Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 159 (2004) 75–125
Artificial Intelligence
www.elsevier.com/locate/artint

Updating beliefs with incomplete observations

Gert de Cooman [a], Marco Zaffalon [b,*]

[a] *Ghent University, Systems Research Group, Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium*
[b] *IDSIA, Galleria 2, 6928 Manno (Lugano), Switzerland*

Received 8 July 2003; accepted 14 May 2004

Abstract

Currently, there is renewed interest in the problem, raised by Shafer in 1985, of updating probabilities when observations are incomplete (or set-valued). This is a fundamental problem in general, and of particular interest for Bayesian networks. Recently, Grünwald and Halpern have shown that commonly used updating strategies fail in this case, except under very special assumptions. In this paper we propose a new method for updating probabilities with incomplete observations. Our approach is deliberately conservative: we make no assumptions about the so-called incompleteness mechanism that associates complete with incomplete observations. We model our ignorance about this mechanism by a vacuous lower prevision, a tool from the theory of imprecise probabilities, and we use only coherence arguments to turn prior into posterior (updated) probabilities. In general, this new approach to updating produces lower and upper posterior probabilities and previsions (expectations), as well as partially determinate decisions. This is a logical consequence of the existing ignorance about the incompleteness mechanism. As an example, we use the new updating method to properly address the apparent paradox in the 'Monty Hall' puzzle. More importantly, we apply it to the problem of classification of new evidence in probabilistic expert systems, where it leads to a new, so-called *conservative updating rule*. In the special case of Bayesian networks constructed using expert knowledge, we provide an exact algorithm to compare classes based on our updating rule, which has linear-time complexity for a class of networks wider than polytrees. This result is then extended to the more general framework of credal networks, where computations are often much harder than with Bayesian nets. Using an example, we show that our rule appears to provide a solid basis for reliable updating with incomplete observations, when no strong assumptions about the incompleteness mechanism are justified.

* Corresponding author.
E-mail addresses: gert.decooman@ugent.be (G. de Cooman), zaffalon@idsia.ch (M. Zaffalon).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2004.05.006

Keywords: Incomplete observations; Updating probabilities; Imprecise probabilities; Coherent lower previsions; Vacuous lower previsions; Naive updating; Conservative updating; Bayesian networks; Credal networks; Puzzles; Credal classification

1. Introduction

Suppose you are given two Boolean random variables, C and A. $C = 1$ represents the presence of a disease and $A = 1$ is the positive result of a medical test. You know that $p(C = 0, A = 0) = 0.99$ and that $p(C = 1, A = 1) = 0.01$, so the test allows you to make a sure diagnosis. However, it may happen that, for some reason, the result of the test is missing. What should your diagnosis be in this case? You might be tempted to say that the posterior probability of $C = 0$, conditional on a missing value of A, is simply $p(C = 0|A \in \{0, 1\}) = p(C = 0) = 0.99$, and that the diagnosis is 'no disease' with high probability. After all, this looks like a straightforward application of Kolmogorov's *definition* of conditional probability, which appears in many textbooks: $P(B|E) = P(B \cap E)/P(E)$, for generic events B and E, with $P(E) > 0$.

Unfortunately, it turns out that the above inference is wrong unless a condition known in the literature as *MAR* (*missing at random*) is satisfied. MAR states that the probability that a measurement for A is missing, is the same both when conditioned on $A = 0$ and when conditioned on $A = 1$, or, in other words, that there is no systematic reason for the missing values of A [25].

The example above is a special case of the more general problem of updating probabilities with observations that are *incomplete*, or set-valued: it could be argued that the fact that a measurement for A is missing corresponds to a set-valued observation of $\{0, 1\}$ for A rather than the *complete* or point-valued observations 0 or 1. The difficulty we are facing is then how to update p with such incomplete observations. To our knowledge, this problem was given serious consideration for the first time in 1985 by Shafer [33]. Rather than taking traditional conditioning as a definition, Shafer derived it from more primitive notions showing that the right way to update probabilities with incomplete observations requires knowledge of what we shall call the *incompleteness mechanism* (called *protocol* in Shafer's paper), i.e., the mechanism that is responsible for turning a complete observation into an incomplete one. Shafer's result tells us that neglecting the incompleteness mechanism leads to a naive application of conditioning (also called *naive conditioning* or *naive updating* in the following) that is doomed to failure in general. This is evident when one addresses well-known puzzles by naive conditioning, such as the three prisoners problem and the Monty Hall puzzle. What the implications are in practise for more realistic applications of probability theory, was partially addressed by Shafer when he observed that "we do not always have protocols in practical problems". In the example above, for instance, we may not know which is the probability that a measurement A is missing conditional on $A = 0$ and conditional on $A = 1$ (such a conditional probability is a specification of the protocol, or incompleteness mechanism). We may not even know whether the two probabilities are equal ...

Surprisingly, Shafer's thesis seems to have been largely overlooked for many years.[1] Kolmogorov's influential formalisation of probability theory [22] may have contributed in this respect: the way the definition of conditioning is presented seems to suggest that one may be totally uninformed about the incompleteness mechanism, and still be allowed to correctly update beliefs after receiving some evidence E. That is, it seems to suggest that naive updating is always the correct way to update beliefs. Actually, the definition produces correct results when MAR does not hold only if the underlying possibility space is built in such a way as to also model the incompleteness mechanism. Apart from the influence of Kolmogorov's formalisation, we might identify the unclear practical implications of Shafer's work as another reason for its being considered by many as something of a statistical curiosity.

The situation has changed recently, when an interesting paper by Grünwald and Halpern [14] kindled a renewed interest in the subject. In that work, strong arguments are presented for the following two theses: (i) the incompleteness mechanism may be unknown, or difficult to model; and (ii) the condition of *coarsening at random* (or CAR [13], a condition more general than MAR), which guarantees that naive updating produces correct results, holds rather infrequently. These two points taken together do raise a fundamental issue in probability theory, which also presents a serious problem for applications: how should one update beliefs when little or no information is available about the incompleteness mechanism?

In the above example, the mechanism might very well be such that A cannot be observed if and only if A has the value 0, and then $C = 0$ would be a certain conclusion. But it might equally well be the case that A cannot be observed if $A = 1$, in which case $C = 1$ would be certain. Of course, all the intermediate randomised cases might also be possible. It follows that the posterior probability of $C = 0$ can, for all we know, lie anywhere in the interval $[0, 1]$, and our ignorance does not allow us to say that one value is more likely than another. In other words, this probability is *vacuous*. Thus, knowing that the value of A is missing, produces complete ignorance about this probability and, as a result, total indeterminacy about the diagnosis: we have no reason to prefer $C = 0$ over $C = 1$, or *vice versa*. All of this is a necessary and logical consequence of our ignorance about the incompleteness mechanism. We cannot get around this indeterminacy, unless we go back to the medical test and gather more relevant information about how it may produce missing values.

Generally speaking, we believe that the first step to answer the question above is to recognise that there may indeed be ignorance about the incompleteness mechanism, and to allow for such ignorance in our models. This is the approach that we take in this paper. In Section 3, we make our model as conservative as possible by representing the ignorance about the incompleteness mechanism by a *vacuous lower prevision*, a tool from the theory of imprecise probabilities [37]. Because we are aware that readers may not be familiar with imprecise probability models, we present a brief discussion in Section 2, with pointers to the relevant literature.[2] Loosely speaking, the vacuous lower prevision

[1] But see the discussion in [37, Section 6.11], which has been a source of inspiration for the present work; and some papers by Halpern et al. [15,16].

[2] See also [39] for a gentle and less dense introduction to imprecise probabilities with emphasis on artificial intelligence.

is equivalent to the set of all distributions, i.e., it makes all incompleteness mechanisms possible *a priori*. Our basic model follows from this as a necessary consequence, using the rationality requirement of *coherence*. This coherence is a generalisation to its imprecise counterpart of the requirements of rationality in precise, Bayesian, probability theory [9]. We illustrate how our basic model works by addressing the Monty Hall puzzle, showing that the apparent paradox vanishes if the knowledge that is actually available about the incompleteness mechanism is modelled properly.

We then apply our method for dealing with incomplete observations to the special case of a classification problem, where objects are assigned to classes on the basis of the values of their attributes. The question we deal with in Section 4, is how classification should be done when values for some of the attributes are missing. We derive a new updating rule that allows us to deal with such missing data without making unwarranted assumptions about the mechanism that produces these missing values. We regard this so-called *conservative updating rule* as a significant step toward a general solution of the updating problem. Our rule leads to an imprecise posterior, and as we argued above, it may lead to inferences that are partially indeterminate. It may for instance happen that, due to the fact that certain of the attribute values are missing, our method will assign an object to a number of (optimal) classes, rather than to a single class, and that it does not express any preference between these optimal classes. This generalised way of doing classification is also called *credal classification* in [41]. As we have argued above, we have to accept that this is the best our system can do, given the information that is incorporated into it. If we want a more precise classification, we shall have to go back and find out more about the mechanism that is responsible for the fact that some attributes are missing. But, given the characteristics of our approach, any such additional information will lead to a new classification that refines ours, but can never contradict it, i.e., assign an object to a class that was not among our optimal classes in the first place.

In Section 5, we then apply the updating rule for classification problems to Bayesian networks. We regard a Bayesian net as a tool that formalises expert knowledge and is used to classify new evidence, i.e., to select certain values of a class variable given evidence about the attribute values. We develop an exact algorithm for credal classification with Bayesian nets that makes pairwise comparison of classes in linear time in the size of the input, when the class node together with its Markov blanket is a singly connected graph. Extension to the general case is provided by an approach analogous to *loop cutset conditioning*. Section 6.1 applies the algorithm to an artificial problem and clarifies the differences with naive updating. There are two important implications of the algorithmic complexity achieved with Bayesian nets: the algorithm makes the new rule immediately available for applications; and it shows that it is possible for the power of robust, conservative, modelling to go hand in hand with efficient computation, even for some multiply connected networks. This is enforced by our next result: the extension of the classification algorithm to *credal networks*, in Section 7, with the same complexity. Credal networks are a convenient way to specify partial prior knowledge. They extend the formalism of Bayesian networks by allowing a specification in terms of sets of probability measures. Credal nets allow the inherent imprecision in human knowledge to be modelled carefully and expert systems to be developed rapidly. Such remarkable

advantages have been partially overshadowed so far by the computational complexity of working in the more general framework of credal nets. Our result shows that in many realistic scenarios, the computational effort with credal networks is the same as that required by Bayesian nets. This may open up a wealth of potential applications for credal networks.

The concluding Section 8 discusses directions and open issues for future research. Additional, technical results have been gathered in the appendices.

2. Basic notions from the theory of imprecise probabilities

The theory of coherent lower previsions (sometimes also called the theory of *imprecise probabilities*[3]) [37] is an extension of the Bayesian theory of (precise) probability [7,9]. It intends to model a subject's uncertainty by looking at his dispositions toward taking certain actions, and imposing requirements of rationality, or consistency, on these dispositions.

To make this more clear, consider a random variable X that may take values in a finite[4] set $\mathcal{X}$. A *gamble* f on the value of X, or more simply, a gamble on $\mathcal{X}$, is a real-valued function on $\mathcal{X}$. It associates a (possibly negative) reward[5] $f(x)$ with any value x the random variable X may assume. If a subject is uncertain about what value X assumes in $\mathcal{X}$, he will be disposed to accept certain gambles, and to reject others, and we may model his uncertainty by looking at which gambles he accepts (or rejects).

In the Bayesian theory of uncertainty (see for instance [9]), it is assumed that a subject can always specify a *fair price*, or *prevision*, $P(f)$ for f, whatever the information available to him. $P(f)$ is the unique real number such that the subject (i) accepts the gamble $f - p$, i.e., accepts to buy the gamble f for a price p, for all $p < P(f)$; and (ii) accepts the gamble $q - f$, i.e., accepts to sell the gamble f for a price q, for all $q > P(f)$. In other words, it is assumed that for essentially any real number r, the available information allows the subject to decide which of the following two options he prefers: buying f for price r, or selling f for that price.

It has been argued extensively [34,37] that, especially if little information is available about X, there may be prices r for which a subject may have no real preference between these two options, or in other words, that on the basis of the available information he remains *undecided* about whether to buy f for price r or to sell it for that price: he may not be disposed to do either. If, as the Bayesian theory requires, the subject *should* choose between these two actions, his choice will then not be based on any real preference: it will be arbitrary, and not a realistic reflection of the subject's dispositions, based on the available information.

[3] Other related names found in the literature are: indeterminate probabilities, interval (or interval-valued) probabilities, credal sets, ...

[4] For simplicity, we shall only deal with variables with a *finite* number of possible values in this paper.

[5] In order to make things as simple as possible, we shall assume that these rewards are expressed in units of some predetermined *linear* utility.

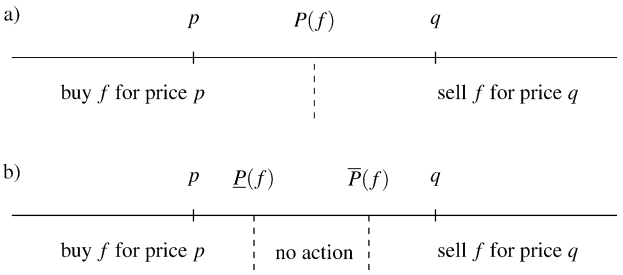

Fig. 1. Buying and selling a gamble f in (a) the Bayesian theory, and (b) in imprecise probability theory.

2.1. Coherent lower and upper previsions

The theory of imprecise probabilities remedies this by allowing a subject to specify two numbers: $\underline{P}(f)$ and $\overline{P}(f)$. The subject's *lower prevision* $\underline{P}(f)$ for f is the greatest real number p such that he is disposed to buy the gamble f for all prices strictly smaller than p, and his *upper prevision* $\overline{P}(f)$ for f is the smallest real number q such that he is disposed to sell f for all prices strictly greater than q. For any r between $\underline{P}(f)$ and $\overline{P}(f)$, the subject does not express a preference between buying or selling f for price r (see Fig. 1).

Since selling a gamble f for price r is the same thing as buying $-f$ for price $-r$, we have the following conjugacy relationship between lower and upper previsions

$$\overline{P}(f) = -\underline{P}(-f). \tag{1}$$

This tells us that whatever we say about upper previsions can always be reformulated in terms of lower previsions. We shall therefore concentrate on lower previsions. It will for the purposes of this paper suffice to consider lower previsions $\underline{P}$ that are defined on the set $\mathcal{L}(\mathcal{X})$ of all gambles on $\mathcal{X}$, i.e., $\underline{P}$ is considered as a function that maps any gamble f on $\mathcal{X}$ to the real number $\underline{P}(f)$.

An *event* A is a subset of $\mathcal{X}$, and it will be identified with its *indicator* I_A, which is a gamble assuming the value one on A and zero elsewhere. We also denote $\underline{P}(I_A)$ by $\underline{P}(A)$ and call it the *lower probability* of the event A. It is the supremum rate for which the subject is disposed to bet on the event A. Similarly, the *upper probability* $\overline{P}(A) = \overline{P}(I_A) = 1 - \underline{P}(\text{co}\, A)$ is one minus the supremum rate for which the subject is disposed to bet against A, i.e., to bet on the complementary event co A. Thus, events are special gambles, and lower and upper probabilities are special cases of lower and upper previsions. We use the more general language of gambles, rather than the more common language of events, because Walley [37] has shown that in the context of imprecise probabilities, the former is much more expressive and powerful.[6] For this reason, we consider 'lower prevision' to be the primary notion, and 'lower probability' to be derived from it; and we follow de Finetti's [9] and Walley's [37] example in using the same symbol P for both (lower) *p*revisions and (lower) *p*robabilities. Standard probabilistic

[6] We shall see in Section 2.2 that for precise probabilities, both languages turn out to be equally expressive.

practice would have us use the symbols E for *e*xpectation/prevision and P for *p*robability here.[7]

Since lower previsions represent a subject's dispositions to act in certain ways, they should satisfy certain criteria that ensure that these dispositions are rational. *Coherence* is the strongest such rationality requirement that is considered in the theory of imprecise probabilities. For a detailed definition and motivation, we refer to [37]. For the purposes of the present discussion, it suffices to mention that a lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X})$ is coherent if and only if it satisfies the following properties, for all gambles f and g on $\mathcal{X}$, and all non-negative real numbers λ:

($\underline{P}1$) $\min_{x\in\mathcal{X}} f(x) \leqslant \underline{P}(f)$ [accepting sure gains];
($\underline{P}2$) $\underline{P}(f+g) \geqslant \underline{P}(f) + \underline{P}(g)$ [super-additivity];
($\underline{P}3$) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$ [positive homogeneity].

Observe that for a coherent $\underline{P}$, we have that $\overline{P}(f) \geqslant \underline{P}(f)$ for all $f \in \mathcal{L}(\mathcal{X})$.

2.2. Linear previsions

It follows from the behavioural interpretation of lower and upper previsions that if $\underline{P}(f) = \overline{P}(f)$ for some gamble f, then this common value is nothing but the fair price, or prevision, $P(f)$ of f, as discussed in the previous section. A *linear prevision* P on $\mathcal{L}(\mathcal{X})$ is defined as a real-valued map on $\mathcal{L}(\mathcal{X})$ that is coherent when interpreted as a lower prevision, and *self-conjugate* in the sense that $P(f) = -P(-f)$ for all gambles f, so the conjugate upper prevision of P is also given by P. This implies that a linear prevision P should satisfy the following properties, for all gambles f and g on $\mathcal{X}$, and all real numbers λ:

($P1$) $\min_{x\in\mathcal{X}} f(x) \leqslant P(f) \leqslant \max_{x\in\mathcal{X}} f(x)$;
($P2$) $P(f+g) = P(f) + P(g)$;
($P3$) $P(\lambda f) = \lambda P(f)$.

This follows at once from the characterisation ($\underline{P}1$)–($\underline{P}3$) of a coherent lower prevision, and the conjugacy relationship (1). Thus, linear previsions turn out to be exactly the same thing as de Finetti's coherent previsions [7,9]. They are the so-called *precise* probability models, which turn out to be special cases of the more general coherent imprecise probability models. Any linear prevision P is completely determined by its so-called *mass function* p, defined by $p(x) = P(\{x\})$, since it follows from the axioms ($P2$) and ($P3$) that for any gamble f,

$$P(f) = \sum_{x\in\mathcal{X}} f(x)p(x)$$

is the expectation of f associated with the mass function p. We denote the set of all linear previsions on $\mathcal{L}(\mathcal{X})$ by $\mathcal{P}(\mathcal{X})$.

[7] Instead, we shall reserve the symbol E for natural extension.

2.3. Sets of linear previsions

With any lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X})$, we can associate its set of dominating linear previsions:

$$\mathcal{M}(\underline{P}) = \big\{P \in \mathcal{P}(\mathcal{X}):\ \big(\forall f \in \mathcal{L}(\mathcal{X})\big)\big(\underline{P}(f) \leqslant P(f)\big)\big\}.$$

Observe that this set $\mathcal{M}(\underline{P})$ is convex and closed.[8] It turns out that the lower prevision $\underline{P}$ is coherent if and only if $\mathcal{M}(\underline{P}) \neq \emptyset$, and if moreover $\underline{P}$ is the lower envelope of $\mathcal{M}(\underline{P})$: for all gambles f on $\mathcal{X}$,[9]

$$\underline{P}(f) = \inf\big\{P(f):\ P \in \mathcal{M}(\underline{P})\big\}.$$

Conversely, the lower envelope $\underline{P}$ of any non-empty subset $\mathcal{M}$ of $\mathcal{P}(\mathcal{X})$, defined by $\underline{P}(f) = \inf\{P(f):\ P \in \mathcal{M}\}$ for all $f \in \mathcal{L}(\mathcal{X})$, is a coherent lower prevision. Moreover $\mathcal{M}(\underline{P}) = \overline{\mathrm{CH}}(\mathcal{M})$, where $\overline{\mathrm{CH}}(\mathcal{M})$ is the convex closure (i.e., the topological closure of the convex hull) of $\mathcal{M}$ [37, Chapters 2 and 3]. This tells us that working with coherent lower previsions is equivalent to working with convex closed sets of linear previsions. It also tells us that a coherent lower prevision $\underline{P}$ is also the lower envelope of the set $\mathrm{ext}(\mathcal{M}(\underline{P}))$ of the set of extreme points of $\mathcal{M}(\underline{P})$.

This brings us to the so-called *Bayesian sensitivity analysis interpretation* of a lower prevision $\underline{P}$ or a set of linear previsions $\mathcal{M}$. On this view, a subject's uncertainty should always be described by some ideal probability measure, or equivalently, by some linear prevision P_T. We could call this the *assumption of ideal precision*. Due to lack of time, resources or elicitation, we may not be able to uniquely identify P_T, but we may often specify a set $\mathcal{M}$ such that we are certain that $P_T \in \mathcal{M}$, or equivalently, a lower prevision $\underline{P}$ such that $\underline{P} \leqslant P_T$. On this view, any conclusions or inferences we derive from the available information must be *robust*: they must be valid for all possible candidates $P \in \mathcal{M}$ for the ideal prevision P_T. Although we emphatically do not make the assumption of ideal precision in this paper, we shall see that many of the results we derive, are compatible with it, i.e., they can also be given a Bayesian sensitivity analysis interpretation.

2.4. Vacuous lower previsions

There is a class of coherent lower previsions that deserves special attention. Consider a non-empty subset B of $\mathcal{X}$. Then the *vacuous lower prevision $\underline{P}_B$ relative to B* is defined by

$$\underline{P}_B(f) = \min_{x \in B} f(x) \tag{2}$$

for all gambles f on $\mathcal{X}$. Verify that $\underline{P}_B$ is a coherent lower prevision, and moreover

$$\mathcal{M}(\underline{P}_B) = \big\{P \in \mathcal{P}(\mathcal{X}):\ P(B) = 1\big\}.$$

[8] We only consider the topology of point-wise convergence on $\mathcal{P}(\mathcal{X})$. If we identify linear previsions with their mass functions, which can in turn be identified with elements of the unit simplex in $\mathbb{R}^n$, where n is the cardinality of $\mathcal{X}$, this topology is also the relativisation to this unit simplex of the usual Euclidean (metric) topology on $\mathbb{R}^n$.

[9] Since $\mathcal{M}(\underline{P})$ is convex and closed, this infimum is actually achieved, and it can be replaced by a minimum.

This tells us that $\underline{P}_B$ is the smallest (and therefore most conservative) coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X})$ that satisfies $\underline{P}(B) = 1$ (and therefore $\overline{P}(B) = P(B) = 1$). $\underline{P}(B) = 1$ means that it is *practically certain* to the subject that X assumes a value in B, since he is disposed to bet at all non-trivial odds on this event. Thus, in the context of the theory of lower probabilities, $\underline{P}_B$ is the appropriate model for the piece of information that 'X assumes a value in B' *and nothing more*: any other coherent lower prevision $\underline{P}$ that satisfies $\underline{P}(B) = 1$ dominates $\underline{P}_B$, and therefore represents stronger behavioural dispositions than those required by coherence and this piece of information alone. Also observe that

$$\operatorname{ext}\big(\mathcal{M}(\underline{P}_B)\big) = \{P_x\colon\ x \in B\},$$

where P_x is the (degenerate) linear prevision on $\mathcal{L}(\mathcal{X})$ all of whose probability mass lies in x, defined by $P_x(f) = f(x)$ for all gambles f on $\mathcal{X}$. $\underline{P}_B$ is therefore the lower envelope of this set of (degenerate) linear previsions, as is also apparent from Eq. (2).

2.5. Marginal lower previsions

Now consider another random variable Y that may assume values in a finite set $\mathcal{Y}$. A coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ is a model for a subject's uncertainty about the values that the joint random variable (X, Y) assumes in $\mathcal{X} \times \mathcal{Y}$. We can associate with $\underline{P}$ the so-called *marginal* lower prevision $\underline{P}_Y$ on $\mathcal{L}(\mathcal{Y})$, defined as follows:

$$\underline{P}_Y(g) = \underline{P}(g')$$

for all $g \in \mathcal{L}(\mathcal{Y})$, where the gamble g' on $\mathcal{X} \times \mathcal{Y}$ is defined by $g'(x, y) = g(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In what follows, we shall identify g and g', and simply write $\underline{P}(g)$ rather than $\underline{P}(g')$. The marginal $\underline{P}_X$ on $\mathcal{L}(\mathcal{X})$ is defined similarly.

The marginal $\underline{P}_Y$ is the corresponding model for the subject's uncertainty about the value that Y assumes in $\mathcal{Y}$, irrespective of what value X assumes in $\mathcal{X}$.

If P is in particular a linear prevision, its marginal P_Y is a linear prevision too, and its mass function p_Y is given by the well-known formula

$$p_Y(y) = P\big(\mathcal{X} \times \{y\}\big) = \sum_{x \in \mathcal{X}} p(x, y).$$

2.6. Conditional lower previsions and separate coherence

Consider any gamble h on $\mathcal{X} \times \mathcal{Y}$ and any value $y \in \mathcal{Y}$. A subject's *conditional lower prevision* $\underline{P}(h|Y = y)$, also denoted as $\underline{P}(h|y)$, is the highest real number p for which the subject would buy the gamble h for any price strictly lower than p, if he knew in addition that the variable Y assumes the value y (and nothing more!).

We shall denote by $\underline{P}(h|Y)$ the *gamble* on $\mathcal{Y}$ that assumes the value $\underline{P}(h|Y = y) = \underline{P}(h|y)$ in $y \in \mathcal{Y}$. We can for the purposes of this paper assume that $\underline{P}(h|Y)$ is defined for all gambles h on $\mathcal{X} \times \mathcal{Y}$, and we call $\underline{P}(\cdot|Y)$ a conditional lower prevision on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$. Observe that $\underline{P}(\cdot|Y)$ maps any gamble h on $\mathcal{X} \times \mathcal{Y}$ to the gamble $\underline{P}(h|Y)$ on $\mathcal{Y}$.

Conditional lower previsions should of course also satisfy certain rationality criteria. $\underline{P}(\cdot|Y)$ is called *separately coherent* if for all $y \in \mathcal{Y}$, $\underline{P}(\cdot|y)$ is a coherent lower prevision on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$, and if moreover $\underline{P}(\mathcal{X} \times \{y\}|y) = 1$. This last condition is natural since it

simply expresses that if the subject knew that $Y = y$, he would be disposed to bet at all non-trivial odds on the event that $Y = y$.

It is a consequence of separate coherence that for all h in $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ and all $y \in \mathcal{Y}$,

$$\underline{P}(h|y) = \underline{P}\big(h(\cdot, y)|y\big).$$

This implies that a separately coherent $\underline{P}(\cdot|Y)$ is completely determined by the values $\underline{P}(f|Y)$ that it assumes in the gambles f on $\mathcal{X}$ alone. We shall use this very useful property repeatedly throughout the paper.

2.7. Joint coherence and the Generalised Bayes Rule

If besides the (separately coherent) conditional lower prevision $\underline{P}(\cdot|Y)$ on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$, the subject has also specified a coherent (unconditional) lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$, then $\underline{P}$ and $\underline{P}(\cdot|Y)$ should in addition satisfy the consistency criterion of *joint coherence*. This criterion is discussed and motivated at great length in [37, Chapter 6]. For our present purposes, it suffices to mention that $\underline{P}$ and $\underline{P}(\cdot|Y)$ are jointly coherent if and only if

$$\underline{P}\big(I_{\mathcal{X}\times\{y\}}\big[h - \underline{P}(h|y)\big]\big) = 0 \quad \text{for all } y \in \mathcal{Y} \text{ and all } h \in \mathcal{L}(\mathcal{X} \times \mathcal{Y}). \tag{GBR}$$

If $\underline{P}$ is a linear prevision P, this can be rewritten as $P(hI_{\mathcal{X}\times\{y\}}) = \underline{P}(h|y)P(\mathcal{X} \times \{y\})$, and if $p_Y(y) = P_Y(\{y\}) = P(\mathcal{X} \times \{y\}) > 0$ it follows that $\underline{P}(\cdot|y)$ is the precise (linear) prevision given by Bayes' rule:

$$\underline{P}(h|y) = P(h|y) = \frac{P(hI_{\mathcal{X}\times\{y\}})}{P(\mathcal{X} \times \{y\})},$$

or equivalently, in terms of mass functions: if $p_Y(y) > 0$ then $p(x|y) = p(x, y)/p_Y(y)$. For this reason, the joint coherence condition given above is also called the *Generalised Bayes Rule* (GBR, for short). It can be shown [37, Theorem 6.4.1] that if $\underline{P}(\mathcal{X} \times \{y\}) > 0$, then $\underline{P}(h|y)$ is uniquely determined by this condition, or in other words: it is the unique solution of the following equation in μ:

$$\underline{P}\big(I_{\mathcal{X}\times\{y\}}[h - \mu]\big) = 0.$$

Equivalently, we then have that

$$\underline{P}(h|y) = \inf\left\{\frac{P(hI_{\mathcal{X}\times\{y\}})}{P(\mathcal{X} \times \{y\})}\colon\ P \in \mathcal{M}(\underline{P})\right\},$$

i.e., the uniquely coherent conditional lower prevision is obtained by applying Bayes' rule to every linear prevision in $\mathcal{M}(\underline{P})$, and then taking the lower envelope. For this reason, this procedure for obtaining a conditional from a joint lower prevision is also called *divisive conditioning* by Seidenfeld et al. [17,32].

2.8. Natural and regular extension

If $\underline{P}(\mathcal{X} \times \{y\}) > 0$, then the conditional lower prevision $\underline{P}(\cdot|y)$ is uniquely determined by the unconditional lower prevision $\underline{P}$. But this is no longer necessarily the case if $\underline{P}(\mathcal{X} \times \{y\}) = 0$ (something similar holds in the Bayesian theory for precise previsions

P if $p_Y(y) = P(\mathcal{X} \times \{y\}) = 0$). The smallest, or most conservative, conditional lower prevision $\underline{E}(\cdot|Y)$ that is jointly coherent with the joint lower prevision $\underline{P}$ is called the *natural extension* of $\underline{P}$ to a conditional lower prevision. For any gamble h on $\mathcal{X} \times \mathcal{Y}$ and y in $\mathcal{Y}$, it is uniquely determined by the GBR if $\underline{P}(\mathcal{X} \times \{y\}) > 0$, and by

$$\underline{E}(h|y) = \min_{x \in \mathcal{X}} h(x, y)$$

if $\underline{P}(\mathcal{X} \times \{y\}) = 0$, i.e., $\underline{E}(\cdot|y)$ is then the *vacuous* lower prevision relative to the set $\mathcal{X} \times \{y\}$.

In certain cases, it may be felt that natural extension is too conservative when $\underline{P}(\mathcal{X} \times \{y\}) = 0$. The following procedure, called *regular extension*, allows us to associate with any coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ another (separately coherent) conditional lower prevision $\underline{R}(\cdot|Y)$ that is jointly coherent with $\underline{P}$:

(RE1) if $\overline{P}(\mathcal{X} \times \{y\}) > 0$, then $\underline{R}(h|y)$ is the greatest solution of the following inequality in μ:

$$\underline{P}\big(I_{\mathcal{X} \times \{y\}}[h - \mu]\big) \geqslant 0;$$

(RE2) if $\overline{P}(\mathcal{X} \times \{y\}) = 0$, then $\underline{R}(\cdot|y)$ is the vacuous lower prevision relative to $\mathcal{X} \times \{y\}$:

$$\underline{R}(h|y) = \min_{x \in \mathcal{X}} h(x, y);$$

where h is any gamble on $\mathcal{X} \times \mathcal{Y}$. Regular extension coincides with natural extension unless $\underline{P}(\mathcal{X} \times \{y\}) = 0$ and $\overline{P}(\mathcal{X} \times \{y\}) > 0$, in which case natural extension is vacuous and regular extension can be much less conservative. We shall see examples of this in the following sections. The regular extension $\underline{R}(\cdot|Y)$ is the smallest, or most conservative, conditional lower prevision that is coherent with the joint $\underline{P}$ and satisfies an additional regularity condition. It is the appropriate conditioning rule to use if a subject accepts precisely those gambles h for which $\underline{P}(h) \geqslant 0$ and $\overline{P}(h) > 0$ (see [37, Appendix J] for more details). It is especially interesting because it has a nice interpretation in terms of sets of linear previsions: if $\overline{P}(\mathcal{X} \times \{y\}) > 0$ it can be shown quite easily that

$$\underline{R}(h|y) = \inf\left\{\frac{P(hI_{\mathcal{X} \times \{y\}})}{P(\mathcal{X} \times \{y\})}\colon\ P \in \mathcal{M}(\underline{P}) \text{ and } P\big(\mathcal{X} \times \{y\}\big) > 0\right\}.$$

Thus, $\underline{R}(h|y)$ can be obtained by applying Bayes' rule (whenever possible) to the precise previsions in $\mathcal{M}(\underline{P})$, and then taking the infimum. Regular extension therefore seems the right way to update lower previsions on the Bayesian sensitivity analysis interpretation as well. It has been called *Bayesian updating* of coherent lower previsions by for instance Jaffray [18]. Regular extension is also used for updating in one of the more successful imprecise probability models, namely Walley's Imprecise Dirichlet Model [38], where using natural extension would lead to completely vacuous inferences. Also see [5,10,36, 37,39] for more information about this type of updating.

2.9. Marginal extension

It may also happen that besides a (separately coherent) conditional lower prevision $\underline{P}(\cdot|Y)$ on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ (or equivalently, through separate coherence, on $\mathcal{L}(\mathcal{X})$), we also have a

coherent marginal lower prevision $\underline{P}_Y$ on $\mathcal{L}(\mathcal{Y})$ modelling the available information about the value that Y assumes in $\mathcal{Y}$.

We can then ask ourselves whether there exists a coherent lower prevision $\underline{P}$ on all of $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ that (i) has marginal $\underline{P}_Y$, and (ii) is jointly coherent with $\underline{P}(\cdot|Y)$. It turns out that this is always possible. In fact, we have the following general theorem (a special case of [37, Theorem 6.7.2]), which is easily proved using the results in the discussion above.

Theorem 1 (Marginal extension theorem). *Let $\underline{P}_Y$ be a coherent lower prevision on $\mathcal{L}(\mathcal{Y})$, and let $\underline{P}(\cdot|Y)$ be a separately coherent conditional lower prevision on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$. Then the smallest* (*most conservative*) *coherent lower prevision on $\mathcal{L}(\mathcal{X} \times \mathcal{Y})$ that has marginal $\underline{P}_Y$ and that is jointly coherent with $\underline{P}(\cdot|Y)$ is given by*

$$\underline{P}(h) = \underline{P}_Y\big(\underline{P}(h|Y)\big) \tag{3}$$

for all gambles h on $\mathcal{X} \times \mathcal{Y}$.

For a linear marginal P_Y and a conditional linear prevision $P(\cdot|Y)$, we again recover well-known results: the marginal extension is the linear prevision $P = P_Y(P(\cdot|Y))$. In terms of mass functions, the marginal extension of the marginal $p_Y(y)$ and the conditional $p(y|x)$ is given by $p(x, y) = p(x|y)p_Y(y)$. Walley has shown [37, Section 6.7] that marginal extension also has a natural Bayesian sensitivity analysis interpretation in terms of sets of linear previsions: for any gamble h on $\mathcal{X} \times \mathcal{Y}$, we have that

$$\begin{aligned}\underline{P}(h) &= \underline{P}_Y\big(\underline{P}(h|Y)\big)\\ &= \inf\big\{P_Y\big(P(h|Y)\big):\ P_Y \in \mathcal{M}(\underline{P}_Y) \text{ and } (\forall y \in \mathcal{Y})\big(P(\cdot|y) \in \mathcal{M}\big(\underline{P}(\cdot|y)\big)\big)\big\}.\end{aligned} \tag{4}$$

The marginal extension of $\underline{P}_Y$ and $\underline{P}(\cdot|Y)$ can in other words be obtained by forming the marginal extension for their compatible, dominating linear previsions, and then taking the infimum. In this infimum, the sets $\mathcal{M}(\underline{P}_Y)$ and $\mathcal{M}(\underline{P}(\cdot|y))$ can be replaced by their sets of extreme points.

2.10. Decision making

Suppose we have two actions a and b, whose outcome depends on the actual value that the variable X assumes in $\mathcal{X}$. Let us denote by f_a the gamble on $\mathcal{X}$ representing the uncertain utility resulting from action a: a subject who takes action a receives $f_a(x)$ units of utility if the value of X turns out to be x. Similar remarks hold for the gamble f_b.

If the subject is uncertain about the value of X, it is not immediately clear which of the two actions he should prefer.[10] But let us assume that he has modelled his uncertainty by a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X})$. Then he *strictly prefers* action a to action b, which we denote as $a > b$, if he is willing to pay some strictly positive amount in order to exchange the (uncertain) rewards of b for those of a. Using the behavioural definition of the lower prevision $\underline{P}$, this can be written as

$$a > b \quad \Leftrightarrow \quad \underline{P}(f_a - f_b) > 0. \tag{5}$$

[10] Unless f_a point-wise dominates f_b or *vice versa*, which we shall assume is not the case.

If $\underline{P}$ is a linear prevision P, this is equivalent to $P(f_a) > P(f_b)$: the subject strictly prefers the action with the highest expected utility. It is easy to see that $\underline{P}(f_a - f_b) > 0$ can also be written as

$$\big(\forall P \in \mathcal{M}(\underline{P})\big)\big(P(f_a) > P(f_b)\big).$$

In other words, $a > b$ if and only if action a yields a higher expected utility than b for every linear prevision compatible with the subject's model $\underline{P}$. This means that $>$ also has a reasonable Bayesian sensitivity analysis interpretation. We shall say that a subject *marginally prefers* a over b if $\underline{P}(f_a - f_b) \geqslant 0$, i.e., when he is willing to exchange f_b for f_a in return for any strictly positive amount of utility.

If we now have some finite set of actions K, and an associated set of uncertain rewards $\{f_a\colon a \in K\}$, then it follows from the coherence of the lower prevision $\underline{P}$ that the binary relation $>$ on K is a strict partial order, i.e., it is transitive and irreflexive. Optimal actions a are those elements of K that are *undominated*, i.e., to which no other actions b in K are strictly preferred: $(\forall b \in K)(b \not> a)$, or equivalently, after some manipulations,

$$(\forall b \in K)\big(\overline{P}(f_a - f_b) \geqslant 0\big).$$

We shall call such actions $\underline{P}$-*maximal* (in K). If $\underline{P}$ is a linear prevision P, the P-maximal actions are simply those actions a in K with the highest expected utility $P(f_a)$.

Two actions a and b are called *equivalent* to a subject, which we denote as $a \approx b$, if he is disposed to (marginally) exchange any of them for the other, i.e., if both $\underline{P}(f_a - f_b) \geqslant 0$ and $\underline{P}(f_b - f_a) \geqslant 0$, or equivalently,

$$a \approx b \quad \Leftrightarrow \quad \overline{P}(f_a - f_b) = \underline{P}(f_a - f_b) = \underline{P}(f_b - f_a) = \overline{P}(f_b - f_a) = 0.$$

When $\underline{P}$ is a linear prevision P, this happens precisely when $P(f_a) = P(f_b)$, i.e., when both actions have the same expected utility.

When $\underline{P}$ is imprecise, two actions a and b may be *incomparable*: they are neither equivalent, nor is either action strictly preferred over the other. This happens when both $\underline{P}(f_a - f_b) \leqslant 0$ and $\underline{P}(f_b - f_a) \leqslant 0$ and at least one of these inequalities is strict. This means that the subject has no preference (not even a marginal one) for one action over the other; he is undecided. Note that this cannot happen for precise previsions.

Any two $\underline{P}$-maximal actions are either equivalent (they always are when $\underline{P}$ is precise), or incomparable, meaning that the information present in the model $\underline{P}$ does not allow the subject to choose between them. It is an essential feature of imprecise probability models that they allow for this kind of indecision.

3. Incomplete observations

We are now ready to describe our basic model for dealing with incomplete observations. It is a general model that describes a situation where we want to measure, or determine, the value of a certain variable X, but for some reason can do so only in an imperfect manner: we perform some kind of measurement whose outcome is O, but this does not allow us to completely determine the value of X.

Let us give a few concrete examples to make this more clear. Suppose we want to measure the voltage (X) across a resistor, but the read-out (O) of our digital voltage meter rounds this voltage to the next millivolt (mV). So if, say, we read that $O = 12$ mV, we only know that the voltage X belongs to the interval (11 mV, 12 mV].

In the example in the Introduction, $X = A$ is the result of the medical test. If we know the result x of the test, then we say that we observe $O = x$. But if the test result is missing, we could indicate this by saying that $O = *$ (or any other symbol to denote that we do not get a test result 0 or 1). In that case, we only know that X belongs to the set $\{0, 1\}$.

In the well-known *three-prisoner problem*, three prisoners a, b and c are waiting to be executed when it is decided that one of them, chosen randomly, is to be set free. The warden tells prisoner a the name of one of the other two convicts, who has not been reprieved. The question is then if what the warden tells a gives him more information about whether he will be executed or not. This can also be seen as a case of an incomplete observation: the variable X identifies which prisoner is to be reprieved, and the observation O is what the warden tells prisoner a. If for instance a is reprieved, then the warden will name either b or c: we then know that O can take any value in the set $\{b, c\}$. Conversely, if the warden names prisoner b, so $O = b$, then all we know is that the variable X can take any value in $\{a, c\}$, so again X is not completely determined by the observation O. We shall see other concrete examples further in this section as well as in the next section.

Let us now present a formal mathematical model that represents the features that are common to problems of this type. We consider a random variable X that may assume values in a *finite* set $\mathcal{X}$. Suppose that we have some model for the available information about what value X will assume in $\mathcal{X}$, which takes the form of a coherent lower prevision $\underline{P}_0$ defined on $\mathcal{L}(\mathcal{X})$.

We now receive additional information about the value of X by observing the value that another random variable O (the observation) assumes in a *finite* set of possible values $\mathcal{O}$. Only, these observations are *incomplete* in the following sense: the value of O does not allow us to identify the value of X uniquely. In fact, the only information we have about the relationship between X and O is the following: if we know that X assumes the value x in $\mathcal{X}$, then we know that O must assume a value o in a *non-empty* subset $\Gamma(x)$ of $\mathcal{O}$, *and nothing more*. This idea of modelling incomplete observations through a so-called *multi-valued map* Γ essentially goes back to Strassen [35].

If we observe the value o of O, then we know something more about X: it can then only assume values in the set

$$\{o\}^* = \{x \in \mathcal{X}\colon\ o \in \Gamma(x)\}$$

of those values of X that *may* produce the observation $O = o$. We shall henceforth assume that $\{o\}^* \neq \emptyset$ for all $o \in \mathcal{O}$: observations o for which $\{o\}^* = \emptyset$, cannot be produced by any x in $\mathcal{X}$, and they can therefore be eliminated from the set $\mathcal{O}$ without any further consequences.

Unless $\{o\}^*$ is a singleton, the observation $O = o$ does not allow us to identify a unique value for X; it only allows us to restrict the possible values of X to $\{o\}^*$. This is even the case if there is some possible value of X for which o is the only compatible observation, i.e., if the set

$$\{o\}_* = \{x \in \mathcal{X}\colon\ \Gamma(x) = \{o\}\}$$

is non-empty: the set $\{o\}^*$ includes $\{o\}_*$ and may still contain more than one element.

The question we want to answer in this section, then, is how we can use this new piece of information that $O = o$ to coherently update the prior lower prevision $\underline{P}_0$ on $\mathcal{L}(\mathcal{X})$ to a posterior lower prevision $\underline{P}(\cdot|O = o) = \underline{P}(\cdot|o)$ on $\mathcal{L}(\mathcal{X})$.

In order to do this, we need to model the available information about the relationship between X and O, i.e., about the so-called *incompleteness mechanism* that turns the values of X into their incomplete observations O. In the special case that the marginal $\underline{P}_0$ is a (precise) linear prevision P_0 (with mass function p_0), it is often assumed that this mechanism obeys the CAR condition, mentioned in the Introduction:

$$p(o|x) = p(o|y) > 0 \tag{CAR}$$

for all $o \in \mathcal{O}$ and all x and y in $\{o\}^*$ such that $p_0(x) > 0$ and $p_0(y) > 0$ (see [13,14] for an extensive discussion and detailed references). It is in other words assumed that the probability of observing $O = o$ is not affected by the specific values x of X that may actually lead to this observation o. After a few manipulations involving Bayes' rule, we derive from the CAR assumption that quite simply

$$p(x|o) = \begin{cases} \dfrac{p_0(x)}{P_0(\{o\}^*)} = p_0(x|\{o\}^*) & \text{if } x \in \{o\}^*, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This means that if we make the CAR assumption about the incompleteness mechanism, then using the so-called *naive updating rule* (6) is justified.

For imprecise priors $\underline{P}_0$, this result can be generalised as follows for observations o such that $\underline{P}_0(\{o\}^*) > 0$. Observe that Theorem 2 has an immediate Bayesian sensitivity analysis interpretation.

Theorem 2. *Assume that $p(o|x) = p(o|y) > 0$ for all $o \in \mathcal{O}$ and all x and y in $\{o\}^*$ such that $\overline{P}_0(\{x\}) > 0$ and $\overline{P}_0(\{y\}) > 0$. Let $o \in \mathcal{O}$ be such that $\underline{P}_0(\{o\}^*) > 0$. Then the conditional lower prevision $\underline{P}(\cdot|o)$ is uniquely determined by coherence, and given by*

$$\underline{P}(f|o) = \inf\left\{ \frac{P(f I_{\{o\}^*})}{P(\{o\}^*)} \colon\ P \in \mathcal{M}(\underline{P}_0) \right\} = \inf\left\{ P\big(f|\{o\}^*\big) \colon\ P \in \mathcal{M}(\underline{P}_0) \right\}$$

for all gambles f on $\mathcal{X}$.

Proof. Let $N = \{x \in \mathcal{X} \colon\ \overline{P}_0(\{x\}) = 0\}$. Then it follows from the coherence of $\underline{P}_0$ that $\overline{P}_0(N) = 0$. Moreover, for any gamble f on $\mathcal{X}$, it follows from the coherence of $\underline{P}_0$ that $\underline{P}_0(f) = \underline{P}_0(f I_{\mathrm{co}\, N})$: $\underline{P}_0(f)$ only depends on the values that f assumes outside N. Moreover, our generalised CAR assumption identifies, for all x outside N, a conditional *linear* prevision $P(\cdot|x)$ on $\mathcal{L}(\mathcal{O})$, and hence, by separate coherence, on $\mathcal{L}(\mathcal{X} \times \mathcal{O})$. We may therefore write, with some abuse of notation,[11] for the marginal extension $\underline{P}$ of $\underline{P}_0$ and $P(\cdot|X)$:

$$\underline{P}(h) = \underline{P}_0\big(P(h|X)\big),$$

[11] The abuse consists in assuming that the conditional lower previsions $\underline{P}(\cdot|x)$ are linear also for x in N, which we can do because we have just shown that the value of the marginal extension does not depend on them.

for all gambles h on $\mathcal{X} \times \mathcal{O}$. It follows from coherence arguments (see [37, Section 6.7.3]) that $\underline{P}$ is the *only* joint lower prevision with marginal $\underline{P}_0$ that is jointly coherent with $P(\cdot|X)$. It also follows readily from the generalised CAR assumption that for the conditional mass function, $p(o|x) = L_o I_{\{o\}^*}(x)$ for all x outside N, where L_o is some strictly positive real number that only depends on o, not on x. Consequently,

$$\underline{P}\big(\mathcal{X} \times \{o\}\big) = \underline{P}_0\big(p(o|X)\big) = \underline{P}_0(L_o I_{\{o\}^*}) = L_o \underline{P}_0\big(\{o\}^*\big) > 0,$$

where the inequality follows from the assumptions. It now follows from the discussion in Sections 2.7 and 2.8 that $\underline{P}(\cdot|o)$ is uniquely determined from the joint $\underline{P}$ by coherence, and given by

$$\underline{P}(f|o) = \inf\left\{ \frac{P(P(f I_{\mathcal{X}\times\{o\}}|X)}{P(P(\mathcal{X} \times \{o\}|X))} \colon\ P \in \mathcal{M}(\underline{P}_0) \text{ and } P\big(P\big(\mathcal{X} \times \{o\}|X\big)\big) > 0 \right\}$$

for all gambles f on $\mathcal{X}$. The proof is complete if we consider that for all $P \in \mathcal{M}(\underline{P}_0)$, $P(N) = 0$, whence with obvious notations, also using separate coherence,

$$\begin{aligned} P\big(P(f I_{\mathcal{X}\times\{o\}}|X)\big) &= \sum_{x\in\mathcal{X}\setminus N} p(x) P\big(f I_{\mathcal{X}\times\{o\}}|x\big) = \sum_{x\in\mathcal{X}\setminus N} p(x) P\big(f(x) I_{\{o\}}|x\big) \\ &= \sum_{x\in\mathcal{X}\setminus N} p(x) f(x) p(o|x) = \sum_{x\in\mathcal{X}\setminus N} p(x) f(x) L_o I_{\{o\}^*}(x) \\ &= L_o P(f I_{\{o\}^*}), \end{aligned}$$

and similarly

$$P\big(P\big(\mathcal{X} \times \{o\}|X\big)\big) = P\big(p(o|X)\big) = L_o P\big(\{o\}^*\big) > 0,$$

where the inequality follows from $P(\{o\}^*) \geqslant \underline{P}(\{o\}^*) > 0$. □

However, Grünwald and Halpern [14] have argued convincingly that CAR is a very strong assumption, which will only be justified in very special cases.

Here, we want to refrain from making such unwarranted assumptions in general: we want to find out what can be said about the posterior $\underline{P}(\cdot|O)$ if *no* assumptions are made about the incompleteness mechanism, apart from those present in the definition of the multi-valued map Γ given above. This implies that anyone making additional assumptions (such as CAR) about the incompleteness mechanism will find results that are compatible but stronger, i.e., will find a posterior (lower) prevision that will point-wise dominate ours.

We proceed as follows. We have argued in Section 2.4 that the appropriate model for the piece of information that 'O assumes a value in $\Gamma(x)$' is the vacuous lower prevision $\underline{P}_{\Gamma(x)}$ on $\mathcal{L}(\mathcal{O})$ relative to the set $\Gamma(x)$. This means that we can model the relationship between X and O through the following (vacuous) conditional lower prevision $\underline{P}(\cdot|X)$ on $\mathcal{L}(\mathcal{O})$, defined by

$$\underline{P}(g|x) = \underline{P}_{\Gamma(x)}(g) = \min_{o\in\Gamma(x)} g(o) \tag{7}$$

for any gamble g on $\mathcal{O}$. We have argued in Section 2.6 that there is a unique separately coherent conditional lower prevision that extends this to gambles on the space $\mathcal{X} \times \mathcal{O}$: for any gamble h in $\mathcal{L}(\mathcal{X} \times \mathcal{O})$,

$$\underline{P}(h|x) = \min_{o\in\Gamma(x)} h(x, o). \tag{8}$$

Eq. (7) also has an interesting Bayesian sensitivity analysis interpretation. The coherent lower prevision $\underline{P}(\cdot|x)$ is the lower envelope of the set

$$\mathcal{M}\big(\underline{P}(\cdot|x)\big) = \big\{P(\cdot|x)\colon\ P\big(\Gamma(x)|x\big) = 1\big\}$$

of all linear previsions on $\mathcal{L}(\mathcal{O})$ that assign probability one to the event $\Gamma(x)$, i.e., for which it is certain that $O \in \Gamma(x)$. On the Bayesian sensitivity analysis interpretation, each such linear prevision $P(\cdot|x)$ represents a so-called *random incompleteness mechanism* (or a protocol, in Shafer's terminology [33]): a random mechanism that chooses an incomplete observation o from the set $\Gamma(x)$ of observations compatible with state x, with probability $p(o|x)$. The set $\mathcal{M}(\underline{P}(\cdot|x))$ contains all possible such random incompleteness mechanisms, and its lower envelope $\underline{P}(\cdot|x)$ models that we have no information at all about which random incompleteness mechanism is active.

Using Walley's marginal extension theorem (see Theorem 1 in Section 2.9), the smallest (unconditional) lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X} \times \mathcal{O})$ that extends $\underline{P}_0$ and is jointly coherent with the conditional lower prevision $\underline{P}(\cdot|X)$ is given by

$$\underline{P}(h) = \underline{P}_0\big(\underline{P}(h|X)\big)$$

for all gambles h on $\mathcal{X} \times \mathcal{O}$.[12] In order to find the posterior lower prevision, we can now apply the technique of regular extension, discussed in Section 2.8. It yields the smallest (most conservative) posterior lower prevision $\underline{R}(\cdot|O)$ that is jointly coherent with $\underline{P}$ (and therefore with $\underline{P}_0$ and $\underline{P}(\cdot|\mathcal{X})$) and satisfies an additional regularity condition. We have argued in Sections 2.8 and 2.9 that it also seems the right way to obtain a posterior lower prevision on the Bayesian sensitivity analysis interpretation.

Theorem 3. *Let $o \in \mathcal{O}$ and let f be any gamble on $\mathcal{X}$. If $\overline{P}_0(\{o\}^*) > 0$, then*

$$\underline{R}(f|o) = \max\big\{\mu\colon\ \underline{P}_0\big(I_{\{o\}_*}\max\{f-\mu,0\} + I_{\{o\}^*}\min\{f-\mu,0\}\big) \geqslant 0\big\}.$$

If $\overline{P}(\{o\}^*) = 0$ then $\underline{R}(f|o) = \min_{x\in\mathcal{X}} f(x)$.

Proof. The discussion in Section 2.8 tells us to look at the value of $\overline{P}(\mathcal{X} \times \{o\}) = \overline{P}_0(\overline{P}(\mathcal{X} \times \{o\})|X)$. Observe that for any $x \in \mathcal{X}$, by Eq. (8),

$$\overline{P}\big(\mathcal{X} \times \{o\}|x\big) = \max_{p\in\Gamma(x)} I_{\mathcal{X}\times\{o\}}(x,p) = I_{\{o\}^*}(x),$$

whence $\overline{P}(\mathcal{X} \times \{o\})|X) = I_{\{o\}^*}$ and consequently $\overline{P}(\mathcal{X} \times \{o\}) = \overline{P}_0(\{o\}^*)$. If $\overline{P}(\mathcal{X} \times \{o\}) = \overline{P}_0(\{o\}^*) = 0$ then the discussion in Section 2.8 tells us that $\underline{R}(\cdot|o)$ is indeed the vacuous lower prevision on $\mathcal{L}(\mathcal{X})$ (relative to the set $\mathcal{X}$). If $\overline{P}(\mathcal{X} \times \{o\}) = \overline{P}_0(\{o\}^*) > 0$, then we know that, by definition, $\underline{R}(f|o)$ is the greatest solution of the following inequality in μ:

$$\underline{P}\big(I_{\mathcal{X}\times\{o\}}[f-\mu]\big) \geqslant 0.$$

But for any $x \in \mathcal{X}$, we find that

[12] See [27] for a more general discussion with more mathematical detail.

$$\begin{aligned}\underline{P}\big(I_{\mathcal{X}\times\{o\}}[f-\mu]|x\big) &= \min_{p\in\Gamma(x)} I_{\mathcal{X}\times\{o\}}(x,p)\big[f(x)-\mu\big]\\ &= \begin{cases} f(x)-\mu & \text{if } x\in\{o\}_*\\ \min\{0, f(x)-\mu\} & \text{if } x\in\{o\}^* \text{ and } x\notin\{o\}_*\\ 0 & \text{if } x\notin\{o\}^*\end{cases}\\ &= I_{\{o\}_*}(x)\max\big\{f(x)-\mu,0\big\}+I_{\{o\}^*}(x)\min\big\{f(x)-\mu,0\big\},\end{aligned}$$

whence indeed

$$\underline{P}\big(I_{\mathcal{X}\times\{o\}}[f-\mu]\big)=\underline{P}_0\big(I_{\{o\}_*}\max\{f-\mu,0\}+I_{\{o\}^*}\min\{f-\mu,0\}\big).$$

This concludes the proof. □

It also follows from this proof and the discussion in Section 2.8, that the natural—as opposed to the regular—extension $\underline{E}(\cdot|o)$ is vacuous whenever $\underline{P}(\mathcal{X}\times\{o\})=\underline{P}_0(\{o\}_*)=0$, and that $\underline{E}(h|o)$ is the unique solution of the equation

$$\underline{P}_0\big(I_{\{o\}_*}\max\{f-\mu,0\}+I_{\{o\}^*}\min\{f-\mu,0\}\big)=0$$

in μ whenever $\underline{P}_0(\{o\}_*)>0$ (in which case regular and natural extension coincide). We shall see later that there are interesting cases where $\{o\}_*$ is empty, and where the natural extension $\underline{E}(\cdot|o)$ is therefore the vacuous lower prevision relative to $\mathcal{X}$. But this seems needlessly imprecise, as we know from the observation $O=o$ that X should belong to the set $\{o\}^*$ of those values that can produce the observation o, which may be a proper subset of $\mathcal{X}$. We shall see in Theorem 4 that regular extension produces results that are more intuitively acceptable in this respect.

Let us now apply the results of Theorem 3 to a puzzle of some standing in probability theory: the Monty Hall puzzle (see for instance [14] for further discussion and references). We mention in passing that it is very closely related to the three prisoners problem, introduced at the beginning of the section, an that it can be dealt with in an almost identical manner.

3.1. The Monty Hall puzzle

In the Monty Hall game show, there are three doors. One of these doors leads to a car, and the remaining doors each have a goat behind them. You indicate one door, and the show's host—let us call him Monty—now opens one of the other doors, which has a goat behind it. After this observation, should you choose to open the door that is left, rather than the one you indicated initially?

To solve the puzzle, we reformulate it using our language of incomplete observations. Label the doors from 1 to 3, and assume without loss of generality that you picked door 1. Let the variable X refer to the door hiding the car, then clearly $\mathcal{X}=\{1,2,3\}$. Observe that there is a precise prior prevision P_0 determined by $P_0(\{1\})=P_0(\{2\})=P_0(\{3\})=\frac{1}{3}$. The observation variable O refers to the door that Monty opens, and consequently $\mathcal{O}=\{2,3\}$ is the set of doors Monty can open. If the car is behind door 1, Monty can choose between opening doors 2 and 3, so $\Gamma(1)=\{2,3\}$, and similarly, $\Gamma(2)=\{3\}$ and $\Gamma(3)=\{2\}$. Since we know nothing at all about how Monty will choose between the options open to him,

we should model the available information about the relation between X and O by the conditional lower prevision $\underline{P}(\cdot|X)$ given by Eq. (8): for any gamble h on $\mathcal{X}\times\mathcal{O}$,

$$\underline{P}(h|1)=\min\bigl\{h(1,2),h(1,3)\bigr\},\quad \underline{P}(h|2)=h(2,3),\quad \underline{P}(h|3)=h(3,2).$$

Applying the marginal extension theorem to the marginal P_0 and the conditional lower prevision $\underline{P}(\cdot|X)$, we find the following joint lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X}\times\mathcal{O})$:

$$\underline{P}(h)=\frac{1}{3}\min\bigl\{h(1,2),h(1,3)\bigr\}+\frac{1}{3}h(2,3)+\frac{1}{3}h(3,2),$$

for all gambles h on $\mathcal{X}\times\mathcal{O}$.

Assume without loss of generality that Monty opens door 2. What can we say about the updated lower prevision $\underline{R}(f|2)$ when f is any gamble on $\mathcal{X}$? Since $\underline{P}(\mathcal{X}\times\{2\})=\frac{1}{3}>0$, we can use the GBR to find the (uniquely!) coherent $\underline{R}(f|2)$ as the unique solution of the following equation in μ:

$$\underline{P}\bigl(I_{\mathcal{X}\times\{2\}}[f-\mu]\bigr)=\frac{1}{3}\min\bigl\{f(1)-\mu,0\bigr\}+\frac{1}{3}\bigl[f(3)-\mu\bigr]=0.$$

It is easy to see that

$$\underline{R}(f|2)=\frac{1}{2}f(3)+\frac{1}{2}\min\bigl\{f(3),f(1)\bigr\}.$$

We are now ready to solve the puzzle. Which of the two actions should we choose: stick to our initial choice and open door 1 (action a), or open door 3 instead (action b). In Table 1 we see the possible outcomes of each action for the three possible values of X. If the gamble f_a on $\mathcal{X}$ represents the uncertain utility received from action a, and similarly for f_b, then we are interested in the gamble f_b-f_a, which represents the uncertain utility from exchanging action a for action b. The possible values for this gamble are also given in Table 1, where Δ denotes the difference in utility between a car and a goat, which is assumed to be strictly positive. Then we find that

$$\underline{R}(f_b-f_a|2)=\frac{1}{2}\Delta+\frac{1}{2}\min\{\Delta,-\Delta\}=0$$

and

$$\underline{R}(f_a-f_b|2)=-\frac{1}{2}\Delta+\frac{1}{2}\min\{\Delta,-\Delta\}=-\Delta.$$

This implies that, with the notions and notations established in Section 2.10, $a\not\succ b$, $b\not\succ a$, and $a\not\approx b$: the available information does not allow us to say which of the two actions, sticking to door 1 (action a) or choosing door 3 (action b), is to be strictly preferred;

Table 1
Possible outcomes in the Monty hall puzzle

	1	2	3
a	car	goat	goat
b	goat	goat	car
f_b-f_a	$-\Delta$	0	Δ

and neither are these actions equivalent. They are incomparable, and we should remain undecided on the basis of the information available in the formulation of the puzzle.

The same conclusion can also be reached in the following way. Suppose first that Monty has decided on beforehand to always open door 3 when the car is behind door 1. Since he has actually opened door 2, the car cannot be behind door 1, and it must therefore be behind door 3. In this case, action b is clearly strictly preferable to action a. Next, suppose that Monty has decided on beforehand to always open door 2 when the car is behind door 1. Since he actually opens door 2, there are two equally likely possibilities, namely that the car is behind door 1 or behind door 3. Both actions a and b now have the same expected utility (zero), and none of them is therefore strictly preferable to the other. Since both possibilities are consistent with the available information, we cannot infer any (robust) strict preference of one action over the other. A similar analysis was made by Halpern [15].

Observe that since $\underline{R}(f_b - f_a|2) = 0$, you *almost-prefer* b to a, in the sense that you are disposed to exchange f_a for f_b in return for any strictly positive amount. In the slightly more involved case that Monty could also decide not to open any door (denote this observation by 0), we now have $\mathcal{O} = \{0, 2, 3\}$, $\Gamma(1) = \{0, 2, 3\}$, $\Gamma(2) = \{0, 3\}$ and $\Gamma(3) = \{0, 2\}$. Consequently, $\{2\}_* = \emptyset$ and $\{2\}^* = \{1, 3\}$, and a similar analysis as before (see in particular Theorem 4 below) tells us that the updated lower prevision is given by $\underline{R}(f|2) = \min\{f(1), f(3)\}$, and we get $\underline{R}(f_b - f_a|2) = \underline{R}(f_a - f_b|2) = -\Delta$: now neither option is even almost-preferred, let alone strictly preferred, over the other.

3.2. When naive updating is justified

We are now in a position to take a closer look at the issue of when using the naive updating rule (6) can be justified, even if nothing is known about the incompleteness mechanism.

We start with a precise prior prevision P_0 on $\mathcal{L}(\mathcal{X})$ and consider an incomplete observation $o \in \mathcal{O}$. We shall assume that $\{o\}_*$ is non-empty[13] and that the mass function p_0 is strictly positive on all elements of $\{o\}^*$. In this case, it follows from the discussion in Section 2.7 and the proof of Theorem 3 that the posterior lower prevision after observing o is *uniquely* determined by coherence, and equal to the regular extension $\underline{R}(\cdot|o)$.

We shall see from the following discussion that using the naive posterior $P_0(\cdot|\{o\}^*)$ is still justified, even if we know nothing at all about the incompleteness mechanism, if and only if

$$\{o\}_* = \{o\}^*, \tag{NAIVE-OK}$$

i.e., if all the states that *may* produce observation o can *only* produce observation o.

First of all, if (NAIVE-OK) holds, it follows immediately from Theorem 3 and the assumptions that

$$\underline{R}(f|o) = \frac{P_0(f I_{\{o\}^*})}{P_0(\{o\}^*)} = P_0\big(f|\{o\}^*\big),$$

[13] If $\{o\}_* = \emptyset$ then the vacuous lower prevision $\underline{P}(\cdot|o)$ relative to $\mathcal{X}$ is coherent with the joint $\underline{P}$, and naive updating will not be justified, as it produces a precise posterior.

indeed yielding the same result as naive updating does [see Eq. (6)].

We now show that (NAIVE-OK) is also necessary. If our regular extension (and therefore coherence) produces the same result as naive updating does, this implies that $\underline{R}(\cdot|o)$ is a linear prevision. So we have that for any gamble f on $\mathcal{X}$, $\overline{R}(f|o) = -\underline{R}(-f|o)$. It then follows from Theorem 3, after some elementary manipulations, that for each gamble f there is a unique μ such that

$$\begin{aligned} &P_0\big(I_{\{o\}_*} \max\{f-\mu,0\} + I_{\{o\}^*} \min\{f-\mu,0\}\big) \\ &\quad = P_0\big(I_{\{o\}_*} \min\{f-\mu,0\} + I_{\{o\}^*} \max\{f-\mu,0\}\big) = 0. \end{aligned}$$

Let x be any element of $\{o\}_*$. Choose in particular $f = I_{\{x\}}$, then it follows that

$$P_0\big(I_{\{o\}_*}[I_{\{x\}} - \mu]\big) = P_0\big(I_{\{o\}^*}[I_{\{x\}} - \mu]\big) = 0,$$

or equivalently

$$\mu = \frac{p_0(x)}{P_0(\{o\}_*)} = \frac{p_0(x)}{P_0(\{o\}^*)},$$

whence $P_0(\{o\}_*) = P_0(\{o\}^*)$, since it follows from our assumptions that $p_0(x) > 0$. Again, since p_0 is assumed to be strictly positive on all elements of $\{o\}^*$, Eq. (NAIVE-OK) follows.

Observe that if Eq. (NAIVE-OK) holds, then all states x in $\{o\}^*$ can only lead to observation o, whence $p(o|x) = 1$, so the CAR condition is forced to hold, but in a very trivial way. In the same vein, it follows from Eq. (NAIVE-OK) and Eq. (8) that for all x in $\{o\}^*$, $\underline{P}(f|x) = f(o)$, so $\underline{P}(\cdot|x)$ is a precise conditional prevision, whose mass function satisfies $p(o|x) = 1$ for all x in $\{o\}^*$.

Our conclusion is that when the incompleteness mechanism is unknown, *naive updating is never justified*, except in those trivial situations where CAR *cannot* fail to hold. It is striking that Grünwald and Halpern obtain essentially the same conclusion using a rather different approach: compare Eq. (NAIVE-OK) to Proposition 4.1 in [14].

3.3. When an observation is not a necessary consequence

To conclude this general discussion of incomplete observations, we shall consider an important special case where nearly all reference to the prior is obliterated[14] from the posterior: we want to find $\underline{R}(\cdot|o)$ for an observation $O = o$ that is not a necessary consequence of any value of X, i.e.,

$$\{o\}_* = \big\{x \in \mathcal{X}\colon\ \Gamma(x) = \{o\}\big\} = \emptyset. \tag{A1}$$

We make the additional assumption that each state of the world compatible with observation o has positive upper probability, i.e.,

$$\overline{P}_0\big(\{x\}\big) > 0 \quad \text{for all } x \in \{o\}^*. \tag{A2}$$

[14] This is essentially due to the fact that updating requires us to condition on a set with zero lower prior probability. Observe that also in the case of precise probabilities, coherence imposes a very weak link between a prior and a posterior obtained after observing a set of zero prior probability. See also Section 2.8.

Under these conditions the regular extension $\underline{R}(\cdot|o)$ does not depend on the prior $\underline{P}_0$, and only retains the information present in the multi-valued map Γ, as the following theorem states. We also want to observe that using natural rather than regular extension here, would lead to a posterior that is vacuous with respect to all of $\mathcal{X}$, which would make us lose even the information present in Γ.

Theorem 4. *If $o \in \mathcal{O}$ satisfies Assumption* (A1) *and $\underline{P}_0$ satisfies Assumption* (A2)*, then $\underline{R}(\cdot|o)$ is the vacuous lower prevision $\underline{P}_{\{o\}^*}$ on $\mathcal{L}(\mathcal{X})$ relative to $\{o\}^*$:*

$$\underline{R}(f|o) = \underline{P}_{\{o\}^*}(f) = \min_{x:\, o\in\Gamma(x)} f(x)$$

for all f in $\mathcal{L}(\mathcal{X})$.

Proof. We apply the results of Theorem 3. Since it follows from Assumption (A2) and the coherence of $\underline{P}_0$ that $\overline{P}_0(\{o\}^*) > 0$, we consider the gamble

$$f_\mu = I_{\{o\}^*} \min\{f-\mu, 0\} + I_{\{o\}_*} \max\{f-\mu, 0\} = I_{\{o\}_*} \min\{f-\mu, 0\}$$

on $\mathcal{X}$, where the last equality follows from Assumption (A1). Then, we know that

$$\underline{R}(f|o) = \max\bigl\{\mu\colon\, \underline{P}_0(f_\mu) \geqslant 0\bigr\} = \max\bigl\{\mu\colon\, \underline{P}_0\bigl(I_{\{o\}_*} \min\{f-\mu, 0\}\bigr) \geqslant 0\bigr\}.$$

Let $\lambda = \min_{x:\, o\in\Gamma(x)} f(x) = \min_{x\in\{o\}^*} f(x)$. If $\mu \leqslant \lambda$ then $f(x) - \mu < 0$ implies $f(x) - \lambda < 0$ whence $x \notin \{o\}^*$. Consequently f_μ is identically zero, whence $\underline{P}_0(f_\mu) = 0$. Assume therefore that $\mu > \lambda$. It remains to prove that $\underline{P}_0(f_\mu) < 0$. Observe that there is some x_0 in $\{o\}^*$ such that $f(x_0) = \lambda$. If f is constant, and therefore equal to λ, on $\{o\}^*$, we find that $f_\mu = -[\mu - \lambda] I_{\{o\}^*}$, whence

$$\underline{P}_0(f_\mu) = -[\mu - \lambda]\overline{P}_0\bigl(\{o\}^*\bigr) < 0,$$

also taking into account that Assumption (A2) implies $\overline{P}_0(\{o\}^*) > 0$. If f is not constant on $\{o\}^*$, let x_1 be an element of $\{o\}^*$ such that f assumes no values between $f(x_0)$ and $f(x_1)$ on $\{o\}^*$, and let $A_0 = \{x \in \{o\}^*\colon\, f(x) = f(x_0)\}$. Assume that $\lambda < \mu < f(x_1)$, then for all $x \in \{o\}^*$ it follows from $f(x) < \mu$ that $x \in A_0$ and therefore $f(x) = f(x_0) = \lambda$. Consequently, $f_\mu = -[\mu - \lambda] I_{A_0}$, whence

$$\underline{P}_0(f_\mu) = -[\mu - \lambda]\overline{P}_0(A_0) < 0,$$

since it follows from Assumption (A2) and the coherence of $\underline{P}_0$ that $\overline{P}_0(A_0) > 0$. Since we can also deduce from the coherence of $\underline{P}_0$ that $\underline{P}_0(f_\mu)$ is non-increasing in μ, the result follows. □

It is illustrative to prove this theorem in an alternative manner, using sets of linear previsions.

Alternative proof using sets of linear previsions. A *selection s* for the multi-valued map Γ is a function from $\mathcal{X}$ to $\mathcal{O}$ that associates with each $x \in \mathcal{X}$ a compatible observation $s(x) \in \Gamma(x)$. Denote by $S(\Gamma)$ the set of all possible selections:

$$S(\Gamma) = \bigl\{s \in \mathcal{O}^{\mathcal{X}}\colon\, (\forall x \in \mathcal{X})\bigl(s(x) \in \Gamma(x)\bigr)\bigr\}.$$

For any s in $S(\Gamma)$, define the conditional linear prevision $P_s(\cdot|X)$ on $\mathfrak{L}(\mathcal{O})$ by $P_s(\cdot|x) = P_{s(x)}$ for all $x \in \mathcal{X}$, where $P_{s(x)}$ is the (degenerate) linear prevision on $\mathfrak{L}(\mathcal{O})$ all of whose probability mass lies in $s(x)$, defined by $P_{s(x)}(g) = g(s(x))$ for all gambles g on $\mathcal{O}$. Then clearly,

$$\{P_s(\cdot|X)\colon\ s \in S(\Gamma)\}$$

is precisely the set of all conditional linear previsions $P(\cdot|X)$ such that

$$P(\cdot|x) \in \operatorname{ext}\big(\mathcal{M}\big(\underline{P}(\cdot|x)\big)\big)$$

for all $x \in \mathcal{X}$, and consequently, following the discussion in Sections 2.8 and 2.9, it is easily seen that

$$\underline{R}(f|o) = \inf\left\{\frac{P_0(P_s(fI_{\mathcal{X}\times\{o\}}|X))}{P_0(P_s(\mathcal{X}\times\{o\}|X))}\colon\ P_0 \in \mathcal{M}(\underline{P}_0),\ s \in S(\Gamma),\ P_0\big(P_s\big(\mathcal{X}\times\{o\}|X\big)\big) > 0\right\}.$$

Now for any x in $\mathcal{X}$, also using separate coherence,

$$P_s\big(\mathcal{X}\times\{o\}|x\big) = I_{\mathcal{X}\times\{o\}}\big(x, s(x)\big) = I_{\{o\}}\big(s(x)\big) = I_{s^{-1}(\{o\})}(x),$$

whence $P_0(P_s(\mathcal{X}\times\{o\}|X)) = P_0(s^{-1}(\{o\}))$, where $s^{-1}(\{o\}) = \{x \in \mathcal{X}\colon\ s(x) = o\} \subseteq \{o\}^*$. Similarly,

$$P_s(fI_{\mathcal{X}\times\{o\}}|x) = f(x)I_{\mathcal{X}\times\{o\}}\big(x, s(x)\big) = f(x)I_{\{o\}}\big(s(x)\big) = f(x)I_{s^{-1}(\{o\})}(x),$$

whence $P_0(P_s(fI_{\mathcal{X}\times\{o\}}|X)) = P_0(fI_{s^{-1}(\{o\})})$. Consequently,

$$\begin{aligned}\underline{R}(f|o) &= \inf\left\{\frac{P_0(fI_{s^{-1}(\{o\})})}{P_0(s^{-1}(\{o\}))}\colon\ P_0 \in \mathcal{M}(\underline{P}_0),\ s \in S(\Gamma),\ P_0\big(s^{-1}\big(\{o\}\big)\big) > 0\right\}\\ &= \inf\big\{P_0\big(f|s^{-1}\big(\{o\}\big)\big)\colon\ P_0 \in \mathcal{M}(\underline{P}_0),\ s \in S(\Gamma),\ P_0\big(s^{-1}\big(\{o\}\big)\big) > 0\big\}. \qquad (9)\end{aligned}$$

Now consider any $x \in \{o\}^*$, whence $o \in \Gamma(x)$. Consequently, there is a selection $s \in S(\Gamma)$ such that $s(x) = o$. Moreover, Assumption (A1) tells us that we can let $s(y) \neq o$ for all $y \neq x$. Indeed, this is guaranteed if for all $y \neq x$ there is some p in $\Gamma(y)$ different from o, so that we can let $s(y) = p$. If this condition did not hold, then there would be some $y \neq x$ such that $p = o$ for all $p \in \Gamma(y)$, i.e., $\Gamma(y) = \{o\}$, whence $y \in \{o\}_*$, which contradicts Assumption (A1). Now for such s it holds that $s^{-1}(\{o\}) = \{x\}$, and consequently $P_0(s^{-1}(\{o\})) = P_0(\{x\})$ and $P_0(fI_{s^{-1}(\{o\})}) = f(x)P_0(\{x\})$ for all $P_0 \in \mathcal{M}(\underline{P}_0)$. But Assumption (A2) tells us that there is at least one P_0 in $\mathcal{M}(\underline{P}_0)$ for which $P_0(\{x\}) > 0$, and it therefore follows from Eq. (9) that $\underline{R}(f|o) \leqslant f(x)$, and consequently $\underline{R}(f|o) \leqslant \min_{x\in\{o\}^*} f(x)$. To prove the converse inequality, use Eq. (9) and observe that for all $s \in S(\Gamma)$ and $P_0 \in \mathcal{M}(\underline{P}_0)$ such that $P_0(s^{-1}(\{o\})) > 0$,

$$\frac{P_0(fI_{s^{-1}(\{o\})})}{P_0(s^{-1}(\{o\}))} \geqslant \min_{x\in s^{-1}(\{o\})} f(x) \geqslant \min_{x\in\{o\}^*} f(x),$$

since the left-hand side is some convex combination of the $f(x)$ for x in $s^{-1}(\{o\})$, and since $s^{-1}(\{o\}) \subseteq \{o\}^*$. □

The selections $s \in S(\Gamma)$ in this proof are essentially the deterministic incompleteness mechanisms. They model that for any state x, the observation $s(x) \in \Gamma(x)$ is selected with probability one: $p_s(s(x)|x) = 1$.

4. Missing data in a classification problem

In order to illustrate the practical implications of our model for the incompleteness mechanism, let us show how it can be applied in classification problems, where objects have to be assigned to a certain class on the basis of the values of their attributes.

4.1. The basic classification problem

Let in such a problem $\mathcal{C}$ be the set of possible classes that we want to assign objects to. Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be the sets of possible values for the n attributes on the basis of which we want to classify the objects. We denote their Cartesian product by

$$\mathcal{X} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n.$$

We consider a *class variable* C, which is a random variable in $\mathcal{C}$, and *attribute variables* A_k, which are random variables in $\mathcal{A}_k$ $(k = 1, \ldots, n)$. The n-tuple $X = (A_1, \ldots, A_n)$ is a random variable in $\mathcal{X}$, and is called the *attributes variable*. The available information about the relationship between class and attribute variables is specified by a (prior) lower prevision $\underline{P}_0$ on $\mathcal{L}(\mathcal{C} \times \mathcal{X})$, or equivalently,[15] by a marginal lower prevision $\underline{P}_0$ on $\mathcal{L}(\mathcal{X})$ and a conditional lower prevision $\underline{P}_0(\cdot|X)$ on $\mathcal{L}(\mathcal{C})$.

To see how classification is performed, let us first look at the case that $\underline{P}_0$ is a linear prevision P_0, or equivalently, a precise probability measure. If the attributes variable X assumes a value x in $\mathcal{X}$, then the available information about the values of the class variable C is given by the conditional linear prevision $P_0(\cdot|x)$. If, on the basis of the observed value x of the attributes variable X, we decide that some c' in $\mathcal{C}$ is the right class, then we can see this as an action with an uncertain reward $f_{c'}$, whose value $f_{c'}(c)$ depends on the value c that C actually assumes. An *optimal class* c_{opt} is one that maximises the expected reward $P_0(f_{c'}|x)$: $P_0(f_{c_{\text{opt}}}|x) \geqslant P_0(f_{c'}|x)$ for all $c' \in \mathcal{C}$. As a common example, if we let $f_{c'} = I_{\{c'\}}$, then $P_0(f_{c'}|x) = p_0(c'|x)$, and this procedure associates the most probable class with each value x of the attributes.

How can this be generalised to the more general case that $\underline{P}_0$ is not a linear prevision? If the attributes variable X assumes a value x in $\mathcal{X}$, then the available information about the values of the class variable C is given by the conditional lower prevision $\underline{P}_0(\cdot|x)$. The discussion in Section 2.10 then tells us that the lower prevision $\underline{P}_0(\cdot|x)$ induces a strict preference $>$ on the set of classes $\mathcal{C}$ by

$$c' > c'' \quad \Leftrightarrow \quad \underline{P}_0(f_{c'} - f_{c''}|x) > 0.$$

An optimal class c_{opt} is now one that is *undominated*, i.e., such that for all $c' \in \mathcal{C}$:

$$\overline{P}_0(f_{c_{\text{opt}}} - f_{c'}|x) \geqslant 0.$$

[15] This is, provided that $\underline{P}_0(\mathcal{C} \times \{x\}) > 0$ for all $x \in \mathcal{X}$.

Observe that this reduces to the previously mentioned maximum expected utility condition $P_0(f_{c_{\text{opt}}}|x) \geqslant P_0(f_{c'}|x)$ when $\underline{P}_0(\cdot|x)$ is a precise, or linear, prevision.

To make this more clear, let us consider a medical domain, where classification is used to make a diagnosis. In this case, the classes are possible diseases and each attribute variable represents a measure with random outcome. For example, attribute variables might represent medical tests, or information about the patient, such as age, gender, life style, etc. We can regard the specific instance of the vector of attribute variables for a patient as a profile by which we characterise the person under examination. The relationship between diseases and profiles is given by a joint mass function on the class and the attribute variables. This induces a linear prevision P_0 on $\mathfrak{L}(\mathcal{C} \times \mathcal{X})$, according to Section 2. A diagnosis is then obtained by choosing the most probable disease given, or conditional on, a profile.

In the case of a linear, or precise, $P_0(\cdot|x)$, if there is more than one optimal class, all these classes are equivalent, as they have the same expected reward. But as we have explained in Section 2.10, this is no longer necessarily so for imprecise $\underline{P}_0(\cdot|x)$. Among the optimal, undominated classes, there may be classes c' and c'' that are not equivalent but *incomparable*: the information in $\underline{P}_0(\cdot|x)$ does not allow us to choose between c' and c'', and for all we know, both are possible candidates for the class that the object is assigned to. This implies that if we classify using an imprecise model $\underline{P}_0(\cdot|x)$, the best we can often do, is assign a *set* of possible, optimal classes to an object with attributes x. In our medical example, a given profile would then lead to a number of optimal candidate diagnoses, none of which is considered to be better than (or even as good as) the others. Classifiers that allow for such set-valued classification are called *credal classifiers* [41].

4.2. Dealing with missing data

Now it may also happen that for a patient some of the attribute variables cannot be measured, i.e., they are missing, e.g., when for some reason a medical test cannot be done. In this case the profile is incomplete and we can regard it as the set of all the complete profiles that are consistent with it. As the above classification procedure needs profiles to be complete, the problem that we are now facing, is how we should update our confidence about the possible diseases given a set-profile.

In more general terms, we observe or measure the value a_k of some of the attribute variables A_k, but not all of them. If a measurement is lacking for some attribute variable A_ℓ, it can in principle assume any value in $\mathcal{A}_\ell$. This means that we can associate with any attribute variable A_k a so-called *observation variable* O_k. This is a random variable taking values in the set

$$\mathcal{O}_k = \mathcal{A}_k \cup \{*\},$$

whose elements are either the possible values of A_k, or a new element $*$ which denotes that the measurement of A_k is missing.

Attribute variables A_k and their observations O_k are linked in the following way: with each possible value $a_k \in \mathcal{A}_k$ of A_k there corresponds the following set of corresponding possible values for O_k:

$$\Gamma_k(a_k) = \{a_k, *\} \subseteq \mathcal{O}_k. \tag{10}$$

This models that whatever value a_k the attribute variable A_k assumes, there is some mechanism, called the *missing data mechanism*, that either produces the (exact) observation a_k, or the observation $*$, which indicates that a value for A_k is missing. For the attributes variable X we then have that with each possible value $x = (a_1, \ldots, a_n)$ there corresponds a set of corresponding possible values for the *observations variable* $O = (O_1, \ldots, O_n)$:

$$\Gamma(x) = \Gamma_1(a_1) \times \cdots \times \Gamma_n(a_n) \subseteq \mathcal{O},$$

where $\mathcal{O} = \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$. To summarise, we have defined a multi-valued map $\Gamma : \mathcal{X} \to \wp(\mathcal{O})$, whose interpretation is the following: if the actual value of the attributes variable X is x, then due to the fact that, for some reason or another, measurements for some attributes may be missing, the observations O must belong to $\Gamma(x)$.

So, in general, we observe some value $o = (o_1, \ldots, o_n)$ of the variable O, where o_k is either the observed value for the kth attribute, or $*$ if a value for this attribute is missing. In order to perform classification, we therefore need to calculate a coherent updated lower prevision $\underline{P}(\cdot|O = o)$ on $\mathfrak{L}(\mathcal{C})$. This is what we now set out to do.

In order to find an appropriate updated lower prevision $\underline{P}(\cdot|o)$, we need to model the available information about the relationship between X and O, i.e., about the missing data mechanism that produces incomplete observations O from attribute values X.

We have arrived at a special case of the model described in the previous section, and our so-called missing data mechanism is a particular instance of the incompleteness mechanism described there. In this special case, it is easy to verify that the general CAR assumption, discussed previously, reduces to what is known in the literature as the MAR assumption [25]: the probability that values for certain attributes are missing, is not affected by the specific values that these attribute variables assume. MAR finds appropriate justification in some statistical applications, e.g., special types of survival analysis. However, there is strongly motivated criticism about the unjustified wide use of MAR in statistics, and there are well-developed methods based on much weaker assumptions [26].

As in the previous section, we want to refrain from making strong assumptions about the mechanism that is behind the generation of missing values, apart from what little is already implicit in the definition of the multi-valued map Γ. We have argued before that the information in Γ, i.e., about the relationship between X and O, can be represented by the following conditional lower prevision $\underline{P}(\cdot|X)$ on $\mathfrak{L}(\mathcal{X} \times \mathcal{O})$:

$$\underline{P}(h|x) = \min_{o \in \Gamma(x)} h(x, o), \tag{11}$$

for all gambles h on $\mathcal{X} \times \mathcal{O}$ and all $x \in \mathcal{X}$.

We make the following additional *irrelevance assumption*: for all gambles f on $\mathcal{C}$,

$$\underline{P}(f|x, o) = \underline{P}_0(f|x) \quad \text{for all } x \in \mathcal{X} \text{ and } o \in \Gamma(x). \tag{MDI}$$

Assumption (MDI) states that, conditional on the attributes variable X, the observations variable O is irrelevant to the class, or in other words that the incomplete observations $o \in \Gamma(x)$ can influence our beliefs about the class only indirectly through the value x of the attributes variable X. We shall discuss this assumption in more detail at the end of this section.

Summarising, we now have a coherent lower prevision $\underline{P}_0$ on $\mathcal{L}(\mathcal{X})$, a separately coherent conditional lower prevision $\underline{P}(\cdot|X)$ on $\mathcal{L}(\mathcal{X}\times\mathcal{O})$, and a separately coherent conditional lower prevision $\underline{P}(\cdot|X,O)$ on $\mathcal{L}(\mathcal{C}\times\mathcal{X}\times\mathcal{O})$, determined from $\underline{P}_0(\cdot|X)$ through the irrelevance assumption (MDI).[16] We can now apply a generalisation of Walley's Marginal Extension Theorem (see Theorem A.1 in Appendix A), to find that the smallest coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{C}\times\mathcal{X}\times\mathcal{O})$ that has marginal $\underline{P}_0$ and is jointly coherent with $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,O)$, is given by

$$\underline{P}(h)=\underline{P}_0\big(\underline{P}\big(\underline{P}(h|X,O)|X\big)\big), \tag{12}$$

for all gambles h on $\mathcal{C}\times\mathcal{X}\times\mathcal{O}$.

We can now use regular extension to obtain the conditional lower prevision $\underline{R}(\cdot|O)$ on $\mathcal{L}(\mathcal{C})$. It yields the smallest (most conservative) posterior lower prevision that is jointly coherent with $\underline{P}$ (and therefore with $\underline{P}_0$, $\underline{P}(\cdot|\mathcal{X})$ and $\underline{P}(\cdot|\mathcal{X}\times\mathcal{O})$) and satisfies an additional regularity condition. Here too, it leads to the right way to obtain a posterior lower prevision on the Bayesian sensitivity analysis interpretation. Again, observe that using natural rather than regular extension would lead to a *completely* vacuous posterior on $\mathcal{C}$.

Theorem 5 (Conservative updating rule). *Assume that the irrelevance assumption* (MDI) *holds. Let o be any element of $\mathcal{O}$. Then $\{o\}_*=\emptyset$. If $\overline{P}_0(\{x\})>0$ for all $x\in\{o\}^*$, then for any gamble f on $\mathcal{C}$:*

$$\underline{R}(f|o)=\min_{x:\,o\in\Gamma(x)}\underline{P}_0(f|x). \tag{13}$$

Proof. Consider any $x=(a_1,\ldots,a_n)$ in $\mathcal{X}$. Since, by Eq. (10), $\Gamma_k(a_k)=\{a_k,*\}$, we find that $\Gamma(x)$ can never be a singleton, whence indeed

$$\{o\}_*=\big\{x\in\mathcal{X}:\ \Gamma(x)=\{o\}\big\}=\emptyset.$$

In order to calculate the regular extension $\underline{R}(f|o)$, the discussion in Section 2.8 tells us that we need to know the value of $\overline{P}(\mathcal{C}\times\mathcal{X}\times\{o\})$. Taking into account separate coherence, we find that for all (x,p) in $\mathcal{X}\times\mathcal{O}$,

$$\overline{P}\big(\mathcal{C}\times\mathcal{X}\times\{o\}|x,p\big)=\overline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}(\cdot,x,p)|x,p\big)=I_{\{o\}}(p)\underline{P}(\mathcal{C}|x,p)=I_{\{o\}}(p),$$

whence $\overline{P}(\mathcal{C}\times\mathcal{X}\times\{o\}|X,O)=I_{\{o\}}$. Consequently, we find for all $x\in\mathcal{X}$ that

$$\overline{P}\big(\overline{P}\big(\mathcal{C}\times\mathcal{X}\times\{o\}|X,O\big)|x\big)=\max_{p\in\Gamma(x)}I_{\{o\}}(p)=\begin{cases}1 & \text{if } o\in\Gamma(x)\\ 0 & \text{otherwise}\end{cases}=I_{\{o\}^*}(x),$$

whence $\overline{P}(\overline{P}(\mathcal{C}\times\mathcal{X}\times\{o\}|X,O)|X)=I_{\{o\}^*}$, and therefore, by Eq. (12),

$$\overline{P}\big(\mathcal{C}\times\mathcal{X}\times\{o\}\big)=\overline{P}_0\big(\overline{P}\big(\overline{P}\big(\mathcal{C}\times\mathcal{X}\times\{o\}|X,O\big)|X\big)\big)=\overline{P}_0\big(\{o\}^*\big)>0,$$

where the last inequality follows from the assumptions. Since $\overline{P}(\mathcal{C}\times\mathcal{X}\times\{o\})>0$, we can calculate the regular extension as

$$\underline{R}(f|o)=\max\big\{\mu:\ \underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]\big)\geqslant 0\big\}.$$

[16] Actually, the irrelevance assumption (MDI) does not determine $\underline{P}(\cdot|X,O)$ completely, but we shall see that this is of no consequence for finding the posterior $\underline{R}(\cdot|O)$.

Again using separate coherence, we find that for all (x, p) in $\mathcal{X} \times \mathcal{O}$,

$$\begin{aligned}\underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]|x,p\big) &= \underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}(\cdot,x,p)[f-\mu]|x,p\big)\\ &= I_{\{o\}}(p)\underline{P}(f-\mu|x,p) = I_{\{o\}}(p)\big[\underline{P}(f|x,p)-\mu\big],\end{aligned}$$

whence $\underline{P}(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]|X,O) = I_{\{o\}}[\underline{P}(f|X,O)-\mu]$. Consequently, we find that for all $x \in \mathcal{X}$, using Eq. (11) and the irrelevance assumption (MDI),

$$\begin{aligned}&\underline{P}\big(\underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]|X,O\big)|x\big)\\ &\quad= \min_{p\in\Gamma(x)} I_{\{o\}}(p)\big[\underline{P}(f|x,p)-\mu\big] = \min_{p\in\Gamma(x)} I_{\{o\}}(p)\big[\underline{P}_0(f|x)-\mu\big]\\ &\quad= \begin{cases}\min\{0,\underline{P}_0(f|x)-\mu\} & \text{if } o\in\Gamma(x)\\ 0 & \text{otherwise}\end{cases}\\ &\quad= I_{\{o\}^*}(x)\min\big\{\underline{P}_0(f|x)-\mu,0\big\},\end{aligned}$$

where we used the fact that $\{o\}_* = \emptyset$. Consequently, $\underline{P}(\underline{P}(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]|X,O)|X) = I_{\{o\}^*}\min\{\underline{P}_0(f|X)-\mu,0\}$, and therefore, by Eq. (12),

$$\begin{aligned}\underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]\big) &= \underline{P}_0\big(\underline{P}\big(\underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]|X,O\big)|X\big)\big)\\ &= \underline{P}_0\big(I_{\{o\}^*}\min\big\{\underline{P}_0(f|X)-\mu,0\big\}\big),\end{aligned}$$

whence

$$\begin{aligned}\underline{R}(f|o) &= \max\big\{\mu\colon \underline{P}\big(I_{\mathcal{C}\times\mathcal{X}\times\{o\}}[f-\mu]\big)\geqslant 0\big\}\\ &= \max\big\{\mu\colon \underline{P}_0\big(I_{\{o\}^*}\min\{\underline{P}_0(f|X)-\mu,0\}\big)\geqslant 0\big\}.\end{aligned}$$

A course of reasoning similar to the one in the proof of Theorem 4 now tells us that indeed

$$\underline{R}(f|o) = \min_{x\in\{o\}^*}\underline{P}_0(f|x)$$

[replace the gamble f on $\mathcal{X}$ in that proof by the gamble $\underline{P}_0(f|X)$]. □

4.3. The conservative updating rule

Let us now denote by E that part of the attributes variable X that is instantiated, for which actual values are available. We denote its value by e. Let R denote the other part, for whose components values are missing. We shall denote the set of its possible values by $\mathcal{R}$, and a generic element of that set by r. Observe that for every $r \in \mathcal{R}$, the attributes vector (e, r) is a possible *completion* of the incomplete observation $o = (e, *)$ (with some abuse of notation) to a complete attributes vector. Moreover, $\{o\}^* = \{e\} \times \mathcal{R}$. We deduce from Theorem 5 that the updated lower prevision $\underline{R}(\cdot|e,*)$ is then given by

$$\underline{R}(f|e,*) = \min_{r\in\mathcal{R}}\underline{P}_0(f|e,r) \tag{CUR}$$

for all gambles f on $\mathcal{C}$, provided that $\overline{P}_0(\{(e,r)\}) > 0$ for all $r \in \mathcal{R}$, which we shall assume to be the case. We shall call (CUR) the *conservative updating rule*.

We shall discuss the case that $\underline{P}_0$ and $\underline{P}_0(\cdot|X)$ are imprecise in Section 7. But let us first, for the remainder of this section, and in Sections 5 and 6, assume that $\underline{P}_0$ and $\underline{P}_0(\cdot|X)$

are precise. Observe that even in this case, the posterior $\underline{R}(\cdot|e,*)$ is imprecise. How can we use this imprecise posterior to perform classification? We shall only discuss the simplest case: we associate a reward function $f_c = I_{\{c\}}$ with each class c in $\mathcal{C}$, and we look for those classes c that are undominated elements of the strict partial order $>$ on $\mathcal{C}$, defined by

$$\begin{aligned} c' > c'' \quad &\Leftrightarrow \quad \underline{R}(I_{\{c'\}} - I_{\{c''\}}|e,*) > 0 \\ &\Leftrightarrow \quad \min_{r\in\mathcal{R}} P_0(I_{\{c'\}} - I_{\{c''\}}|e,r) > 0 \\ &\Leftrightarrow \quad (\forall r \in \mathcal{R})\big(p_0(c'|e,r) > p_0(c''|e,r)\big) \\ &\Leftrightarrow \quad \min_{r\in\mathcal{R}} \frac{p_0(c'|e,r)}{p_0(c''|e,r)} > 1, \end{aligned} \tag{14}$$

where we have used (CUR), and where $p_0(\cdot|e,r)$ denotes the mass function of $P_0(\cdot|e,r)$. Since for all r in $\mathcal{R}$, it is also assumed that $p_0(e,r) > 0$, we can apply Bayes' rule to rewrite this as

$$c' > c'' \quad \Leftrightarrow \quad \min_{r\in\mathcal{R}} \frac{p_0(c',e,r)}{p_0(c'',e,r)} > 1. \tag{15}$$

Eq. (14) is interesting: it tells us that $c' > c''$ if c' is strictly preferred to c'' under all the possible completions (e,r) of the observed data $(e,*)$, i.e., if the strict preference is *robust* under all these possible completions.

Classification is then done by assigning an object with observed attributes $(e,*)$ to the *set* of optimal, undominated classes for the strict preference $>$. Among these optimal classes, there may be classes c' and c'' that are equivalent:

$$(\forall r \in \mathcal{R})\big(p_0(c'|e,r) = p_0(c''|e,r)\big),$$

i.e., that are equally probable under all possible completions (e,r) of $(e,*)$. Otherwise they are incomparable, which means that $p_0(c'|e,r_1) \geqslant p_0(c''|e,r_1)$ for some completion (e,r_1) and $p_0(c'|e,r_2) \leqslant p_0(c''|e,r_2)$ for another completion (e,r_2), where one of these inequalities will be strict. For such incomparable classes, the fact that observations are missing is responsible for our inability to make a choice between them.

In the case of the earlier medical example, e denotes the part of the profile that is known for a patient and the same incomplete profile can be regarded as the set $\{(e,r)|r \in \mathcal{R}\}$ of complete profiles that are consistent with it. The conservative updating rule tells us that in order to update our beliefs on the possible diseases given the incomplete profile, we have to consider all the complete profiles consistent with it, which leads us to lower and upper probabilities and previsions. As we explained above, this will generally give rise only to partial classifications. That is, in general we shall only be able to exclude some of the possible diseases given the evidence. This *may* lead to the identification of a single disease, but only when the conditions justify precision.

The conservative updating rule is a significant result: it provides us with the correct updating rule to use with an unknown incompleteness mechanism; and it shows that robust, conservative inference can be achieved by relying only on the original prior model of domain uncertainty.

It also is a conceptually simple rule, as it involves taking all the possible completions of the missing attributes. It is not, therefore, very surprising that the use of analogous

procedures has already been advocated in the context of robust statistical inference (see for instance [26,31,40]). These focus on the problem of *learning* a model from an incomplete sample, which is then simply regarded as the set of all the complete samples that are consistent with it. But we are not aware of anyone proposing (and justifying) the same intuitive principle for updating beliefs when observations are incomplete. Perhaps the reluctance to change firmly entrenched beliefs about the more traditional naive updating has played a role in this. In contradistinction with the previous work on learning models, we are indeed proposing a new (coherent) rule for *updating beliefs*.

4.4. Some comments on the irrelevance assumption

Let us end this section with a discussion of the irrelevance assumption (MDI), but placed in a context more general than classification. (Additional technical comments on Assumption (MDI) in the case that $\underline{P}_0$ and $\underline{P}_0(\cdot|X)$ are precise, are given in Appendix B.)

Assume that we are studying the relation between *observations* X and *conclusions* C, in the sense that observing the value x of X in $\mathcal{X}$ changes our beliefs about which value C assumes in $\mathcal{C}$. Due to some reason, we cannot observe the value of X, but there is an incompleteness mechanism that produces an incomplete version O of X. In this general context, Assumption (MDI) tells us that if we have a precise observation $X = x$, then the additional knowledge of what incomplete observation $O = o$ is generated by x, will not affect our beliefs about the conclusion C. In other words, if we know the value of the precise observation, then knowing what incomplete observation it produces, becomes completely superfluous. This can be easily reformulated in the more specific context of classification discussed above: if we know the value of all the attributes, then knowing that some of the attributes fail to be measured will be irrelevant to the classification.

We feel that this is precisely what characterises problems of missing data, or of incomplete observations: when something that can be missing is actually measured, the problem of missing data disappears. Let us consider the opposite case, where the bare fact that an attribute is not measured is directly relevant to predicting the class. This fact should then become part of the classification model by making a new attribute out of it, and treating it accordingly, so that this should not be regarded as a problem of missing information. Stated differently, once the model properly includes all the factors that are relevant to predicting the class, (MDI) follows naturally.

Regarding the relationship between assumption CAR/MAR and our irrelevance assumption (MDI), it is not difficult to prove that if the former is satisfied (even in the case of an imprecise prior discussed in Theorem 2) then the latter holds automatically. This is not surprising as the CAR/MAR assumption identifies a subset of a much larger class of incomplete observation (and missing data) problems, which are characterised in general by (MDI). Note, however, that although one implies the other, they do refer to different things. In the context of classification, MAR states that any incomplete observation o is equally likely to have been produced by all the attribute vectors x that may produce it, i.e., there is no compatible attribute vector x that yields observation o with a higher probability $p(o|x)$ than any other compatible attribute vector. MAR therefore says something about the mechanism that produces observations o from attribute vectors x, i.e., about the *the missing data mechanism* itself. Our irrelevance condition (MDI), on the other hand,

states that if we know the attribute vector precisely, then knowing in addition what observation o is produced will not affect the classification. In other words, we assume that the classification only depends on the attributes, and *not on the missing data mechanism*.

CAR/MAR is much stronger than our irrelevance assumption, but it is worth pointing out that there are cases where making the MAR assumption is completely justified, and where, consequently, our approach leads to results that are much too weak. We give one notable example: the case of an attribute that we know is always missing. In this case the missing data mechanism clearly satisfies the MAR assumption: the probability of outcome $*$ is one, irrespective of the actual value of the attribute. MAR then tells us that we can discard this attribute variable, or 'marginalise it out', as is the usual practice. We should therefore not apply the conservative updating rule. We advocate using our rule only when nothing is known about the incompleteness mechanism, and this clearly is not the case here.

It may useful to extend the discussion to statistical inference, even if, strictly speaking, this goes beyond the scope of our present work. In particular, it is well-known (see for instance [26, Proposition 2.1]) that the CAR/MAR assumption cannot be tested statistically, in the sense that we cannot use incomplete observations to check whether it is reasonable. It does not seem to be possible to test Assumption (MDI) either, for essentially the same reasons. To understand this, let us, for the sake of simplicity, look at the case of precise probabilities: it should be tested whether or not $p(c|x,o) = p(c|x)$ for all classes c (with obvious notations). The problem is that the precise observation x is always hidden to us; we can only see the incomplete observation o. So in a statistical inference setting only $p(c,o)$ and not $p(c,x,o)$ would be accessible via the data, and we would not be able to perform the test. Therefore, there appears to exist a fundamental limitation of statistical inference in the presence of missing data: the actually observed data seem not to allow us to test our assumptions about the missing data mechanism, but nevertheless our inferences rely heavily on the specific assumptions that we make about it! This is one of the reasons why we are advocating that only those assumptions should be imposed that are weak enough to be tenable. On our view, (MDI) is a good candidate.

5. Classification in expert systems with Bayesian networks

One popular way of doing classification in complex real-world domains involves using *Bayesian networks*. These are precise probabilistic models defined by a directed acyclic graph and a collection of conditional mass functions [29].

A generic node Z in the graph is identified with a random variable taking values in a finite set $\mathcal{Z}$ (we use 'node' and 'variable' interchangeably, and we reserve the same symbol for both). Each variable Z holds a collection of conditional mass functions $p_0^{Z|\pi_Z}$, one for each possible joint value π_Z of its direct predecessor nodes (or *parents*) Π_Z. The generic conditional mass function $p_0^{Z|\pi_Z}$ assigns the probability $P_0(\{z\}|\pi_Z) = p_0(z|\pi_Z)$ to a value $z \in \mathcal{Z}$ (we drop the superscript when we refer to actual probabilities).

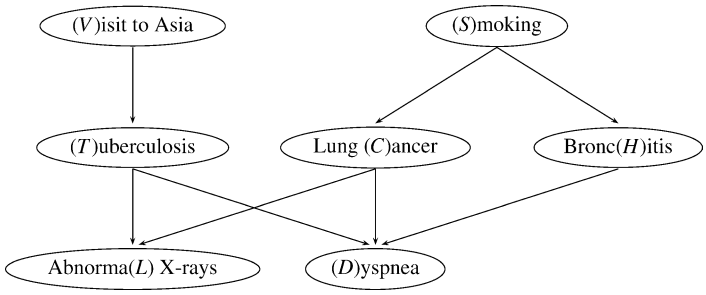

Fig. 2. The 'Asia' Bayesian network.

Table 2
Asia example: probabilities for each variable (first column) in the graph conditional on the values of the parent variables

$V = v'$	0.01							
$S = s'$	0.5							
	v'	v''						
$T = t'$	0.05	0.01						
	s'	s''						
$C = c'$	0.1	0.01						
	s'	s''						
$H = h'$	0.6	0.3						
	$t'c'$	$t'c''$	$t''c'$	$t''c''$				
$L = l'$	0.98	0.98	0.98	0.05				
	$t'c'h'$	$t'c'h''$	$t'c''h'$	$t'c''h''$	$t''c'h'$	$t''c'h''$	$t''c''h'$	$t''c''h''$
$D = d'$	0.9	0.7	0.9	0.7	0.9	0.7	0.8	0.1

Fig. 2 displays the well-known example of Bayesian network called 'Asia'.[17] This models an artificial medical problem by means of cause-effect relationships between random variables, e.g., $S \to C$ (each variable is denoted for short by the related letter between parentheses in Fig. 2). The variables are binary and for any given variable, for instance V, its two possible values are denoted by v' and v'', for the values 'yes' and 'no', respectively. The conditional probabilities for the variables of the model are reported in Table 2.

Bayesian nets satisfy the *Markov condition*: every variable is stochastically independent of its non-descendant non-parents given its parents. Let us consider a generic Bayesian network with nodes C, $A_1, \ldots, A_n$ (for consistency with the notation in Section 4). From the Markov condition, it follows that the joint mass function p_0 is given by

$$p_0(c, a_1, \ldots, a_n) = p_0(c|\pi_C) \prod_{i=1}^{n} p_0(a_i|\pi_{A_i}) \quad \forall (c, a_1, \ldots, a_n) \in \mathcal{C} \times \mathcal{X}, \tag{16}$$

[17] The network presented here is equivalent to the traditional one, although it is missing a logical OR node.

where the values of the parent variables are those consistent with $(c, a_1 \ldots, a_n)$. Hence, a Bayesian network is equivalent to a joint mass function over the variables of the graph. We assume that such a joint mass function assigns positive probability to any event.

Bayesian nets play an important role in the design of expert systems. In this case, domain experts are supposed to provide both the qualitative graphical structure and the numerical values for the probabilities, thus implicitly defining an overall model of the prior uncertainty for the domain of interest. Users can then query the expert system for updating the marginal prior probability of C to a posterior probability according to the available evidence $E = e$, i.e., a set of nodes with known values. In the Asia net, one might ask for the updated probability of lung cancer ($C = c'$), given that a patient is a smoker ($S = s'$) and has abnormal X-rays ($L = l'$), aiming ultimately at making the proper diagnosis for the patient. This kind of updating is very useful as it enables users to do classification, along the lines given in Section 4.

5.1. On updating probabilities with Bayesian networks

Updating the uncertainty for the class variable in a Bayesian net is subject to the considerations concerning incomplete observations in the preceding sections, as generally the evidence set E will not contain all the attributes. To address this problem, one can assume that MAR holds and correspondingly use the naive updating rule to get the posterior $p_0(c|\{e\} \times \mathcal{R})$, but we have already pointed out that this approach is likely to be problematical in real applications. Nevertheless, assuming MAR seems to be the most popular choice with Bayesian nets and the literature presents plenty of algorithmic developments dealing with this case.

Peot and Shachter [30] are a notable exception. In their paper, they explicitly report that "the current practice for modelling missing observations in interactive Bayesian expert systems is incorrect". They show this by focusing on the medical domain where there exists a systematic (i.e., non-MAR) incompleteness mechanism originated by the user of the expert system and also by the patient himself. Indeed, there is a bias in reporting, and asking for, symptoms that are present instead of symptoms that are absent; and a bias to report, and ask for, urgent symptoms over the others. Peot and Shachter tackle this problem by proposing a model of the incompleteness mechanism for the specific situation under study. Explicitly modelling the missing data mechanism is in fact another way to cope with the problem of incomplete observations, perhaps involving the same Bayesian net. The net would then also comprise the nodes O_k, $k = 1, \ldots, n$, for the incomplete observations; and the posterior probability of interest would become $p(c|o)$. Unfortunately, this approach presents serious practical difficulties. Modelling the mechanism can be as complex as modelling the prior uncertainty. Furthermore, it can be argued that in contrast with domain knowledge (e.g., medical knowledge), the way information can be accessed depends on the particular environment where a system will be used; and this means that models of the missing data mechanism will probably not be re-usable, and therefore costly.

These considerations support adopting a robust approach that can be effectively implemented, like the one we proposed in Section 4. It is also useful to stress that our approach has quite general applicability. The conservative updating rule, for example, is

perfectly suited to addressing Peot and Shachter's problem, as the biases they deal with are easily shown to satisfy the irrelevance condition (MDI).

We next develop an algorithm that exploits (CUR) to perform reliable classification with Bayesian networks.

6. An algorithm to classify incomplete evidence with Bayesian networks

In this section we develop an algorithm to perform classification with Bayesian networks by using the conservative updating rule (CUR). As discussed in Section 2.10 and later at the end of Section 4, it is important to realise first that conservative updating will not always allow two classes to be compared, i.e., (CUR) generally produces only a partial order on the classes.

As a consequence, the classification procedure consists in comparing each pair of classes by strict preference (which we shall also call *credal dominance*, in accordance with [41]) and in discarding the dominated ones. The system will then output a set of *possible*, optimal classes. In the following we address the issue of efficient computation of the credal dominance test. Let c' and c'' be two classes in $\mathcal{C}$. We shall use Eq. (15) to test whether c' credal-dominates c''.

Let π' and π'' denote values of the parent variables consistent with the completions (c', e, r) and (c'', e, r), respectively. If a node's parents do not contain C, let π denote the value of the parent variables consistent with (e, r). With some abuse of notation, we shall treat the vector R of those attributes for which measurements are missing, in the following as a set. Furthermore, without loss of generality, let $A_1, \ldots, A_m$, $m \leqslant n$, be the *children* (i.e., the direct successor nodes) of C, and $K = \{1, \ldots, m\}$. We shall denote C in the following also as A_0. For each $i = 0, \ldots, m$, let $\Pi^+_{A_i} = \Pi_{A_i} \cup \{A_i\}$. Consider the functions $\phi_{A_i} : \times_{j:\, A_j \in \Pi^+_{A_i} \cap R} \mathcal{A}_j \to \mathbb{R}^+$ $(i = 0, \ldots, m)$, with values equal to $p_0(a_i|\pi'_{A_i})/p_0(a_i|\pi''_{A_i})$ for $i \in K$, and equal to $p_0(c'|\pi_C)/p_0(c''|\pi_C)$ for $i = 0$. We use the symbol μ to denote the minima of the ϕ-functions, in the following way:

$$\mu_{A_0} = \min_{\substack{a_j \in \mathcal{A}_j, \\ A_j \in \Pi^+_C \cap R}} \frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)}, \tag{17}$$

$$\mu_{A_i} = \min_{\substack{a_j \in \mathcal{A}_j, \\ A_j \in \Pi^+_{A_i} \cap R}} \frac{p_0(a_i|\pi'_{A_i})}{p_0(a_i|\pi''_{A_i})}, \quad i \in K. \tag{18}$$

Consider the *Markov blanket* of C, that is, the set of nodes consisting of the parents of C, its children, and the parents of the children of C. Denote by B^+ the union of C with its Markov blanket. We shall refer to B^+ both as a set of nodes and as a subgraph, depending on the context. Initially we focus on networks for which B^+ is singly connected (the overall network can still be multiply connected). We have the following result.

Theorem 6. *Consider a Bayesian network with nodes C, $A_1, \ldots, A_n$, for which B^+ is singly connected. Let $c', c'' \in \mathcal{C}$. Then c' credal-dominates c'' if and only if $\prod_{i=0}^m \mu_{A_i} > 1$.*

Proof. Rewrite the minimum in Eq. (15) as follows:

$$\min_{r\in\mathcal{R}} \frac{p_0(c',e,r)}{p_0(c'',e,r)} = \min_{r\in\mathcal{R}}\left[\frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)}\prod_{i\in K}\frac{p_0(a_i|\pi'_{A_i})}{p_0(a_i|\pi''_{A_i})}\prod_{j\notin K}\frac{p_0(a_j|\pi_{A_j})}{p_0(a_j|\pi_{A_j})}\right]$$

$$= \min_{\substack{a_j\in\mathcal{A}_j,\\ A_j\in B^+\cap R}}\left[\frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)}\prod_{i\in K}\frac{p_0(a_i|\pi'_{A_i})}{p_0(a_i|\pi''_{A_i})}\right]. \tag{19}$$

This shows that the variables that do not belong to B^+ can be discarded in order to test credal dominance. Now recall that every function ϕ_{A_i} [that is, every ratio in Eq. (19)] depends only on the variables in $\Pi^+_{A_i}\cap R$. Given that B^+ is singly connected, we have that only ϕ_{A_i} depends on the variables in $\Pi^+_{A_i}\cap R$. Let us show the last statement by contradiction, by assuming that another function ϕ_{A_k} ($k\in\{0,\ldots,m\}\setminus\{i\}$) depends on a variable in $\Pi^+_{A_i}\cap R$. There are two cases, either the variable in $\Pi^+_{A_i}\cap R$ is A_i or it is a parent of A_i, say U.

In the first case, neither A_i nor A_k coincide with the class variable C: A_i does not coincide with C because no ϕ-function depends on C; in order for ϕ_{A_k} to depend on A_i, A_i must be a parent of A_k, so A_i is not a child of A_k, whence A_k cannot coincide with C. But A_i being a parent of A_k would create the undirected loop C–A_i–A_k–C, making B^+ multiply connected. This case is impossible.

Consider now the second case when ϕ_{A_k} depends on U. In this case U must be a parent of A_k, besides being a parent of A_i. Note that U does not coincide with C because no ϕ-function depends on C. As before, these conditions imply that B^+ should be multiply connected. In the case that A_k coincides with C, the loop is U–C–A_i–U. If C coincides with A_i, the loop is U–C–A_k–U. When neither A_k nor A_i coincide with C, the loop is U–A_k–C–A_i–U. In every case we have a contradiction.

Since the variables in $\Pi^+_{A_i}\cap R$ appear only in the argument of ϕ_{A_i}, they can be minimised out locally to A_i, obtaining μ_{A_i}. (Observe that μ_{A_i} is a number because only the variables in $\Pi^+_{A_i}\cap R$ are in the argument of ϕ_{A_i}.) Then the thesis follows immediately. □

Theorem 6 renders the solution of the credal-dominance test very easy when B^+ is singly connected,[18] with overall computational complexity linear in the size of the input, i.e., B^+ (more precisely, the input is the Bayesian network restricted to B^+). It is useful to emphasise that the theorem works also for networks in which B^+ is multiply connected, provided that the evidence $E=e$ makes B^+ become singly connected. Indeed it is well known with Bayesian networks that the arcs leaving evidence nodes can be removed while preserving the value $p_0(c|e)$ ($c\in\mathcal{C}$) represented by the network. This result extends to credal dominance because it is computed by $\min_{r\in\mathcal{R}}[p_0(c'|e,r)/p_0(c''|e,r)]$ and because $p_0(c|e,r)$ is preserved by dropping the arcs leaving E, for each $c\in\mathcal{C}$ and $r\in\mathcal{R}$.

Now we move to the case that B^+ is multiply connected, and show how the ideas behind the traditional way of dealing with multiply connected networks, called *conditioning*, can

[18] This corrects the invalid claim, made in an earlier version of this paper [6], that the complexity is linear for all networks.

be applied here as well. Conditioning [29] works by instantiating a subset of nodes called the *loop cutset*. The removal of the arcs leaving the loop cutset creates a singly connected net. The computation is then carried out on the singly connected net as many times as there are joint states of the variables in the cutset, and the results are eventually summarised to obtain the result related to the multiply connected net.

With credal dominance, the situation is analogous. We assume that the arcs leaving evidence nodes in B^+ have been removed, and that a loop cutset is given that opens the remaining loops (recall that, according to the above observation, the loops are opened by the cutset also where credal dominance is concerned). Call R_1 the loop cutset, and let R_2 be the set of nodes such that $R = R_1 \cup R_2$. Rewrite the test of credal dominance as

$$\min_{r_1 \in \mathcal{R}_1} \left[\min_{r_2 \in \mathcal{R}_2} \frac{p_0(c'|e, r_1, r_2)}{p_0(c''|e, r_1, r_2)} \right].$$

The inner minimisation is computed by Theorem 6 on the graph B^+ made singly connected by dropping the arcs leaving $E \cup R_1$. The outer minimisation is a simple enumeration of the states of the loop cutset, which takes exponential time in general.

From the viewpoint of worst-case computation complexity, the situation is similar to the computation of the updating. However, the computation of credal dominance will be easier in the cases where B^+ does not coincide with the entire network. Furthermore, since B^+ can be singly connected even when the network is multiply connected, the computation will be linear also on some multiply connected nets.

6.1. An example

Let us consider the Asia net, where we choose C as the class and set the evidence to $L = l'$ and $S = s'$. We want to test whether c' credal-dominates c''.

Dropping the arcs leaving S, we obtain a new network in which B^+ is $\{C, L, D, T, H\}$. B^+ is multiply connected, and we select $\{T\}$ as loop cutset. We start by considering the case $T = t'$. We must compute μ_D, μ_L, and μ_C. We have:

$$\mu_D = \min_{d \in \mathcal{D}, h \in \mathcal{H}} \frac{p_0(d|t', c', h)}{p_0(d|t', c'', h)} = \min\left\{ \frac{0.9}{0.9}, \frac{0.7}{0.7}, \frac{0.1}{0.1}, \frac{0.3}{0.3} \right\} = 1,$$

$$\mu_L = \frac{p_0(l'|t', c')}{p_0(l'|t', c'')} = \frac{0.98}{0.98} = 1,$$

$$\mu_C = \frac{p_0(c'|s')}{p_0(c''|s')} = \frac{0.1}{0.9} = \frac{1}{9},$$

and their product is $1/9$. In the case $T = t''$, we obtain the following values,

$$\mu_D = \min_{d \in \mathcal{D}, h \in \mathcal{H}} \frac{p_0(d|t'', c', h)}{p_0(d|t'', c'', h)} = \min\left\{ \frac{0.9}{0.8}, \frac{0.7}{0.1}, \frac{0.1}{0.2}, \frac{0.3}{0.9} \right\} = \frac{1}{3},$$

$$\mu_L = \frac{p_0(l'|t', c')}{p_0(l'|t', c'')} = \frac{0.98}{0.05} = \frac{98}{5},$$

$$\mu_C = \frac{p_0(c'|s')}{p_0(c''|s')} = \frac{0.1}{0.9} = \frac{1}{9},$$

with product equal to $98/135 \simeq 0.726$. The minimum of the products obtained with the two values for T is just $1/9$, so that c'' is undominated.

Testing whether c'' credal-dominates c' is very similar and leads to $45/686$ as the value of the test, so c' is undominated as well. In this situation, the system suspends judgement, i.e., it outputs both the classes, as there is not enough information to allow us to choose between the two. This can be seen also by computing the posterior interval of probability for c' by the conservative updating rule, which leads to $[0.1, 0.934]$. The width of this interval quantifies the mentioned lack of information. All of this should be contrasted with naive updating, which produces $p_0(c'|l', s') \simeq 0.646$, and leads us to diagnose cancer.

It is useful to better analyse the reasons for the indeterminate output of the proposed system. Given our assumptions, the system cannot exclude that the available evidence is part of a more complete piece of evidence where $T = t'$, $D = d'$, and $H = h'$. If this were the case, then c'' would be nine times as probable *a posteriori* as c', and we should diagnose no cancer. However, the system cannot exclude either that the more complete evidence would be $T = t''$, $D = d'$, and $H = h''$. In this case, the ratio of the posterior probability of c' to that of c'' would be $686/45$, leading us to the opposite diagnosis.

Of course when the evidence is strong enough, the proposed system does produce determinate conclusions. For instance, the evidence $L = l'$, $S = s'$ and $T = t'$ will make the system exclude the presence of cancer.

7. Working with credal networks

Credal networks provide a convenient way of specifying prior knowledge using the theory of coherent lower previsions. They extend the formalism of Bayesian networks by allowing sets of mass functions [2,11], or equivalently, sets of linear previsions. These are also called *credal sets* after Levi [24]. We recall that a credal set is equivalent to a coherent lower prevision, as pointed out in Section 2.3.

A *credal network* is a pair composed of a directed acyclic graph and a collection of conditional credal sets[19] (i.e., a collection of conditional lower previsions). We intend the graph to code strong independences. Two variables Z_1 and Z_2 are said to be *strongly independent* when every vertex in the credal set of joint mass functions for (Z_1, Z_2), satisfies stochastic independence of Z_1 and Z_2. That is, for every extreme mass function p in the credal set, and for all the possible pairs $(z_1, z_2) \in \mathcal{Z}_1 \times \mathcal{Z}_2$, it holds that $p(z_1|z_2) = p(z_1)$ and $p(z_2|z_1) = p(z_2)$.[20] Each variable Z in the net holds a collection of conditional lower previsions, denoted by $\underline{P}_0^{Z|\pi_Z}$, one for each possible joint value π_Z of the node Z's parents Π_Z. With some abuse of notation,[21] let $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ be the credal set of mass functions for the linear previsions dominating $\underline{P}_0^{Z|\pi_Z}$. $p_0^{Z|\pi_Z} \in \mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ assigns

[19] In this context, as in [2], we restrict ourselves to credal sets with a finite number of extreme points.

[20] See also [28] for a complete account of different strong independence concepts and [3] for a deep analysis of strong independence.

[21] In preceding sections, the symbol $\mathcal{M}$ was used to denote the dominating set of linear previsions. We use the same symbol here as there is one-to-one correspondence between linear previsions and mass functions (see Section 2.3).

the probability $p_0(z|\pi_Z)$ to a value $z \in \mathcal{Z}$. In the following we assume that each of these mass functions assigns positive probability to any event. Given the equivalence between lower probability functions and credal sets, we can regard each node of the net to hold a collection of conditional, so-called *local*, credal sets. Actually, the usual approach of specifying the conditional lower previsions for the nodes precisely amounts to providing the local credal sets directly. This is commonly done by *separately specifying* these credal sets [12,37], something that we also assume here: this implies that selecting a mass function from a credal set does not influence the possible choices in others. This assumption is natural within a Bayesian sensitivity analysis interpretation of credal nets.

Credal nets satisfy a generalised version of the Markov condition called the *strong Markov condition*: each variable is strongly independent of its non-descendant non-parents given its parents. This leads immediately to the definition of the *strong extension* [3] of a credal net. This is the most conservative lower prevision $\underline{P}_0$ on $\mathcal{L}(\mathcal{C} \times \mathcal{X})$ that coherently extends the nodes' conditional lower previsions, subject to the strong Markov condition. Let the nodes of the network be C (i.e., A_0), $A_1, \ldots, A_n$, as before. It is well known that the credal set equivalent to $\underline{P}_0$ is

$$\mathcal{M}(\underline{P}_0) = \mathrm{CH}\big\{p_0 \text{ factorising as in Eq. (16): } p_0^{A_i|\pi_{A_i}} \in \mathcal{M}\big(\underline{P}_0^{A_i|\pi_{A_i}}\big), \; i = 0, \ldots, n\big\}, \tag{20}$$

where CH denotes the convex hull operation. In other words, $\mathcal{M}(\underline{P}_0)$ is the convex hull of the set of all the joint mass functions that factorise according to Eq. (16), obtained by selecting conditional mass functions from the local credal sets of the net in all the possible ways. The strong extension is an imprecise prior defined by means of the composition of local information. From yet another viewpoint, the credal set $\mathcal{M}(\underline{P}_0)$ makes a Bayesian sensitivity analysis interpretation of credal nets very natural: working with a credal net can equivalently be regarded as working simultaneously with the set of all Bayesian nets consistent with $\mathcal{M}(\underline{P}_0)$.

The credal set $\mathcal{M}(\underline{P}_0)$ can have a huge number of extreme mass functions. Indeed, the computation of lower and upper probabilities with strong extensions is NP-hard [12][22] also when the graph is a polytree. Polytrees are directed acyclic graphs with the characteristic that forgetting the direction of arcs, the resulting graph has no undirected cycles. This should be contrasted with Bayesian networks for which common computations take polynomial time with polytrees. Indeed, the difficulty of computation with credal nets has severely limited their use so far, even though credal nets have the great advantage over Bayesian nets of not requiring the model probabilities to be specified precisely. This is a key point for faithfully modelling human knowledge, which also allows expert systems to be developed quickly.

In the following we extend Theorem 6 to credal nets, showing that conservative updating allows classification with credal nets to be realised with the same complexity needed for

[22] However, it should be observed that Ferreira da Rocha and Cozman's result is proved for the subset of polytrees in which the local credal sets are convex hulls of degenerate mass functions that assign all the mass to one elementary event. As such, it does not tell us anything about the complexity of working with the case of polytrees whose credal sets are made up of mass functions that assign positive probability to any event.

Bayesian nets. This appears to be an important result, with implications for the practical usability of credal nets in modelling knowledge.

Below we reuse the definition $\Pi^+_{A_i}$ given in Section 6, we again denote by B^+ the union of C with its Markov blanket, and we refer to C also by A_0. Consider the following quantities:

$$p^{C|\pi_C}_{0*} = \operatorname{argmin}_{p_0^{C|\pi_C} \in \mathcal{M}(\underline{P}_0^{C|\pi_C})} \frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)}, \tag{21}$$

and, for each $i \in K$,

$$\underline{p}_0(a_i|\pi'_{A_i}) = \min_{p_0^{A_i|\pi'_{A_i}} \in \mathcal{M}(\underline{P}_0^{A_i|\pi'_{A_i}})} p_0(a_i|\pi'_{A_i}), \tag{22}$$

$$\overline{p}_0(a_i|\pi''_{A_i}) = \max_{p_0^{A_i|\pi''_{A_i}} \in \mathcal{M}(\underline{P}_0^{A_i|\pi''_{A_i}})} p_0(a_i|\pi''_{A_i}), \tag{23}$$

as well as the functions $\underline{\phi}_{A_i} : \times_{j:\, A_j \in \Pi^+_{A_i} \cap R} \mathcal{A}_j \to \mathbb{R}^+$ $(i = 0, \ldots, m)$, with values equal to $\underline{p}(a_i|\pi'_{A_i})/\overline{p}(a_i|\pi''_{A_i})$ for $i \in K$, and equal to $p_{0*}(c'|\pi_C)/p_{0*}(c''|\pi_C)$ for $i = 0$. We use the symbol $\underline{\mu}$ to denote the minima of the $\underline{\phi}$-functions, as follows:

$$\underline{\mu}_{A_0} = \min_{\substack{a_j \in \mathcal{A}_j,\\ A_j \in \Pi^+_C \cap R}} \frac{p_{0*}(c'|\pi_C)}{p_{0*}(c''|\pi_C)}, \tag{24}$$

$$\underline{\mu}_{A_i} = \min_{\substack{a_j \in \mathcal{A}_j,\\ A_j \in \Pi^+_{A_i} \cap R}} \frac{\underline{p}_0(a_i|\pi'_{A_i})}{\overline{p}_0(a_i|\pi''_{A_i})}, \quad i \in K. \tag{25}$$

We have the following result.

Theorem 7. *Consider a credal net with nodes C, $A_1, \ldots, A_n$, for which B^+ is singly connected. Let $c', c'' \in \mathcal{C}$. Then c' credal-dominates c'' if and only if $\prod_{i=0}^m \underline{\mu}_{A_i} > 1$.*

Proof. A credal net can equivalently be regarded as a set of Bayesian nets, as is apparent from Eq. (20). Accordingly, for credal dominance to hold with a credal net, it is necessary that it holds for all the joint mass functions consistent with the strong extension. This can be tested by solving the following double minimisation problem:

$$\min_{p_0 \in \mathcal{M}(\underline{P}_0)} \min_{r \in \mathcal{R}} \frac{p_0(c', e, r)}{p_0(c'', e, r)} \tag{26}$$

$$= \min_{p_0^{C|\pi_C} \in \mathcal{M}(\underline{P}_0^{C|\pi_C})} \min_{\substack{p_0^{A_k|\pi'_{A_k}} \in \mathcal{M}(\underline{P}_0^{A_k|\pi'_{A_k}}),\\ p_0^{A_k|\pi''_{A_k}} \in \mathcal{M}(\underline{P}_0^{A_k|\pi''_{A_k}}),\\ k \in K}} \min_{\substack{a_j \in \mathcal{A}_j,\\ A_j \in B^+ \cap R}} \left[\frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)} \prod_{i \in K} \frac{p_0(a_i|\pi'_{A_i})}{p_0(a_i|\pi''_{A_i})} \right] \tag{27}$$

$$= \min_{\substack{a_j \in \mathcal{A}_j, \\ A_j \in B^+ \cap R}} \left\{ \min_{p_0^{C|\pi_C} \in \mathcal{M}(\underline{P}_0^{C|\pi_C})} \left[\frac{p_0(c'|\pi_C)}{p_0(c''|\pi_C)} \right] \times \prod_{i \in K} \frac{\min_{p_0^{A_i|\pi'_{A_i}} \in \mathcal{M}(\underline{P}_0^{A_i|\pi'_{A_i}})} p_0(a_i|\pi'_{A_i})}{\max_{p_0^{A_i|\pi''_{A_i}} \in \mathcal{M}(\underline{P}_0^{A_i|\pi''_{A_i}})} p_0(a_i|\pi''_{A_i})} \right\} \tag{28}$$

$$= \min_{\substack{a_j \in \mathcal{A}_j, \\ A_j \in B^+ \cap R}} \left[\frac{p_{0*}(c'|\pi_C)}{p_{0*}(c''|\pi_C)} \prod_{i \in K} \frac{\underline{p}(a_i|\pi'_{A_i})}{\overline{p}(a_i|\pi''_{A_i})} \right], \tag{29}$$

where the passage from (26) to (27) is due to (19) and (20);[23] and the following passage is possible thanks to the characteristic of separate specification of credal sets in the credal network. Note that expression (29) resembles expression (19) of Theorem 6. In fact, the proof of Theorem 6 below expression (28) applies here as well: $\underline{\phi}_{A_i}$ depends only on the variables in $\Pi_{A_i}^+ \cap R$ and only $\underline{\phi}_{A_i}$ depends on them. As in Theorem 6, the thesis follows immediately since the variables in $\Pi_{A_i}^+ \cap R$ can then be minimised out locally to A_i, obtaining $\underline{\mu}_{A_i}$. □

Theorem 7 renders the solution of the credal dominance test for credal networks very easy when B^+ is singly connected. However, in order to have a better idea of the computational complexity, one has to carefully examine the complexity of solving problems (21)–(23). This is what we set out to do in the following.

Let again Z be a generic variable in the network. We consider three common ways of specifying the local credal sets of the net.

1. In the first case, the conditional[24] credal set $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ for the variable Z is specified via linear constraints on the probabilities $p_0(z|\pi_Z)$, $z \in \mathcal{Z}$. That is, in this representation the vector of probabilities $p_0(z|\pi_Z)$, $z \in \mathcal{Z}$, can take every value in a closed and bounded space described by linear constraints on the variables $p_0(z|\pi_Z)$, i.e., in a *polytope*.
2. In the second case, we assume that $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ is the convex hull of a set of mass functions directly provided by the modeller.
3. Finally, we consider the case when $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ is provided by specifying intervals of probability for the elementary events $(z|\pi_Z)$, $z \in \mathcal{Z}$. This is a special case of case 1 where the only constraints allowed on the probabilities $p_0(z|\pi_Z)$ are bounds, except for $\sum_{z \in \mathcal{Z}} p_0(z|\pi_Z) = 1$. Without loss of generality, we assume that the probability intervals are *reachable* [4]. This holds if and only if $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ is non-empty and the intervals are tight, i.e., for each lower and upper bound there is a mass function

[23] Actually, the passage is also based on the fact that the minimum of (26) is achieved at an extreme point of $\mathcal{M}(\underline{P}_0)$. This is well-known with credal networks and is pointed out formally by Theorems 5 and 7 in reference [11].

[24] The situation with root nodes is analogous.

in $\mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ at which the bound is attained. Reachable intervals produce a coherent lower prevision $\underline{P}_0^{Z|\pi_Z}$ that is *2-monotone* [4]. For 2-monotone lower previsions it holds that, given any two mutually exclusive events $\mathcal{Z}', \mathcal{Z}'' \subseteq \mathcal{Z}$, there is a mass function $p_{0+}^{Z|\pi_Z} \in \mathcal{M}(\underline{P}_0^{Z|\pi_Z})$ for which $\underline{p}_0(\mathcal{Z}'|\pi_Z) = p_{0+}(\mathcal{Z}'|\pi_Z)$ and $\overline{p}_0(\mathcal{Z}''|\pi_Z) = p_{0+}(\mathcal{Z}''|\pi_Z)$. We shall use this property in the following.

Observe that the representations in cases 1 and 2 are fully general as any credal set can be represented by one or by the other. In the following we consider that all the local credal set of the net are specified either as in case 1 or 2 or 3. We do not consider mixed cases, which should be easy to work out once the 'pure' cases have been addressed.

Let us now focus on the complexity of testing credal dominance in case 1. Let S be the size of the largest local credal set in the network. The size is defined as the dimension of the constraints-variables matrix that describes the linear domain. Let $\mathrm{O}(L(S))$ be the complexity to solve a linear minimisation problem of size S. Note that this is a polynomial-time complexity [19]. We have that each minimisation in Eqs. (22)–(23) takes time $\mathrm{O}(L(S))$ at most. This holds also for the minimisation in (21) which can be converted to a linear minimisation problem by a result from Charnes and Cooper [1]. Note that each of the mentioned minimisations must be repeated for all the joint states of the variables in $\Pi_{A_i}^+ \cap R$, whose number is upper bounded by the states of those in Π_{A_i}. Denoting by H the worst-case number of states of the variables in Π_{A_i} obtained by letting i vary from 0 to m, we have that the overall computational complexity for problems (22)–(23) is $\mathrm{O}(H \cdot L(S))$ at most. We can regard this part as a pre-processing step of the test of credal dominance. Once the pre-processing is over, the set of minimisations in Eqs. (24)–(25) takes linear time in the size of B^+ as in the case of Bayesian networks.

Case 2 presents a lower *overall* complexity for testing credal dominance. In fact, the minimisations in Eqs. (21)–(23) can be solved simply by enumerating the mass functions that make up each credal set. These mass functions are specified directly by the modeller, i.e., they are an input of the problem. For this reason the overall complexity of testing credal dominance is linear in the size of B^+.

The final case of probability intervals is also easily solved. With respect to Eqs. (22)–(23), $\underline{p}_0(a_i|\pi'_{A_i})$ and $\overline{p}_0(a_i|\pi''_{A_i})$ are just the left and the right extreme of the probability intervals for $(a_i|\pi'_{A_i})$ and $(a_i|\pi''_{A_i})$, respectively, so no computation is needed for them. As far as Eq. (21) is concerned, we have that the minimum of $p_0(c'|\pi_C)/p_0(c''|\pi_C)$ taken with respect to the mass functions in $\mathcal{M}(\underline{P}_0^{C|\pi_C})$ is equal to $\underline{p}_0(c'|\pi_C)/\overline{p}_0(c''|\pi_C)$ by the property mentioned at the end of case 3. Again, $\underline{p}_0(c'|\pi_C)$ and $\overline{p}_0(c''|\pi_C)$ are readily available as an input of the problem. Overall, the complexity of testing credal dominance is linear in the size of B^+ in this case as well.

So far we have treated the case when B^+ is singly connected. The extension to the general case is completely analogous to that already developed for Bayesian networks, basically because the arcs leaving evidence nodes can be dropped in credal networks, too. The reason is that a credal net can be regarded as a set of Bayesian nets, and the mentioned property applies to all the Bayesian nets in the set. More precisely, assume, as in the description at the end of Section 6, that a loop cutset is given that together with E

can open all the loops in B^+. Call R_1 the loop cutset, and let R_2 be the set of nodes such that $R = R_1 \cup R_2$. Re-write the test of credal dominance for credal networks as

$$\min_{p_0 \in \mathcal{M}(\underline{P}_0)} \min_{r \in \mathcal{R}} \frac{p_0(c'|e,r)}{p_0(c''|e,r)}$$

$$= \min_{p_0 \in \mathcal{M}(\underline{P}_0)} \left\{ \min_{r_1 \in \mathcal{R}_1} \left[\min_{r_2 \in \mathcal{R}_2} \frac{p_0(c'|e,r_1,r_2)}{p_0(c''|e,r_1,r_2)} \right] \right\} \tag{30}$$

$$= \min_{r_1 \in \mathcal{R}_1} \left\{ \min_{p_0 \in \mathcal{M}(\underline{P}_0)} \left[\min_{r_2 \in \mathcal{R}_2} \frac{p_0(c'|e,r_1,r_2)}{p_0(c''|e,r_1,r_2)} \right] \right\}. \tag{31}$$

Eq. (30) makes it clear that for each selected mass function $p_0 \in \mathcal{M}(\underline{P}_0)$, the minimum in square brackets can be obtained on the graph B^+ that is made singly connected by dropping the arcs leaving $E \cup R_1$. Of course this property continues to hold in the next expression. When we consider the part in braces in (31), that is, also the variations of p_0, we are focusing on the singly connected credal net, with graph B^+, obtained from the multiply connected one dropping the arcs leaving $E \cup R_1$. Hence, expression (31) shows that the inner double minimisation can be computed by Theorem 7. The outer minimisation is the usual enumeration of the states of the loop cutset.

It turns out that the complexity of testing credal dominance when B^+ is multiply connected is the same both for credal and Bayesian networks. This is an important result, as the complexity to work with credal networks is usually much harder than that needed with Bayesian nets.

8. Conclusions

It seems to us that updating probabilities with incomplete observations presents an important problem for research in uncertain reasoning, and is a pervasive issue in applications. It has been clearly pointed out in the literature that the commonly used CAR assumption about the incompleteness mechanism is often unjustified, and more generally, that it may happen in practical applications that little or no knowledge about the incompleteness mechanism is available. In those cases, naive updating is simply inappropriate.

This paper has addressed the problem of updating probabilities when strong assumptions about the incompleteness mechanism cannot be justified, thus filling an important gap in literature. It has done so by deliberately choosing the conservative point of view of not assuming any knowledge about the incompleteness mechanism. A new so-called conservative updating method follows as a logical consequence, using only arguments of coherence. We used it to derive a new coherent updating rule for probabilistic expert systems. By focusing on expert systems based on Bayesian nets, we have shown that this conservative updating leads to efficient classification of new evidence for a wide class of networks, so the new developments can be exploited immediately in real environments. Furthermore, the related algorithm can be implemented easily and does not require changes in pre-existing knowledge bases, so that existing expert systems can be upgraded to make our robust, conservative, inferences with minimal changes.

We want to stress here that the proposed conservative updating strategy is different in one important respect from the more traditional ones: it generally leads only to partially determined inferences and decisions, and ultimately to systems that can recognise the limits of their knowledge, and suspend judgement when these limits are reached. As necessary consequences of our refusal to make unwarranted assumptions, we believe that these limitations are important characteristics of the way systems ought to operate in the real world. A system that, in a certain state, cannot support any decision on the basis of its knowledge base, will induce a user to look for further sources of information external to the system. In contrast, systems that may make arbitrary choices without making that evident, will wrongly lead a user to think that also these choices are well motivated.

We also believe it is important to stress here that it is difficult to avoid partial indeterminacy in real applications. Realistic states of partial knowledge about the incompleteness mechanism, other than the total ignorance modelled here, should in principle also be modelled by a (non-vacuous) coherent lower prevision, which may again lead to indeterminacy except in very special cases, such as when enough information is available to justify modelling the incompleteness mechanism by a precise probability model. For analogous reasons, domain knowledge should most likely be modelled by a coherent lower prevision, too. In practise this can be done by moving from Bayesian to credal networks. It appears that this step has not really been taken so far, probably because of the computational complexity of working in the more general framework of credal networks. This paper shows that the classification complexity is unchanged by moving from Bayesian to credal networks, in the realistic scenarios that involve a state of ignorance about the incompleteness mechanism. We hope that this encouraging result may contribute to credal networks receiving due credit also as practical modelling tools.

With respect to future research, we believe an important issue is the development of models able to take advantage of intermediate states of knowledge about the incompleteness mechanism, to the extent of making stronger inferences and decisions. With regard to Bayesian and credal nets, one could for instance think of partitioning the set of attributes in those for which MAR holds and the rest for which the mechanism is unknown. Such hybrid modelling seems to provide a good compromise between generality and flexibility.

Acknowledgements

The authors are grateful to Peter Walley for initial stimulating discussions on the topic of the paper in August 2001. They would also like to thank two anonymous referees for their help in making this paper more readable, and for pointing out a mistaken claim about computational complexity which is now corrected.

This research is partially supported by research grant G.0139.01 of the Flemish Fund for Scientific Research (FWO), and by the Swiss NSF grant 2100-067961.

Appendix A. Extending Walley's Marginal Extension Theorem

This appendix is devoted to the proof of an important theorem, needed in Section 4. It is a generalisation to three random variables of Walley's Marginal Extension Theorem, discussed in Section 2.9 (see Theorem 1). Because the proof is rather technical, and it uses results and notions not explained in the main text, we have decided to discuss it separately.

We consider three random variables X, Y and Z taking values in the respective non-empty and finite spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$.

Theorem A.1. *Consider a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\mathcal{X})$, a separately coherent conditional lower prevision $\underline{P}(\cdot|X)$ on $\mathcal{L}(\mathcal{X}\times\mathcal{Y})$, and a separately coherent conditional lower prevision $\underline{P}(\cdot|X,Y)$ on $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. Then the smallest coherent lower prevision on $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ that has marginal $\underline{P}$ and is jointly coherent with $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$, is given by*

$$\underline{Q}(h)=\underline{P}\big(\underline{P}\big(\underline{P}(h|X,Y)|X\big)\big) \tag{A.1}$$

for all gambles h on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$.

Proof. Lemma A.2 tells us that $\underline{Q}$ is a indeed a coherent lower prevision that has marginal $\underline{P}$. To prove that $\underline{Q}$, $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ are jointly coherent, Walley's Reduction Theorem [37, Theorem 7.1.5] tells us that we need only prove that $\underline{Q}$, $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ are weakly coherent, and that $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ are jointly coherent. This is done in Lemmas A.3 and A.4, respectively. Finally, in Lemma A.5 we prove that any other coherent lower prevision on $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ that has marginal $\underline{P}$ and is jointly coherent with $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$, dominates $\underline{Q}$.

Lemma A.2. *The lower prevision $\underline{Q}$ defined on $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ by Eq.* (A.1) *is coherent and has marginal $\underline{P}$.*

Proof. It is easily verified that $\underline{Q}$ satisfies the axioms ($\underline{P}1$)–($\underline{P}3$) of a coherent lower prevision, because the coherent $\underline{P}$, and the separately coherent $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ do so. It remains to show that $\underline{Q}$ has marginal $\underline{P}$. Consider any gamble f on $\mathcal{X}$. If follows from the separate coherence of $\underline{P}(\cdot|X,Y)$ that $\underline{P}(f|X,Y)=f$ and consequently, from the separate coherence of $\underline{P}(\cdot|X)$ that $\underline{P}(\underline{P}(f|X,Y)|X)=\underline{P}(f|X)=f$, whence indeed $\underline{Q}(f)=\underline{P}(\underline{P}(\underline{P}(f|X,Y)|X))=\underline{P}(f)$. □

Lemma A.3. *$\underline{Q}$, $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ are weakly coherent.*

Proof. Following the discussion in [37, Section 7.1.4], we must prove that

(a) $\max[G(f)+G(g|X)+G(h|X,Y)-G(f_0)]\geqslant 0$;
(b) $\max[G(f)+G(g|X)+G(h|X,Y)-G(g_0|x_0)]\geqslant 0$;
(c) $\max[G(f)+G(g|X)+G(h|X,Y)-G(h_0|x_0,y_0)]\geqslant 0$;

for all f, f_0, h, h_0 in $\mathscr{L}(\mathscr{X}\times\mathscr{Y}\times\mathscr{Z})$, all g, g_0 in $\mathscr{L}(\mathscr{X}\times\mathscr{Y})$, all x_0 in $\mathscr{X}$ and all y_0 in $\mathscr{Y}$, where we use the notations $G(f)=f-\underline{Q}(f)$, $G(g|X)=g-\underline{P}(g|X)$, $G(h|X,Y)=h-\underline{P}(h|X,Y)$, $G(g_0|x_0)=I_{\{x_0\}}[g-\underline{P}(g|x_0)]$ and $G(h_0|x_0,y_0)=I_{\{(x_0,y_0)\}}[h_0-\underline{P}(h_0|x_0,y_0)]$.

To prove that (a) holds, recall from Lemma A.2 that $\underline{Q}$ is a coherent lower prevision, whence (see for instance [37, Section 2.6.1] for properties of coherent lower previsions)

$$\begin{aligned}
&\max\big[G(f)+G(g|X)+G(h|X,Y)-G(f_0)\big]\\
&\quad\geqslant \overline{Q}\big(G(f)+G(g|X)+G(h|X,Y)-G(f_0)\big)\\
&\quad\geqslant \underline{Q}\big(G(f)+G(g|X)+G(h|X,Y)\big)-\underline{Q}\big(G(f_0)\big)\\
&\quad\geqslant \underline{Q}\big(G(f)\big)+\underline{Q}\big(G(g|X)\big)+\underline{Q}\big(G(h|X,Y)\big)-\underline{Q}\big(G(f_0)\big).
\end{aligned}$$

Now, again using the coherence of $\underline{Q}$, we find that $\underline{Q}(G(f))=\underline{Q}(f-\underline{Q}(f))=\underline{Q}(f)-\underline{Q}(f)=0$ and similarly $\underline{Q}(G(f_0))=0$. Moreover, it follows from the separate coherence of $\underline{P}(\cdot|X,Y)$ that for all (x,y) in $\mathscr{X}\times\mathscr{Y}$

$$\begin{aligned}
\underline{P}\big(G\big(h|X,Y\big)|x,y\big)&=\underline{P}\big(h-\underline{P}(h|X,Y)|x,y\big)\\
&=\underline{P}\big(h-\underline{P}(h|x,y)|x,y\big)=\underline{P}(h|x,y)-\underline{P}(h|x,y)=0,
\end{aligned}$$

whence $\underline{P}(G(h|X,Y)|X,Y)=0$ and consequently $\underline{Q}(G(h|X,Y))=0$. Similarly, it follows from the separate coherence of $\underline{P}(\cdot|X,Y)$ that $\underline{P}(G(g|X)|X,Y)=G(g|X)$, and from the separate coherence of $\underline{P}(\cdot|X)$ that for all x in $\mathscr{X}$,

$$\begin{aligned}
\underline{P}\big(\underline{P}\big(G(g|X)|X,Y\big)|x\big)&=\underline{P}\big(G(g|X)|x\big)=\underline{P}\big(g-\underline{P}(g|X)|x\big)\\
&=\underline{P}\big(g-\underline{P}(g|x)|x\big)=\underline{P}(g|x)-\underline{P}(g|x)=0,
\end{aligned}$$

whence $\underline{P}(\underline{P}(G(g|X)|X,Y)|X)=0$ and consequently also $\underline{Q}(G(g|X))=0$. It follows that (a) is indeed verified.

An argument similar to the one above tells us that (b) will hold if we can prove that $\underline{Q}(G(g_0|x_0))=0$. Now it follows from the separate coherence of $\underline{P}(\cdot|X,Y)$ that, since $G(g_o|x_0)\in\mathscr{L}(\mathscr{X}\times\mathscr{Y})$, $\underline{P}(G(g_o|x_0)|X,Y)=G(g_o|x_0)$, whence, using the separate coherence of $\underline{P}(\cdot|X)$,

$$\begin{aligned}
\underline{P}\big(\underline{P}\big(G(g_o|x_0)|X,Y\big)|X\big)&=\underline{P}\big(G(g_o|x_0)|X\big)\\
&=\underline{P}\big(I_{\{x_0\}}\big[g(x_0,\cdot)-\underline{P}\big(g_o(x_0,\cdot)|x_0\big)\big]|X\big)\\
&=I_{\{x_0\}}\big[\underline{P}\big(g(x_0,\cdot)|X\big)-\underline{P}\big(g_o(x_0,\cdot)|x_0\big)\big]=0,
\end{aligned}$$

whence indeed $\underline{Q}(G(g_0|x_0))=0$.

Similarly, (c) will be verified if we can prove that $\underline{Q}(G(h_0|x_0,y_0))=0$. Now it follows from the separate coherence of $\underline{P}(\cdot|X,Y)$ that

$$\begin{aligned}
\underline{P}\big(G(h_0|x_0,y_0)|X,Y\big)&=\underline{P}\big(I_{\{(x_0,y_0)\}}\big[h-\underline{P}(h|x_0,y_0)\big]|X,Y\big)\\
&=I_{\{(x_0,y_0)\}}\big[\underline{P}(h|X,Y)-\underline{P}(h|x_0,y_0)\big]=0,
\end{aligned}$$

whence indeed $\underline{Q}(G(h_0|x_0,y_0))=0$. □

Lemma A.4. *Separately coherent conditional lower previsions $\underline{P}(\cdot|X)$ on $\mathscr{L}(\mathscr{X}\times\mathscr{Y})$ and $\underline{P}(\cdot|X,Y)$ on $\mathscr{L}(\mathscr{X}\times\mathscr{Y}\times\mathscr{Z})$ are always jointly coherent.*

Proof. We use the discussion of joint coherence in [37, Section 7.1.4]. Consider arbitrary g in $\mathcal{L}(\mathcal{X}\times\mathcal{Y})$ and h in $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ and the corresponding sets

$$S(g)=\big\{\{x\}\times\mathcal{Y}\times\mathcal{Z}\colon\ g(x,\cdot)\neq 0\big\}\quad\text{and}\quad S(h)=\big\{\{x\}\times\{y\}\times\mathcal{Z}\colon\ h(x,y,\cdot)\neq 0\big\}.$$

First of all, consider any x_0 in $\mathcal{X}$ and g_0 in $\mathcal{L}(\mathcal{X}\times\mathcal{Y})$, then we must show that there is some B in

$$S(g)\cup S(h)\cup\big\{\{x_0\}\times\mathcal{Y}\times\mathcal{Z}\big\}$$

such that (if we also take into account the separate coherence of $\underline{P}(\cdot|X,Y)$ and $\underline{P}(\cdot|X)$)

$$\max_{(x,y,z)\in B}\big[g(x,y)-\underline{P}\big(g(x,\cdot)|x\big)+h(x,y,z)-\underline{P}\big(h(x,y,\cdot)|x,y\big)-I_{\{x_0\}}(x)\big(g_0(x,y)-\underline{P}\big(g_0(x,\cdot)|x\big)\big)\big]\geqslant 0.$$

We choose $B=\{x_0\}\times\mathcal{Y}\times\mathcal{Z}$, and prove that the corresponding supremum

$$S=\max_{y\in\mathcal{Y}}\max_{z\in\mathcal{Z}}\big[g(x_0,y)-\underline{P}\big(g(x_0,\cdot)|x_0\big)+h(x_0,y,z)-\underline{P}\big(h(x_0,y,\cdot)|x_0,y\big)-\big(g_0(x_0,y)-\underline{P}\big(g_0(x_0,\cdot)|x_0\big)\big)\big]\geqslant 0.$$

Now, since it follows from the coherence of the lower prevision $\underline{P}(\cdot|x_0,y)$ that

$$\max_{z\in\mathcal{Z}}\big[h(x_0,y,z)-\underline{P}\big(h(x_0,y,\cdot)|x_0,y\big)\big]\geqslant 0$$

for all $y\in\mathcal{Y}$, we see that indeed

$$S\geqslant\max_{y\in\mathcal{Y}}\big[g(x_0,y)-\underline{P}\big(g(x_0,\cdot)|x_0\big)-\big(g_0(x_0,y)-\underline{P}\big(g_0(x_0,\cdot)|x_0\big)\big)\big]\geqslant 0,$$

where the last inequality follows from the coherence of the lower prevision $\underline{P}(\cdot|x_0)$.

As a second step, consider any (x_0,y_0) in $\mathcal{X}\times\mathcal{Y}$ and h_0 in $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, then we must show that there is some B in

$$S(g)\cup S(h)\cup\big\{\{x_0\}\times\{y_0\}\times\mathcal{Z}\big\}$$

such that [if we also take into account the separate coherence of $\underline{P}(\cdot|X,Y)$ and $\underline{P}(\cdot|X)$]

$$\max_{(x,y,z)\in B}\big[g(x,y)-\underline{P}\big(g(x,\cdot)|x\big)+h(x,y,z)-\underline{P}\big(h(x,y,\cdot)|x,y\big)-I_{\{(x_0,y_0)\}}(x,y)\big(h_0(x,y,z)-\underline{P}\big(h_0(x,y,\cdot)|x,y\big)\big)\big]\geqslant 0.$$

If $g(x_1,\cdot)\neq 0$ for some $x_1\neq x_0$, then we choose $B=\{x_1\}\times\mathcal{Y}\times\mathcal{Z}$, and similar arguments as in the first step of the proof lead us to conclude that the corresponding supremum

$$\max_{y\in\mathcal{Y}}\max_{z\in\mathcal{Z}}\big[g(x_1,y)-\underline{P}\big(g(x_1,\cdot)|x_1\big)+h(x_1,y,z)-\underline{P}\big(h(x_1,y,\cdot)|x_1,y\big)\big]$$

is indeed non-negative. Assume therefore that $g(x,\cdot)=0$ for all $x\neq x_0$. Then there are two possibilities left. Either $g(x_0,\cdot)=0$, whence $g=0$. Then we choose $B=\{x_0\}\times\{y_0\}\times\mathcal{Z}$, and it follows from the coherence of the lower prevision $\underline{P}(\cdot|x_0,y_0)$ that for the corresponding supremum;

$$\max_{z\in\mathcal{Z}}\big[h(x_0,y_0,z)-\underline{P}\big(h(x_0,y_0,\cdot)|x_0,y_0\big)$$
$$-\big(h_0(x_0,y_0,z)-\underline{P}\big(h_0(x_0,y_0,\cdot)|x_0,y_0\big)\big)\big]\geqslant 0.$$

Or $g(x_0)\neq 0$ and then we choose $B=\{x_0\}\times\mathcal{Y}\times\mathcal{Z}$, and it follows, in a similar way as in the first step of the proof, that the corresponding supremum

$$\max_{y\in\mathcal{Y},z\in\mathcal{Z}}\big[g(x_0,y)-\underline{P}\big(g(x_0,\cdot)|x_0\big)+h(x_0,y,z)-\underline{P}\big(h(x_0,y,\cdot)|x_0,y\big)$$
$$-I_{\{y_0\}}(y)\big(h_0(x_0,y,z)-\underline{P}\big(h_0(x_0,y,\cdot)|x_0,y\big)\big)\big]$$

is again non-negative. □

Lemma A.5. *Any coherent lower prevision $\underline{Q}'$ on $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$ that has marginal $\underline{P}$ and is jointly coherent with $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$, dominates $\underline{Q}$.*

Proof. Consider any h in $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, then we have to prove that $\underline{Q}'(h)\geqslant\underline{Q}(h)$. Since $\underline{Q}'$ jointly coherent with $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$, it follows that $\underline{Q}'$, $\underline{P}(\cdot|X)$ and $\underline{P}(\cdot|X,Y)$ are weakly coherent (see [37, Section 7.1.4]), and consequently we have for any h_0, h_1 and g in $\mathcal{L}(\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$, and any f in $\mathcal{L}(\mathcal{X}\times\mathcal{Y})$ that

$$\max\big[h_1-\underline{Q}'(h_1)+f-\underline{P}(f|X)+g-\underline{P}(g|X,Y)-\big(h_0-\underline{Q}'(h_0)\big)\big]\geqslant 0.$$

If we choose $h_0=g=h$, $f=\underline{P}(h|X,Y)$ and $h_1=\underline{P}(\underline{P}(h|X,Y)|X)$, this reduces to

$$\underline{Q}'(h)\geqslant\underline{Q}'\big(\underline{P}\big(\underline{P}(h|X,Y)|X\big)\big)$$

and since $\underline{P}(\underline{P}(h|X,Y)|X)$ is a gamble on $\mathcal{X}$, and $\underline{Q}'$ has marginal $\underline{P}$, we find that

$$\underline{Q}'\big(\underline{P}\big(\underline{P}(h|X,Y)|X\big)\big)=\underline{P}\big(\underline{P}\big(\underline{P}(h|X,Y)|X\big)\big)=\underline{Q}(h),$$

whence indeed $\underline{Q}'(h)\geqslant\underline{Q}(h)$. □

Appendix B. Additional discussion of the irrelevance condition (MDI)

This appendix provides additional discussion of the irrelevance assumption (MDI) in Section 4. We use the notations established there. We shall restrict ourselves to the case that the lower prevision $\underline{P}_0$ and the conditional lower prevision $\underline{P}_0(\cdot|X)$ are precise.

It turns out that if we make Assumption (MDI), coherence guarantees that another type of irrelevance is satisfied, as the following theorem makes clear.

Theorem B.1. *Assume we have a linear prevision P_0 on $\mathcal{L}(\mathcal{X})$, and a linear conditional prevision $P_0(\cdot|X)$ on $\mathcal{L}(\mathcal{C}\times\mathcal{X})$. Also assume that the irrelevance condition* (MDI) *holds. Then for all x in $\mathcal{X}$ and c in $\mathcal{C}$ such that $p_0(c,x)=p_0(x)\,p_0(c|x)>0$, and for all gambles f on $\mathcal{O}$, the conditional lower prevision $\underline{P}(f|c,x)$ is uniquely determined by coherence, and given by*

$$\underline{P}(f|c,x)=\underline{P}(f|x)=\min_{o\in\Gamma(x)}f(o).$$

Proof. Let us first consider $\underline{P}(\{c\} \times \{x\} \times \mathcal{O})$. For any $y \in \mathcal{X}$ and $p \in \mathcal{O}$, we have, by separate coherence, that

$$\underline{P}\big(\{c\} \times \{x\} \times \mathcal{O}|y, p\big) = \underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}(\cdot, y, p)|y, p\big) = I_{\{x\}}(y)\underline{P}\big(\{c\}|y, p\big),$$

whence $\underline{P}(\{c\} \times \{x\} \times \mathcal{O}|X, O) = I_{\{x\}}\underline{P}(\{c\}|X, O)$. Consequently, for all $y \in \mathcal{X}$, using separate coherence, Eq. (11) and the irrelevance condition (MDI),

$$\begin{aligned}\underline{P}\big(\underline{P}\big(\{c\} \times \{x\} \times \mathcal{O}|X, O\big)|y\big) &= \underline{P}\big(I_{\{x\}}(y)\underline{P}\big(\{c\}|y, O\big)|y\big)\\ &= I_{\{x\}}(y) \min_{p\in\Gamma(x)} \underline{P}\big(\{c\}|x, p\big)\\ &= I_{\{x\}}(y) \min_{p\in\Gamma(x)} P_0\big(\{c\}|x\big) = I_{\{x\}}(y)p_0(c|x),\end{aligned}$$

whence $\underline{P}(\underline{P}(\{c\} \times \{x\} \times \mathcal{O}|X, O)|X) = I_{\{x\}}p_0(c|x)$, and therefore,

$$\begin{aligned}\underline{P}\big(\{c\} \times \{x\} \times \mathcal{O}\big) &= P_0\big(\underline{P}\big(\underline{P}\big(\{c\} \times \{x\} \times \mathcal{O}|X, O\big)|X\big)\big)\\ &= P_0\big(I_{\{x\}}p_0(c|x)\big) = p_0(x)p_0(c|x) = p_0(c, x).\end{aligned}$$

The material in Section 2.7 then tells us that whenever $\underline{P}(\{c\} \times \{x\} \times \mathcal{O}) = p_0(c, x) > 0$, $\underline{P}(f|c, x)$ is uniquely determined by coherence as the unique solution of the following equation in μ:

$$\underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}[f - \mu]\big) = 0. \tag{B.1}$$

Now, for any $y \in \mathcal{X}$ and $p \in \mathcal{O}$, we have, by separate coherence, that

$$\begin{aligned}\underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}[f - \mu]|y, p\big) &= \underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}(\cdot, y, p)[f - \mu]|y, p\big)\\ &= I_{\{x\}}(y)\underline{P}\big(I_{\{c\}}[f - \mu]|y, p\big),\end{aligned}$$

whence $\underline{P}(I_{\{c\}\times\{x\}\times\mathcal{O}}[f - \mu]|X, O) = I_{\{x\}}\underline{P}(I_{\{c\}}[f - \mu]|X, O)$. Consequently, for all $y \in \mathcal{X}$, using separate coherence, Eq. (11) and the irrelevance assumption (MDI),

$$\begin{aligned}&\underline{P}\big(\underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}[f - \mu]|X, O\big)|y\big)\\ &\quad= \underline{P}\big(I_{\{x\}}(y)\underline{P}\big(I_{\{c\}}[f - \mu]|y, O\big)|y\big) = I_{\{x\}}(y) \min_{o\in\Gamma(x)} \underline{P}\big(I_{\{c\}}\big[f(o) - \mu\big]|x, o\big)\\ &\quad= I_{\{x\}}(y) \min_{o\in\Gamma(x)} P_0\big(I_{\{c\}}\big[f(o) - \mu\big]|x\big) = I_{\{x\}}(y) \min_{o\in\Gamma(x)} \big[f(o) - \mu\big]p_0(c|x)\\ &\quad= I_{\{x\}}(y)p_0(c|x)\big[\min_{o\in\Gamma(x)} f(o) - \mu\big] = I_{\{x\}}(y)p_0(c|x)\big[\underline{P}(f|x) - \mu\big],\end{aligned}$$

whence $\underline{P}(\underline{P}(I_{\{c\}\times\{x\}\times\mathcal{O}}|X, O)|X) = I_{\{x\}}p_0(c|x)[\underline{P}(f|x) - \mu]$, and therefore,

$$\begin{aligned}\underline{P}\big(I\{c\} \times \{x\} \times \mathcal{O}[f - \mu]\big) &= P_0\big(\underline{P}\big(\underline{P}\big(I_{\{c\}\times\{x\}\times\mathcal{O}}[f - \mu]|X, O\big)|X\big)\big)\\ &= P_0\big(I_{\{x\}}p_0(c|x)\big[\underline{P}(f|x) - \mu\big]\big)\\ &= p_0(x)p_0(c|x)\big[\underline{P}(f|x) - \mu\big]\\ &= p_0(c, x)\big[\underline{P}(f|x) - \mu\big].\end{aligned}$$

If $p_0(c, x) > 0$, it follows that the unique solution of Eq. (B.1) is indeed given by $\mu = \underline{P}(f|x)$. □

This theorem tells us that for a linear prior P_0, the irrelevance assumption (MDI) implies, through arguments of coherence, that conditional on the attributes X, the class C is irrelevant to the observations O, i.e., if we know that $X = x$, then the additional knowledge that $C = c$ does not change our beliefs about the value of O.

We now intend to show that the above statement does not imply (MDI). Let us, to this effect, start with a linear prevision P_0 on $\mathscr{L}(\mathcal{C} \times \mathcal{X})$, and assume that for all gambles f on $\mathcal{O}$, and all (c, x) in $\mathcal{C} \times \mathcal{X}$ such that $p_0(c, x) = p_0(x)p_0(c|x) > 0$:

$$\underline{P}(f|c, x) = \underline{P}(f|x) = \min_{o \in \Gamma(x)} f(o). \tag{I′}$$

We can now use Walley's marginal extension theorem (see Theorem 1 in Section 2.9) to combine the marginal linear prevision P_0 on $\mathscr{L}(\mathcal{C} \times \mathcal{X})$ and the conditional lower prevision $\underline{P}(\cdot|C, X)$ on $\mathscr{L}(\mathcal{O})$—or, through separate coherence, on $\mathscr{L}(\mathcal{C} \times \mathcal{X} \times \mathcal{O})$—into a joint lower prevision $\underline{Q}$ on $\mathscr{L}(\mathcal{C} \times \mathcal{X} \times \mathcal{O})$ defined by

$$\underline{Q}(h) = P_0\big(\underline{P}(h|C, X)\big)$$

for all gambles h on $\mathcal{C} \times \mathcal{X} \times \mathcal{O}$. The following theorem tells us that Assumption (MDI) is effectively stronger than Assumption (I′).

Theorem B.2. *Assume that* (I′) *holds. Consider a separately coherent conditional lower prevision* $\underline{P}(\cdot|X, O)$ *on* $\mathscr{L}(\mathcal{C} \times \mathcal{X} \times \mathcal{O})$. *If this conditional lower prevision satisfies* (MDI), *i.e.,*

$$\underline{P}(f|x, o) = P_0(f|x)$$

for all $f \in \mathscr{L}(\mathcal{C})$ *, for all* $x \in \mathcal{X}$ *such that* $p_0(x) > 0$, *and for all* $o \in \Gamma(x)$, *then it cannot be jointly coherent with the joint lower prevision* $\underline{Q}$ *on* $\mathscr{L}(\mathcal{C} \times \mathcal{X} \times \mathcal{O})$.

Proof. Let $x \in \mathcal{X}$ such that $p_0(x) > 0$ and let $o \in \Gamma(x)$. Consider an arbitrary gamble f on $\mathcal{C}$ that is not almost everywhere constant on $\mathcal{C}$ with respect to the linear prevision $P_0(\cdot|x)$ (which is uniquely determined from P_0 through coherence). The theorem is proved if we can show that

$$\underline{Q}\big(\big[f - \underline{P}(f|x, o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}\big) < 0.$$

By separate coherence and Assumption (I′), we find for any $c \in \mathcal{C}$ and $y \in \mathcal{X}$ that

$$\begin{aligned}
&\underline{P}\big(\big(\big[f - \underline{P}(f|x, o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}\big)|c, y\big)\\
&\quad = \underline{P}\big(\big(\big[f(c) - \underline{P}(f|x, o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}(c, y, \cdot)\big)|c, y\big)\\
&\quad = I_{\{x\}}(y) \min_{p \in \Gamma(x)} \big[f(c) - \underline{P}(f|x, o)\big]I_{\{o\}}(p)\\
&\quad = I_{\{x\}}(y)I_{\{o\}^*}(x) \min\big\{f(c) - \underline{P}(f|x, o), 0\big\}\\
&\quad = I_{\{x\}}(y)I_{\{o\}^*}(x) \min\big\{f(c) - P_0(f|x), 0\big\}
\end{aligned}$$

where the last equality follows from the assumptions of the theorem. Consequently,

$$\underline{P}\big(\big(\big[f - \underline{P}(f|x, o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}\big)|C, X\big) = I_{\{x\}}I_{\{o\}^*}(x) \min\big\{f - P_0(f|x), 0\big\}$$

and we find that

$$\begin{aligned}\underline{Q}\big(\big[f-\underline{P}(f|x,o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}\big) &= P_0\Big(\big(\underline{P}\big(\big[f-\underline{P}(f|x,o)\big]I_{\mathcal{C}\times\{x\}\times\{o\}}\big)|C,X\big)\Big)\\ &= P_0\big(I_{\{x\}}I_{\{o\}^*}(x)\min\big\{f-P_0(f|x),0\big\}\big)\\ &= I_{\{o\}^*}(x)P_0\big(I_{\{x\}}\min\big\{f-P_0(f|x),0\big\}\big)\\ &= I_{\{o\}^*}(x)p_0(x)P_0\big(\min\big\{f-P_0(f|x),0\big\}|x\big)<0,\end{aligned}$$

where the inequality follows from $x\in\{o\}^*$, $p_0(x)>0$, and Lemma B.3. □

Lemma B.3. *Let P be a linear prevision on $\mathcal{L}(\mathcal{C})$. Then for all gambles f on $\mathcal{C}$ that are not almost everywhere constant* (*with respect to P*)*, and for all real μ, we have that*

$$P\big(\min\{f-\mu,0\}\big)\geqslant 0 \quad\Rightarrow\quad \mu<P(f).$$

Proof. Let f be a gamble that is not constant almost everywhere, i.e., f is not constant on the set $D_p=\{c\in\mathcal{C}\colon\ p(c)>0\}$, where we denote by p the mass function of P. It clearly suffices to show that $P(\min\{f-P(f),0\})<0$. Assume, *ex absurdo*, that $P(\min\{f-P(f),0\})\geqslant 0$. Since the gamble $\min\{f-P(f),0\}$ on $\mathcal{C}$ is non-positive, this implies that $P(\min\{f-P(f),0\})=0$, and this can only happen if $p(c)=P(\{c\})=0$ for all $c\in\mathcal{C}$ such that $f(c)<P(f)$. Consequently, $P(f)\leqslant f(c)$ for all $c\in D_p$, whence $P(f)\leqslant\min_{c\in D_p}f(c)$. But since $P(f)$ is a non-trivial convex mixture of the $f(c)$ for all $c\in D_p$, and since f is not constant on D_p, we also know that $P(f)>\min_{c\in D_p}f(c)$, a contradiction. □

References

[1] A. Charnes, W.W. Cooper, Programming with linear fractional functionals, Naval Research Logistic Quarterly 9 (1962) 181–186.

[2] F.G. Cozman, Credal networks, Artificial Intelligence 120 (2000) 199–233.

[3] F.G. Cozman, Separation properties of sets of probabilities, in: C. Boutilier, M. Goldszmidt (Eds.), Uncertainty in Artificial Intelligence (Proceedings of the Sixteenth Conference), Morgan Kaufmann, San Francisco, CA, 2000, pp. 107–115.

[4] L.M. de Campos, J.F. Huete, S. Moral, Probability intervals: a tool for uncertain reasoning, Internat. J. Uncertainty, Fuzziness and Knowledge-Based Systems 2 (1994) 167–196.

[5] L.M. de Campos, M.T. Lamata, S. Moral, The concept of conditional fuzzy measures, Internat. J. Intelligent Systems 5 (1990) 237–246.

[6] G. de Cooman, M. Zaffalon, Updating with incomplete observations, in: C. Meek, U. Kjærulff (Eds.), Uncertainty in Artificial Intelligence (Proceedings of the Nineteenth Conference), Morgan Kaufmann, San Francisco, CA, 2003, pp. 142–150.

[7] B. de Finetti, La prévision: ses lois logiques, ses sources subjectives, Ann. Inst. H. Poincaré 7 (1937) 1–68. English translation in [23].

[8] B. de Finetti, Teoria delle Probabilità, Einaudi, Torino, 1970.

[9] B. de Finetti, Theory of Probability, Wiley, Chichester, 1974–1975, English Translation of [8], two volumes.

[10] R. Fagin, J.Y. Halpern, A new approach to updating beliefs, in: P.P. Bonissone, M. Henrion, L.N. Kanal, J.F. Lemmer (Eds.), Uncertainty in Artificial Intelligence, vol. 6, North-Holland, Amsterdam, 1991, pp. 347–374.

[11] E. Fagiuoli, M. Zaffalon, 2U: an exact interval propagation algorithm for polytrees with binary variables, Artificial Intelligence 106 (1998) 77–107.

[12] J.C. Ferreira da Rocha, F.G. Cozman, Inference with separately specified sets of probabilities in credal networks, in: A. Darwiche, N. Friedman (Eds.), Uncertainty in Artificial Intelligence (Proceedings of the Eighteenth Conference), Morgan Kaufmann, San Francisco, CA, 2002, pp. 430–437.

[13] R. Gill, M. Van der Laan, J. Robins, Coarsening at random: characterisations, conjectures and counter-examples, in: D.-Y. Lin (Ed.), Proceedings of the First Seattle Conference on Biostatistics, Springer, Berlin, 1997, pp. 255–294.
[14] P.D. Grünwald, J.Y. Halpern, Updating probabilities, J. Artificial Intelligence Res. (2003) 243–278.
[15] J.Y. Halpern, A logical approach to reasoning about uncertainty: a tutorial, in: X. Arrazola, K. Korta, F.J. Pelletier (Eds.), Discourse, Interaction, and Communication, Kluwer, Dordrecht, 1998, pp. 141–155.
[16] J.Y. Halpern, M. Tuttle, Knowledge, probability, and adversaries, J. ACM 40 (4) (1993) 917–962.
[17] T. Herron, T. Seidenfeld, L. Wasserman, Divisive conditioning: further results on dilation, Philos. Sci. 64 (1997) 411–444.
[18] J.-Y. Jaffray, Bayesian updating and belief functions, IEEE Trans. Systems Man Cybernet. 22 (1992) 1144–1152.
[19] L.G. Khachian, A polynomial algorithm for linear programming, Soviet Math. Dokl. 20 (1979) 191–194. English translation of [20].
[20] L.G. Khachian, A polynomial algorithm for linear programming, Dokl. Akad. Nauk SSSR 244 (1979) 1093–1096. In Russian.
[21] A.N. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin, 1933.
[22] A.N. Kolmogorov, Foundations of Probability, Chelsea, New York, 1950. English translation of [21].
[23] H.E. Kyburg Jr., H.E. Smokler (Eds.), Studies in Subjective Probability, Wiley, New York, 1964, Second edition (with new material) 1980.
[24] I. Levi, The Enterprise of Knowledge, MIT Press, London, 1980.
[25] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 1987.
[26] C. Manski, Partial Identification of Probability Distributions, Springer, Berlin, 2003.
[27] E. Miranda, G. de Cooman, I. Couso, Imprecise probabilities induced by multi-valued mappings, in: Proceedings of the Ninth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002, Annecy, France, July 1–5, 2002), Gutenberg, 2002, pp. 1061–1068.
[28] S. Moral, A. Cano, Strong conditional independence for credal sets, Ann. Math. Artificial Intelligence 35 (1–4) (2002) 295–321.
[29] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Mateo, CA, 1988.
[30] M.A. Peot, R.D. Shachter, Learning from what you don't observe, in: G.F. Cooper, S. Moral (Eds.), Uncertainty in Artificial Intelligence (Proceedings of the Fourteenth Conference), Morgan Kaufmann, San Francisco, CA, 1998, pp. 439–446.
[31] M. Ramoni, P. Sebastiani, Robust learning with missing data, Machine Learning 45 (2) (2001) 147–170.
[32] T. Seidenfeld, L. Wasserman, Dilation for sets of probabilities, Ann. Statist. 21 (1993) 1139–1154.
[33] G. Shafer, Conditional probability, Internat. Statist. Rev. 53 (1985) 261–277.
[34] C.A.B. Smith, Consistency in statistical inference and decision, J. Roy. Statist. Soc. Ser. A 23 (1961) 1–37.
[35] V. Strassen, Meßfehler und Information, Z. Wahr. Verw. Geb. 2 (1964) 273–305.
[36] P. Walley, Coherent lower (and upper) probabilities, Technical Report, University of Warwick, Coventry, 1981. Statistics Research Report 22.
[37] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.
[38] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. Roy. Statist. Soc. Ser. B 58 (1996) 3–57. With discussion.
[39] P. Walley, Measures of uncertainty in expert systems, Artificial Intelligence 83 (1996) 1–58.
[40] M. Zaffalon, Exact credal treatment of missing data, J. Statistical Planning and Inference 105 (1) (2002) 105–122.
[41] M. Zaffalon, The naive credal classifier, J. Statistical Planning and Inference 105 (2002) 5–21.