\varepsilon(\gamma, i) - q_\varepsilon(S) &= \sum_{i \in S} z_\varepsilon(\gamma, i) - \sum_{i \in S} z_\varepsilon(\pi, i) \\ &= (1 - \varepsilon) \cdot \underbrace{\left(\sum_{i \in S} x_\Gamma(\gamma, i) - \sum_{i \in S} x_\Gamma(\pi, i) \right)}_{>0} + \varepsilon \cdot \left(\sum_{i \in S} y(\gamma, i) - \sum_{i \in S} y(\pi, i) \right), \end{aligned}$$

and observe that the limit of this expression with ε tending to zero is positive. Therefore, after considering all pairs (γ, S) , $\gamma \in \Gamma$, $S \notin \mathcal{T}_\gamma$, (7) is ensured for sufficiently small $|\varepsilon|$.

Thus, there exists $0 < \varepsilon$ such that both $r := (1 - \varepsilon) \cdot m + \varepsilon \cdot t$ and $s := (1 + \varepsilon) \cdot m - \varepsilon \cdot t$ belong to $E_\ell(N)$. Clearly, $m = \frac{1}{2} \cdot r + \frac{1}{2} \cdot s$. Neither r nor s is a multiple of m as otherwise the linearity of the one-to-one correspondence $y \in \mathbb{R}^{\Gamma \times N} \leftrightarrow t \in \mathbb{R}^{P(N)}$ implies that the line L contains the zero vector in $\mathbb{R}^{\Gamma \times N}$, which is not the case.

References

- A. Antonucci and F. Cuzzolin. Credal set approximation by lower probabilities: application to credal networks. In E. Hüllermeier, R. Kruse, and F. Hoffman, editors, *Lecture Notes in AI 6178: IPMU 2010*, pages 716–725. Springer, 2010.
- P. Csóka, P. J.-J. Herings, and L. A. Kóczy. Balancedness conditions for exact games. *Mathematical Methods of Optimization Research*, 74:44–52, 2011.
- J. Derkx and J. Kuipers. On the number of extreme points of the core of a transferable utility game. In P. Borm and H. Peters, editors, *Chapters in Game Theory in Honour of Stef Tijs*, pages 83–97. Kluwer, 2002.
- J. Kuipers, D. Vermeulen, and M. Voorneveld. A generalization of the Shapley-Ichiishi result. *International Journal of Game Theory*, 39:585–602, 2010.
- S. Maass. Continuous linear representation of coherent lower previsions. In J. M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *Proceedings in Informatics 18: ISIPTA'03*, pages 372–382. Carleton Scientific, 2003.
- E. Quaeghebeur and G. de Cooman. Extreme lower probabilities. *Fuzzy Sets and Systems*, 159: 2163–2175, 2008.
- J. Rosenmüller. *Game Theory: Stochastics, Information, Strategies and Cooperation*. Kluwer, Boston, 2000.
- M. Studený. *Probabilistic Conditional Independence Structures*. Springer, London, 2005.
- M. Studený and T. Kroupa. Core-based criterion for extreme supermodular functions. *Discrete Applied Mathematics*, 20:122–151, 2016.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
- A. Wallner. Maximal number of vertices of polytopes defined by f-probabilities. In F. G. Cozman, R. Nau, and T. Seidenfeld, editors, *Proceedings of ISIPTA'05*, pages 126–139, 2005.
- G. M. Ziegler. *Lectures on Polytopes*. Springer, New York, 1995.

A Note on Imprecise Monte Carlo over Credal Sets via Importance Sampling

Matthias C. M. Troffaes

Durham University (United Kingdom)

MATTHIAS.TROFFAES@DURHAM.AC.UK

Abstract

This brief paper is an exploratory investigation of how we can apply sensitivity analysis over importance sampling weights in order to obtain sampling estimates of lower previsions described by a parametric family of distributions. We demonstrate our results on the imprecise Dirichlet model, where we can compare with the analytically exact solution. We discuss the computational limitations of the approach, and propose a simple iterative importance sampling method in order to overcome these limitations. We find that the proposed method works pretty well, at least in the example studied, and we discuss some further possible extensions.

Keywords: importance sampling; lower prevision; Monte Carlo; optimisation.

1. Introduction

Various sensible approaches to sampling for lower previsions can be found in the literature. Some of these are:

- two-level Monte Carlo sampling, where first one samples distributions over the (extreme points of the) credal set, and then samples from these distributions,
- sampling random sets, and then evaluating the resulting belief function ([Moral and Wilson, 1996](#)), and
- perform importance sampling from a reference distribution, and then solve an optimisation problem over the importance sampling weights ([O'Neill, 2009](#); [Fetz and Oberguggenberger, 2015](#); [Zhang and Shields, 2016](#)).

The first is inefficient, and only provides a non-conservative solution. The second is more efficient, but requires a large number of optimisation problems to be solved (one for each sample), and requires a suitable belief function approximation to be identified if one wants to apply this to arbitrary lower previsions. The third can be quite effective. For example, [de Angelis et al. \(2015\)](#) have successfully used sensitivity analysis over importance sampling weights with respect to the mean parameter of a normal distribution. [Fetz and Oberguggenberger \(2015\)](#) used importance sampling over both the mean and the variance parameters of a normal distribution using a 2-dimensional grid. A case study comparing a wide range of techniques, specifically aimed at reliability analysis, can be found in [Oberguggenberger et al. \(2009\)](#). Here, we are interested in seeing whether importance sampling can be performed over larger parameter spaces and distributions with non-trivial normalisation constants, using standard high-dimensional optimisation procedures.

Importance sampling in imprecise probability has been studied already in the '90s; see for example [Moral and Wilson \(1996\)](#); [Cano et al. \(1996\)](#); [Hernández and Moral \(1997\)](#) for some early works. In this paper, we follow [O'Neill \(2009\)](#), and look specifically at how we can use sensitivity analysis over the importance sampling weights directly in order to obtain sampling estimates, without needing to draw large numbers of samples, and without needing to solve large numbers

of optimisation problems. Unlike O'Neill (2009), however, we do not just look at Bayesian sensitivity analysis, and admit arbitrary sets of distributions in our theoretical treatment. Also unlike for instance O'Neill (2009); de Angelis et al. (2015); Fetz and Oberguggenberger (2015); Zhang and Shields (2016), in this paper, we use self-normalised importance sampling instead of standard importance sampling, as we find that this drastically speeds up calculations.

The main contribution of this paper is a simple yet novel (as far as we know) iterative importance sampling method that requires far less computational power compared to standard importance sampling methods for sensitivity analysis, in the sense that far smaller samples can be used, and that far smaller optimisation problems need to be solved. The key novelty is the idea of iteratively changing the importance sampling distribution itself, in order to ensure that the final answer has an effective sample size that is as close as possible to the actual sample size.

No novel theory is proved in the paper, however we do demonstrate the method on a fully worked example. This leads us to conjecture that convergence of the technique can be established under certain circumstances.

Section 2 reviews the basic theory behind importance sampling. Section 3 looks at how sensitivity analysis can be applied on importance sampling. An example of this approach is discussed in section 4, and various issues are identified. Section 5 describes a simple way of addressing some of these issues. The example is revisited in section 6. Section 7 concludes the paper with a discussion and some further ideas for future research.

2. Importance Sampling

In this section, we review the basic ideas behind importance sampling. For the theory behind the results that are presented here, we refer to Owen (2013, Chapter 9).

Assume we have an i.i.d. sample x_1, \dots, x_n drawn from a strictly positive probability density function q . Throughout the entire paper, we will consider many different probability density functions, but the sample x_1, \dots, x_n will always be one drawn from q . Assume we have a real-valued function $f(x)$, and we would like to calculate the expectation of f with respect to some other probability density function p .

In case $p = q$, by the central limit theorem, an approximate 95% confidence interval for the expectation of f with respect to q is then given by $\hat{\mu} \pm 1.96\hat{\sigma}/\sqrt{n}$ where

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n f(x_i) \quad \hat{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^n (f(x_i) - \hat{\mu})^2 \quad (1)$$

Can we use the same sample x_1, \dots, x_n drawn from q to get an estimate for the expectation of f with respect to $p \neq q$? The following equality gives a clue as to how we might do that:

$$\int f(x)p(x)dx = \int \frac{p(x)}{q(x)}f(x)q(x)dx = \int w_p(x)f(x)q(x)dx \quad (2)$$

where $w_p = p/q$. So, the expectation of f with respect to p is the same as the expectation of $w_p f$ with respect to q , and therefore an approximate 95% confidence interval for the expectation of f with respect to p is then given by $\hat{\mu}_p \pm 1.96\hat{\sigma}_p/\sqrt{n}$ where

$$\hat{\mu}_p := \frac{1}{n} \sum_{i=1}^n w_p(x_i)f(x_i) \quad \hat{\sigma}_p^2 := \frac{1}{n-1} \sum_{i=1}^n (w_p(x_i)f(x_i) - \hat{\mu}_p)^2 \quad (3)$$

This estimate is called the *importance sampling estimate*.

Often, the normalisation constant of the densities is unknown, or is slow to compute, and we only know $w'_p = cp/q$ for some unknown value of c . In this case, we can use the *self-normalised importance sampling estimate*:

$$\hat{\mu}_p := \frac{\sum_{i=1}^n w'_p(x_i) f(x_i)}{\sum_{i=1}^n w'_p(x_i)} \quad \hat{\sigma}_p^2 := \frac{1}{n-1} \frac{\frac{1}{n} \sum_{i=1}^n w'_p(x_i)^2 (f(x_i) - \hat{\mu}_p)^2}{\left(\frac{1}{n} \sum_{i=1}^n w'_p(x_i)\right)^2} \quad (4)$$

Although $\hat{\sigma}_p^2$ gives an indication of the quality of the estimate, one must be wary that $\hat{\sigma}_p^2$ is by itself only an approximation of the true error. An additional diagnostic to consider is the effective sample size, which can be calculated as follows:

$$n_p := \frac{\left(\sum_{i=1}^n w'_p(x_i)\right)^2}{\sum_{i=1}^n w'_p(x_i)^2} \quad (5)$$

Note that there are many different ways to define effective sample size and even more ways to define diagnostics for importance sampling. What matters for this paper is that a low n_p is bad, and that $n_p \simeq n$ is good. For an in-depth discussion about diagnostics for importance sampling, we refer to [Owen \(2013, Section 9.3\)](#).

3. Sensitivity Analysis

Importance sampling has many different uses, including variance reduction, numerical integration, and Bayesian inference. In this paper, we aim to study importance sampling in order to do inference over sets of distributions.

A key observation is that we can use importance sampling in order to estimate the lower prevision of a gamble f . [O'Neill \(2009\)](#) studied this technique already in a Bayesian setting. Here, we present the theory generally for an arbitrary set of probability density functions.

Say we have some set \mathcal{M} of probability density functions. The *lower prevision* of f is then defined as

$$\underline{E}(f) := \min_{p \in \mathcal{M}} \int f(x)p(x)dx \quad (6)$$

where we assume that the minimum is achieved, for simplicity of presentation. But we know that $\hat{\mu}_p \pm 1.96\hat{\sigma}_p/\sqrt{n}$ provides a confidence interval for the integral on the right hand side. So, if

$$p^* := \arg \min_{p \in \mathcal{M}} \hat{\mu}_p \quad (7)$$

then $\hat{\mu}_{p^*} \pm 1.96\hat{\sigma}_{p^*}/\sqrt{n}$ provides a 95% confidence interval for \underline{E} provided that p^* is equal to, or close enough to, the density that minimises the expectation in eq. (6). The key observation here is that we only need a single sample x_1, \dots, x_n , and that the optimisation procedure operates on the weights only.

One issue with this method is that $\hat{\sigma}_{p^*}$ can be very large. So, the method will only work if $\hat{\sigma}_p$ remains reasonably bounded. From the literature on importance sampling for variance reduction, we know that good choices for q are those that are proportional to $|f|p$ ([Owen, 2013, Chapter 9, p. 6](#)). So, in case \mathcal{M} covers a wide range of distributions p , it may be hard to identify a single

sampling distribution q . [Zhang and Shields \(2016\)](#), Section 3) discuss ways of choosing optimal sampling distributions for credal sets.

A second problem is that, in general, there is no single sampling distribution q that can guarantee a good effective sample size for all p in \mathcal{M} . Consequently, with this approach, even if we try to choose q optimally, the effective sample size at p^* can still become extremely low.

A third problem is that p^* as determined by eq. (7) may not be close at all to the density that minimises the expectation in eq. (6), especially when the effective sample size is low. In that case, $\hat{\mu}_{p^*} \pm 1.96\hat{\sigma}_{p^*}/\sqrt{n}$ may not provide a very accurate confidence interval on E . [O'Neill \(2009\)](#), Section 7) derived some explicit statistical bounds on the absolute and relative errors, but these bounds only cover standard (not self-normalising) importance sampling.

4. Example

As a first example, we demonstrate the use of importance sampling for sensitivity analysis on the imprecise Dirichlet model, similar to the one studied in [O'Neill \(2009\)](#).

Denote the k -dimensional unit simplex by Δ . Consider an unknown parameter $x \in \Delta$, say, modelling the probabilities of some multinomial process. Consider the following class of probability density functions on x :

$$p(x | t) = \frac{\Gamma(s)}{\prod_{j=1}^k \Gamma(st_j)} \prod_{j=1}^k x_j^{st_j-1} \quad (8)$$

with hyperparameters $s > 0$ and $t \in \Delta$ —these are Dirichlet distributions. We are interested in finding the lower expectation of some function $f(x)$, over all $t \in \mathcal{T} \subseteq \Delta$ and with $s = 2$ fixed.

Note that in our notation, we will parameterise everything in terms of t rather than in terms of p . So $w_t := w_{p(\cdot|t)}$, $\hat{\mu}_t := \hat{\mu}_{p(\cdot|t)}$, $n_t := n_{p(\cdot|t)}$, and so on.

For $q(x)$, we take the Dirichlet distribution with uniform $\tilde{t}_j = 1/k$ and with the same value for $\tilde{s} = 2$. An alternative option is to take $\tilde{s} = \alpha k$ with $0 < \alpha < 1$, say $\alpha = 1/2$. This will incur a bias for sampling towards the extremes, i.e. make the tails heavier. Experimentally, we observed that increasing the variance of the reference distribution can increase the effective sample size.

In order to apply importance sampling, we need to calculate the weight function. The unnormalised weights are:

$$w_t(x) = p(x | t)/q(x) \propto \prod_{j=1}^k x_j^{st_j - \tilde{s}\tilde{t}_j} = w'_t(x) \quad (9)$$

In this case, we have a very simple closed analytical expression for $w'_t(x)$. Note that we could also use $w_t(x)$ directly, however evaluating the normalisation constants requires several evaluations of the Gamma function, and slows down the optimisation procedure considerably. The optimisation problem for the lower expectation can be written as

$$t^* = \arg \min_{t \in \mathcal{T}} \frac{\sum_{i=1}^n w'_t(x_i) f(x_i)}{\sum_{i=1}^n w'_t(x_i)} \quad (10)$$

As a numerical example, we take $k = 5$, $\mathcal{T} = \{t \in \Delta : t_j \geq 0.1\}$, and $f(x) = x_1 + 2x_2 + 5x_3 + 4x_4 - 3x_5$. In this case, we know that the exact expectation of f , for fixed t , is given by

$$E(f) = t_1 + 2t_2 + 5t_3 + 4t_4 - 3t_5. \quad (11)$$

So, the lower prevision of f over all $t \in \mathcal{T}$ is clearly achieved for $t^* = (0.1, 0.1, 0.1, 0.1, 0.6)$, and is given by

$$\underline{E}(f) = 0.1 + 2 \times 0.1 + 5 \times 0.1 + 4 \times 0.1 - 3 \times 0.6 = -0.6 \quad (12)$$

The next table summarizes our simulation results for $\tilde{s} = k/2 = 2.5$:

n	5	50	500	5000
$\hat{\mu}_{t^*}$	1.50	0.13	-0.85	-0.29
$\hat{\sigma}_{t^*}$	0.11	3.18	10.83	10.74
$\hat{\sigma}_{t^*}/\sqrt{n}$	0.048	0.45	0.48	0.15
n_{t^*}	1.104	15.016	6.061	141.67
t_1^*	0.1	0.1	0.17	0.1
t_2^*	0.57	0.1	0.1	0.1
t_3^*	0.1	0.1	0.1	0.1
t_4^*	0.1	0.1	0.1	0.1
t_5^*	0.13	0.6	0.53	0.6

The code was implemented in R. The `constrOptim` function was used to do the actual optimisation, through the downhill simplex method. The cases $n = 5$ and $n = 50$ give a result instantly, for $n = 500$, the simulation took about 10 seconds, and for $n = 5000$, the simulation took about 200 seconds. The bottleneck is clearly the optimisation procedure. We emphasize that we have not tried to write the fastest possible code, and there might still be good opportunities for optimisation.

Unsurprisingly, the $n = 5$ case is quite bad: t^* is completely off, and the stimate is completely off the chart. Also the error is underestimated substantially, due to the very small effective sample size. The $n = 50$ case fares better. Interestingly, t^* is fully correctly identified. However, the effective sample size is not too high, and the actual estimate is still quite far off, due to the variance once more being underestimated.

Intriguingly, the $n = 500$ case has a lower effective sample size than the $n = 50$ case, and a worse t^* . Nevertheless, the estimate is reasonably correct, and at least the actual value lies inside the 95% confidence interval in this case. The $n = 5000$ case gives the correct estimate for t^* , and again the actual value lies just at the edge of the 95% confidence interval.

5. Iterated Importance Sampling

We have seen that a single importance sampling distribution q may not provide a good effective sample size across the entire set of distributions \mathcal{M} , even if n is quite large. For instance, in the numerical example, with $n = 500$ we still only had $n_{t^*} \simeq 6$, and with $n = 5000$ we had only $n_{t^*} \simeq 141$.

What we conclude from this is that plain sensitivity analysis over our importance sampling does not work very well, even in simple cases. Next we discuss some extensions of the proposed procedure in order to make it work.

Even though the estimates are quite bad, our numerical experimentation shows that the correct t^* , or nearly correct t^* , can be identified already with lower n . So, rather than increasing n in order to guarantee a high n_{t^*} , one idea is to iterate the procedure so that $q(x)$ eventually converges to $p(x|t^*)$ where t^* is the actual optimal choice. If $q(x)$ is equal to $p(x|t^*)$, then all weights are identical, and $n = n_{t^*}$. Also, in this case, it turns out that the optimisation in eq. (10) runs very quickly, because we are already near the optimal solution.

Here is how we might implement this in practice:

1. Set t to some reasonable initial value.
2. Generate sample from $q(x) := p(x | t)$.
3. Find optimal t_* through eq. (10).
4. Check if n_{t^*} is close to n . If yes, stop.
5. Set $t = t_*$, and return to item 2.

One suggestion is to take the same value for n through each step, however a case could be made for choosing a lower value for n , and then simply to repeat the final step of the procedure for a large value of n in order to obtain a final accurate estimate. Another option might be to increase the value of n as the algorithm converges closer to the correct t^* .

6. Example Revisited

Let us apply the proposed iterative procedure on our Dirichlet example. For simplicity, we chose a fixed value of $n = 141$; this corresponds roughly to our earlier n_{t^*} when $n = 5000$, so provides a good basis for comparison of computational efficiency. The next table summarises the results:

iteration	1	2	3
$\hat{\mu}_{t^*}$	0.062	-0.39	-0.63
$\hat{\sigma}_{t^*}$	4.28	2.00	1.76
$\hat{\sigma}_{t^*}/\sqrt{n}$	0.36	0.17	0.15
n_{t^*}	21.60	105.93	141.00
t_1^*	0.16	0.1	0.1
t_2^*	0.1	0.1	0.1
t_3^*	0.1	0.1	0.1
t_4^*	0.1	0.1	0.1
t_5^*	0.54	0.6	0.6

The entire simulation took only 6 seconds, compared to 200 seconds from before for the same effective sample size.

We see that the simulation converges in just 3 steps. In the first step, we get fairly close to the correct t^* , even though the effective sample size $n_{t^*} \simeq 22$ is pretty low. The second step uses this t^* to draw samples, and as this distribution is much closer to the actual optimal distribution, the effective sample size increases substantially. In this step, we also identify the correct value for t^* . The last step uses the correct distribution for sampling, and gets a full effective sample size.

We also ran the simulations using standard (not self-normalised) importance sampling. In that case, the entire simulation took 86 seconds, which is almost a factor 15 slower than the self-normalised version. Undoubtedly this is due to the computational expense of calculating the normalisation constant during the optimisation. Unless the normalisation constant is trivial, self-normalised importance sampling will outperform standard importance sampling for sensitivity analysis over the weights. In addition, the self-normalised estimator has also better consistency properties, even though it has a higher theoretical variance (Owen, 2013, Section 9.2).

7. Discussion and Conclusion

We have described how sensitivity analysis over importance sampling can be used to estimate lower previsions. The key observation that makes this possible is that importance sampling allows us to

estimate means not just from the distribution that we are sampling from, but from an entire neighbourhood of distributions around the sampling distribution. Through straightforward optimisation over the importance sampling weights, we can therefore estimate lower previsions without having to, say, draw samples from all extreme points of the credal set. The technique is simple, seems largely unknown in the community, and is readily applicable for medium sized problems.

We saw that a naive application of sensitivity analysis around the weights may not work very well, due to poor effective sample sizes especially when the optimal distribution is far away from the sampling distribution. We suggested simple yet novel solution for this problem: an iterative procedure which naturally moves the sampling distribution towards the optimal distribution. We demonstrated how this led to a much quicker estimate with far less computational power required.

Whilst the procedure that we have described will work well for medium sized problems, we foresee that for really large scale problems, the effective sample size may still be too limited to ensure that the optimal distribution can be identified at all. In such cases, perhaps the credal set could scale throughout the algorithm, in order to ensure a reasonable effective sample size, and therefore to help convergence of the algorithm.

Another idea is to use importance sampling to explore only a very small region of \mathcal{M} , but then to use the resulting derivative information to move q in the right direction. A problem with this however is that the derivatives obtained are quite noisy, and in practice we have not found a good way of using these noisy derivatives to ensure convergence.

Obviously, this note only gave an initial exploration of what is possible with sensitivity analysis over the importance sampling weights. It would be interesting to try out these methods on large scale problems. Moreover, it would be great to develop theoretical guarantees and diagnostics for convergence. Finally, it would be interesting to see if the importance sampling as described could be integrated into Markov chain Monte Carlo methods for full robust Bayesian inference over large sets of priors.

Acknowledgements

The author would like to thank Jonathan Rougier, Louis Aslett, Ullrika Sahlin, and Rasmus Bååth for stimulating discussions on the topic of importance sampling for imprecise probability. The author is also grateful to the reviewers for their extremely useful and constructive comments, and in particular for the suggestion of various references relevant to the topic.

References

- J. E. Cano, L. D. Hernández, and S. Moral. Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning*, 15(1):77–92, 1996.
- M. de Angelis, E. Patelli, and M. Beer. Advanced line sampling for efficient robust reliability analysis. *Structural Safety*, 52, Part B:170–182, 2015. ISSN 0167-4730. doi:[10.1016/j.strusafe.2014.10.002](https://doi.org/10.1016/j.strusafe.2014.10.002).
- T. Fetz and M. Oberguggenberger. Imprecise random variables, random sets, and Monte Carlo simulation. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, 2015. ISSN 2214-6702. doi:[10.1017/etds.2015.10](https://doi.org/10.1017/etds.2015.10).

- tions, pages 137–146, 2015. URL <http://www.sipta.org/isipta15/data/paper/12.pdf>.
- L. D. Hernández and S. Moral. Mixing exact and importance sampling propagation algorithms in dependence graphs. *International Journal of Intelligent Systems*, 12(8):553–576, Aug. 1997.
- S. Moral and N. Wilson. Importance sampling algorithms for the calculation of Dempster-Shafer belief. In *Proceedings of IPMU-96 Conference*, volume 3, pages 1337–1344, 1996.
- M. Oberguggenberger, J. King, and B. Schmelzer. Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. *International Journal of Approximate Reasoning*, 50(4):680–693, 2009. ISSN 0888-613X. doi:[10.1016/j.ijar.2008.09.004](https://doi.org/10.1016/j.ijar.2008.09.004).
- B. O'Neill. Importance sampling for Bayesian sensitivity analysis. *International Journal of Approximate Reasoning*, 50(2):270–278, 2009. ISSN 0888-613X. doi:[10.1016/j.ijar.2008.03.015](https://doi.org/10.1016/j.ijar.2008.03.015).
- A. B. Owen. *Monte Carlo theory, methods and examples*. 2013. URL <http://statweb.stanford.edu/~owen/mc/>.
- J. Zhang and M. D. Shields. Efficient propagation of imprecise probabilities. In *7th International Workshop on Reliable Engineering Computing*, pages 197–209, 2016. URL <http://rec2016.rub.de/papers.html>.

Imprecise Swing Weighting for Multi-Attribute Utility Elicitation Based on Partial Preferences

Matthias C. M. Troffaes

Durham University (United Kingdom)

MATTHIAS.TROFFAES@DURHAM.AC.UK

Ullrika Sahlin

Lund University (Sweden)

ULLRIKA.SAHLIN@CEC.LU.SE

Abstract

We describe a novel approach to multi-attribute utility elicitation which is both general enough to cover a wide range of problems, whilst at the same time simple enough to admit reasonably straightforward calculations. We allow both utilities and probabilities to be only partially specified, through bounding. We still assume marginal utilities to be precise. We derive necessary and sufficient conditions under which our elicitation procedure is consistent. As a special case, we obtain an imprecise generalization of the well known swing weighting method for eliciting multi-attribute utility functions. An example from ecological risk assessment demonstrates our method.

Keywords: utility; partial preference; consistency; uniqueness; multi-attribute; elicitation; imprecise; robust; swing weighting.

1. Introduction

In many decision problems where outcomes feature multiple attributes, additive multi-attribute utility functions are a popular choice due to their simplicity (Clemen and Reilly, 2001). They split the joint utility function into a weighted sum of marginal utility functions. Elicitation of the joint can then be split into two elicitation procedures: one for each of the marginals, and one for the weights.

A reoccurring issue is the precision of the attribute weights. Indeed, whilst marginal utility functions on separate attributes are often quite easily elicited, the way in which these attributes should be weighed against each other is much harder to quantify precisely. Such decision problems appear in different applications (Yemshanov et al., 2013; Hermerén et al., 2014). So, even if an additive form can be assumed, the weights themselves might still be subject to imprecision due to incomplete preferences between multi-attribute lotteries. Hermerén et al. (2014) suggest different types of value uncertainty and conclude that more work is needed to understand the cause of this uncertainty, in order to understand how to treat it. The only applied example we identified is Yemshanov et al. (2013), who considered uncertainty in weights by a multidimensional efficiency frontier analysis, which treat each attribute separately. Here we are interested in the elicitation of these weights.

Swing weighting (von Winterfeldt and Edwards, 1986) is a simple and popular method for eliciting the weights of an additive multi-attribute utility function. Unfortunately, the standard treatment of swing weighting uses ‘scores’ which are not usually directly interpreted in terms of preferences over lotteries, giving it the impression of a heuristic rather than a normative method. Moreover, swing weighting forces completeness of preferences between multi-attribute lotteries.

In this paper, we generalise swing weighting so bounds on the weights of the joint utility function can be elicited normatively. This is an important step to further widen the applicability of utility theory in problems where the consequences of decisions have multiple aspects that cannot be easily

weighed against each other. As an extra bonus, we also derive a normative interpretation of the standard swing weighting procedure. The resulting problems, when both probabilities and utilities are allowed to be imprecise, require quadratic programming, for which standard algorithms exist.

We are of course aware that decision theory has been generalised to deal with arbitrary partial preferences in their full generality (Seidenfeld et al., 1995). However, such theories can be technically difficult to work with due to the fact that they lead to non-convex sets of utilities and probabilities. Various special cases have been studied that do allow convex analysis to be used for elicitation, modelling, and inference (Williams, 1975, 2007; Levi, 1980; Walley, 1991). These works do not explicitly try to deal with multiple attributes. The contribution of this paper can be seen as a practical approach towards multi-attribute decision problems where marginal utilities are still precise, but where we wish to be a bit more cautious about modelling preferences across attributes. It can be seen as a simple generalisation of Walley (1991) to the multi-attribute case.

The idea of generalising swing weighting to allow for partial preferences is not new either; see for instance Mustajoki et al. (2005); Riabacke et al. (2009); Gomes et al. (2011); Riabacke et al. (2012); Danielson et al. (2014) and references therein. Those works generally focus on reducing the cognitive requirements on decision makers, and propose specific models for eliciting attribute weights, but without relating the elicitation directly to preferences between multi-attribute gambles. Instead, in this paper, we develop a general mathematical framework for eliciting attribute weights in a directly operational way through preferences between multi-attribute gambles. We thereby generalise the interval swing weighting method proposed by Mustajoki et al. (2005) (at least in the cases where the reference attribute is either the worst or the best attribute). The theory that we develop can be adapted to a wide range of situations, and possibly could accommodate cognitive limitations in a more flexible way, although we will not fully explore this in this paper.

The paper is structured as follows. Section 2 introduces the notation and explains the assumptions that we make throughout the paper. Section 3 briefly describes how marginal utility functions can be elicited, and serves as an introduction to the idea of utility elicitation. Section 4 reviews the standard swing weighting procedure, and provides a simple normative interpretation of swing weighting in terms of lotteries. Section 5 generalises the swing weighting procedure to allow imprecise weights, and section 6 identifies necessary and sufficient conditions for this elicitation procedure to be consistent. Section 7 provides a fully worked example of our method, using an example from ecological risk assessment. Section 8 concludes the paper.

2. Notation and Assumptions

Let $\mathcal{R} := \mathcal{A}_1 \times \dots \times \mathcal{A}_n$ be a finite set of rewards, each reward $r = (a_1, \dots, a_n)$ comprising of n attributes. A *lottery* ℓ on \mathcal{R} is a probability mass function over \mathcal{R} , and is interpreted as a random reward with precisely known probabilities. The set of all lotteries over \mathcal{R} is denoted by $L(\mathcal{R})$.

Note that at this stage, we are not yet interested in modelling uncertainty. Rather, we will use lotteries in order to elicit a subject's attitudes towards rewards. For modelling uncertainty, one might consider *horse lotteries*, which for our purpose would be mappings from some finite possibility space Ω to $L(\mathcal{R})$. This follows the traditional approach (Anscombe and Aumann, 1963; Seidenfeld et al., 1995). In this paper, we do not consider horse lotteries, and focus purely on the utility aspect. That said, in section 7, we will demonstrate how uncertainty can be incorporated in an example.

So, we wish to model our preferences between lotteries over our multi-attribute rewards. A *utility function* on \mathcal{R} is any function $U: \mathcal{R} \rightarrow \mathbb{R}$. We lift U to $L(\mathcal{R})$ in the usual way:

$$U(\ell) := \sum_{r \in \mathcal{R}} \ell(r)U(r). \quad (1)$$

Note that U satisfies

$$U(\alpha\ell_1 + (1 - \alpha)\ell_2) = \alpha U(\ell_1) + (1 - \alpha)U(\ell_2) \quad (2)$$

for all $\alpha \in [0, 1]$. The standard approach assumes that our preferences form a complete preorder \succeq on $L(\mathcal{R})$ and can be represented through a utility function U , where

$$\ell_1 \succeq \ell_2 \iff U(\ell_1) \geq U(\ell_2) \quad (3)$$

for all ℓ_1 and $\ell_2 \in L(\mathcal{R})$. This representation can be directly motivated from some simple assumptions on \succeq ([Herstein and Milnor, 1953](#)).

However, in many applications, preferences between rewards are inherently incomplete, in the sense that there may be lotteries between which we cannot state any preference. We will assume that our preferences form a preorder \succeq on $L(\mathcal{R})$ (so we drop completeness), and can be represented through a *set* \mathcal{U} of utility functions $U: L(\mathcal{R}) \rightarrow \mathbb{R}$:

$$\ell_1 \succeq \ell_2 \iff \forall U \in \mathcal{U}: U(\ell_1) \geq U(\ell_2) \quad (4)$$

for all ℓ_1 and $\ell_2 \in L(\mathcal{R})$. Elicitation is then concerned with finding a procedure for identifying \mathcal{U} .

In cases where rewards are comprised of multiple attributes, in standard utility theory, it is customary to split the elicitation problem into two parts:

1. Elicit *marginal utility functions* $U_i: \mathcal{A}_i \rightarrow \mathbb{R}$ for each $i \in \{1, \dots, n\}$.
2. Assume that the joint utility function can be written as a particular function of the marginal utility functions, and elicit the parameters of that function.

The simplest of these joint forms is the *additive form*:

$$U(a_1, \dots, a_n) = \sum_{i=1}^n k_i U_i(a_i) \quad (5)$$

Again, this form can be directly motivated from some simple assumptions on \succeq ([Keeney and Raiffa, 1993](#)). Although those assumptions are quite restrictive and are easily criticised, the simplicity of the additive form, having only n parameters, make it one of the most commonly used models for multi-attribute utility in practical applications.

3. Elicitation of Marginal Utility

To introduce the idea of utility elicitation, and for the sake of completeness, we mention a simple standard method for eliciting the marginal utility functions U_i ([Clemen and Reilly, 2001](#)):

1. Identify a worst reward \underline{a}_i and a best reward \bar{a}_i in \mathcal{A}_i .
2. For every other reward a_i in \mathcal{A}_i , find $\alpha(a_i)$ so that the subject is indifferent between (i) getting a_i with certainty and (ii) getting \underline{a}_i with probability $1 - \alpha(a_i)$ or \bar{a}_i with probability $\alpha(a_i)$:

$$a_i \simeq (1 - \alpha(a_i))\underline{a}_i \oplus \alpha(a_i)\bar{a}_i \quad (6)$$

where \oplus denotes the combination of rewards into lotteries, so $(1 - \alpha)r_1 \oplus \alpha r_2$ is the lottery ℓ with $\ell(r_1) = 1 - \alpha$, $\ell(r_2) = \alpha$, and $\ell(r) = 0$ for all other rewards. We also denoted indifference by \simeq : $\ell_1 \simeq \ell_2 \iff (\ell_1 \succeq \ell_2 \text{ and } \ell_2 \succeq \ell_1)$.

3. Set $U_i(\underline{a}_i) := 0$, $U_i(\bar{a}_i) := 1$, and $U_i(a_i) := \alpha(a_i)$ for every other reward a_i in \mathcal{A}_i .

Naturally, an interesting question relates to how we can relax this elicitation procedure to allow for incomplete preferences in the marginals. As we shall see in section 5, allowing incompleteness in both marginals and in the weights introduces non-linear constraints. So, for practical reasons, in this paper, we only investigate incompleteness in the weights, and assume marginals to be fully precise.

4. Swing Weighting

For eliciting the weights k_i in the joint utility function of eq. (5), various methods exist, but a simple and effective method is *swing weighting* (von Winterfeldt and Edwards, 1986):

1. Score the following $n + 1$ rewards:

	reward	score
$r_0 := (\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n)$	0	
$r_1 := (\bar{a}_1, \underline{a}_2, \dots, \underline{a}_n)$	s_1	
$r_2 := (\underline{a}_1, \bar{a}_2, \dots, \underline{a}_n)$	s_2	
⋮	⋮	
$r_n := (\underline{a}_1, \underline{a}_2, \dots, \bar{a}_n)$	s_n	

where the worst score is 0 (always assigned to the worst reward), the best score is 100, and the other scores indicate the “relative improvement” from the worst reward.

2. Set

$$k_i := \frac{s_i}{\sum_{i=1}^n s_i}. \quad (7)$$

Note that this formula hinges on the assumption that all marginal utility functions are renormalized to the $[0, 1]$ interval—this is the case if we use the marginal method as described earlier.

Although we find swing weighting a straightforward and mathematically elegant method for eliciting the weights, what is missing is an interpretation directly in terms of preferences over lotteries. In fact, this is very easy to do, but much to our surprise it is not mentioned anywhere in the literature as far as we could find:

1. Consider again the rewards r_0, \dots, r_n as constructed above. Clearly r_0 is the worst reward.
2. Identify the best of these rewards. Without loss of generality, we may assume that this is r_n (we can always permute the order of the attributes if need be).
3. For all $i \in \{1, \dots, n\}$, find α_i such that

$$r_i \simeq (1 - \alpha_i)r_0 \oplus \alpha_i r_n. \quad (8)$$

(Note that $\alpha_n = 1$.)

4. Set

$$k_i := \frac{\alpha_i}{\sum_{i=1}^n \alpha_i} \quad (9)$$

It is easy to see that this choice of k_i is the only choice that is compatible with eq. (8). So, we can interpret the swing weighting scores directly in terms of probabilities, which we find more appealing. This also puts the method on a firm normative basis.

5. Imprecise Swing Weighting

A common criticism against the swing weighting method (and, in fact, also against the marginal method that we presented) is that all lotteries considered in the elicitation involve extremes only. We therefore adapt the swing weighting method to allow for more flexible comparisons, not just focusing on extremes.

Remember that we are dropping the completeness assumption, and therefore that we are interested in identifying a set \mathcal{U} of utility functions, rather than a single utility function. To do so, we will view all weights k_i as parameters (so we have n parameters), and we will represent \mathcal{U} through a collection of constraints on these parameters:

1. Consider *any* joint rewards r_0, \dots, r_n such that for all $j \in \{1, \dots, n-1\}$ we have that

$$r_0 \preceq r_j \preceq r_n \quad (10)$$

2. For all $j \in \{1, \dots, n-1\}$, find the largest $\underline{\alpha}_j$ and smallest $\bar{\alpha}_j$ such that

$$(1 - \underline{\alpha}_j)r_0 \oplus \underline{\alpha}_j r_n \preceq r_j \preceq (1 - \bar{\alpha}_j)r_0 \oplus \bar{\alpha}_j r_n \quad (11)$$

3. Let u_j denote the vector of marginal utilities for r_j , i.e. if $r_j = (a_1, \dots, a_n)$ then $u_j = (U_1(a_1), \dots, U_n(a_n))$. Let k denote the vector (k_1, \dots, k_n) . With this notation, impose

$$\forall j \in \{1, \dots, n-1\}: \quad (u_j - (1 - \underline{\alpha}_j)u_0 - \underline{\alpha}_j u_n) \cdot k \geq 0 \quad (12a)$$

$$\forall j \in \{1, \dots, n-1\}: \quad (u_j - (1 - \bar{\alpha}_j)u_0 - \bar{\alpha}_j u_n) \cdot k \leq 0 \quad (12b)$$

$$1 \cdot k = 1 \quad (12c)$$

The last constraint is simply another way of writing $\sum_{i=1}^n k_i = 1$, and fixes the multiplicative scaling of the joint utility function.

To see that the other two constraints indeed represent the elicited preferences, note that eq. (11) is equivalent to

$$(1 - \underline{\alpha}_j)U(r_0) + \underline{\alpha}_j U(r_n) \leq U(r_j) \leq (1 - \bar{\alpha}_j)U(r_0) + \bar{\alpha}_j U(r_n) \quad (13)$$

and note that $U(r_j) = u_j \cdot k$.

These inequalities are quadratic in the marginal utilities and in the weights. However, if the marginal utilities are precise, then we have a simple set of linear constraints on the weights k_j .

Naturally, we also recover swing weighting as a special case. In the imprecise case however it is important to realise that we cannot always take the rewards as in the standard swing weighting method. We already argued that this might be a bad idea due to the focus on extremes, however it may also cause a problem because the method requires that there is a single best attribute—we may not have such best attribute if we allow for incompleteness.

6. Consistency and Uniqueness

The procedure that we described works for any choice of rewards r_j . Naturally, a good choice of rewards r_j should ensure that the constraints obtained admit a solution for all possible choices of $0 \leq \underline{\alpha}_j \leq \bar{\alpha}_j \leq 1$. In fact, we also would like this solution to be unique in the precise case (i.e. when $\underline{\alpha}_j = \bar{\alpha}_j$ for all j), so that we can at least in principle allow a complete elicitation of preferences if possible. Both of these desirata are satisfied if:

- (i) $u_0 \leq u_n$, and
- (ii) the system

$$\forall j \in \{1, \dots, n-1\}: \quad (u_j - (1 - \alpha_j)u_0 - \alpha_j u_n) \cdot k = 0 \quad (14a)$$

$$1 \cdot k = 1 \quad (14b)$$

has a unique solution, regardless our choice of $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$.

Note that $u_0 \leq u_n$ guarantees that $(u_j - (1 - \alpha_j)u_0 - \alpha_j u_n)$ is a decreasing function of α_j , so in this case it is ensured that, say, if we solve eqs. (12b) and (12c) with equalities everywhere, then the inequality in eq. (12a) is automatically satisfied; in other words, eq. (12) is consistent.

We will henceforth assume that $u_0 \leq u_n$, and focus on the uniqueness of the solution of eq. (14).

Theorem 1 Consider any $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$. If the matrix

$$\begin{bmatrix} u_1 - (1 - \alpha_1)u_0 - \alpha_1 u_n \\ u_2 - (1 - \alpha_2)u_0 - \alpha_2 u_n \\ \vdots \\ u_{n-1} - (1 - \alpha_{n-1})u_0 - \alpha_{n-1} u_n \\ 1 \end{bmatrix} \quad (15)$$

has full rank, then eq. (14) has a unique solution.

The following theorem provides much quicker check for uniqueness, in case u_0 is constant (note that u_0 being constant is a standard feature of the usual swing weighting procedure).

Theorem 2 Consider any $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$. Assume that u_0 is constant, and that the vectors $(u_1, \dots, u_{n-1}, 1)$ are linearly independent. Let λ_j be the coefficients that decompose u_n as a linear combination of $(u_1, \dots, u_{n-1}, 1)$, i.e.

$$u_n = \lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \quad (16)$$

Then eq. (14) has a unique solution if and only if

$$\sum_{j=1}^{n-1} \alpha_j \lambda_j \neq 1 \quad (17)$$

In particular, when $\lambda_1 \leq 0, \dots, \lambda_{n-1} \leq 0$, then eq. (14) has a unique solution, regardless our choice of $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$.

Proof We need to show that the matrix

$$\begin{bmatrix} u_1 - (1 - \alpha_1)u_0 - \alpha_1 u_n \\ u_2 - (1 - \alpha_2)u_0 - \alpha_2 u_n \\ \vdots \\ u_{n-1} - (1 - \alpha_{n-1})u_0 - \alpha_{n-1} u_n \\ 1 \end{bmatrix} \quad (18)$$

has full rank. Because u_0 is constant, and so is the final row, this matrix has full rank if and only if

$$\begin{bmatrix} u_1 - \alpha_1 u_n \\ u_2 - \alpha_2 u_n \\ \vdots \\ u_{n-1} - \alpha_{n-1} u_n \\ 1 \end{bmatrix} \quad (19)$$

has full rank.

Because the $(u_1, \dots, u_{n-1}, 1)$ are linearly independent, we can write u_n as a linear combination of these vectors:

$$u_n = \lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \quad (20)$$

So, our matrix can be written as

$$\begin{bmatrix} u_1 - \alpha_1 u_n \\ u_2 - \alpha_2 u_n \\ \vdots \\ u_{n-1} - \alpha_{n-1} u_n \\ 1 \end{bmatrix} = \begin{bmatrix} u_1 - \alpha_1 \left(\lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \right) \\ u_2 - \alpha_2 \left(\lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \right) \\ \vdots \\ u_{n-1} - \alpha_{n-1} \left(\lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \right) \\ 1 \end{bmatrix} \quad (21)$$

$$= \begin{bmatrix} 1 - \alpha_1 \lambda_1 & -\alpha_1 \lambda_2 & \dots & -\alpha_1 \lambda_{n-1} & -\alpha_1 \lambda_n \\ -\alpha_2 \lambda_1 & 1 - \alpha_2 \lambda_2 & \dots & -\alpha_2 \lambda_{n-1} & -\alpha_2 \lambda_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\alpha_{n-1} \lambda_1 & -\alpha_{n-1} \lambda_2 & \dots & 1 - \alpha_{n-1} \lambda_{n-1} & -\alpha_{n-1} \lambda_n \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ 1 \end{bmatrix} \quad (22)$$

which has full rank if both matrices on the right hand side have full rank. The second matrix has full rank by assumption. The first matrix has full rank if and only if

$$\begin{bmatrix} 1 - \alpha_1 \lambda_1 & -\alpha_1 \lambda_2 & \dots & -\alpha_1 \lambda_{n-1} \\ -\alpha_2 \lambda_1 & 1 - \alpha_2 \lambda_2 & \dots & -\alpha_2 \lambda_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_{n-1} \lambda_1 & -\alpha_{n-1} \lambda_2 & \dots & 1 - \alpha_{n-1} \lambda_{n-1} \end{bmatrix} \quad (23)$$

has full rank. This matrix can be written as

$$I - \alpha \lambda^T \quad (24)$$

where $\alpha = (\alpha_1, \dots, \alpha_{n-1})$ and $\lambda = (\lambda_1, \dots, \lambda_{n-1})$. This has full rank if and only if its determinant is non-zero. We now use Sylvester's determinant identity:

$$\det(I - \alpha \lambda^T) = \det(1 - \lambda^T \alpha) = 1 - \sum_{j=1}^{n-1} \alpha_j \lambda_j \quad (25)$$

We arrive at the desired result. ■

This theorem applies for instance if we use the joint rewards as in standard swing weighting:

reward	u_j
$r_0 := (\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n)$	$(0, 0, \dots, 0)$
$r_1 := (\bar{a}_1, \underline{a}_2, \dots, \underline{a}_n)$	$(1, 0, \dots, 0)$
$r_2 := (\underline{a}_1, \bar{a}_2, \dots, \underline{a}_n)$	$(0, 1, \dots, 0)$
\vdots	\vdots
$r_n := (\underline{a}_1, \underline{a}_2, \dots, \bar{a}_n)$	$(0, 0, \dots, 1)$

Note that $u_0 \leq u_n$, as required. Also, u_0 is constant (zero), and $(u_1, \dots, u_{n-1}, 1)$ are linearly independent: the theorem applies. Because

$$u_n = 1 - \sum_{j=1}^{n-1} u_j, \quad (26)$$

it follows that $\lambda_j = -1$ for all $j \in \{1, \dots, n-1\}$. The condition for uniqueness is satisfied.

We can also consider the case where u_n is constant:

Theorem 3 Consider any $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$. Assume that u_n is constant, and that the vectors $(u_1, \dots, u_{n-1}, 1)$ are linearly independent. Let λ_j be the coefficients that decompose u_0 as a linear combination of $(u_1, \dots, u_{n-1}, 1)$, i.e.

$$u_0 = \lambda_n + \sum_{j=1}^{n-1} \lambda_j u_j \quad (27)$$

Then eq. (14) has a unique solution if and only if

$$\sum_{j=1}^{n-1} (1 - \alpha_j) \lambda_j \neq 1 \quad (28)$$

In particular, when $\lambda_1 \leq 0, \dots, \lambda_{n-1} \leq 0$, then eq. (14) has a unique solution, regardless our choice of $\alpha_1, \dots, \alpha_{n-1} \in [0, 1]$.

The proof is almost identical to the proof of the previous theorem, and hence is left as an exercise to the reader. We will use this variant in the example below.

7. Example

We now provide a fully worked example to see the theory at work. In addition to imprecise utilities, we will also admit imprecise probabilities.

Following Bohman and Edsman (2013), we are interested in an ecological management decision, namely the eradication of an invasive species that has been observed in a water system. The following management decisions were identified:

- I Do nothing.
- II Mechanical removal.
- III Drain the system on water and remove of individuals by hand.
- IV Drain the system of water, dredge and sieve the masses to identify and remove individuals.
- V Use a decomposable biocide in combination with drainage to increase the biocide concentration.
- VI Increase pH in combination with drainage and removal by hand.

The decision comprises several attributes. Each decision was scored according to attributes identified as relevant by a group of experts. For each of these attributes, a discrete scale ranging from 1 to 4 was constructed, where 1 corresponds to the worst outcome, and 4 corresponds to the best outcome. We will interpret these scores as marginal utility functions.

Besides the attributes, the experts also bounded the probability that the method is successful in eradication. Note that we are using hypothetical values here, these values are not actual expert judgements, and only serve to demonstrate the methodology. Note also that in the actual problem, there is considerable uncertainty about whether the invasive species is present at all. For simplicity, in this example, we assume that the alien species is present with certainty.

The following table lists all attributes considered, as well as the interpretation of the scores for each attribute and for each management decision, and the expert assessments for the attribute scores, in case of success:

Attribute	Worst (score 1)	Best (score 4)	Decision d					
			I	II	III	IV	V	VI
Biotic impact	High	Low	4	4	3	3	2	1
Longevity of impacts	Long	Short	4	4	3	3	1	2
Experience	Little	High	4	3	1	4	1	1
Feasibility	Difficult	Easy	4	4	2	3	1	2
Cost	High	Low	4	4	3	1	2	3

In case of failure to eradicate the invasive species, the scores for biotic impact and longevity of impacts drop to their worst values:

Attribute	Worst (score 1)	Best (score 4)	Decision d					
			I	II	III	IV	V	VI
Biotic impact	High	Low	1	1	1	1	1	1
Longevity of impacts	Long	Short	1	1	1	1	1	1
Experience	Little	High	4	3	1	4	1	1
Feasibility	Difficult	Easy	4	4	2	3	1	2
Cost	High	Low	4	4	3	1	2	3

Bounds on the probability of successful eradication of the species under the different management decisions are:

Probability	Decision d					
	I	II	III	IV	V	VI
$\frac{p_d}{\bar{p}_d}$	0	0.05	0.3	0.4	1.0	0.7
	0	0.25	0.5	0.7	1.0	0.8

The joint expected utility of decision d can be written as:

$$\sum_{j=1}^n k_j (\theta U_{1j}(a_{jd}) + (1 - \theta) U_{2j}(a_{jd})) \quad (29)$$

where U_{1j} are the marginal utilities as listed in the first table, and U_{2j} are the marginal utilities as listed in the second table (both after rescaling to 0–1).

Because the decision affects the probability of successful management (i.e. we have act-state dependence), we will treat the problem using interval dominance.

For eliciting the weights k_j of the joint utility function, we will use a variant of swing weighting, and we will consider the following joint rewards (directly expressed in terms of marginal utilities, rescaled to 0–1):

rewards
$u_0 := (2/3, 1, 1, 1, 1)$
$u_1 := (1, 2/3, 1, 1, 1)$
$u_2 := (1, 1, 2/3, 1, 1)$
$u_3 := (1, 1, 1, 2/3, 1)$
$u_4 := (1, 1, 1, 1, 2/3)$
$u_5 := (1, 1, 1, 1, 1)$

These rewards are more natural from an ecological risk perspective compared to the rewards considered by the original swing weighting method: they consider only small changes from a normal state, instead of extremes, and are thus easier to compare (regardless of any imprecision in preferences).

Note that $u_0 \leq u_5$ as required for consistency. Also note that u_5 is constant, so we can apply theorem 3. We see that

$$u_0 = 14/3 - \sum_{j=1}^4 u_j. \quad (30)$$

Consequently all $\lambda_j = -1$ in theorem 3, and so the condition for uniqueness is always satisfied.

We consider biotic impact to be the most important attribute, so clearly we have that $r_0 \preceq r_j \preceq r_5$ for all $j \in \{0, \dots, 5\}$. We also assess that

$$0.8r_0 \oplus 0.2r_5 \preceq r_1 \preceq 0.7r_0 \oplus 0.3r_5 \quad (31)$$

$$0.5r_0 \oplus 0.5r_5 \preceq r_2 \preceq 0.4r_0 \oplus 0.6r_5 \quad (32)$$

$$0.3r_0 \oplus 0.7r_5 \preceq r_3 \preceq 0.1r_0 \oplus 0.9r_5 \quad (33)$$

$$0.2r_0 \oplus 0.8r_5 \preceq r_4 \preceq 0.1r_0 \oplus 0.9r_5 \quad (34)$$

With these assessments, eq. (12) becomes

$$((1, 2/3, 1, 1, 1) - 0.8(2/3, 1, 1, 1, 1) - 0.2) \cdot k \geq 0 \quad (35a)$$

$$((1, 1, 2/3, 1, 1) - 0.5(2/3, 1, 1, 1, 1) - 0.5) \cdot k \geq 0 \quad (35b)$$

$$((1, 1, 1, 2/3, 1) - 0.3(2/3, 1, 1, 1, 1) - 0.7) \cdot k \geq 0 \quad (35c)$$

$$((1, 1, 1, 1, 2/3) - 0.2(2/3, 1, 1, 1, 1) - 0.8) \cdot k \geq 0 \quad (35d)$$

$$((1, 2/3, 1, 1, 1) - 0.7(2/3, 1, 1, 1, 1) - 0.3) \cdot k \leq 0 \quad (35e)$$

$$((1, 1, 2/3, 1, 1) - 0.4(2/3, 1, 1, 1, 1) - 0.6) \cdot k \leq 0 \quad (35f)$$

$$((1, 1, 1, 2/3, 1) - 0.1(2/3, 1, 1, 1, 1) - 0.9) \cdot k \leq 0 \quad (35g)$$

$$((1, 1, 1, 1, 2/3) - 0.1(2/3, 1, 1, 1, 1) - 0.9) \cdot k \leq 0 \quad (35h)$$

$$1 \cdot k = 1 \quad (35i)$$

So, for each decision, we need to minimize and maximize the joint utility expressed in eq. (29), subject to the above constraints and subject to $\underline{p}_d \leq \theta \leq \bar{p}_d$. The constraints are all linear, and the objective function is quadratic, hence this is a quadratic programming problem. Because θ itself only appears linearly and is constrained separately, it suffices to consider only the extreme values for θ . Consequently, for each decision, we must merely solve two linear programs: one for $\theta = \underline{p}_d$ and one for $\theta = \bar{p}_d$.

Using `scipy` (Jones et al., 2001–), we find the following bounds:

Decision	Lower Utility	Upper Utility
I	0.25	0.37
II	0.23	0.47
III	0.18	0.31
IV	0.38	0.57
V	0.14	0.17
VI	0.11	0.17

Options I, III, V, and VI are dominated by option IV so should not be considered. Either option II (mechanical removal) or IV (drain the system of water, dredge and sieve), could be considered.

For the sake of completeness, we also present the bounds on the attribute weights, resulting from eq. (35):

Attribute	Lower Weight	Upper Weight
Biotic impact	0.36	0.43
Longevity of impacts	0.26	0.33
Experience	0.15	0.21
Feasibility	0.04	0.12
Cost	0.04	0.08

8. Conclusions

In this paper, we provided an imprecise generalisation of the swing weighting method for eliciting multi-attribute utility functions. The proposed method enables us to cover a wider range of problems where preference can only be partially specified, whilst at the same time still admitting straightforward calculations. We studied the consistency of the elicitation procedure, and found simple conditions under which consistency is always guaranteed. We demonstrated our method using a real example concerning the management of an invasive species featuring substantial uncertainty in the management outcomes and ambiguity in the preferences over different impacts. In this example, we allowed both utilities and probabilities to be only partially specified, through bounding.

We do note that our approach is still limited in that we will assume that all marginal utility functions are precise. Relaxing this is possible but leads to fully non-linear optimisation, and more work is needed to identify whether such treatment can be feasible in practice. Naturally, another limitation is that we only discussed *additive* multi-attribute utility functions.

Another open end is that we have assumed that our preferences over horse lotteries are representable by a convex set of weights along with a convex set of probability mass functions. Whilst such representation is appealing mathematically (inference becomes a quadratic programming problem), it would be interesting to have an axiomatic treatment from first principles (as in [Seidenfeld et al. \(1995\)](#)) identifying the conditions under which such treatment is feasible.

Acknowledgements

U. Sahlin has been funded by the Swedish research council FORMAS project number 219-2013-1271. The research presented in this paper is a contribution to the strategic research area Biodiversity and Ecosystems in a Changing Climate, BECC, funded by the Swedish government.

References

- F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34(1):199–205, Mar. 1963.
- P. Bohman and L. Edsman. Marmorkräftan i Märstaån. Riskanalys och åtgärdsförslag. Aqua Reports 2013:17, Sveriges lantbruksuniversitet, Drottningholm, 2013.
- R. T. Clemen and T. Reilly. *Making Hard Decisions*. Duxbury, 2001.
- M. Danielson, L. Ekenberg, A. Larsson, and M. Riabacke. Weighting under ambiguous preferences and imprecise differences in a cardinal rank ordering process. *International Journal of Computational Intelligence Systems*, 7:105–112, 2014. doi:[10.1080/18756891.2014.853954](https://doi.org/10.1080/18756891.2014.853954).
- L. F. A. M. Gomes, L. A. D. Rangel, and M. d. R. Leal Junior. Treatment of uncertainty through the interval smart/swing weighting method: a case study. *Pesquisa Operacional*, 31(3):467–485, 12 2011. ISSN 0101-7438. doi:[10.1590/S0101-74382011000300004](https://doi.org/10.1590/S0101-74382011000300004).
- G. Hermerén, I. Brinck, J. Persson, and N.-E. Sahlin. *Value uncertainty and value instability in decision-making*, pages 100–110. Liber Amicorum Pascal Engel. University of Geneva, 2014.
- I. N. Herstein and J. Milnor. An axiomatic approach to measurable utility. *Econometrica*, 21(2): 291–297, Apr. 1953.
- E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL <http://www.scipy.org/>. [Online; accessed 2017-02-20].
- R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, 1993. ISBN 0-521-44185-4.
- I. Levi. *The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability, and Chance*. MIT Press, Cambridge, 1980.
- J. Mustajoki, R. P. Hämäläinen, and A. Salo. Decision support by interval SMART/SWING: Incorporating imprecision in the SMART and SWING methods. *Decision Sciences*, 36(2):317–339, 2005. ISSN 1540-5915. doi:[10.1111/j.1540-5414.2005.00075.x](https://doi.org/10.1111/j.1540-5414.2005.00075.x).
- M. Riabacke, M. Danielson, L. Ekenberg, and A. Larsson. *A Prescriptive Approach for Eliciting Imprecise Weight Statements in an MCDA Process*, pages 168–179. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04428-1. doi:[10.1007/978-3-642-04428-1_15](https://doi.org/10.1007/978-3-642-04428-1_15).
- M. Riabacke, M. Danielson, and L. Ekenberg. State-of-the-art prescriptive criteria weight elicitation. *Advances in Decision Sciences*, 2012. doi:[10.1155/2012/276584](https://doi.org/10.1155/2012/276584).
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. A representation of partially ordered preferences. *The Annals of Statistics*, 23:2168–2217, 1995.
- D. von Winterfeldt and W. Edwards. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, 1986.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

IMPRECISE SWING WEIGHTING FOR MULTI-ATTRIBUTE UTILITY ELICITATION

- P. M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex, 1975.
- P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007. doi:[10.1016/j.ijar.2006.07.019](https://doi.org/10.1016/j.ijar.2006.07.019).
- D. Yemshanov, F. H. Koch, Y. Ben-Haim, M. Downing, F. Sapiro, and M. Siltanen. A new multi-criteria risk mapping approach based on a multiattribute frontier concept. *Risk Analysis*, 33(9):1694–1709, 2013. ISSN 1539-6924. doi:[10.1111/risa.12013](https://doi.org/10.1111/risa.12013).

Exchangeable Choice Functions

Arthur Van Camp

Gert de Cooman

IDLab, Ghent University

Ghent (Belgium)

ARTHUR.VANCAMP@UGENT.BE

GERT.DECOOMAN@UGENT.BE

Abstract

We investigate how to model exchangeability with choice functions. Exchangeability is a structural assessment on a sequence of uncertain variables. We show how such assessments constitute a special kind of indifference assessment, and how this idea leads to a counterpart of de Finetti's Representation Theorem, both in a finite and a countable context.

Keywords: exchangeability; choice functions; indifference; sets of desirable gambles; representation.

1. Introduction

In this paper, we study how to model exchangeability, a structural assessment for uncertainty models that is important for inference purposes, in the framework of choice functions, an interesting approach to modelling uncertainty. This work builds on earlier results by [De Cooman et al. \(2009\)](#); [De Cooman and Quaeghebeur \(2012\)](#).

Choice functions are related to the fundamental problem in decision theory: how to make a choice from within a set of available options. In their book, [von Neumann and Morgenstern \(1944\)](#) provide an axiomatisation of choice based on a pairwise comparison between options. Later on, many authors ([Arrow, 1951](#); [Uzawa, 1956](#); [Rubin, 1987](#)) generalised this idea and proposed a theory of choice functions based on choice between more than two elements. One of the aspects of [Rubin's \(1987\)](#) theory is that, between any pair of options, the agent either prefers one of them or is indifferent between them, so two options are never incomparable. However, for instance when the available information does not allow for a complete comparison of the options, the agent may be undecided between two options without being indifferent between them; this will for instance typically be the case when there is little or no relevant information available. This is one of the motivations for a theory of imprecise probabilities ([Walley, 1991](#)), where incomparability and indifference are distinguished. [Kadane et al. \(2004\)](#) and [Seidenfeld et al. \(2010\)](#) generalise [Rubin's \(1987\)](#) axioms to allow for incomparability.

Exchangeability is a structural assessment on a sequence of uncertain variables. Loosely speaking, making a judgement of exchangeability means that the order in which the variables are observed is considered irrelevant. This irrelevancy will be modelled through an indifference assessment. The first detailed study of exchangeability was given by [de Finetti \(1937\)](#). We refer to the paper by [De Cooman and Quaeghebeur \(2012, Sec. 1\)](#) for a brief historical overview.

In Sec. 2, we recall the necessary tools for modelling indifference with choice functions. Next, in Sec. 3, we derive de Finetti-like Representation Theorems for a finite sequence that is exchangeable. We take this one step further in Sec. 4, where we consider a countable sequence and de-

rive a representation theorem for such sequences. To compare with earlier work (De Cooman and Quaeghebeur, 2012), we also provide representation theorems for sets of desirable gambles.¹

2. Choice Functions, Desirability and Indifference

Consider a real vector space \mathcal{V} , provided with the vector addition and scalar multiplication. Elements u of \mathcal{V} are intended as abstract representations of *options* amongst which a subject can express his preferences, by specifying, as we will see below, choice functions. Mostly, options will be real-valued maps on the possibility space, interpreted as uncertain rewards, and therefore also called *gambles*. The set of all gambles on the possibility space \mathcal{X} will be denoted by $\mathcal{L}(\mathcal{X})$. However, we will define choice functions on \mathcal{V} rather than on $\mathcal{L}(\mathcal{X})$, because, as we will see later, we will need to define choice functions on *equivalence classes* of gambles, which are no longer gambles themselves, but still constitute a vector space.² Given any subset O of \mathcal{V} , we will define the *linear hull* $\text{span}(O) := \{\sum_{k=1}^n \lambda_k u_k : n \in \mathbb{N}, \lambda_k \in \mathbb{R}, u_k \in O\} \subseteq \mathcal{V}$ and the *positive hull* $\text{posi}(O) := \{\sum_{k=1}^n \lambda_k u_k : n \in \mathbb{N}, \lambda_k \in \mathbb{R}_{>0}, u_k \in O\} \subseteq \text{span}(O)$, where $\mathbb{R}_{>0}$ is the set of all (strictly) positive real numbers. Furthermore, for any λ in $\mathbb{R}_{>0}$ and u in \mathcal{V} , we let $\lambda O + \{v\} := \{\lambda u + v : u \in O\}$. A subset O of \mathcal{V} is called a *convex cone* if it is closed under positive finite linear combinations, i.e. if $\text{posi}(O) = O$. A convex cone \mathcal{K} is called *proper* if $\mathcal{K} \cap -\mathcal{K} = \{0\}$. With any proper convex cone $\mathcal{K} \subseteq \mathcal{V}$, we associate an ordering $\leq_{\mathcal{K}}$ on \mathcal{V} as follows: $u \leq_{\mathcal{K}} v \Leftrightarrow v - u \in \mathcal{K}$ for any u and v in \mathcal{V} . For any u and v in \mathcal{V} , we write $u <_{\mathcal{K}} v$ if $u \leq_{\mathcal{K}} v$ and $u \neq v$. We collect all the options u for which $0 <_{\mathcal{K}} u$ in $\mathcal{V}_{>0}$. When we work with gambles, then $\mathcal{V} = \mathcal{L}(\mathcal{X})$ and the ordering will be the standard one \leq , given by $f \leq g \Leftrightarrow (\forall x \in \mathcal{X}) f(x) \leq g(x)$. We collect the positive gambles—gambles f for which $0 < f$ —in $\mathcal{L}(\mathcal{X})_{>0}$. Then \leq corresponds to $\leq_{\mathcal{K}}$ where we let $\mathcal{K} := \mathcal{L}(\mathcal{X})_{>0} \cup \{0\}$.

We denote by $\mathcal{Q}(\mathcal{V})$ the set of all non-empty *finite* subsets of \mathcal{V} . Elements of $\mathcal{Q}(\mathcal{V})$ are the option sets amongst which a subject can choose his preferred options.

A choice function C on \mathcal{V} is a map $C: \mathcal{Q} \rightarrow \mathcal{Q} \cup \{\emptyset\}: O \mapsto C(O)$ such that $C(O) \subseteq O$. Not every such map represents rational beliefs; only the coherent ones are considered to do so. We call a choice function C on \mathcal{V} *coherent*³ if for all O, O_1 and O_2 in $\mathcal{Q}(\mathcal{V})$, u and v in \mathcal{V} , and λ in $\mathbb{R}_{>0}$:

- C_1 . $C(O) \neq \emptyset$;
- C_2 . if $u < v$ 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 $O_1 + \{u\} \subseteq C(O_2 + \{u\})$.

Consider two isomorphic vector spaces \mathcal{V} and \mathcal{W} , a linear order isomorphism ϕ between \mathcal{V} and \mathcal{W} , and a choice function C on \mathcal{V} . Define the choice function C' on \mathcal{W} as $u \in C(O) \Leftrightarrow \phi(u) \in C'(\phi(O))$ for all O in $\mathcal{Q}(\mathcal{V})$ and u in O . Then, because ϕ is a bijection, C satisfies Axioms C_1 and C_3 if and only if C' does; furthermore, because ϕ is order preserving, C satisfies Axiom C_2 if and only if C' does; and finally, because ϕ is linear, C satisfies Axiom C_4 if and only if C' does: such isomorphisms preserve coherence.

1. Due to page constraints, the proofs are not included in the paper. Readers interested in verifying the main proofs can access them through [arXiv:0801.0980](https://arxiv.org/abs/0801.0980).

2. This also allows us to connect our approach with the theory of coherent choice functions by Seidenfeld et al. (2010), where the authors define their choice function on *horse lotteries* instead of gambles. We intend to report on this later.

3. Our rationality axioms are based on those by Seidenfeld et al. (2010), slightly modified for use with sets of desirable options.

A set of desirable options (or gambles) $D \subseteq \mathcal{V}$ is essentially the restriction to pairwise comparison of a choice function: $D = \{u \in \mathcal{V} \setminus \{0\} : \{u\} = C(\{0, u\})\}$. We call D coherent if $0 \notin D$, $\mathcal{V}_{>0} \subseteq D$, $u \in D \Rightarrow \lambda u \in D$, and $u, v \in D \Rightarrow u + v \in D$ for all u and v in \mathcal{V} and λ in $\mathbb{R}_{>0}$. D is coherent if the choice function C it is based on, is coherent.

Since, as we will see, an exchangeability assessment amounts to a specific indifference assessment, we recall how to model such assessments (Van Camp et al., 2017, Sec. 5). Next to $C(O)$ —the options that the agent *strictly* prefers from O —or D —the options that he *strictly* prefers to 0—we consider the options in $I \subseteq \mathcal{V}$, which the agent considers to be *equivalent to the zero option*. We call a set of indifferent options I coherent if, for all u and v in \mathcal{V} and λ in \mathbb{R} :

- I₁. $0 \in I$;
- I₂. if $u \in \mathcal{V}_{>0} \cup \mathcal{V}_{<0}$ then $u \notin I$;
- I₃. if $u \in I$ then $\lambda u \in I$;
- I₄. if $u, v \in I$ then $u + v \in I$.

We collect all options that are indifferent to an option u in \mathcal{V} into the *equivalence class* $[u] := \{v \in \mathcal{V} : v - u \in I\} = \{u\} + I$. The set of all these equivalence classes is the *quotient space* $\mathcal{V}/I := \{[u] : u \in \mathcal{V}\}$, a linear space itself. We provide it with the natural ordering inherited from \mathcal{V} : $\tilde{u} \leq \tilde{v} \Leftrightarrow (\exists u \in \tilde{u}, v \in \tilde{v}) u \leq v$, for all \tilde{u} and \tilde{v} in \mathcal{V}/I .

In the remainder of this section, we will recall some of the results by Van Camp et al. (2017), needed for this paper. Consider any coherent set of indifferent options I . A choice function C is called *compatible* with I if there is some *representing* choice function C' on \mathcal{V}/I such that $C(O) = \{u \in O : [u] \in C'(O/I)\}$ for all O in $\mathcal{Q}(\mathcal{V})$. In that case, C' is uniquely determined by $C'(O/I) = C(O)/I$ for all O in $\mathcal{Q}(\mathcal{V})$, and, moreover, C is coherent if and only if C' is. Equivalently, we find the following useful characterisation: C is compatible with I if and only if $0 \in C(O) \Leftrightarrow u \in C(O)$ for all u in I and $O \supseteq \{0, u\}$ in $\mathcal{Q}(\mathcal{V})$, which corresponds to the definition of indifference given by Seidenfeld (1988).

For desirability, compatibility with a coherent set of indifferent options I is defined as follows. We call a set of desirable gambles D *compatible* with I if $D + I \subseteq D$, and this is equivalent to $D = \bigcup D'$ where $D' \subseteq \mathcal{V}/I$ is the *representing* set of desirable options. In that case, D' is uniquely given by $D' = D/I$ —so $D = \bigcup_{u \in D} [u]$ —and, moreover, D is coherent if and only if D' is.

3. Finite Exchangeability

Consider $n \in \mathbb{N}$ uncertain variables X_1, \dots, X_n taking values in a non-empty set \mathcal{X} . The possibility space of the uncertain sequence (X_1, \dots, X_n) is \mathcal{X}^n .

We denote by $x = (x_1, \dots, x_n)$ an arbitrary element of \mathcal{X}^n . For any n in \mathbb{N} we call \mathcal{P}_n the group of all permutations π of the index set $\{1, \dots, n\}$. There are $|\mathcal{P}_n| = n!$ such permutations. With any such permutation π , we associate a permutation of \mathcal{X}^n , also denoted by π , and defined by $(\pi x)_k := x_{\pi(k)}$ for every k in $\{1, \dots, n\}$, or in other words, $\pi(x_1, \dots, x_n) = (x_{\pi(1)}, \dots, x_{\pi(n)})$. Similarly, we lift π to a permutation π^t on $\mathcal{L}(\mathcal{X}^n)$ by letting $\pi^t f := f \circ \pi$, so $(\pi^t f)(x) = f(\pi x)$ for all x in \mathcal{X}^n . Observe that π^t is a linear permutation of the vector space $\mathcal{L}(\mathcal{X}^n)$ of all gambles on \mathcal{X}^n .

If a subject assesses that the sequence of variables X in \mathcal{X}^n is exchangeable, this means that he is indifferent between any gamble f on \mathcal{X}^n and its permuted variant $\pi^t f$, for all π in \mathcal{P}_n . This leads us to the following proposal for the corresponding set of indifferent gambles:

$$I_{\mathcal{P}_n} := \text{span}\{f - \pi^t f : f \in \mathcal{L}(\mathcal{X}^n), \pi \in \mathcal{P}_n\}. \quad (1)$$

Definition 1 A choice function C on $\mathcal{L}(\mathcal{X}^n)$ is called (finitely) exchangeable if it is compatible with $I_{\mathcal{P}_n}$. Similarly, a set of desirable gambles $D \subseteq \mathcal{L}(\mathcal{X}^n)$ is called (finitely) exchangeable if it is compatible with $I_{\mathcal{P}_n}$.

Of course, so far, we do not yet know whether this notion of exchangeability is well-defined: indeed, we do not know yet whether $I_{\mathcal{P}_n}$ is a *coherent* set of indifferent gambles. In the next section, we will show that this is indeed the case.

3.1 Count Vectors

Let us now provide the tools necessary to prove that $I_{\mathcal{P}_n}$ is a coherent set of indifferent gambles, as introduced by [De Cooman et al. \(2009\)](#) and [De Cooman and Quaeghebeur \(2012\)](#).

The *permutation invariant atoms* $[x] := \{\pi x : x \in \mathcal{X}^n\}$, x in \mathcal{X}^n are the smallest permutation invariant subsets of \mathcal{X}^n . We introduce the *counting map* $T: \mathcal{X}^n \rightarrow \mathcal{N}^n: x \mapsto T(x)$ where $T(x)$ is called the *count vector* of x . It is the \mathcal{X} -tuple with components $T_z(x) := |\{k \in \{1, \dots, n\} : x_k = z\}|$ for all z in \mathcal{X} , so $T_z(x)$ is the number of times that z occurs in the sequence x_1, \dots, x_n . The range of T —the set \mathcal{N}^n —is called the set of possible count vectors and is given by $\mathcal{N}^n := \{m \in \mathbb{Z}_{\geq 0}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} m_x = n\}$. Applying any permutation to x leaves its result under the counting map unchanged. For any x in \mathcal{X}^n , if $m = T(x)$ then $[x] = \{y \in \mathcal{X}^n : T(y) = m\}$, so the permutation invariant atom $[x]$ is completely determined by the count vector m of all its elements, and is therefore also denoted by $[T(x)] = [m]$. Remark that $\{[m] : m \in \mathcal{N}^n\}$ partitions \mathcal{X}^n into disjoint parts with constant count vectors, and that $|[m]| = \binom{n}{m} := \frac{n!}{\prod_{z \in \mathcal{X}} m_z!}$.

In order to extend the idea of the count vectors for use with gambles, let us define the *set of all permutation invariant gambles* as $\mathcal{L}_{\mathcal{P}_n}(\mathcal{X}^n) := \{f \in \mathcal{L}(\mathcal{X}^n) : (\forall \pi \in \mathcal{P}_n) \pi^t f = f\} \subseteq \mathcal{L}(\mathcal{X}^n)$, and a special transformation $\text{inv}_{\mathcal{P}_n}$ of the linear space $\mathcal{L}(\mathcal{X}^n)$

$$\text{inv}_{\mathcal{P}_n}: \mathcal{L}(\mathcal{X}^n) \rightarrow \mathcal{L}(\mathcal{X}^n): f \mapsto \text{inv}_{\mathcal{P}_n}(f) := \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} \pi^t f,$$

which, as we will see, is closely linked with $\mathcal{L}_{\mathcal{P}_n}(\mathcal{X}^n)$ ([De Cooman and Quaeghebeur, 2012](#); [Van Camp et al., 2017](#)).

Proposition 2 $\text{inv}_{\mathcal{P}_n}$ is a linear transformation of $\mathcal{L}(\mathcal{X}^n)$, and

- (i) $\text{inv}_{\mathcal{P}_n} \circ \pi^t = \text{inv}_{\mathcal{P}_n} = \pi^t \circ \text{inv}_{\mathcal{P}_n}$ for all π in \mathcal{P} ;
- (ii) $\text{inv}_{\mathcal{P}_n} \circ \text{inv}_{\mathcal{P}_n} = \text{inv}_{\mathcal{P}_n}$;
- (iii) $\text{kern}(\text{inv}_{\mathcal{P}_n}) = I_{\mathcal{P}_n}$;
- (iv) $\text{rng}(\text{inv}_{\mathcal{P}_n}) = \mathcal{L}_{\mathcal{P}_n}(\mathcal{X}^n)$.

So we see that $\text{inv}_{\mathcal{P}_n}$ is a linear projection operator that maps any gamble to a permutation invariant counterpart.

As shown by [De Cooman and Quaeghebeur \(2012\)](#), the linear projection operator $\text{inv}_{\mathcal{P}_n}$ renders a gamble insensitive to permutation by replacing it with the uniform average of all its permutations. As a result, it assumes the same value for all gambles that can be related to each other through some permutation: $\text{inv}_{\mathcal{P}_n}(f) = \text{inv}_{\mathcal{P}_n}(g)$ if $f = \pi^t g$ for some π in \mathcal{P}_n , for all f and g in $\mathcal{L}(\mathcal{X}^n)$. Furthermore, for any f in $\mathcal{L}(\mathcal{X}^n)$, its transformation $\text{inv}_{\mathcal{P}_n}(f)$ is permutation invariant and therefore constant on the permutation invariant atoms $[m]$: $(\text{inv}_{\mathcal{P}_n}(f))(x) = (\text{inv}_{\mathcal{P}_n}(f))(y)$ if $[x] = [y]$, for all x and y in \mathcal{X}^n . We can use the properties of $\text{inv}_{\mathcal{P}_n}$ to prove that $I_{\mathcal{P}_n}$ is suitable for the definition of exchangeability.

Proposition 3 For any n in \mathbb{N} , the set $I_{\mathcal{P}_n}$, defined in Eq. (1), is a coherent set of indifferent gambles.

Since $I_{\mathcal{P}_n}$ is coherent, exchangeability is well-defined, and by the discussion in Sec. 2, the representing choice function C' is defined on $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, and, similarly, the representing set of desirable gambles $D' \subseteq \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$. So we can focus on the quotient space and its elements, exchangeable equivalent classes of gambles.

But before we do that, it will pay to further explore the notions we have introduced thus far.

Consider any f in $\mathcal{L}(\mathcal{X}^n)$. What is the constant value that $\text{inv}_{\mathcal{P}_n}(f)$ assumes on a permutation invariant atom $[m]$? To answer this question, consider any x in $[m]$, then $(\text{inv}_{\mathcal{P}_n}(f))(x) = \frac{1}{n!} \sum_{\pi \in \mathcal{P}_n} f(\pi x) = \frac{1}{n!} \frac{|\mathcal{P}_n|}{|[m]|} \sum_{y \in \{\pi x : \pi \in \mathcal{P}_n\}} f(y) = \frac{1}{\binom{n}{m}} \sum_{y \in [x]} f(y) = \frac{1}{\binom{n}{m}} \sum_{y \in [m]} f(y)$ where we used the fact that $|\mathcal{P}_n| = n!$ and $|[m]| = \binom{n}{m}$, whence $\text{inv}_{\mathcal{P}_n} = \sum_{m \in \mathcal{N}^n} H_n(\cdot|m) \mathbb{I}_{[m]}$, where $H_n(\cdot|m)$ is the linear expectation operator associated with the uniform distribution on the invariant atom $[m]$:

$$H_n(f|m) := \frac{1}{\binom{n}{m}} \sum_{y \in [m]} f(y) \text{ for all } f \text{ in } \mathcal{L}(\mathcal{X}^n) \text{ and } m \text{ in } \mathcal{N}^n.$$

It characterises a (multivariate) hyper-geometric distribution (Johnson et al., 1997), associated with random sampling without replacement from an urn with n balls of types \mathcal{X} , whose composition is characterised by the count vector m .

The result of subjecting a gamble f on \mathcal{X}^n to the map

$$H_n : \mathcal{L}(\mathcal{X}^n) \rightarrow \mathcal{L}(\mathcal{N}^n) : f \mapsto H_n(f) := H_n(f|\cdot)$$

is the gamble $H_n(f)$ on \mathcal{N}^n that assumes the value $\frac{1}{\binom{n}{m}} \sum_{y \in [m]} f(y)$ in every m in \mathcal{N}^n .

3.2 Exchangeable Equivalent Classes of Gambles

We already know that exchangeable choice functions are represented by choice functions on the quotient space $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, and similarly for sets of desirable gambles. In the quest for an elegant representation theorem, we thus need to focus on the quotient space $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and its elements, which are exchangeable equivalent classes of gambles.

In this section we investigate how the representation of permutation invariant gambles helps us find a representation for exchangeable choice functions. To that end, the representation will use equivalence classes $[f] := \{f\} + I_{\mathcal{P}_n}$ of gambles, for any f in $\mathcal{L}(\mathcal{X}^n)$. Recall that the quotient space $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} := \{[f] : f \in \mathcal{L}(\mathcal{X}^n)\}$ is a linear space itself, with additive identity $[0] = I_{\mathcal{P}_n}$, and therefore any element \tilde{f} of $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ is invariant under addition of $I_{\mathcal{P}_n}$: $\tilde{f} + I_{\mathcal{P}_n} = \tilde{f}$. Elements of $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ will be generically denoted by \tilde{f} or \tilde{g} .

Proposition 4 Consider any f and g in $\mathcal{L}(\mathcal{X}^n)$. Then $[f] = [g]$ if and only if $H_n(f) = H_n(g)$.

Therefore, it makes sense to introduce the map \tilde{H}_n :

$$\tilde{H}_n : \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} \rightarrow \mathcal{L}(\mathcal{N}^n) : \tilde{f} \mapsto H_n(f) \text{ for any } f \text{ in } \tilde{f}. \quad (2)$$

Then Proposition 4 guarantees that elements of $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ are characterised using \tilde{H}_n , in the sense that $\tilde{f} = \{f \in \mathcal{L}(\mathcal{X}^n) : H_n(f) = \tilde{H}_n(\tilde{f})\}$ for all \tilde{f} in $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$.

The map \tilde{H}_n takes as an argument an equivalence class of gambles, and maps it to some representing gamble on the count vectors. It will be useful later on to consider the inverse map \tilde{H}_n^{-1} :

$$\tilde{H}_n^{-1} : \mathcal{L}(\mathcal{N}^n) \rightarrow \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} : f \mapsto \left[\sum_{m \in \mathcal{N}^n} f(m) \mathbb{I}_{[m]} \right]. \quad (3)$$

Proposition 5 *The maps \tilde{H}_n as defined in Eq. (2) and \tilde{H}_n^{-1} as defined in Eq. (3) are each other's inverses.*

The importance of Prop. 5 lies in the fact that now, \tilde{H}_n is a bijection between $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and $\mathcal{L}(\mathcal{N}^n)$, and therefore, exchangeable equivalence classes of gambles are in a one-to-one correspondence with gambles on count vectors.

$$\begin{array}{ccc} \mathcal{L}(\mathcal{X}^n) & \xrightarrow{\quad H_n \quad} & \mathcal{L}(\mathcal{N}^n) \\ [-] \downarrow & \nearrow \tilde{H}_n & \\ \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} & & \end{array}$$

The commuting diagram shows the surjections $[-]: \mathcal{L}(\mathcal{X}^n) \rightarrow \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}: f \mapsto [f]$ and H_n (indicated with a single arrow), and the bijection \tilde{H}_n (indicated with a double arrow). Since the representing choice function C' is defined from C through $[-]$ —working point-wise on sets—this already suggests that C' can be transformed into a choice function on $\mathcal{L}(\mathcal{N}^n)$. To prove that they preserve coherence, there is only one missing link: the map \tilde{H}_n should be linear and preserve the ordering between $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and $\mathcal{L}(\mathcal{N}^n)$. Therefore, to define the ordering \leq on $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, as usual, we let \leq be inherited by the ordering \leq on $\mathcal{L}(\mathcal{X}^n)$:

$$\tilde{f} \leq \tilde{g} \Leftrightarrow (\exists f \in \tilde{f}, \exists g \in \tilde{g}) f \leq g$$

for all \tilde{f} and \tilde{g} in $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, turning $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ into an ordered linear space. It turns out that this vector ordering on $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ can be represented elegantly using \tilde{H}_n :

Proposition 6 *Consider any \tilde{f} and \tilde{g} in $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, then $\tilde{f} \leq \tilde{g}$ if and only if $\tilde{H}_n(\tilde{f}) \leq \tilde{H}_n(\tilde{g})$.*

Props. 5 and 6 imply that H_n is a linear order isomorphism.

3.3 A Representation Theorem

Now that we have found a linear order isomorphism \tilde{H}_n between $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and $\mathcal{L}(\mathcal{N}^n)$, we are ready to represent coherent and exchangeable choice functions.

Theorem 7 (Finite Representation) *Consider any choice function C on $\mathcal{L}(\mathcal{X}^n)$. Then C is exchangeable if and only if there is a unique representing choice function \tilde{C} on $\mathcal{L}(\mathcal{N}^n)$ such that*

$$C(O) = \{f \in O : H_n(f) \in \tilde{C}(H_n(O))\} \text{ for all } O \text{ in } \mathcal{Q}(\mathcal{L}(\mathcal{X}^n)).$$

Furthermore, in that case, \tilde{C} is given by $\tilde{C}(H_n(O)) = H_n(C(O))$ for all O in $\mathcal{Q}(\mathcal{L}(\mathcal{X}^n))$. Finally, C is coherent if and only if \tilde{C} is.

Similarly, consider any set of desirable gambles $D \subseteq \mathcal{L}(\mathcal{X}^n)$. Then D is exchangeable if and only if there is a unique representing set of desirable gambles $\tilde{D} \subseteq \mathcal{L}(\mathcal{N}^n)$ such that $D = \bigcup \tilde{H}_n^{-1}(\tilde{D})$. Furthermore, in that case, \tilde{D} is given by $\tilde{D} = H_n(D)$. Finally, D is coherent if and only if \tilde{D} is.

The number of occurrences of any outcome in a sequence (x_1, \dots, x_n) is fixed by its count vector m in \mathcal{N}^n . If we impose an exchangeability assessment on it, then we see, using Theorem 7, that the joint model on \mathcal{X}^n is characterised by a model on $\mathcal{L}(\mathcal{N}^n)$. So an exchangeable choice function C essentially represents preferences between urns with n balls of types \mathcal{X} with different compositions m : the choice $C(O)$ between the gambles in O is based upon the composition m .

3.4 Finite Representation in Terms of Polynomials

In Sec. 4, we will prove a similar representation theorem for infinite sequences. Since it no longer makes sense to *count* in such sequences, we first need to find an equivalent representation theorem in terms of something that does not depend on counts. More specifically, we need, for every n in \mathbb{N} another order-isomorphic linear space to $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, that allows for embedding: the linear space for $n_1 < n_2$ must be a subspace of the one for n_2 .

All the maps in this section have been introduced by [De Cooman et al. \(2009\)](#) and [De Cooman and Quaeghebeur \(2012\)](#). We use their ideas and work with polynomials on the \mathcal{X} -simplex $\Sigma_{\mathcal{X}} := \{\theta \in \mathbb{R}^{\mathcal{X}} : \theta \geq 0, \sum_{x \in \mathcal{X}} \theta_x = 1\}$. We consider the special subset $\mathcal{V}(\Sigma_{\mathcal{X}})$ of $\mathcal{L}(\Sigma_{\mathcal{X}})$: $\mathcal{V}(\Sigma_{\mathcal{X}})$ are the *polynomial gambles* h on $\Sigma_{\mathcal{X}}$, which are those gambles that are the restriction to $\Sigma_{\mathcal{X}}$ of a multivariate polynomial p on $\mathbb{R}^{\mathcal{X}}$, in the sense that $h(\theta) = p(\theta)$ for all θ in $\Sigma_{\mathcal{X}}$. We call p then a representation of h . It will be useful to introduce a notation for polynomial gambles with fixed degree n in \mathbb{N} : $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ is the collection of all polynomial gambles that have at least one representation whose degree is not higher than n . Both $\mathcal{V}(\Sigma_{\mathcal{X}})$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ are linear subspaces of $\mathcal{L}(\Sigma_{\mathcal{X}})$, and, as wanted, for $n_1 \leq n_2$, $\mathcal{V}^{n_1}(\Sigma_{\mathcal{X}})$ is a subspace of $\mathcal{V}^{n_2}(\Sigma_{\mathcal{X}})$.

Some special polynomial gambles are the *Bernstein gambles*:

Definition 8 (Bernstein gambles) Consider any n in \mathbb{N} and any m in \mathcal{N}^n . Define the Bernstein basis polynomial B_m on $\mathbb{R}^{\mathcal{X}}$ as $B_m(\theta) := \binom{n}{m} \prod_{x \in \mathcal{X}} \theta_x^{m_x}$ for all θ in $\mathbb{R}^{\mathcal{X}}$. The restriction to $\Sigma_{\mathcal{X}}$ is called a Bernstein gamble, which we also denote as B_m .

As shown by [De Cooman and Quaeghebeur \(2012\)](#) and also by [De Bock et al. \(2016\)](#), the set of all Bernstein gambles constitutes a basis for the linear space $\mathcal{V}^n(\Sigma_{\mathcal{X}})$:

Proposition 9 Consider any n in \mathbb{N} . The set of Bernstein gambles $\{B_m : m \in \mathcal{N}^n\}$ constitutes a basis for the linear space $\mathcal{V}^n(\Sigma_{\mathcal{X}})$.

As we have seen, to preserve coherence between two ordered linear spaces, we need a linear order isomorphism. So we wonder whether there is one between $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$. In Sec. 3.2 we have seen that there is one between $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ and $\mathcal{L}(\mathcal{N}^n)$, namely \tilde{H}_n . Therefore, it suffices to find one between $\mathcal{L}(\mathcal{N}^n)$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$. Consider the map

$$\text{CoM}_n : \mathcal{L}(\mathcal{N}^n) \rightarrow \mathcal{V}^n(\Sigma_{\mathcal{X}}) : r \mapsto \sum_{m \in \mathcal{N}^n} r(m) B_m.$$

Before we can establish that CoM_n is a linear order isomorphism, we need to provide the linear space $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ with an order \leq_B^n . We use the proper cone $\{0\} \cup \text{posi}(\{B_m : m \in \mathcal{N}^n\})$ to define the order \leq_B^n :

$$h_1 \leq_B^n h_2 \Leftrightarrow h_2 - h_1 \in \{0\} \cup \text{posi}(\{B_m : m \in \mathcal{N}^n\}) \text{ for all } h_1 \text{ and } h_2 \text{ in } \mathcal{V}^n(\Sigma_{\mathcal{X}}).$$

The following proposition is proved by [De Cooman and Quaeghebeur \(2012\)](#).

Proposition 10 Consider any n in \mathbb{N} . Then the map CoM_n is a linear order isomorphism between the ordered linear spaces $\mathcal{L}(\mathcal{N}^n)$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$.

The linear order isomorphism CoM_n helps us to define a linear order isomorphism between the linear spaces $\mathcal{L}(\mathcal{X}^n)$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$, a final tool needed for a representation theorem in terms of polynomial gambles. Indeed, consider for the map $M_n := \text{CoM}_n \circ H_n$:

$$M_n : \mathcal{L}(\mathcal{X}^n) \rightarrow \mathcal{V}^n(\Sigma_{\mathcal{X}}) : f \mapsto M_n(f|\theta),$$

where $M_n(f|\theta) := \sum_{m \in \mathcal{N}^n} \sum_{y \in [m]} f(y) \prod_{x \in \mathcal{X}} \theta_x^{m_x}$ is the expectation of f associated with the multinomial distribution whose parameters are n and θ . We introduce its version

$$\tilde{M}_n := \text{CoM}_n \circ \tilde{H}_n, \quad (4)$$

mapping $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$ to $\mathcal{V}^n(\Sigma_{\mathcal{X}})$. There is an immediate connection between M_n and \tilde{M}_n : they are both compositions of two linear order isomorphisms, and are therefore linear order isomorphisms themselves. Due to Prop. 4, considering any \tilde{f} in $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, M_n is constant on \tilde{f} , and the value it takes on any element of \tilde{f} is exactly $\tilde{M}_n(\tilde{f})$.

$$\begin{array}{ccccc} & & \text{CoM}_n & & \\ \mathcal{L}(\mathcal{N}^n) & \xleftarrow{H_n} & & \xrightarrow{M_n} & \mathcal{V}^n(\Sigma_{\mathcal{X}}) \\ & \searrow & & \swarrow & \\ & \mathcal{L}(\mathcal{X}^n) & \xrightarrow{[\cdot]} & & \\ & \swarrow & & \downarrow & \searrow \\ & & \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} & & \tilde{M}_n \end{array}$$

The commuting diagram shows the surjections $[\cdot]$, H_n and M_n , and the bijections \tilde{H}_n , \tilde{M}_n and CoM_n . It shows that both $\mathcal{L}(\mathcal{N}^n)$ and $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ are order-isomorphic to $\mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n}$, so they are both suitable to define a representing choice function on. In Theorem 7, we used the space $\mathcal{L}(\mathcal{N}^n)$. Here, we will use the other equivalent space $\mathcal{V}^n(\Sigma_{\mathcal{X}})$.

Theorem 11 (Finite Representation) Consider any choice function C on $\mathcal{L}(\mathcal{X}^n)$. Then C is exchangeable if and only if there is a unique representing choice function \tilde{C} on $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ such that

$$C(O) = \{f \in O : M_n(f) \in \tilde{C}(M_n(O))\} \text{ for all } O \text{ in } \mathcal{Q}(\mathcal{L}(\mathcal{X}^n)).$$

Furthermore, in that case, \tilde{C} is given by $\tilde{C}(M_n(O)) = M_n(C(O))$ for all O in $\mathcal{Q}(\mathcal{L}(\mathcal{X}^n))$. Finally, C is coherent if and only if \tilde{C} is.

Similarly, consider any set of desirable gambles $D \subseteq \mathcal{L}(\mathcal{X}^n)$. Then D is exchangeable if and only if there is a unique representing set of desirable gambles $\tilde{D} \subseteq \mathcal{V}^n(\Sigma_{\mathcal{X}})$ such that $D = \bigcup \tilde{M}_n^{-1}(\tilde{D})$. Furthermore, in that case, \tilde{D} is given by $\tilde{D} = M_n(D)$. Finally, D is coherent if and only if \tilde{D} is.

4. Countable Exchangeability

In the previous section, we assumed a finite sequence X_1, \dots, X_n to be exchangeable, and inferred representation theorems. In this section, we will consider the countable sequence X_1, \dots, X_n, \dots to be exchangeable, and derive representation theorems for such assessments. We will call $\mathcal{X}^{\mathbb{N}} := \bigtimes_{j \in \mathbb{N}} \mathcal{X}$, the set of all possible countable sequences where each variable takes values in \mathcal{X} .

First, we will need a way to relate gambles on different domains. Let f be some gamble on \mathcal{X}^n , and let f^* be its *cylindrical extension*, defined as

$$f^*(x_1, \dots, x_n, \dots) := f(x_1, \dots, x_n) \text{ for all } (x_1, \dots, x_n, \dots) \text{ in } \mathcal{X}^{\mathbb{N}}.$$

Formally, f^* belongs to $\mathcal{L}(\mathcal{X}^{\mathbb{N}})$ while f belongs to $\mathcal{L}(\mathcal{X}^n)$. However, they contain the same information, and therefore, are indistinguishable from a behavioural point of view. In this paper, we will identify f with its cylindrical extension f^* . Using this convention, we can for instance identify $\mathcal{L}(\mathcal{X}^n)$ with a subset of $\mathcal{L}(\mathcal{X}^{\mathbb{N}})$, and, as an other example, for any $\mathcal{A} \subseteq \mathcal{L}(\mathcal{X}^{\mathbb{N}})$, regard $\mathcal{A} \cap \mathcal{L}(\mathcal{X}^n)$ as those gambles in \mathcal{A} that depend upon the first n variables only.

4.1 Marginalisation

Using the notational convention we have just discussed, we can very easily define what marginalisation means for choice functions. Given any choice function C on $\mathcal{L}(\mathcal{X}^{\mathbb{N}})$ and any n in \mathbb{N} , its \mathcal{X}^n -marginal C_n is determined by $C_n(O) := C(O)$ for all O in $\mathcal{Q}(\mathcal{L}(\mathcal{X}^n))$.

Similarly, given any set of desirable gambles $D \subseteq \mathcal{L}(\mathcal{X}^{\mathbb{N}})$ and any n in \mathbb{N} , its \mathcal{X}^n -marginal D_n is defined by $D_n := D \cap \mathcal{L}(\mathcal{X}^n)$.

Coherence is preserved under marginalisation [it is an immediate consequence of the definition; see, amongst others, (De Cooman and Miranda, 2012, Proposition 6) for sets of desirable gambles].

Proposition 12 *Consider any coherent choice function C on $\mathcal{L}(\mathcal{X}^{\mathbb{N}})$ and any coherent set of desirable gambles $D \subseteq \mathcal{L}(\mathcal{X}^{\mathbb{N}})$. Then for every n in \mathbb{N} , their \mathcal{X}^n -marginals C_n and D_n are coherent.*

4.2 Gambles of Finite Structure

Before we can explain what it means to assess a countable sequence to be exchangeable, we need to realise that now there are infinitely many variables. From an operational point of view, it will be impossible to describe choosing between gambles that depend upon an infinite number of variables. Indeed, since we can never observe the actual outcome in a finite time, gambles will never be actually paid off, and hence every assessment is essentially without any risk. But, it does make operational and behavioural sense to consider choices between gambles of *finite structure*: gambles that each depend on a finite number of variables only. See (De Bock et al., 2016, Sec. 3.2) for a discussion.

Definition 13 (Gambles of finite structure) *We will call any gamble that depends only on a finite number of variables a gamble of finite structure. We collect all such gambles in $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$:*

$$\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}}) := \{f \in \mathcal{L}(\mathcal{X}^{\mathbb{N}}) : (\exists n \in \mathbb{N}) f \in \mathcal{L}(\mathcal{X}^n)\} = \bigcup_{n \in \mathbb{N}} \mathcal{L}(\mathcal{X}^n).$$

$\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$ is a linear space, with the usual ordering \leq : for any f and g in $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$, $f \leq g \Leftrightarrow f(x) \leq g(x)$ for all x in $\mathcal{X}^{\mathbb{N}}$.

Due to our finitary context, we can even establish a converse result to Prop. 12, whose proof for the part about sets of desirable gambles can be found in (De Bock et al., 2016, Proposition 4), and for the part about choice functions is a straightforward verification of all the axioms.

Proposition 14 *Consider any choice function C on $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$, and any set of desirable gambles $D \subseteq \bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$. If for every n in \mathbb{N} , its \mathcal{X}^n -marginal C_n on $\mathcal{L}(\mathcal{X}^n)$ is coherent, then C is coherent. Similarly, if for every n in \mathbb{N} , its \mathcal{X}^n -marginal $D_n \subseteq \mathcal{L}(\mathcal{X}^n)$ is coherent, then D is coherent.*

4.3 Set of indifferent gambles

If a subject assesses the sequence of variables X_1, \dots, X_n, \dots to be exchangeable, this means that he is indifferent between any gamble f in $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$ and its permuted variant $\pi^t f$, for any π in \mathcal{P}_n , where n now is the (finite) number of variables that f depends upon: his set of indifferent gambles is

$$I_{\mathcal{P}} := \{f \in \bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}}) : (\exists n \in \mathbb{N}) f \in I_{\mathcal{P}_n}\} = \bigcup_{n \in \mathbb{N}} I_{\mathcal{P}_n}.$$

If we want to use $I_{\mathcal{P}}$ to define countable exchangeability, it must be a coherent set of indifferent gambles.

Proposition 15 *The set $I_{\mathcal{P}}$ is a coherent set of indifferent gambles.*

Countable exchangeability is now easily defined, similar to the definition for the finite case.

Definition 16 *A choice function C on $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$ is called (countably) exchangeable if C is compatible with $I_{\mathcal{P}}$. Similarly, a set of desirable gambles $D \subseteq \bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$ is called (countably) exchangeable if it is compatible with $I_{\mathcal{P}}$.*

This definition is closely related to its finite counterpart.

Proposition 17 *Consider any coherent choice function C on $\bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$. Then C is exchangeable if and only if for every choice of n in \mathbb{N} , the \mathcal{X}^n -marginal C_n of C is exchangeable. Similarly, consider any coherent set of desirable gambles $D \subseteq \bar{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$. Then D is exchangeable if and only if for every choice of n in \mathbb{N} , the \mathcal{X}^n -marginal D_n of D is exchangeable.*

4.4 A Representation Theorem for Countable Sequences

We will look for a similar representation result. However, since we no longer deal with finite sequences of length n , now the representing choice function won't be defined on $\mathcal{V}^n(\Sigma_{\mathcal{X}})$, but instead on $\mathcal{V}(\Sigma_{\mathcal{X}})$.

$$\begin{array}{ccccc} & & \text{CoM}_n & & \\ & \swarrow H_n & & \searrow M_n & \\ \mathcal{L}(\mathcal{N}^n) & & \mathcal{L}(\mathcal{X}^n) & & \mathcal{V}^n(\Sigma_{\mathcal{X}}) \\ \searrow \tilde{H}_n & & \downarrow [-] & & \downarrow \\ & & \mathcal{L}(\mathcal{X}^n)/I_{\mathcal{P}_n} & & \mathcal{V}(\Sigma_{\mathcal{X}}) \end{array}$$

In the commuting diagram, a dashed line represents an embedding: indeed, for every n in \mathbb{N} , $\mathcal{V}^n(\Sigma_{\mathcal{X}})$ is a subspace of $\mathcal{V}(\Sigma_{\mathcal{X}})$. That shows the importance of the polynomial representation.

As we have seen, in order to define coherent choice functions on some linear space, we need to provide it with a vector ordering. Similar to what we did before, we use the proper cone $\{0\} \cup \text{posi}(\{B_m : m \in \mathcal{N}^n, n \in \mathbb{N}\})$ to define the order \leq_B on $\mathcal{V}(\Sigma_{\mathcal{X}})$:

$$h_1 \leq_B h_2 \Leftrightarrow h_2 - h_1 \in \{0\} \cup \text{posi}(\{B_m : m \in \mathcal{N}^n, n \in \mathbb{N}\})$$

for all h_1 and h_2 in $\mathcal{V}(\Sigma_{\mathcal{X}})$.

Keeping Props. 12 and 14 in mind, the following result is not surprising.

Proposition 18 Consider any choice function C' on $\mathcal{V}(\Sigma_{\mathcal{X}})$. Then C' is coherent if and only if for every n in \mathbb{N} the choice function C'_n , given by $C'_n(O) := C'(O)$ for all O in $\mathcal{Q}(\mathcal{V}^n(\Sigma_{\mathcal{X}}))$ is coherent.

Theorem 19 (Countable Representation) Consider any choice function C on $\tilde{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$. Then C is exchangeable if and only if there is a unique representing choice function \tilde{C} on $\mathcal{V}(\Sigma_{\mathcal{X}})$ such that, for every n in \mathbb{N} , the \mathcal{X}^n -marginal C_n of C is determined by

$$C_n(O) = \{f \in O : M_n(f) \in \tilde{C}(M_n(O))\} \text{ for all } O \text{ in } \mathcal{Q}(\mathcal{L}(\mathcal{X}^n)).$$

Furthermore, in that case, \tilde{C} is given by $\tilde{C}(O) := \bigcup_{n \in \mathbb{N}} \tilde{C}_n(O \cap \mathcal{V}^n(\Sigma_{\mathcal{X}}))$ for all O in $\mathcal{Q}(\mathcal{V}(\Sigma_{\mathcal{X}}))$, with $\tilde{C}_n(M_n(O)) := M_n(C_n(O))$ for every O in $\mathcal{Q}(\mathcal{L}(\mathcal{X}^n))$, and where we let $\tilde{C}_n(\emptyset) := \emptyset$ for notational convenience. Finally, C is coherent if and only if \tilde{C} is.

Similarly, consider any set of desirable gambles $D \subseteq \tilde{\mathcal{L}}(\mathcal{X}^{\mathbb{N}})$. Then D is exchangeable if and only if there is a unique representing $\tilde{D} \subseteq \mathcal{V}(\Sigma_{\mathcal{X}})$ such that, for every n in \mathbb{N} , the \mathcal{X}^n -marginal D_n is given by $D_n = \bigcup \tilde{M}_n^{-1}(\tilde{D} \cap \mathcal{V}^n(\Sigma_{\mathcal{X}}))$. Furthermore, in that case, \tilde{D} is given by $\tilde{D} = \bigcup_{n \in \mathbb{N}} M_n(D_n)$. Finally, D is coherent if and only if \tilde{D} is.

5. Conclusion

We have studied exchangeability and we have found counterparts to de Finetti's finite and countable representation results, in the general setting of choice functions. We have shown that an exchangeability assessment is a particular indifference assessment, where we identified the set of indifferent options. The main idea that made (finite) representation possible is the linear order isomorphism \tilde{H}_n^{-1} between the quotient space and the set of gambles on count vectors, indicating that (finitely) exchangeable choice functions can be represented by a choice function that essentially represents preferences between urns with n balls of types \mathcal{X} with different compositions m . Alternatively, for the countable case, we have shown that there is a polynomial representation.

Choice functions form a belief structure (Van Camp et al., 2017). Therefore, any infimum of coherent choice functions is a coherent choice function itself. Since any infimum of choice functions compatible with some fixed set of indifferent options I , is compatible with I as well (Van Camp et al., 2017), our results indicate that, using choice functions, it is conceptually easy to reason about exchangeable sequences: infima of exchangeable and coherent choice functions will be exchangeable and coherent as well.

A possible future goal is to investigate how exchangeability behaves under updating. In (De Cooman and Quaeghebeur, 2012) it is shown that, for exchangeable sets of desirable gambles, updating can be done directly for the representing set of desirable gambles in the count space. We expect this to be the case for choice functions as well. Other possible extensions are to develop a framework for partial exchangeability, and to model other structural judgements, such as an irrelevance assessment.

Acknowledgments

Arthur Van Camp's research was partly funded by Banco Santander via Campus de Excelencia Internacional. Gert de Cooman's research was partly funded through project number 3G012512 of the Research Foundation Flanders (FWO). The authors would like to thank Enrique Miranda and three anonymous referees for their valuable comments.

References

- K. J. Arrow. *Social choice and individual values*. Cowles Foundation Monographs Series. Yale University Press, 1951.
- J. De Bock, A. Van Camp, M. A. Diniz, and G. de Cooman. Representation theorems for partially exchangeable random variables. *Fuzzy Sets and Systems*, 284:1–30, 2016. doi:[10.1016/j.fss.2014.10.027](https://doi.org/10.1016/j.fss.2014.10.027).
- G. de Cooman and E. Miranda. Irrelevance and independence for sets of desirable gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012. doi:[10.1613/jair.3770](https://doi.org/10.1613/jair.3770).
- G. de Cooman and E. 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](https://doi.org/10.1016/j.ijar.2010.12.002). Precisely imprecise: A collection of papers dedicated to Henry E. Kyburg, Jr.
- G. de Cooman, E. Quaeghebeur, and E. Miranda. Exchangeable lower previsions. *Bernoulli*, 15(3):721–735, 2009. doi:[10.3150/09-BEJ182](https://doi.org/10.3150/09-BEJ182). URL <http://hdl.handle.net/1854/LU-498518>.
- B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l’Institut Henri Poincaré*, 7:1–68, 1937. English translation in ([Kyburg Jr. and Smokler, 1964](#)).
- N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, 1997.
- J. B. Kadane, M. J. Schervish, and T. Seidenfeld. A Rubinesque theory of decision. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 45:45–55, 2004. doi:[10.1214/lnms/1196285378](https://doi.org/10.1214/lnms/1196285378). URL <http://www.jstor.org/stable/4356297>.
- H. E. Kyburg Jr. and H. E. Smokler, editors. *Studies in Subjective Probability*. Wiley, New York, 1964. Second edition (with new material) 1980.
- H. 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. doi:[10.1524/strm.1987.5.12.47](https://doi.org/10.1524/strm.1987.5.12.47).
- T. Seidenfeld. Decision theory without “independence” or without “ordering”. *Economics and Philosophy*, 4:267–290, Oct. 1988. doi:[10.1017/S0266267100001085](https://doi.org/10.1017/S0266267100001085).
- T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172(1):157–176, 2010. doi:[10.1007/s11229-009-9470-7](https://doi.org/10.1007/s11229-009-9470-7).
- H. Uzawa. Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics*, 8:35–40, 1956. doi:[10.1007/BF02863564](https://doi.org/10.1007/BF02863564).
- A. Van Camp, G. de Cooman, E. Miranda, and E. Quaeghebeur. Coherent choice functions, desirability and indifference. *Fuzzy sets and systems*, 2017. Submitted for publication.
- J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1944. ISBN 0691041830. Third edition 1953.
- P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Computing Minimax Decisions with Incomplete Observations

Thijs van Ommen

*Universiteit van Amsterdam
Amsterdam (The Netherlands)*

T.VANOMMEN@UVA.NL

Abstract

Decision makers must often base their decisions on incomplete (coarse) data. Recent research has shown that in a wide variety of coarse data problems, minimax optimal strategies can be recognized using a simple probabilistic condition. This paper develops a computational method to find such strategies in special cases, and shows what difficulties may arise in more general cases.

Keywords: coarse data; incomplete observations; minimax decision making.

1. Introduction

Suppose that we are faced with a decision where the loss we will incur depends on the outcome x of a random experiment. While the distribution of x is known, our observation of x is incomplete: we only have access to a *coarse* observation y , a set that we know includes x but may also include other elements. An infamous example of this problem is the Monty Hall puzzle ([Selvin, 1975](#)).

Example 1 *In a game show, a car is hidden uniformly at random behind one of three doors. The contestant picks a door; we will assume the middle one. But now the quizmaster steps in and opens one of the two remaining doors, revealing a goat behind it. Should the contestant switch to the remaining door, or stick with his initial guess?*

We will make the standard assumptions that the quizmaster always opens a door, always with a goat behind it. Then this is an instance of the incomplete data problem, where we will either observe $y = \{\text{left, middle}\}$ (if the quizmaster opens the rightmost door) or $y = \{\text{middle, right}\}$ (if he opens the leftmost door). It is well known—but quite surprising—that it is wrong to conclude the remaining doors now each have probability 1/2 of hiding the car. But then what probability distribution (or set of distributions) should we use to base our decision on?

A key issue here is that we do not know the *coarsening mechanism*, the random process that maps the true outcome x to the set y we observe. A common assumption about this mechanism is *coarsening at random (CAR)*, which says that for each set y , the probability that the coarsening mechanism reports y is the same no matter which outcome $x \in y$ is the true outcome ([Heitjan and Rubin, 1991](#)). But this is a strong assumption that often fails to hold in practice; in fact, in the Monty Hall puzzle, it can never hold ([Grünwald and Halpern, 2003](#); [Gill and Grünwald, 2008](#)).

An approach that avoids any assumptions on the coarsening mechanism is to model the problem using the credal set \mathcal{P} of all joint distributions P on (x, y) that are (a) consistent with the known distribution of x , and (b) satisfy $P(x, y) = 0$ for $x \notin y$. This (convex) set \mathcal{P} represents both our aleatory uncertainty about x and our epistemic uncertainty about its relation with y . To then incorporate an observation y , the generalized Bayes rule can be used; [De Cooman and Zaffalon \(2004\)](#) apply this approach to coarse data problems. The resulting posterior on the outcomes exhibits *dilation* ([Seidenfeld and Wasserman, 1993](#)): the prior was a precise distribution, but the posterior

may be a large set of distributions. If we want to be sure that the true distribution of x given y is included in this set, then this phenomenon is unavoidable. However, it may lead to suboptimal decisions, as described by [Augustin \(2003\)](#), [Grünwald and Halpern \(2011\)](#), and others.

By formulating a strategy before making an observation, the effect of dilation on decisions can be avoided ([Seidenfeld, 2004](#)). This approach has been investigated for coarse data problems by [Van Ommen et al. \(2016\)](#), who found that for many situations, minimax strategies are characterized by the *RCAR condition* (which looks like the CAR condition, but with x and y reversed).

To apply these results in practice, we would like efficient computational methods to find RCAR strategies. How difficult this is depends largely on the family of possible observations. In this paper, we describe a computational method for a restricted class of such families. This reveals a relation between minimax optimal strategies and statistical independence. We also point out the various computational difficulties that may occur in larger classes of families: there finding an exact solution may involve a combinatorial search, or solving polynomial (rather than linear) equations.

This paper is structured as follows. In Section 2, we summarize the relevant results of [Van Ommen et al. \(2016\)](#). Section 3 introduces the main tool: homogeneous induced colourings. These may not exist for all families of possible observations, which leads to a categorization of such families. A computational procedure, and its limitations, are described in Section 4. Section 5 interprets this procedure for the families where it is guaranteed to work. Section 6 concludes.

The contents of Sections 3 to 5 are adapted from Chapter 7 of PhD thesis ([Van Ommen, 2015](#)).

2. Optimality of RCAR Strategies

This section summarizes the main results from [Van Ommen et al. \(2016\)](#). We consider coarse data decision problems with finite outcome space \mathcal{X} . The decision maker must pick an action based on a coarse observation $y \subseteq \mathcal{X}$, which we call a *message*. The choice of action will depend not only on the received message, but on the entire set of messages that the coarsening mechanism might produce: the *message structure* $\mathcal{Y} \subset 2^{\mathcal{X}}$ (in the Monty Hall example, this is $\mathcal{Y} = \{\{\text{left, middle}\}, \{\text{middle, right}\}\}$). It will also depend on the (known) distribution p of the outcomes, which we assume to be nowhere zero, and on the loss function $L : \mathcal{X} \times \mathcal{A} \rightarrow [0, \infty]$.¹ We assume throughout that L satisfies the technical conditions in ([Van Ommen et al., 2016](#), Theorem 3); these are in particular satisfied for all finite L , and also for logarithmic loss $L(x, Q) = -\log Q(x)$.

The decision problem is modelled as a zero-sum game between a quizmaster and a contestant. The (imaginary) quizmaster picks as strategy a joint distribution P on $\mathcal{X} \times \mathcal{Y}$ from the credal set $\mathcal{P} = \{P \mid \sum_y P(x, y) = p_x \text{ for all } x, P(x, y) = 0 \text{ for all } y \in \mathcal{Y}, x \notin y\}$. Simultaneously, the contestant picks as strategy a function $A : \mathcal{Y} \rightarrow \mathcal{A}$. The two players seek to maximize resp. minimize the expected loss

$$\sum_{x,y} P(x, y) L(x, A(y)) = \sum_x p_x \sum_{y \in \mathcal{Y}, y \ni x} P(y \mid x) L(x, A(y)),$$

where the second expression reflects that the quizmaster's influence is limited to $P(y \mid x)$, with x always sampled from the fixed marginal p . Strategies achieving this maximum/minimum are called *worst-case optimal*. If the action space is rich enough, this game has a Nash equilibrium; then

1. In ([Van Ommen et al., 2016](#)), the decision maker's action space \mathcal{A} is always taken to be the set of distributions on \mathcal{X} , and the actions are interpreted as probability updates. But due to the generality of the loss functions allowed, the same theory can be applied for arbitrary action spaces.

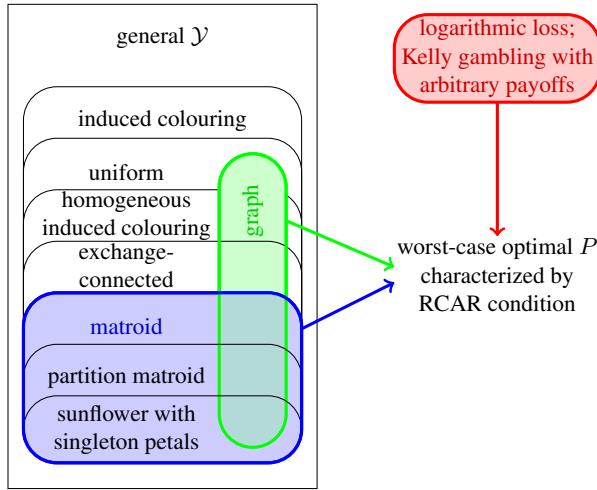


Figure 1: Overview of classes of message structures. As shown by Van Ommen et al. (2016), the RCAR condition characterizes worst-case optimal P for many games, including all graph and matroid games. The other classes shown in this figure are defined and explored in Sections 3 and 5.

given a worst-case (maximin) optimal P for the quizmaster, we can easily determine a worst-case (minimax) optimal A for the contestant.² Thus we focus on finding a worst-case optimal P .

For several classes of games, a strategy P is worst-case optimal if, for some vector $q \in [0, 1]^{\mathcal{X}}$, it satisfies the *RCAR condition*:

$$\begin{aligned} q_x &= P(x \mid y) \text{ for all } y \in \mathcal{Y} \text{ with } P(y) > 0 \text{ and all } x \in y, \text{ and} \\ \sum_{x \in y} q_x &\leq 1 \text{ for all } y \in \mathcal{Y}. \end{aligned} \tag{1}$$

Such a strategy is called an *RCAR strategy*, and the vector q is called an *RCAR vector*. The three classes of games where this holds are illustrated in Figure 1: if L is logarithmic loss or an affine transformation of it (this represents Kelly gambling games with arbitrary payoffs); if \mathcal{Y} is a graph (each message consists of two outcomes); and if \mathcal{Y} is a matroid (defined in (3) below). In the latter two cases, there is also a symmetry condition on the loss function L . What is surprising here is that, as long as we are in one of these cases, we can find a worst-case optimal P without knowing what the loss function is, because the RCAR condition is purely probabilistic and does not depend on L . The rest of this paper deals with the problem of computing an RCAR strategy for a given game.

3. Induced Colourings

Fix a set $\mathcal{Y}' \subseteq \mathcal{Y}$ with $\bigcup_{y \in \mathcal{Y}'} y = \mathcal{X}$, and assume that an RCAR strategy P exists with support $\mathcal{Y}_P := \{y \in \mathcal{Y} \mid P(y) > 0\}$ equal to \mathcal{Y}' . (For example, we may in many cases take $\mathcal{Y}' = \mathcal{Y}$.) We will now consider different properties of \mathcal{Y}_P that may help us find P . The classes of message

² In many cases, $A(y)$ is simply the optimal response to $P(\cdot \mid y)$ for each $y \in \mathcal{Y}$. This is not always well-defined; a general solution is given by Theorem 7 of (Van Ommen et al., 2016), using Theorem 3 to determine λ^* .

structures defined by these properties, and the inclusion relations between them that we establish here, are shown graphically in Figure 1, and examples are given in Figure 2.

Consider the system of linear equations

$$\sum_{x \in y} q_x = 1 \text{ for all } y \in \mathcal{Y}'. \quad (2)$$

If an RCAR strategy $P \in \mathcal{P}$ exists with support \mathcal{Y}' and RCAR vector q , then q is positive and satisfies (2). The converse is not true: if $\mathcal{Y}' \neq \mathcal{Y}$, then (1) additionally imposes inequalities on messages $y \in \mathcal{Y} \setminus \mathcal{Y}'$. We will start our search for RCAR strategies by examining the solutions of (2). (A similar system is studied in the CAR literature, where it plays a role in characterizing message structures that admit a CAR coarsening mechanism; see Grünwald and Halpern (2003); Jaeger (2005); Gill and Grünwald (2008). Since we study RCAR rather than CAR, the roles of outcomes and messages are reversed here.)

Define a *colouring* as a partition of \mathcal{X} . We say a colouring is *induced* by a set of messages \mathcal{Y}' if the system of linear equations (2) has at least one solution q with $q_x > 0$ for all x , and x, x' are in the same class of the colouring ('have the same colour') if and only if $q_x = q_{x'}$ for all such solutions to that system (in other words, the colour classes are the equivalence classes of this relation on \mathcal{X}). If the system has at least one positive solution, then the colouring induced by \mathcal{Y}' is unique; otherwise, there is no induced colouring.

We say a colouring is *homogeneous on \mathcal{Y}'* if the number of outcomes of each colour is the same for every message in \mathcal{Y}' (for example, if each message consists of one 'red' and two 'blue' outcomes). This is only possible if \mathcal{Y}' is *uniform*: all messages in \mathcal{Y}' have the same size. We are interested in \mathcal{Y}' whose induced colouring is homogeneous. One class of such \mathcal{Y}' is defined in terms of pairs of messages y_1, y_2 that differ by the *exchange* of one outcome, meaning that $|y_1 \setminus y_2| = |y_2 \setminus y_1| = 1$. We call \mathcal{Y}' *exchange-connected* if, for each pair of messages $y^{(a)}, y^{(b)} \in \mathcal{Y}'$, there exists a sequence of messages $y_1, y_2, \dots, y_\ell \in \mathcal{Y}'$ (an *exchange-path*) with $y_1 = y^{(a)}$ and $y_\ell = y^{(b)}$ whose adjacent messages differ by the exchange of one outcome. Finally, \mathcal{Y}' is a *matroid* if it satisfies the *basis exchange* property: for all $y_1, y_2 \in \mathcal{Y}'$ and $x_1 \in y_1 \setminus y_2$,

$$(y_1 \setminus \{x_1\}) \cup \{x_2\} \in \mathcal{Y}' \text{ for some } x_2 \in y_2 \setminus y_1. \quad (3)$$

Figure 2 illustrates these definitions with a few examples. Each table represents a message structure \mathcal{Y} as an incidence matrix: each row represents a message, and (coloured) stars mark the outcomes it contains.

The message structure shown in Figure 2a has no induced colouring: any solution of (2) must have $q_{x_3} = 1 - q_{x_4} = q_{x_5} = 1 - q_{x_1}$ and thus $q_{x_2} = 0$, so there is no positive solution, and it follows that no RCAR strategy P exists with $P(y) > 0$ for all $y \in \mathcal{Y}'$. On the other hand, any uniform game has an induced colouring, because there is at least one solution to (2):

$$q_x = 1/k \quad \text{for all } x \in \mathcal{X}, \quad (4)$$

where k is the size of the game's messages.

Figures 2b and 2c are examples of message structures that do have an induced colouring, but one that is not homogeneous. In both these examples, all outcomes have different colours in the induced colouring, because no pair of outcomes necessarily has the same value of q in a solution of (2). The message structure shown in Figure 2c will be revisited in Example 2 in the next section.

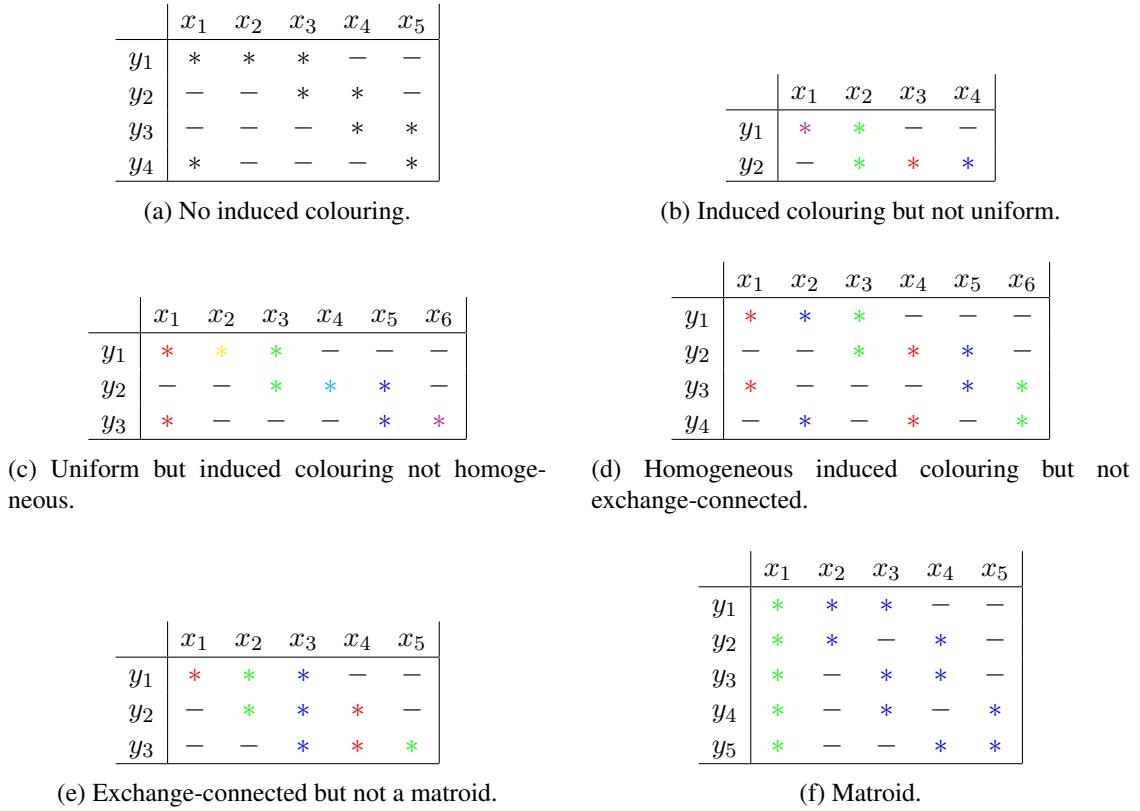


Figure 2: Examples of messages structures and their induced colourings.

The three remaining message structures do have homogeneous induced colourings. Figure 2d shows that it is possible for a message structure to have a homogeneous induced colouring without being exchange-connected. In this message structure, which adds the message y_4 to the structure in Figure 2c, each pair of messages differs by *two* exchanges. Yet the added message changes the induced colouring: for example, $q_{x_1} = q_{x_4}$ follows because by the equalities from (2) on y_1 and y_3 , $1 - q_{x_1} = q_{x_2} + q_{x_3} = q_{x_5} + q_{x_6}$, and by y_2 and y_4 , $1 - q_{x_4} = q_{x_3} + q_{x_5} = q_{x_2} + q_{x_6}$; thus $2 - 2q_{x_1} = 2 - 2q_{x_4} = q_{x_2} + q_{x_3} + q_{x_5} + q_{x_6}$.

The message structure shown in Figure 2e is exchange-connected. For such structures, it is easy to determine the induced (homogeneous) colouring: if messages y_1, y_2 differ by the exchange of one outcome (x_1 for x_2), then any solution of (2) must satisfy $q_{x_1} = q_{x_2}$, so such x_1, x_2 must be the same colour. Any vector q that satisfies all these equalities and satisfies $\sum_{x \in y} q_x = 1$ for any one message $y \in \mathcal{Y}'$ satisfies (2) for all messages in \mathcal{Y}' , so this determines the induced colouring. This colouring is clearly homogeneous on any pair of message that differ by the exchange of one outcome; because exchange-paths exist between all pairs of messages, it follows that the induced colouring of an exchange-connected game is homogeneous.

Finally, the class of matroid games is a subclass of exchange-connected games: (3) requires the existence of not just one, but possibly many different exchange-paths between any pair of messages. Figure 2f gives an example. The structure in Figure 2e is not a matroid: there is no outcome in $y_3 \setminus y_1$ that can be added to $y_1 \setminus \{x_2\} = \{x_1, x_3\}$ to make a message.

The following lemma gives two alternate characterizations of the induced colouring of a matroid. The first of these is in terms of a concept from matroid theory: the colour classes of the induced colouring coincide with the *2-connected components* of the matroid. (We refer to Oxley (2011) for the definition.) We observed above (when discussing Figure 2e) that if messages exist that differ in the exchange of one outcome, then the outcomes being exchanged must be the same colour. The second characterization shows that for matroids, the converse also holds.

Lemma 1 (Matroid colouring) *Given a matroid $(\mathcal{X}, \mathcal{Y})$ and two elements $x_1, x_2 \in \mathcal{X}$, the following statements are equivalent:*

1. *x_1 and x_2 are in the same colour class of the induced colouring of \mathcal{Y} ;*
2. *x_1 and x_2 are in the same 2-connected component of $(\mathcal{X}, \mathcal{Y})$;*
3. *There exist $y_1, y_2 \in \mathcal{Y}$ such that $y_1 \setminus y_2 = \{x_1\}$ and $y_2 \setminus y_1 = \{x_2\}$.*

4. A Computational Procedure for Finding RCAR Strategies

Consider the case that \mathcal{Y}' induces a homogeneous colouring, and assume as before that an RCAR strategy P exists with $\mathcal{Y}_P = \mathcal{Y}'$. Then the corresponding RCAR vector q must be a solution of the linear system (2). Additionally, P must agree with the marginal p . These constraints allow us to compute the vector q directly.

Let S be the set of all outcomes with a particular colour. Then there is some value q_S such that $P(x \mid y) = q_x = q_S$ for all $y \in \mathcal{Y}$, $x \in S \cap y$. Let $k_S = |S \cap y|$ (this is independent of y by homogeneity). We must have

$$k_S q_S = k_S q_S \sum_y P(y) = \sum_y k_S P(y) q_S = \sum_y \sum_{x \in S \cap y} P(y) q_S = \sum_{x \in S} \sum_{y \ni x} P(y) P(x \mid y) = \sum_{x \in S} p_x,$$

so that q_S can be computed by

$$q_S = \frac{1}{k_S} \sum_{x \in S} p_x. \quad (5)$$

A simple case is when the induced colouring assigns the same colour to all outcomes: then we see that as in (4), we get $q_x = 1/k$ for all $x \in \mathcal{X}$, where k is the size of the messages. When a colour consists of just one outcome x (which must then be an element of every message for the colouring to be homogeneous), we find $q_x = p_x$.

If an RCAR strategy P exists with $\mathcal{Y}_P = \mathcal{Y}'$ where \mathcal{Y}' induces a homogeneous colouring, then P must have the vector q determined by (5) as its RCAR vector. However, it may be the case that no such strategy exists. To find P if it exists, we still need to determine the $P(y)$'s. We can find a nonnegative solution or determine that no nonnegative solution exists by solving the following linear programming problem (which we can do in polynomial time):

$$\begin{aligned} & \text{maximize} && \sum_{y \in \mathcal{Y}} r_y \\ & \text{subject to} && \sum_{y \ni x} r_y \leq \frac{p_x}{q_x} \quad \text{for all } x \in \mathcal{X}, \end{aligned} \quad (6)$$

with $r \in \mathbf{R}_{\geq 0}^{\mathcal{Y}}$. If a vector achieving $\sum_{y \in \mathcal{Y}} r_y = 1$ is found, we have a strategy P with r as the marginal on messages ($P(x, y) = q_x r_y$ for all $x \in \mathcal{X}$). If no vector r achieves the value 1, there is no RCAR strategy P satisfying the assumption $\mathcal{Y}_P = \mathcal{Y}'$.

Now we may want to apply this procedure in practice to find an RCAR strategy for a given game. (By Lemma 11 from [Van Ommen et al. \(2016\)](#), such a strategy always exists.)

When doing so we encounter two problems: we need to provide the procedure with an \mathcal{Y}' such that $\bigcup \mathcal{Y}' = \mathcal{X}$, and even if we have an idea about what \mathcal{Y}' to take, it may not have a homogeneous induced colouring. Still, let us investigate what happens if we just guess an \mathcal{Y}' . We will then encounter one of the cases 1, 2a-2c which we now describe. Briefly, in case 1, the procedure cannot be used because it cannot determine q , and in case 2a and 2b it gives an inconclusive result; in case 2c we have success. We now consider each case in detail.

1. \mathcal{Y}' has no homogeneous induced colouring.

In this case, the procedure is not applicable. Indeed, finding an RCAR vector may be a more difficult type of problem, as illustrated by the following example which uses the message structure from Figure 2c. (This example is a uniform game; the class of uniform games is the smallest class among those identified in the previous section that strictly contains the class of games with a homogeneous induced colouring.)

Example 2 (Irrational RCAR vector) Consider the problem with $\mathcal{X} = x_1, \dots, x_6$, $\mathcal{Y} = \{y_1 = \{x_1, x_2, x_3\}, y_2 = \{x_3, x_4, x_5\}, y_3 = \{x_1, x_5, x_6\}\}$, and marginal p and strategy P given by the following table:

P	x_1	x_2	x_3	x_4	x_5	x_6
y_1	1/10	1/10	$\frac{3}{10} - \frac{1}{10}\sqrt{5}$	—	—	—
y_2	—	—	$\frac{1}{10}\sqrt{5} - \frac{1}{10}$	1/5	$\frac{1}{10}\sqrt{5} - \frac{1}{10}$	—
y_3	1/10	—	—	—	$\frac{3}{10} - \frac{1}{10}\sqrt{5}$	1/10
q_x	$\frac{1}{4} + \frac{1}{20}\sqrt{5}$	$\frac{1}{4} + \frac{1}{20}\sqrt{5}$	$\frac{1}{2} - \frac{1}{10}\sqrt{5}$	$\frac{1}{5}\sqrt{5}$	$\frac{1}{2} - \frac{1}{10}\sqrt{5}$	$\frac{1}{4} + \frac{1}{20}\sqrt{5}$
p_x	1/5	1/10	1/5	1/5	1/5	1/10

The strategy P is RCAR, with the vector q that is also shown in the table. We see that the RCAR strategy P and RCAR vector q (both of which are unique) contain irrational numbers, while the marginal p was rational. The solution techniques used in this section (the formula (5) for q and linear optimization for (6)) do not yield irrational results when given rational inputs, so this example shows that these techniques will not suffice in general for games that do not have a homogeneous induced colouring. (General-purpose convex optimization techniques could be used here instead.)

Conclusion: in this case, an RCAR strategy P with $\mathcal{Y}_P = \mathcal{Y}'$ may exist, but it may be not be easy to find. So in general, for such \mathcal{Y}' , we do not know how to efficiently determine if such a P exists.

2. \mathcal{Y}' does have a homogeneous induced colouring.

In this case, we can use (5) to compute a candidate q for the RCAR vector. We distinguish three subcases:

2a. If $\mathcal{Y}' \neq \mathcal{Y}$, there may be a message $y \in \mathcal{Y} \setminus \mathcal{Y}'$ for which $\sum_{x \in y} q_x > 1$.

This may happen because the described procedure ignores the existence of messages not in \mathcal{Y}' . However, the RCAR condition (1) puts an inequality constraint on $\sum_{x \in y} q_x$ even for messages y with $P(y) = 0$. If the vector q computed by (5) does not satisfy this constraint, then q is not an RCAR vector: we chose the wrong \mathcal{Y}' .

2b. No solution r of (6) achieves $\sum_{y \in \mathcal{Y}} r_y = 1$.

This also means that our choice of \mathcal{Y}' was incorrect.

2c. Otherwise, q is an RCAR vector, and together with r determines an RCAR strategy P .

In this case, we can report success.

In cases 2a and 2b, \mathcal{Y}' has a homogeneous induced colouring but we find that no RCAR strategy P exists with $\mathcal{Y}_P = \mathcal{Y}'$. Then we may face two problems. First, it is not clear how we might choose a different \mathcal{Y}' on which to try the procedure next. For small message structures, it may be feasible to try all candidates. For larger structures, the number of possible choices grows exponentially, and a more efficient way of searching would be needed.

The second problem is that in general, \mathcal{Y}' might not induce a homogeneous colouring even though \mathcal{Y} does. For example, if \mathcal{Y} is the message structure shown in Figure 2e, but there is no RCAR strategy P with $\mathcal{Y}_P = \mathcal{Y}$ for our marginal p , we have to conclude that the RCAR strategy must have $\mathcal{Y}_P = \{y_1, y_3\}$ (because this is the only other choice of \mathcal{Y}' that satisfies $\bigcup \mathcal{Y}' = \mathcal{X}$). However, this message structure is no longer exchange-connected, and in fact does not have a homogeneous induced colouring, so that we end up in case 1.

In Section 5, we will see a subclass of matroid games for which the procedure is guaranteed to succeed for the choice $\mathcal{Y}' = \mathcal{Y}$. So for that class of inputs, the procedure discussed here is an efficient algorithm for finding an RCAR strategy (which is worst-case optimal for any loss function by the results of [Van Ommen et al. \(2016\)](#)).

Two efficient algorithms, for graph games and for matroid games, are given in [Van Ommen \(2015, Chapter 8\)](#). These algorithms can also be viewed as instances of the computational procedure in this section: both essentially compute q and r as we did here; then, if $\sum_{y \in \mathcal{Y}} r < 1$, they pick a new set \mathcal{Y}' , guided by properties of the linear optimization problem (6). The choice of \mathcal{Y}' is such that each new \mathcal{Y}' is a subset of the previous \mathcal{Y}' (i.e. no backtracking is needed), and such that case 2a will never occur.

Case 1 will never occur either for these algorithms: the chosen \mathcal{Y}' will always have a homogeneous induced colouring. This happens for different reasons for the two cases of graph and matroid games. These reasons shed light on what makes graphs and matroids special among more general message structures, so we conclude this section by giving brief explanations.

For graphs: Any connected component of a graph is also exchange-connected, and thus induces a homogeneous colouring. While some choices of \mathcal{Y}' may produce a disconnected graph $(\mathcal{X}, \mathcal{Y}')$, each component of this graph will have a homogeneous induced colouring, and the algorithm can be applied to each of these components recursively.

For matroids: On a matroid game, for any RCAR strategy P , \mathcal{Y}_P determines a homogeneous colouring. (This colouring is not induced in the usual sense, but is uniquely determined by the equalities on \mathcal{Y}_P combined with inequalities for $\mathcal{Y} \setminus \mathcal{Y}_P$; see ([Van Ommen et al., 2016](#), proof of Theorem 19) for details.) The conditional probabilities $P(x | y)$ respect this colouring.

5. Partition Matroids

We now describe a class of games for which a worst-case optimal strategy can be completely computed using the procedure from the previous section, because regardless of the marginal p , we can take $\mathcal{Y}' = \mathcal{Y}$ and the procedure will succeed in finding an RCAR strategy.

A message structure \mathcal{Y} is called a *partition matroid* if \mathcal{X} can be partitioned into nonempty sets S_1, \dots, S_k such that \mathcal{Y} consists of all subsets of \mathcal{X} that take one element from each of the sets S_i ([Oxley, 2011](#)). This class forms a subclass of matroids, so if \mathcal{Y} is a partition matroid, it induces a homogeneous colouring. Using Lemma 1, it is easy to see that this colouring is given by the sets S_i . An example of a partition matroid is given in Figure 3a; the matroid we saw in Figure 2f is not a partition matroid.

	x_1	x_2	x_3	x_4	x_5
y_1	*	—	*	—	—
y_2	*	—	—	*	—
y_3	*	—	—	—	*
y_4	—	*	*	—	—
y_5	—	*	—	*	—
y_6	—	*	—	—	*

(a) Partition matroid but not a sunflower.

	x_1	x_2	x_3	x_4	x_5
y_1	*	*	*	—	—
y_2	*	*	—	*	—
y_3	*	*	—	—	*

(b) Sunflower with singleton petals.

Figure 3: More examples of messages structures and their induced colourings.

As an illustration, suppose a shopkeeper sells items of brands x_1 and x_2 , in colours x_3, x_4 and x_5 , and customers buy items based on a preference for either a brand or a colour. The shopkeeper observes a customer buying an item, but would like to know the underlying preference for recommendation purposes. This coarse data problem corresponds to the partition matroid in Figure 3a.

Because a partition matroid induces a homogeneous colouring, we can carry out the procedure described in the previous section to find for each x that $q_x = \sum_{x' \in S_i} p_{x'}$, where S_i is the set containing x . Now a solution for the $P(y)$'s that satisfies $\sum_{y \ni x} P(y)q_x = p_x$ always exists:

$$P(y) = \prod_{x \in y} \frac{p_x}{q_x}.$$

In words, this means that given the true outcome x , it is worst-case optimal for the quizmaster to choose a message by randomly sampling an outcome from each set $S_i \not\ni x$ according to the marginal probabilities conditioned on S_i , and give the message consisting of x and these outcomes. The existence of this strategy shows that, for partition matroid games, the procedure always succeeds in finding a worst-case optimal strategy for the choice $\mathcal{Y}' = \mathcal{Y}$.

Example 1 (continued) *The message structure $\mathcal{Y} = \{\{\text{left}, \text{middle}\}, \{\text{middle}, \text{right}\}\}$ in the Monty Hall puzzle is a partition matroid with sets $\{\text{left}, \text{right}\}$ and $\{\text{middle}\}$. For an arbitrary prior p on the three doors, the RCAR strategy and vector are given by*

$$\begin{aligned} q_{\text{left}} &= q_{\text{right}} = p_{\text{left}} + p_{\text{right}}; & q_{\text{middle}} &= p_{\text{middle}}; \\ P(\{\text{left}, \text{middle}\}) &= p_{\text{left}} / (p_{\text{left}} + p_{\text{right}}); & P(\{\text{middle}, \text{right}\}) &= p_{\text{right}} / (p_{\text{left}} + p_{\text{right}}). \end{aligned}$$

What does a message Y generated by this strategy tell the contestant about the true (random) outcome X ? Clearly, it means that if $X \in S_i$ for some i , then X must be the unique outcome in $Y \cap S_i$. Of course, the contestant does not know which of these sets contains X . Write I for the (random) index of the set containing X . Does Y tell the contestant anything about I ? The answer is no: For each index i , regardless of whether $I = i$, the outcome in $Y \cap S_i$ will be randomly distributed according to the marginal p conditioned on S_i , independently of $Y \cap S_j$ for $j \neq i$. This implies that Y is independent of I . Then for each outcome $x \in Y$, the probability that $X = x$ given message Y equals the probability that $I = i$, where i is the index of the set containing x . These are exactly the probabilities that appear in the RCAR vector q . We know from (Van Ommen et al., 2016, Theorem 19) that the same is true also if the quizmaster is using a worst-case optimal strategy different from the one described above.

In more general message structures, there may be a message that must be excluded from \mathcal{Y}' , so that the worst-case optimal P cannot be computed so easily:

Theorem 2 *If a game induces a homogeneous colouring but is not a partition matroid, then there exist a marginal p and a message $y \in \mathcal{Y}$ such that $P(y) = 0$ for all RCAR strategies P .*

We distinguish one subclass of the class of partition matroid games. A message structure in which the intersection of any two messages is constant is called a *sunflower* (Jukna, 2001). The common intersection is called the *core*, and each set difference between a message and the core is called a *petal*. An example of a *sunflowers with singleton petals* is shown in Figure 3b. The Monty Hall game itself (Example 1) is another example.

If a message structure is a sunflower with singleton petals, it is a partition matroid: each outcome in the core forms a (singleton) class of the partition, and another class contains all the petals. Among partition matroids, sunflowers can be recognized by the property that all of its colour classes except one are singleton outcomes. For this class of games, the strategy P described above is the *unique* RCAR strategy: a strategy P' with $P'(y) \neq P(y)$ for some $y \in \mathcal{Y}$ would disagree with the unique RCAR vector.

The message structure shown in Figure 3a is a partition matroid, but not a sunflower. Because at least two of its colour classes are not singletons, such a message structure contains a cycle of four messages in which neighbouring messages differ by the exchange of one outcome, but the pairs of messages on opposite sides of the cycle differ by two outcomes. (In Figure 3a, there are three such cycles; one is (y_1, y_2, y_5, y_4) .) For this type of game, the strategy P found above can be modified by increasing $P(y)$ for two messages at opposite sides of the cycle, and decreasing it by the same amount for the other two, leaving the conditionals unchanged. Thus P is not the unique RCAR strategy. In fact, RCAR strategies exist with $P(y) = 0$ for some $y \in \mathcal{Y}$. For such a strategy P , we have $\mathcal{Y}_P \subsetneq \mathcal{Y}$, but we do still have $\sum_{x \in y} q_x = 1$ even for messages y with $P(y) = 0$.

6. Conclusion

We have presented an efficient algorithm for finding the minimax optimal strategy in a coarse data problem where the message structure is a partition matroid. While this problem could also be solved using general-purpose convex optimization algorithms, this would be much less efficient. We have also seen how RCAR strategies may be qualitatively different beyond partition matroids, suggesting that in the general case, exact computation of these strategies may be a harder problem.

Acknowledgments

I thank Peter Grünwald and Wouter Koolen for the enjoyable collaboration which led to this paper, and the anonymous reviewers for their valuable feedback. This research was supported by Vici grant 639.073.04 from the Netherlands Organization for Scientific Research (NWO).

Appendix A. Proofs

Proof [Lemma 1] ($2 \Leftarrow 3$) Two elements $x_1 \neq x_2$ of \mathcal{X} are in the same 2-connected component if and only if there is a *circuit* (minimal dependent set) containing both (Oxley, 2011). Since a *basis* $y \in \mathcal{Y}$ is a maximal independent set, $y_1 \cup y_2$ is dependent. Find a circuit $C \subseteq y_1 \cup y_2$; this circuit contains both x_1 and x_2 , as otherwise it would be contained in a basis and thus independent.

($2 \Rightarrow 3$) Let C be a circuit with $\{x_1, x_2\} \subseteq C$; our goal is to find the bases y_1, y_2 , which we will do iteratively. Let y_1 be a basis containing the independent set $C \setminus \{x_2\}$, and y_2 a basis containing $C \setminus \{x_1\}$. While $y_1 \setminus \{x_1\} \neq y_2 \setminus \{x_2\}$, pick any $x'_1 \in y_1 \setminus (y_2 \cup \{x_1\})$ and use basis exchange to find a basis $y' = (y_1 \setminus \{x'_1\}) \cup \{x'_2\}$ for some $x'_2 \in y_2 \setminus y_1$. Note that $x'_2 \neq x_2$, as that would result in $C \subseteq y'$. Replace y_1 by y' and repeat until $y_1 \setminus \{x_1\} = y_2 \setminus \{x_2\}$. This process terminates, as the set difference becomes smaller with each step.

($1 \Leftrightarrow 3$) For exchange-connected message structures, the colour classes are the equivalence classes of the transitive reflexive closure of the relation on \mathcal{X} stated in point 3. For matroids, the equivalence of points 2 and 3 shows that this relation is already transitive. Thus for all $x_1 \neq x_2$, points 1 and 3 are equivalent. ■

Proof [Theorem 2] We will construct a marginal p with the required property by first finding a vector q that is the RCAR vector for some game with the given message structure. We distinguish two cases. If there exists $y' \subset \mathcal{X}$ that is consistent with the homogeneous induced colouring but $y' \notin \mathcal{Y}$, then pick $0 < \epsilon < 1/(k(k-1))$ and set initial values for q as

$$q_x = \begin{cases} \frac{1}{k} + \epsilon & \text{for } x \in y'; \\ \frac{1}{k} - (k-1)\epsilon & \text{otherwise.} \end{cases}$$

Each message contains at least one outcome with the smaller q_x , so $\sum_{x \in y} q_x \leq 1$ for all $y \in \mathcal{Y}$.

Otherwise, if \mathcal{Y} is not a partition matroid there must exist a colour class $C \subseteq \mathcal{X}$ for which the number of outcomes of this colour occurring in a message is at least two. Then pick any $x^+ \in C$ and $0 < \epsilon < 1/k$, and initialize q according to

$$q_x = \begin{cases} \frac{1}{k} + \epsilon & \text{for } x = x^+; \\ \frac{1}{k} - \epsilon & \text{for } x \in C \text{ but } x \neq x^+; \\ \frac{1}{k} & \text{otherwise.} \end{cases}$$

Again we see $\sum_{x \in y} q_x \leq 1$ for all $y \in \mathcal{Y}$.

Starting from the values of q determined above, we apply a greedy algorithm that repeatedly increases q_x for some x until none can be increased further, maintaining $\sum_{x \in y} q_x \leq 1$ for all $y \in \mathcal{Y}$. For the resulting vector q , let P be the joint distribution on x, y with $P(y)$ uniform on

$\{y \in \mathcal{Y} \mid \sum_{x \in y} q_x = 1\}$, $P(x \mid y) = q_x$ for all $x \in y$, and $P(x, y) = 0$ elsewhere. This P is an RCAR strategy for the game with marginal $p_x = \sum_{y \ni x} P(x, y)$, and q is the unique RCAR vector.

In the first case, there must exist some $x^- \in \mathcal{X}$ with $q_{x^-} \leq 1/k$. Let C be the colour class containing x^- , and let x^+ be the unique outcome in $C \cap y'$. In the second case, there must exist some $x^- \in C$ with $q_{x^-} \leq 1/k$. Thus in either case, we have two outcomes x^- and x^+ of the same colour C but with $q_{x^-} \leq 1/k < 1/k + \epsilon \leq q_{x^+}$. Because this contradicts the definition of an induced colouring, there must be a message for which q violates the equality (2). This message must have $P(y) = 0$ in any RCAR strategy for the game with message structure \mathcal{Y} and marginal p . ■

References

- T. Augustin. On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view: A cautionary note on updating imprecise priors. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications (ISIPTA)*, pages 31–45, 2003.
- R. D. Gill and P. D. Grünwald. An algorithmic and a geometric characterization of coarsening at random. *The Annals of Statistics*, 36:2409–2422, 2008.
- P. D. Grünwald and J. Y. Halpern. Updating probabilities. *Journal of Artificial Intelligence Research*, 19:243–278, 2003.
- P. D. Grünwald and J. Y. Halpern. Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research*, 42:393–426, 2011.
- D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19:2244–2253, 1991.
- M. Jaeger. Ignorability in statistical and probabilistic inference. *Journal of Artificial Intelligence Research*, 24:889–917, 2005.
- S. Jukna. *Extremal combinatorics: With applications in computer science*. Springer, Berlin, 2001.
- T. van Ommen. *Better predictions when models are wrong or underspecified*. PhD thesis, Mathematical Institute, Faculty of Science, Leiden, 2015.
- T. van Ommen, W. M. Koolen, T. E. Feenstra, and P. D. Grünwald. Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57, 2016.
- J. Oxley. *Matroid theory*. Oxford University Press, New York, second edition, 2011.
- T. Seidenfeld. A contrast between two decision rules for use with (convex) sets of probabilities: Γ -maximin versus E -admissibility. *Synthese*, 140:69–88, 2004.
- T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21: 1139–1154, 1993.
- S. Selvin. A problem in probability. *The American Statistician*, 29:67, 1975. Letter to the editor.
- G. de Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(12):75–125, 2004. ISSN 0004-3702.

Agreeing to Disagree and Dilation

Jiji Zhang

Lingnan University (Hong Kong)

JIJIZHANG@LN.EDU.HK

Hailin Liu

Sun Yat-sen University, Guangzhou (China)

LIUHLIN3@MAIL.SYSU.EDU.CN

Teddy Seidenfeld

Carnegie Mellon University, Pittsburgh (USA)

TEDDY@STAT.CMU.EDU

Abstract

We consider Geanakoplos and Polemarchakis’s generalization of Aumann’s famous result on “agreeing to disagree”, in the context of imprecise probability. The main purpose is to reveal a connection between the possibility of agreeing to disagree and the interesting and anomalous phenomenon known as dilation. We show that for two agents who share the same set of priors and update by conditioning on every prior, it is impossible to agree to disagree on the lower or upper probability of a hypothesis unless a certain dilation occurs. With some common topological assumptions, the result entails that it is impossible to agree not to have the same set of posterior probabilities unless dilation is present. This result may be used to generate sufficient conditions for guaranteed full agreement in the generalized Aumann-setting for some important models of imprecise priors, and we illustrate the potential with an agreement result involving the density ratio classes. We also provide a formulation of our results in terms of “dilation-averse” agents who ignore information about the value of a dilating partition but otherwise update by full Bayesian conditioning.

Keywords: agreeing to disagree; common knowledge; dilation; imprecise probability.

1. Introduction

In a simple but insightful paper, Aumann (1976) famously showed that two (Bayesian) agents who start with the same (precise) prior cannot agree to disagree on their posteriors of a hypothesis, in the sense that if the posteriors of the hypothesis (as well as the structures of their respective information partitions) are common knowledge, then the posteriors must be equal. This result has been generalized in at least two ways. First, Aumann’s result applies only to those events whose posteriors happen to be common knowledge. Geanakoplos and Polemarchakis (1982) generalized the framework to a communication setting where the agents are invited to repeatedly make their credences public via announcements and update by conditioning on the announced credences, until no new information is conveyed. They showed that for any hypothesis/event, this communication procedure is guaranteed to lead to an agreement on the probability of the hypothesis, if the agents start with the same (precise) prior (and each agent’s information partition is finite).

Second, Kajii and Ui (2005, 2009) and Carvajal and Correia-da-Silva (2010) generalized Aumann’s result in the setting of multiple priors. In this line of work, “agreement” is taken to mean “partial agreement”, in the sense that two sets of probabilities agree if they have a non-empty intersection. These authors established several sufficient conditions under which two agents who (partially) agree on their priors are guaranteed to (partially) agree on their posteriors of a hypothesis if these posteriors are common knowledge.

In this paper, we combine the two more general settings and establish a connection between the possibility of agreeing to disagree and the interesting and anomalous phenomenon known as dilation (Good, 1974; Seidenfeld, 1981; Walley, 1991; Seidenfeld and Wasserman, 1993; Herron et al., 1997). Dilation occurs when conditioning on each element of a partition, the lower and upper probabilities of a hypothesis become more divergent than the unconditional ones. In such a case, for agents who use full Bayesian conditioning as the updating rule, their credences on a hypothesis become less precise or determinate after learning the value of the dilating partition, no matter which value they learn! This counterintuitive phenomenon is often interpreted as a distinctive challenge to the orthodox Bayesian doctrine on the value of information and to the Bayesian merging of opinions, but as far as we know, it has never been discussed in connection to Aumann's result. We shall show that it is the key obstacle to reaching agreements via communicating posteriors by Bayesian agents with imprecise priors.

We will establish the following. After introducing the setting and reviewing the special case of precise probability in Section 2, we show in Section 3 that dilation is the *only* obstacle for agents with the same (imprecise) prior to reaching agreements on lower and upper probabilities of a hypothesis by communicating their posteriors on the hypothesis. Without dilation, the two agents in our setting are guaranteed to end up agreeing on lower and upper probabilities of the hypothesis of interest. An immediate consequence of this result, as we note in Section 4, is that under common topological assumptions, dilation is the *only* obstacle to reaching a full agreement, *full* in the sense that the sets of probability values representing the agents' credences on the hypothesis of interest are identical. This result opens the door to generating sufficient conditions for reaching full consensus in the generalized Aumann-setting by plugging in sufficient conditions for the absence of dilation in common and important models of imprecise probabilities. As an example, we include a corollary about density ratio classes, which are shown to be dilation-immune by Seidenfeld and Wasserman (1993). In Section 5, we provide another perspective on our results and reformulate the theorems in terms of "dilation-averse" agents, who update by full Bayesian conditioning unless the information is about the value of a dilating partition (in which case they ignore the information). For such agents, they are guaranteed to end up agreeing on lower and upper probabilities, and, under some common assumptions, end up fully agreeing.

2. A Procedure of Communicating Posteriors

In Geanakoplos and Polemarchakis (1982)'s setup, two agents share a common measurable space (Ω, \mathcal{A}) and have possibly different information partitions of Ω , \mathcal{P}^1 and \mathcal{P}^2 , which are assumed to be finite. Henceforth we use $i \in \{1, 2\}$ to index the two agents, and when i is used in a statement we always intend that the statement is true for both $i = 1$ and $i = 2$. For any $w \in \Omega$, let $\mathcal{P}^i(w)$ denote the member of \mathcal{P}^i that contains w ; intuitively, $\mathcal{P}^i(w)$ represents agent i 's initial information at state w . Both the space and the partitions are assumed to be common knowledge, in the standard sense of the term used in game theory: some proposition is common knowledge just in case agent i knows it, agent j (where $j = 3 - i$) knows that agent i knows it, agent i knows that agent j knows that agent i knows it, ... and so on. Let $\mathcal{P} = \mathcal{P}^1 \wedge \mathcal{P}^2$ be the meet of the two partitions (i.e., the finest common coarsening of \mathcal{P}^1 and \mathcal{P}^2). As Aumann (1976, p. 1237) explained, at state w , $\mathcal{P}(w)$ — the member of \mathcal{P} that contains w — is the finest event in \mathcal{A} that is common knowledge: any event that is common knowledge is a superset of $\mathcal{P}(w)$. In Geanakoplos and Polemarchakis's setting, common

knowledge may grow as the agents communicate their posteriors of a hypothesis. So we call $\mathcal{P}(w)$ the initial common knowledge and denote it by \mathcal{C}_0 .

Instead of a common precise prior, we assume that the two agents have a common, (possibly) imprecise prior, i.e., a common, non-empty set of priors, denoted by \mathbf{Q} . Let $\mathcal{P}^1 \vee \mathcal{P}^2$ denote the join (i.e., the coarsest common refinement) of \mathcal{P}^1 and \mathcal{P}^2 . We assume that every member of $\mathcal{P}^1 \vee \mathcal{P}^2$ receives a positive probability under every measure in \mathbf{Q} , so that all the relevant conditional probabilities are well defined as ratios of unconditional probabilities. Let $H \in \mathcal{A}$ be a hypothesis of interest. Henceforth by credences or posteriors we mean the agents' credences or posteriors of H . Let $\mathbf{Q}(H)$ denote the set of prior probabilities of H : $\mathbf{Q}(H) = \{p(H) \mid p \in \mathbf{Q}\}$. For any $E \in \mathcal{A}$ such that $p(E) > 0$ for every $p \in \mathbf{Q}$, let $\mathbf{Q}(H|E) = \{p(H|E) = p(H \cap E)/p(E) \mid p \in \mathbf{Q}\}$. Unless otherwise noted (in Section 5), we assume that the agents update their credences by full Bayesian conditioning, where each and every prior in \mathbf{Q} is updated by conditioning.

Suppose the true state is w . At step 0, agent i 's information is $\mathcal{P}^i(w) \cap \mathcal{C}_0 = \mathcal{P}^i(w)$. Thus agent i updates her credence of H to $\mathbf{Q}_0^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w))$. Let $\mathcal{P}_0^i = \{E \in \mathcal{P}^i \mid E \cap \mathcal{C}_0 \neq \emptyset\}$, which is the set of those members of \mathcal{P}^i that are not ruled out by the initial common knowledge.

At step 1, the agents announce $\mathbf{Q}_0^1(H)$ and $\mathbf{Q}_0^2(H)$, respectively.¹ Consider $\mathcal{N}_1^i = \{E \in \mathcal{P}_0^i \mid \mathbf{Q}(H|E) = \mathbf{Q}_0^i(H)\}$. Intuitively, \mathcal{N}_1^i is the set of those members of \mathcal{P}_0^i that are compatible with $\mathbf{Q}_0^i(H)$, and the effect of agent i 's announcement of $\mathbf{Q}_0^i(H)$ is that it becomes common knowledge that $\mathcal{P}^i(w) \in \mathcal{N}_1^i$, or that $w \in \bigcup \mathcal{N}_1^i$ (where $\bigcup \mathcal{N}_1^i$ denotes the union of all the sets in \mathcal{N}_1^i). Therefore, after the announcements at this step, $\mathcal{C}_1 = \bigcup \mathcal{N}_1^1 \cap \bigcup \mathcal{N}_1^2$ becomes common knowledge. Let $\mathcal{P}_1^i = \{E \in \mathcal{N}_1^i \mid E \cap \mathcal{C}_1 \neq \emptyset\}$, which is the set of those members of \mathcal{N}_1^i that are not ruled out by the common knowledge at this step. Clearly $\mathcal{P}_1^i \subseteq \mathcal{N}_1^i \subseteq \mathcal{P}_0^i$ and $\mathcal{C}_1 = \bigcup \mathcal{P}_1^1 \cap \bigcup \mathcal{P}_1^2$. Now, if $\mathcal{P}_1^i = \mathcal{P}_0^i$, or equivalently, if $\mathcal{C}_1 = \mathcal{C}_0$, neither agent learns new information and their credences will stay the same no matter how many more exchanges take place; so the procedure stops. Otherwise, agent i updates credence of H to $\mathbf{Q}_1^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_1)$, and enters the next step.

In general, at step $n+1$, the agents announce $\mathbf{Q}_n^1(H)$ and $\mathbf{Q}_n^2(H)$, respectively. Let

$$\begin{aligned}\mathcal{N}_{n+1}^i &= \{E \in \mathcal{P}_n^i \mid \mathbf{Q}(H|E \cap \mathcal{C}_n) = \mathbf{Q}_n^i(H)\} \\ \mathcal{C}_{n+1} &= \bigcup \mathcal{N}_{n+1}^1 \cap \bigcup \mathcal{N}_{n+1}^2 \\ \mathcal{P}_{n+1}^i &= \{E \in \mathcal{N}_{n+1}^i \mid E \cap \mathcal{C}_{n+1} \neq \emptyset\}.\end{aligned}$$

Again, \mathcal{N}_{n+1}^i is the set of those members of \mathcal{P}_n^i that are compatible with $\mathbf{Q}_n^i(H)$.² Hence, after the announcements at this step, \mathcal{C}_{n+1} becomes common knowledge, and \mathcal{P}_{n+1}^i is the set of those members of \mathcal{N}_{n+1}^i that are not ruled out by \mathcal{C}_{n+1} . Clearly, $\mathcal{P}_{n+1}^i \subseteq \mathcal{N}_{n+1}^i \subseteq \mathcal{P}_n^i$ and $\mathcal{C}_{n+1} = \bigcup \mathcal{P}_{n+1}^1 \cap \bigcup \mathcal{P}_{n+1}^2$. If $\mathcal{P}_{n+1}^i = \mathcal{P}_n^i$, or equivalently, if $\mathcal{C}_{n+1} = \mathcal{C}_n$, neither agent learns new information and the procedure stops; otherwise, agent i updates credence of H to $\mathbf{Q}_{n+1}^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_{n+1})$, and enters the next step.

We will refer to this procedure as the *(Bayesian) procedure of communicating posteriors (of H)*. Obviously, since \mathcal{P}^1 and \mathcal{P}^2 are assumed to be finite, the procedure is guaranteed to stop at step $m+1$ for some $m \geq 0$. Aumann's original setting — where $\mathbf{Q}_0^1(H)$ and $\mathbf{Q}_0^2(H)$ are assumed to be common knowledge at step 0 (i.e., it is assumed that $\mathcal{N}_1^i = \mathcal{P}_0^i$) — is a special case in which the

1. In Geanakoplos and Polemarchakis's design, at each step, agent 2 announces her prior *after* agent 1's announcement, already taking into account whatever information is conveyed in agent 1's announcement. This feature is immaterial, at least for the purpose of this paper.

2. Note that the definition of \mathcal{N}_{n+1}^i also applies to $n = 0$, as for every $E \in \mathcal{P}_0^i$, $E \cap \mathcal{C}_0 = E$.

procedure stops at step 1. In general, the procedure stops at step $m + 1$ if and only if both $\mathbf{Q}_m^1(H)$ and $\mathbf{Q}_m^2(H)$ are already common knowledge at step m (i.e., before they are announced).

We adapt an example from Geanakoplos and Polemarchakis (1982) to illustrate this procedure.

Example 1 Let $\Omega = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$ and \mathcal{A} be the power set of Ω . Let $\mathcal{P}^1 = \{\{w_1, w_2, w_3\}, \{w_4, w_5, w_6\}, \{w_7, w_8, w_9\}\}$ and $\mathcal{P}^2 = \{\{w_1, w_2, w_3, w_4\}, \{w_5, w_6, w_7, w_8\}, \{w_9\}\}$. Let $H = \{w_3, w_4\}$, and suppose the true state of the world is w_1 . For the common set of priors, suppose \mathbf{Q} is a density ratio class (Seidenfeld and Wasserman, 1993; see also Section 4):

$$\mathbf{Q} = \{(q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9) \mid \sum_{1 \leq j \leq 9} q_j = 1, \text{ and } \frac{1}{2} \leq \frac{q_k}{q_l} \leq 2, 1 \leq k, l \leq 9.\}.$$

It is easy to calculate that the lower probability of H is: $\underline{\mathbf{Q}}(H) = \inf_{p \in \mathbf{Q}} p(H) = 1/8$ (obtained at $(1/8, 1/8, 1/16, 1/16, 1/8, 1/8, 1/8, 1/8, 1/8)$), and the upper probability of H is: $\overline{\mathbf{Q}}(H) = \sup_{p \in \mathbf{Q}} p(H) = 4/11$ (obtained at $(1/11, 1/11, 2/11, 2/11, 1/11, 1/11, 1/11, 1/11, 1/11)$). Since \mathbf{Q} is closed and connected, $\mathbf{Q}(H) = [1/8, 4/11]$.

Suppose the two agents in this example carry out the procedure of communicating posteriors. Here is a summary of the execution:

Step 0 $\mathcal{C}_0 = (\mathcal{P}^1 \wedge \mathcal{P}^2)(w_1) = \Omega$. $\mathcal{P}^1(w_1) = \{w_1, w_2, w_3\}$ and $\mathbf{Q}_0^1(H) = \mathbf{Q}(H|\mathcal{P}^1(w_1)) = [1/5, 1/2]$; $\mathcal{P}^2(w_1) = \{w_1, w_2, w_3, w_4\}$ and $\mathbf{Q}_0^2(H) = \mathbf{Q}(H|\mathcal{P}^2(w_1)) = [1/3, 2/3]$.

Step 1 Agent i announces $\mathbf{Q}_0^i(H)$. $\mathcal{N}_1^1 = \{\{w_1, w_2, w_3\}, \{w_4, w_5, w_6\}\}$ (for $\mathbf{Q}(H|\{w_7, w_8, w_9\}) = \{0\} \neq \mathbf{Q}_0^1(H)$.) $\mathcal{N}_1^2 = \{\{w_1, w_2, w_3, w_4\}\}$ (for $\mathbf{Q}(H|\{w_5, w_6, w_7, w_8\}) = \mathbf{Q}(H|\{w_9\}) = \{0\} \neq \mathbf{Q}_0^2(H)$.) Thus $\mathcal{C}_1 = \bigcup \mathcal{N}_1^1 \cap \bigcup \mathcal{N}_1^2 = \{w_1, w_2, w_3, w_4\}$, and $\mathcal{P}_1^i = \mathcal{N}_1^i$. $\mathbf{Q}_1^1(H) = \mathbf{Q}(H|\mathcal{P}^1(w_1) \cap \mathcal{C}_1) = [1/5, 1/2]$; $\mathbf{Q}_1^2(H) = \mathbf{Q}(H|\mathcal{P}^2(w_1) \cap \mathcal{C}_1) = [1/3, 2/3]$.³

Step 2 Agent i announces $\mathbf{Q}_1^i(H)$. $\mathcal{N}_2^1 = \{\{w_1, w_2, w_3\}\}$ (for $\mathbf{Q}(H|\{w_4, w_5, w_6\} \cap \mathcal{C}_1) = \{1\} \neq \mathbf{Q}_1^1(H)$.) $\mathcal{N}_2^2 = \mathcal{P}_1^2$. Thus $\mathcal{C}_2 = \{w_1, w_2, w_3\}$, and $\mathcal{P}_2^i = \mathcal{N}_2^i$. $\mathbf{Q}_2^1(H) = \mathbf{Q}(H|\mathcal{P}^1(w_1) \cap \mathcal{C}_2) = [1/5, 1/2]$; $\mathbf{Q}_2^2(H) = \mathbf{Q}(H|\mathcal{P}^2(w_1) \cap \mathcal{C}_2) = [1/5, 1/2]$.

Step 3 Agent i announces $\mathbf{Q}_2^i(H)$. $\mathcal{N}_3^i = \mathcal{P}_2^i$, and so $\mathcal{C}_3 = \mathcal{C}_2$. The procedure stops.

In this example, the communication ends up making each agent's private information public. This is not always the case, as later examples will illustrate. When (at least one agent's) private information remains private, it is in general possible to agree to disagree. However, in the case of a precise prior, that is, if $\mathbf{Q} = \{\tilde{p}\}$ is a singleton, Geanakoplos and Polemarchakis (1982, Proposition 1) showed that when the procedure stops at step $m + 1$, it is necessarily the case that $\mathbf{Q}_m^1(H) = \mathbf{Q}_m^2(H)$. We present a version of the argument here that will facilitate our subsequent discussion. Suppose the procedure stops at step $m + 1$. It means that $\mathcal{P}_{m+1}^i = \mathcal{P}_m^i$ (for both $i = 1, 2$, as we always intend). This entails, by the definition of \mathcal{P}_{m+1}^i , that

$$\forall E \in \mathcal{P}_m^i, \mathbf{Q}(H|E \cap \mathcal{C}_m) = \mathbf{Q}_m^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m). \quad (1)$$

Since $\mathbf{Q} = \{\tilde{p}\}$, $\mathbf{Q}(H|E \cap \mathcal{C}_m) = \{\tilde{p}(H|E \cap \mathcal{C}_m)\}$ and $\mathbf{Q}_m^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) = \{\tilde{p}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)\}$. It follows that

$$\forall E \in \mathcal{P}_m^i, \tilde{p}(H|E \cap \mathcal{C}_m) = \tilde{p}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m). \quad (2)$$

3. Although $\mathbf{Q}_1^i(H) = \mathbf{Q}_0^i(H)$, the procedure goes on, because some agent still acquires new information in this step.

Since all members of \mathcal{P}_m^i are mutually disjoint, (2) entails that

$$\tilde{p}(H | \bigcup \mathcal{P}_m^i \cap \mathcal{C}_m) = \tilde{p}(H | \mathcal{P}^i(w) \cap \mathcal{C}_m). \quad (3)$$

Recall that $\mathcal{C}_m = \bigcup \mathcal{P}_m^1 \cap \bigcup \mathcal{P}_m^2$. Hence $\bigcup \mathcal{P}_m^i \cap \mathcal{C}_m = \mathcal{C}_m$. It then follows from (3) that

$$\tilde{p}(H | \mathcal{P}^i(w) \cap \mathcal{C}_m) = \tilde{p}(H | \mathcal{C}_m). \quad (4)$$

Therefore, $\tilde{p}(H | \mathcal{P}^1(w) \cap \mathcal{C}_m) = \tilde{p}(H | \mathcal{P}^2(w) \cap \mathcal{C}_m)$; that is, the two agents end up agreeing.

Two comments are in order. First, equation (4) shows that the two agents are driven to the same posterior because when the communication stops, the resulting common knowledge (\mathcal{C}_m) renders each agent's private information ($\mathcal{P}^i(w)$) irrelevant to H (even if $\mathcal{P}^i(w)$ remains private). However, it does not follow that $\mathcal{P}^1(w)$ and $\mathcal{P}^2(w)$ are jointly irrelevant to H given \mathcal{C}_m . As Geanakoplos and Polemarchakis (1982, Proposition 3) observed, the consensus reached via the procedure of communicating posteriors can be different from the consensus that would result from directly exchanging private information. Clearly, they are different if and only if $\mathcal{P}^1(w)$ and $\mathcal{P}^2(w)$ are jointly relevant to H given \mathcal{C}_m , even though each is marginally irrelevant given \mathcal{C}_m (see Example 3 in Section 6).

Second, and more importantly for the purpose of this paper, a crucial step in the above argument is the move from (2) to (3), where what is needed is the following fact: if all members of a (finite) set of events \mathcal{E} are mutually disjoint, and for every $E \in \mathcal{E}$, $\tilde{p}(H | E) = q$, then $\tilde{p}(H | \bigcup \mathcal{E}) = q$. An analogous condition for imprecise probabilities would be the following: if all members of a (finite) set of events \mathcal{E} are mutually disjoint, and for every $E \in \mathcal{E}$, $\mathbf{Q}(H | E) = Q$ (where Q is a set of real numbers), then $\mathbf{Q}(H | \bigcup \mathcal{E}) = Q$. This condition does not hold in general for sets of probabilities.⁴

3. Dilation and Agreeing to Disagree on Lower and Upper Probabilities

We borrow a simple example from Carvajal and Correia-da-Silva (2010) to illustrate the failure of the said condition for sets of probabilities.

Example 2 Let $\Omega = \{w_1, w_2, w_3, w_4\}$ and \mathcal{A} the power set of Ω . Let $\mathcal{P}^1 = \{\{w_1, w_2\}, \{w_3, w_4\}\}$ and $\mathcal{P}^2 = \{\Omega\}$. Suppose $\mathbf{Q} = \{(1/2, 0, 1/2, 0), (0, 1/2, 0, 1/2)\}$; that is, the common set of priors consists of just two probability measures, represented by the two probability vectors.⁵ Let $H = \{w_2, w_3\}$, and suppose the true state of the world is w_1 .

This is an Aumann case in that the agents' posteriors on H are already common knowledge at the beginning; the procedure of communicating posteriors stops at step 1, for $\mathcal{C}_1 = \mathcal{C}_0 = \Omega$. However, $\mathbf{Q}_0^1(H) = \{0, 1\}$ and $\mathbf{Q}_0^2(H) = \{1/2\}$. Not only are the sets non-identical, they are in full

-
- 4. Even in the case of precise probability, it is well known that this condition, as a special case of conglomerability, can fail for finitely but not countably additive probability measures (de Finetti, 1972; Schervish et al., 1984; Hill and Lane, 1985). This does not matter in the setup we are considering, for the partitions are assumed to be finite. However, the original setup in Aumann (1976) seems to allow denumerable infinite partitions, in which case Aumann's result does not necessarily hold for merely finitely additive probabilities. More generally, Schervish et al. (2016) showed that conglomerability can fail in a partition of cardinality κ for a probability measure that is not κ -additive. Thus, if uncountable partitions are allowed, Aumann's result may fail even for countably additive measures.
 - 5. In case readers are concerned that the two probabilities are not positive and are mutually singular, these special features are not essential. We can also use $\mathbf{Q} = \{(1/2 - \epsilon, \epsilon, 1/2 - \epsilon, \epsilon), (\epsilon, 1/2 - \epsilon, \epsilon, 1/2 - \epsilon)\}$, $0 < \epsilon < 1/4$, to make the same point.

disagreement in the sense that they do not even intersect and have different lower and upper probabilities. The agents agree to fully disagree. The condition we highlighted at the end of Section 2 fails dramatically in this case for agent 1's partition \mathcal{P}_1^1 (which is identical to \mathcal{P}^1 in this case): $\mathbf{Q}(H|\{w_1, w_2\}) = \mathbf{Q}(H|\{w_3, w_4\}) = \{0, 1\}$, while $\mathbf{Q}(H|\{w_1, w_2, w_3, w_4\}) = \{1/2\}$.

This dramatic failure of the condition is known as dilation (Good, 1974; Seidenfeld, 1981; Walley, 1991; Seidenfeld and Wasserman, 1993; Herron et al., 1997). No matter which member of \mathcal{P}_1^1 is the case, the resulting conditional probability is less precise than the probability conditional on $\bigcup \mathcal{P}_1^1$. Given a non-empty set of probabilities \mathbf{R} , let $\underline{\mathbf{R}}(A|E) = \inf_{p \in \mathbf{R}} p(A|E)$ denote the lower probability of A conditional on E , and $\overline{\mathbf{R}}(A|E) = \sup_{p \in \mathbf{R}} p(A|E)$ denote the upper probability of A conditional on E . Here is a definition of dilation that suits the present purpose.

Definition 1 (Dilation) Let \mathbf{R} be a non-empty set of probability measures on (Ω, \mathcal{A}) . Let \mathcal{E} be a finite, non-empty set of mutually disjoint events. \mathcal{E} is said to dilate an event A with respect to \mathbf{R} (or $\mathbf{R}(\bullet|\bigcup \mathcal{E})$) if for every $E \in \mathcal{E}$, the interval $[\underline{\mathbf{R}}(A|E), \overline{\mathbf{R}}(A|E)]$ strictly contains the interval $[\underline{\mathbf{R}}(A|\bigcup \mathcal{E}), \overline{\mathbf{R}}(A|\bigcup \mathcal{E})]$.

This is a slight generalization of the standard definition of dilation (Seidenfeld and Wasserman, 1993, p. 1141)⁶, for it considers dilation in a subspace $\bigcup \mathcal{E}$ (the definition reduces to the standard one when $\bigcup \mathcal{E} = \Omega$), but the idea and the anomalous feature are exactly the same. Again, in example 2, \mathcal{P}_1^1 , which happens to be the same as $\{E \cap \mathcal{C}_1 \mid E \in \mathcal{P}_1^1\}$, dilates the hypothesis of interest with respect to the given prior. This is not a coincidence, as Theorem 3 below shows. It is a simple consequence of the following lemma, which is a straightforward generalization of Lemma 1 in Carvajal and Correia-da-Silva (2010; also see Kajii and Ui, 2005, Proposition 3).

Lemma 2 Suppose the procedure of communicating posteriors stops at step $m + 1$. Then

$$\mathbf{Q}(H|\mathcal{C}_m) \subseteq [\underline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m), \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)]$$

for both $i = 1, 2$.

Proof As already mentioned, when the procedure stops at step $m + 1$, we have equation (1), namely,

$$\forall E \in \mathcal{P}_m^i, \mathbf{Q}(H|E \cap \mathcal{C}_m) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m).$$

Consider $i = 1$ first. Let $\mathcal{P}_m^1 = \{E_1, \dots, E_k\}$. Notice that $\{E_1 \cap \mathcal{C}_m, \dots, E_k \cap \mathcal{C}_m\}$ forms a partition of $\bigcup \mathcal{P}_m^1 \cap \mathcal{C}_m$. Hence, for every $p \in \mathbf{Q}$, by the law of total probability

$$p(H|\mathcal{C}_m) = p(H|\bigcup \mathcal{P}_m^1 \cap \mathcal{C}_m) = \sum_{1 \leq j \leq k} p(H|E_j \cap \mathcal{C}_m)p(E_j \cap \mathcal{C}_m|\bigcup \mathcal{P}_m^1 \cap \mathcal{C}_m). \quad (5)$$

Given equation (1), we have that for every $1 \leq j \leq k$, $p(H|E_j \cap \mathcal{C}_m) \in \mathbf{Q}(H|E_j \cap \mathcal{C}_m) = \mathbf{Q}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m)$. It follows that for every $1 \leq j \leq k$,

$$\underline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) \leq p(H|E_j \cap \mathcal{C}_m) \leq \overline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m). \quad (6)$$

6. Herron et al. (1997, p. 412) gave a weaker definition of dilation, requiring only that all conditional intervals contain and some of them *strictly* contain the unconditional interval. This definition (similarly generalized) would work equally well for our purpose. We thank an anonymous referee for this point.

Equation (5) and (6) together entail that

$$p(H|\mathcal{C}_m) \geq \underline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) \sum_{1 \leq j \leq k} p(E_j \cap \mathcal{C}_m | \bigcup \mathcal{P}_m^1 \cap \mathcal{C}_m) = \underline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m), \quad (7)$$

and

$$p(H|\mathcal{C}_m) \leq \overline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) \sum_{1 \leq j \leq k} p(E_j \cap \mathcal{C}_m | \bigcup \mathcal{P}_m^1 \cap \mathcal{C}_m) = \overline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m). \quad (8)$$

Since (7) and (8) hold for every $p \in \mathbf{Q}$, the desired conclusion is established for $i = 1$. The case of $i = 2$ is of course entirely parallel. \blacksquare

Lemma 2 shows that although equation (1) does not entail that $\mathbf{Q}(H|\mathcal{C}_m) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$, it does entail that $\mathbf{Q}(H|\mathcal{C}_m)$ is bounded by the infimum and supremum of $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$. The following theorem is then immediate.

Theorem 3 Suppose the procedure of communicating posteriors stops at step $m + 1$. If for both $i = 1, 2$, $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ does not dilate H , then $\underline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) = \underline{\mathbf{Q}}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m)$ and $\overline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) = \overline{\mathbf{Q}}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m)$.

Proof Lemma 2 entails that for both $i = 1, 2$,

$$\underline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) \leq \underline{\mathbf{Q}}(H|\mathcal{C}_m), \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) \leq \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m).$$

Since $\mathcal{C}_m = \bigcup \mathcal{P}_m^i \cap \mathcal{C}_m$, if either of the inequality is strict, then $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ dilates H , because of equation (1). Therefore, if $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ does not dilate H , then

$$\underline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) = \underline{\mathbf{Q}}(H|\mathcal{C}_m), \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) = \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m).$$

The desired conclusion follows. \blacksquare

Thus, the two agents can agree to disagree on the lower or upper probability of a hypothesis only if a certain dilation takes place. Without dilation, the two agents are guaranteed to reach consensus on lower and upper probabilities by communicating posteriors.

It is worth noting that for the argument for Theorem 3 to go through, it is not necessary to require the agents to communicate their sets of posteriors. It is sufficient to ask them to communicate lower and upper probabilities at each step. Consider the procedure of communicating lower and upper posteriors: at step $n + 1$, agent i announces $\underline{\mathbf{Q}}_n^i(H)$ and $\overline{\mathbf{Q}}_n^i(H)$. Let

$$\begin{aligned} \mathcal{N}_{n+1}^{i\dagger} &= \{E \in \mathcal{P}_n^{i\dagger} \mid \underline{\mathbf{Q}}(H|E \cap \mathcal{C}_n) = \underline{\mathbf{Q}}_n^i(H) \text{ and } \overline{\mathbf{Q}}(H|E \cap \mathcal{C}_n) = \overline{\mathbf{Q}}_n^i(H)\} \\ \mathcal{C}_{n+1}^\dagger &= \bigcup \mathcal{N}_{n+1}^{1\dagger} \cap \bigcup \mathcal{N}_{n+1}^{2\dagger} \\ \mathcal{P}_{n+1}^{i\dagger} &= \{E \in \mathcal{N}_{n+1}^{i\dagger} \mid E \cap \mathcal{C}_{n+1}^\dagger \neq \emptyset\} \end{aligned}$$

If $\mathcal{P}_{n+1}^{i\dagger} = \mathcal{P}_n^{i\dagger}$, or equivalently, if $\mathcal{C}_{n+1}^\dagger = \mathcal{C}_n^\dagger$, the procedure stops; otherwise, agent i updates credence to $\mathbf{Q}_{n+1}^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_{n+1}^\dagger)$, and enters the next step.

As before, this modified procedure is guaranteed to stop at step $m + 1$ for some $m \geq 0$, because \mathcal{P}^1 and \mathcal{P}^2 are assumed to be finite. The version of Lemma 2 on this procedure remains valid, for equation (1) is not necessary for the argument. All that is needed is the weaker condition that

$$\forall E \in \mathcal{P}_m^i, \underline{\mathbf{Q}}(H|E \cap \mathcal{C}_m) = \underline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m) \text{ and } \overline{\mathbf{Q}}(H|E \cap \mathcal{C}_m) = \overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$$

This weaker condition obviously remains true when \mathcal{P}_m^i is replaced by $\mathcal{P}_m^{i\dagger}$ and \mathcal{C}_m by \mathcal{C}_m^\dagger . Hence we also have the following variant of Theorem 3.

Theorem 4 Suppose the procedure of communicating lower and upper posteriors stops at step $m + 1$. If for both $i = 1, 2$, $\{E \cap \mathcal{C}_m^\dagger \mid E \in \mathcal{P}_m^{i\dagger}\}$ does not dilate H , then $\underline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m^\dagger) = \underline{\mathbf{Q}}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m^\dagger)$ and $\overline{\mathbf{Q}}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m^\dagger) = \overline{\mathbf{Q}}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m^\dagger)$.

Proof Extremely similar to that of Theorem 3. ■

Although both procedures result in consensus on lower and upper probabilities in the absence of dilation, in general the agreements they lead to may well be different, for in general communicating lower and upper posteriors conveys less information than communicating the full sets of posteriors.

4. More Agreement Results

Under some common assumptions, however, lower and upper probabilities are sufficient to identify the full set, in which case the two procedures are equivalent and, more importantly, the consensus reached in the absence of dilation will be full consensus. For example, if we follow Carvajal and Correia-da-Silva (2010) to assume that the set of priors is closed and connected (or follow Kajii and Ui (2005) to assume that the set of posteriors is a closed interval), we obtain the following result.

Theorem 5 Suppose \mathbf{Q} is closed and connected (with respect to the total variation topology), and suppose the procedure of communicating posteriors stops at step $m + 1$. If for both $i = 1, 2$, $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ does not dilate H , then $\mathbf{Q}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) = \mathbf{Q}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m)$.

Proof Given the assumption that all the relevant conditional probabilities are well defined as ratios of unconditional probabilities, the mapping from \mathbf{Q} to $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ is continuous.⁷ Hence, since \mathbf{Q} is assumed to be closed and connected, $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ is a closed interval. Thus $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ is identified by $\overline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ and $\underline{\mathbf{Q}}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$. Then Theorem 3 entails that $\mathbf{Q}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) = \mathbf{Q}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m)$. ■

To our knowledge, Theorem 5 is the first attempt to formulate a generalization of Aumann's agreement theorem in the context of imprecise probability that takes agreement to mean full agreement (identical set of posteriors). In addition to revealing a connection to the important phenomenon of dilation, Theorem 5 may also be used to generate sufficient conditions for guaranteed full agreement via communicating posteriors, for important models of imprecise probability, if sufficient conditions for the absence of dilation in those models are known. As a simple example, consider the density ratio classes for finite spaces (Wasserman, 1992; Seidenfeld and Wasserman, 1993).

Definition 6 (Density Ratio Prior) Let $\Omega = \{w_1, \dots, w_n\}$ and \mathcal{A} the power set of Ω . A density ratio prior is defined by

$$\mathbf{D}_{p,k} = \{(q_1, \dots, q_n) \mid \sum_{1 \leq j \leq n} q_j = 1 \text{ and } \frac{q_h}{q_j} \leq k \frac{p_h}{p_j}, \forall 1 \leq h, j \leq n\}$$

where $k \geq 1$ and (p_1, \dots, p_n) is a probability vector such that $p_j > 0$ for all $1 \leq j \leq n$.

7. The mapping is given by: $p \mapsto p(H \cap \mathcal{P}^i(w) \cap \mathcal{C}_m)/p(\mathcal{P}^i(w) \cap \mathcal{C}_m)$, which is obviously continuous, as long as the ratio is always defined.

For instance, Example 1 in Section 2 employs a density ratio prior, where p is the uniform distribution over the 9-atom algebra and $k = 2$.

Corollary 7 *If two agents start with a common density ratio prior and carry out the procedure of communicating posteriors, they are guaranteed to reach the same set of posteriors.*

Proof Seidenfeld and Wasserman (1993, Theorem 4.1) showed that the density ratio priors are dilation-immune in the sense that no finite partition of the sample space dilates any event. Note also that if \mathbf{D} is a density ratio prior on (Ω, \mathcal{A}) , then for every $E \in \mathcal{A}$, $\mathbf{D}(\bullet|E)$ remains a density ratio prior on the space restricted to E , which follows easily from Definition 6. Moreover, a density ratio prior is obviously closed and connected. Then Theorem 5 entails the desired conclusion. ■

Finally, if we consider just *partial* agreement, in the sense of a non-empty intersection of sets of posteriors, we can drop the assumption of connectedness in Theorem 5.

Theorem 8 *Suppose \mathbf{Q} is closed, and suppose the procedure of communicating posteriors stops at step $m + 1$. If for both $i = 1, 2$, $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ does not dilate H , then $\mathbf{Q}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) \cap \mathbf{Q}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m) \neq \emptyset$.*

Proof Since \mathbf{Q} is closed, $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ is also closed, for the mapping from \mathbf{Q} to $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ is continuous. Thus, $\mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_m)$ contains its infimum and supremum. It then follows from Theorem 3 that $\mathbf{Q}(H|\mathcal{P}^1(w) \cap \mathcal{C}_m) \cap \mathbf{Q}(H|\mathcal{P}^2(w) \cap \mathcal{C}_m) \neq \emptyset$. ■

5. Dilation-Averse Updating

The presence of dilation may alarm some agents, who may be inclined to think that they are permitted or even rationally required to ignore information about the value of a partition that dilates the hypothesis of interest (Grünwald and Halpern, 2004). Whether dilating information should be ignored is a matter of debate. For example, Kyburg's (1974) theory of "epistemological", interval-valued probability precludes altogether the possibility of dilation. However, the rule in his theory that is responsible for the impossibility of dilation was forcefully criticized by Levi (1977) on the grounds that it runs afoul of some basic Bayesian tenets even when the theory delivers precise probability values. We do not pretend to resolve the debate here, but we would like to reformulate the main ideas of this paper in terms of dilation-averse agents, which we believe provides a useful perspective to think about our results. An agent participating in the communication procedure is said to be *dilation-averse* if she does not condition on her information about the value of a partition that dilates the hypothesis of interest, but is otherwise happy to update by Bayesian conditioning. Suppose it is common knowledge that the two agents are dilation-averse. To model this situation, the procedure of communicating posteriors should be modified as follows.

At step 0, for each agent i , if $\mathcal{P}_0^i = \{E \in \mathcal{P}^i \mid E \cap \mathcal{C}_0 \neq \emptyset\}$ dilates H , she updates her credence by conditioning on the common knowledge \mathcal{C}_0 : $\mathbf{Q}_0^1(H) = \mathbf{Q}(H|\mathcal{C}_0)$; otherwise, she updates her credence in the standard way: $\mathbf{Q}_0^1(H) = \mathbf{Q}(H|\mathcal{P}^1(w))$.

At step $n + 1$, the agents announce $\mathbf{Q}_n^1(H)$ and $\mathbf{Q}_n^2(H)$, respectively. Consider the set $\tilde{\mathcal{N}}_{n+1}^i = \{E \in \mathcal{P}_n^i \mid \mathbf{Q}(H|E \cap \mathcal{C}_n) = \mathbf{Q}_n^i(H)\}$. It is easy to see that $\tilde{\mathcal{N}}_{n+1}^i = \emptyset$ if and only if there was dilation at step n . Let

$$\mathcal{N}_{n+1}^i = \begin{cases} \mathcal{P}_n^i & \text{if } \tilde{\mathcal{N}}_{n+1}^i = \emptyset, \\ \tilde{\mathcal{N}}_{n+1}^i & \text{otherwise.} \end{cases}$$

That is, when $\tilde{\mathcal{N}}_{n+1}^i = \emptyset$, no new information is conveyed by the announcement of $\mathbf{Q}_n^i(H)$. As before, let $\mathcal{C}_{n+1} = \bigcup \mathcal{N}_{n+1}^1 \cap \bigcup \mathcal{N}_{n+1}^2$, and $\mathcal{P}_{n+1}^i = \{E \in \mathcal{N}_{n+1}^i \mid E \cap \mathcal{C}_{n+1} \neq \emptyset\}$. Clearly, it remains true that $\mathcal{P}_{n+1}^i \subseteq \mathcal{N}_{n+1}^i \subseteq \mathcal{P}_n^i$ and $\mathcal{C}_{n+1} = \bigcup \mathcal{P}_{n+1}^1 \cap \bigcup \mathcal{P}_{n+1}^2$. If $\mathcal{P}_{n+1}^i = \mathcal{P}_n^i$, or equivalently, if $\mathcal{C}_{n+1} = \mathcal{C}_n$, the procedure stops; otherwise, agent i updates credence of H according to whether $\{E \cap \mathcal{C}_{n+1} \mid E \in \mathcal{P}_{n+1}^i\}$ dilates H . If it does not dilate H , the credence is updated to $\mathbf{Q}_{n+1}^i(H) = \mathbf{Q}(H|\mathcal{P}^i(w) \cap \mathcal{C}_{n+1})$; otherwise, the credence is updated to $\mathbf{Q}_{n+1}^i(H) = \mathbf{Q}(H|\mathcal{C}_{n+1})$.

For instance, if the agents in Example 2 are commonly known to be dilation-averse and follow the above procedure, then at step 0, seeing that her partition $\{\{w_1, w_2\}, \{w_3, w_4\}\}$ dilates H , agent 1 will ignore her private information (i.e., $\{w_1, w_2\}$) and go with $\mathbf{Q}_0^1(H) = \mathbf{Q}(H|\mathcal{C}_0) = \{1/2\}$. Then at step 1, $\mathcal{P}_1^i = \mathcal{N}_1^i = \mathcal{P}_0^i$, and the procedure stops (with a consensus).

As the original, Bayesian procedure of communicating posteriors, this dilation-averse procedure will surely stop at step $m + 1$ for some $m \geq 0$. It is then very easy to adapt the arguments for Theorems 3, 5, and 8 to show the following:

Theorem 9 Suppose that the dilation-averse procedure of communicating posteriors stops at step $m + 1$. Then

- 1) $\underline{\mathbf{Q}}_m^1(H) = \underline{\mathbf{Q}}_m^2(H)$ and $\overline{\mathbf{Q}}_m^1(H) = \overline{\mathbf{Q}}_m^2(H)$;
- 2) If \mathbf{Q} is closed, then $\mathbf{Q}_m^1(H) \cap \mathbf{Q}_m^2(H) \neq \emptyset$; and
- 3) If \mathbf{Q} is closed and connected, then $\mathbf{Q}_m^1(H) = \mathbf{Q}_m^2(H)$.

Proof For each i , either $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ dilates H , in which case $\mathbf{Q}_m^i(H) = \mathbf{Q}(H|\mathcal{C}_m)$ by the design of the procedure, or $\{E \cap \mathcal{C}_m \mid E \in \mathcal{P}_m^i\}$ does not dilate H , in which case the argument for Theorem 3 is applicable to derive that $\underline{\mathbf{Q}}_m^i(H) = \mathbf{Q}(H|\mathcal{C}_m)$ and $\overline{\mathbf{Q}}_m^i(H) = \overline{\mathbf{Q}}(H|\mathcal{C}_m)$. Either way we have 1). The derivations of 2) and 3) from 1) are the same as those of Theorems 8 and 5. ■

Therefore, two agents who are commonly known to be dilation-averse cannot agree to disagree on lower or upper probabilities, and, under common assumptions, cannot agree not to fully agree.

6. Concluding Remarks

Like Aumann's original result, the results in this paper are mathematically simple once the framework is set up, but they highlight an interesting connection between the possibility of agreeing to disagree and the phenomenon of dilation. We offered two perspectives to view this connection. For Bayesian agents with a common set of priors, agreeing to disagree on lower or upper posteriors entails the presence of dilation for at least one of them. For dilation-averse (but otherwise Bayesian) agents with a common set of priors, it is impossible to agree to disagree on lower or upper posteriors.

Although the absence of dilation is sufficient for Bayesian agents to reach agreements by communicating posteriors, it is not necessary. Here is a simple example to show this.

Example 3 Let $\Omega = \{w_1, w_2, w_3, w_4\}$ and \mathcal{A} be its power set. Suppose $\mathcal{P}^1 = \{\{w_1, w_2\}, \{w_3, w_4\}\}$ and $\mathcal{P}^2 = \{\{w_1, w_3\}, \{w_2, w_4\}\}$. Let $H = \{w_1, w_4\}$, and suppose the true state of the world is w_1 . Let \tilde{p} be the uniform distribution over the 4-atom algebra, and Λ be the set of all distributions over the 4-atom algebra. Define $\mathbf{Q} = \{(0.8\tilde{p} + 0.2q) \mid q \in \Lambda\}$.⁸

8. This ϵ -contamination model ($\epsilon = 0.2$) can be equivalently specified as the largest set of probability measures on the 4-atom algebra satisfying the constraint that every atom receives a lower probability of 0.2.

Like Example 2, this is an Aumann case, where the posteriors of H are common knowledge without announcements, because $\mathbf{Q}(H|\{w_1, w_2\}) = \mathbf{Q}(H|\{w_3, w_4\}) = [1/3, 2/3]$, and $\mathbf{Q}(H|\{w_1, w_3\}) = \mathbf{Q}(H|\{w_2, w_4\}) = [1/3, 2/3]$. So the procedure of communicating posteriors stops at step 1, and $\mathcal{C}_1 = \mathcal{C}_0 = \Omega$. Dilation does occur, for both agents, because $\mathbf{Q}(H|\mathcal{C}_0) = \mathbf{Q}(H) = [0.4, 0.6]$, which is strictly contained in $[1/3, 2/3]$. Despite the presence of dilations, the two agents will still reach an agreement even if they are not dilation-averse, though the agreement is different from the one dilation-averse agents would reach.

It is also worth noting that this example is a generalization of an example from Geanakoplos and Polemarchakis (1982), which was used to illustrate the fact we mentioned in Section 2, that the consensus resulting from communicating posteriors can be different from the consensus resulting from directly exchanging private information. If both pieces of private information in the example become public, the two agents will converge on a precise, extreme probability.

We close by mentioning two ways our results may be expanded. First, when “agreement” is interpreted as partial agreement, the common prior assumption may also be relaxed to the assumption that priors (partially) agree, i.e., that the two sets of priors have a non-empty intersection. This is, for example, what Carvajal and Correia-da-Silva (2010) assume in their results. Their main agreement result about Bayesian agents (Proposition 1) is that if two Bayesian agents have closed and connected sets of priors that have a non-empty intersection, and both sets of posteriors on a hypothesis are common knowledge, then the sets of posteriors also have a non-empty intersection. This result, just like Aumann’s original result, is straightforwardly generalizable to the setting of communicating posteriors. The more interesting question, in light of our results here, is what purchase the condition of no dilation has in the context of priors that do not fully agree, or to put it differently, whether stronger agreement results are available in this context for dilation-averse agents.

Second, we have only considered the full Bayesian updating rule (and the dilation-averse variant). Other updating rules may be examined in our setting, especially the Dempster-Shafer or maximum likelihood updating considered by Kajii and Ui (2005) and Carvajal and Correia-da-Silva (2010). For Dempster-Shafer updating, Carvajal and Correia-da-Silva’s main agreement result requires each agent’s set of likelihood maximizers as well as their sets of posteriors to be common knowledge, which suggests that in general communication of posteriors alone is not enough to guarantee agreement. One natural idea is to allow also the communication of likelihood maximizers. On the other hand, Seidenfeld (1997) showed that for ϵ -contamination models (Huber, 1973; Berger, 1984) Dempster-Shafer updating is equivalent to Bayesian updating. Therefore, if we can derive a corollary about ϵ -contamination models (in the spirit of Corollary 7) from Theorem 5 and results on dilation in ϵ -contamination models, that will also apply to Dempster-Shafer updating.

Acknowledgements

This research was supported by the Research Grants Council of Hong Kong under the General Research Fund LU13600715, and by a Faculty Research Grant from Lingnan University.

References

- Aumann, R. (1976). Agreeing to disagree. *Ann. Stat.*, 4: 1236-1239.
- Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics* (J. Kadane, ed.), pp. 63-124. North-Holland, Amsterdam.

- Carvajal, A. and Correia-da-Silva, J. (2010). *Agreeing to disagree with multiple priors* (No. 368). Universidade do Porto, Faculdade de Economia do Porto.
- De Finetti, B. (1972). *Probability, Induction, and Statistics*. John Wiley, New York.
- Geanakoplos, J. and Polemarchakis, M. (1982). We can't disagree forever. *J. Econ. Theory*, 26: 363-390.
- Good, I. J. (1974). A little learning can be dangerous. *Br. J. Philos. Sci.*, 25: 340-342.
- Grünwald, P. D. and Halpern, J. Y. (2004). When ignorance is bliss. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 226-234. AUAI Press.
- Herron, T., Seidenfeld, T. and Wasserman, L. (1997). Divisive conditioning: Further results on dilation. *Philosophy of Science*, 64: 411-444.
- Hill, B. M. and Lane, D. (1985). Conglomerability and countable additivity. *Sankhyā: The Indian Journal of Statistics*, Series A: 366-379.
- Huber, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Internat. Statist. Inst.*, 45: 181-191.
- Levi, I. (1977). Direct inference. *The Journal of Philosophy*, 74(1): 5-29.
- Kajii, A. and Ui, T. (2005). Incomplete information games with multiple priors. *JPN Econ. Rev.*, 56: 332-351.
- Kajii, A., and Ui, T. (2009). Interim efficient allocations under uncertainty. *J. Econ. Theory*, 144: 337-353.
- Kyburg H. E. (1974). *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht.
- Schervish, M. J., Seidenfeld, T. and Kadane, J. B. (1984). The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66(2): 205-226.
- Schervish, M. J., Seidenfeld, T. and Kadane, J. B. (2016). Non-conglomerability for countably additive measures that are not κ -additive. *The Review of Symbolic Logic*, forthcoming.
- Seidenfeld, T. (1981). Levi on the dogma of randomization. In *Henry E. Kyburg and Isaac Levi* (R. Bogdan, ed.), pp. 263-291. Reidel, Dordrecht.
- Seidenfeld, T. (1997). Some static and dynamic aspects of robust Bayesian theory. In *Random Sets*, pp. 385-406. Springer, New York.
- Seidenfeld, T. and Wasserman, L. (1993). Dilation for sets of probabilities. *Ann. Stat.*, 21: 1139-1154.
- Wasserman, L. (1992). Invariance properties of density ratio priors. *Ann. Stat.*, 20: 2177-2182.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York.