EPISTEMIC IRRELEVANCE IN CREDAL NETS: THE CASE OF IMPRECISE MARKOV TREES

GERT DE COOMAN, FILIP HERMANS, ALESSANDRO ANTONUCCI, AND MARCO ZAFFALON

ABSTRACT. We focus on credal nets, which are graphical models that generalise Bayesian nets to imprecise probability. We replace the notion of strong independence commonly used in credal nets with the weaker notion of epistemic irrelevance, which is arguably more suited for a behavioural theory of probability. Focusing on directed trees, we show how to combine the given local uncertainty models in the nodes of the graph into a global model, and we use this to construct and justify an exact message-passing algorithm that computes updated beliefs for a variable in the tree. The algorithm, which is linear in the number of nodes, is formulated entirely in terms of coherent lower previsions, and is shown to satisfy a number of rationality requirements. We supply examples of the algorithm's operation, and report an application to on-line character recognition that illustrates the advantages of our approach for prediction. We comment on the perspectives, opened by the availability, for the first time, of a truly efficient algorithm based on epistemic irrelevance.

1. INTRODUCTION

The last twenty years have witnessed a rapid growth of *graphical models* in the fields of artificial intelligence and statistics. These models combine graphs and probability to address complex multivariate problems in a variety of domains, such as medicine, finance, risk analysis, defence, and environment, to name just a few.

Much has been done also on the front of imprecise probability. In particular, *credal nets* [4] have been and still are the subject of intense research. A credal net creates a global model of a domain by combining local uncertainty models using some notion of independence, and then uses this to do inference. The local models represent uncertainty by closed convex sets of probabilities, also called *credal sets*.

The notion of independence used with credal nets in the vast majority of cases is that of *strong independence* (with some exceptions in [8]). Loosely speaking, two variables X, Y are strongly independent if the credal set for (X, Y) can be regarded as originating from a number of precise models in each of which X and Y are stochastically independent. Strong independence is closely related to the *sensitivity analysis* interpretation of credal sets, which regards an imprecise model as arising out of partial ignorance of a precise one.

In the particular case of credal nets, strong independence leads to a mathematical equivalence: a credal net model is equivalent to a model consisting of a set of Bayesian nets, each with the same graph but with different values for the parameters. The sensitivity analysis interpretation is then that there is some (kind of ideal) Bayesian net model of the problem under consideration, and the graph of such a net is known. But, for some reason, the net's parameters are not known precisely, and that is why one considers the set of all the Bayesian nets that are consistent with the partial specification of the parameters. Common causes for the existence of partial knowledge are the cost of, and time constraints on, eliciting parameters, and disagreement amongst a group of experts consulted for that purpose. Non-ignorable missing data can be another reason, in case the parameters are inferred from a data set [29].

The sensitivity analysis interpretation of imprecise-probability models, and hence strong independence, is not always applicable. A notable case arises when one wishes to model an

Key words and phrases. Coherence, credal net, epistemic irrelevance, epistemic independence, strong independence, imprecise Markov tree, separation, hidden Markov model.

expert's beliefs: it is then not always tenable that there should be some ideal Bayesian net that models these beliefs, and that it is only because of our limited resources that we cannot define it precisely. Rather, it seems more reasonable to concede that expert knowledge may be *inherently* imprecise to some extent.[1] This simple observation makes the sensitivity analysis interpretation fail, and hence it makes strong independence an inadequate model, in general, for such a situation.[2]

An alternative and attractive approach to expressing irrelevance that is not committed to the sensitivity analysis interpretation is offered by *epistemic irrelevance* [24]: we say that X is epistemically irrelevant to Y if observing X does not affect our beliefs about Y. In other words, by making an epistemic irrelevance assessment, a subject states that her belief model about Y does (or will) not change after receiving information about X. When the belief model is a precise probability, both epistemic irrelevance and strong independence reduce to the usual (stochastic) independence.[3] But when the model is a set of probabilities, this is no longer the case, because in contradistinction with strong independence, epistemic irrelevance is a property *of this set* that cannot be explained using properties of the precise probabilities in the set. Epistemic irrelevance is defined directly in terms of a subject's belief model (the set of probabilities). For this reason, it is very well suited for a behavioural theory of imprecise probability. Contrary to strong independence, it is not a symmetrical notion: generally speaking, the epistemic irrelevance of X to Y does not entail the epistemic irrelevance of Y to X. It is also weaker than strong independence, in the sense that strong independence implies epistemic irrelevance: sets of probabilities that correspond to assessments of epistemic irrelevance usually include those related to strong independence assessments. It therefore does not lead to overconfident inferences when the sensitivity analysis interpretation is not justified.

At this point, the question we address in this paper should be clear: can we define credal nets based on epistemic irrelevance, and moreover create an exact algorithm to perform efficient inferences with them? We give a fully positive answer to this question in the special case that (i) the graph under consideration is a directed tree, and (ii) the related variables assume finitely many values. The intuitions that showed us the way towards this result originated in previous work done by some of us on imprecise probability trees [9] and imprecise Markov chains [10].

How do we address this problem?

In Section 2, we discuss some preliminary graph-theoretic notions, and define the local uncertainty models that will be used at each node of a tree. These models are formalised through the language of *coherent lower previsions* [24]. We discuss how such local models will give rise to a global uncertainty model, which plays the same role as the joint mass function built by the chain rule in a Bayesian net. Based on the global model, we state the Markov condition that defines the imprecise-probability interpretation of our credal trees. As announced before, this Markov condition involves epistemic irrelevance rather than strong independence.

In Section 3, we take a brief detour to discuss in general terms how to combine marginal models into joint ones using irrelevance assessments, in a way that is as conservative as possible. We do so because the notion of so-called *epistemic independence*, which arises out of a symmetrisation of epistemic irrelevance, has so far been defined in the literature only for the case of two variables. We define and discuss the *independent natural extension* of a number of marginals. This is the most conservative joint model that arises out of the marginals and epistemic independence alone. Moreover, we show that the independent

[1] For a detailed argumentation and exposition of this point of view, we refer to [24, Chapter 5].

[2] Obviously, there will be special cases where strong independence is justified in order to model an expert's knowledge. Moreover, strong independence could provide a good approximation to more accurate models, even when it is not entirely appropriate. This is something that seems to deserve further investigation.

[3] If we ignore issues related to events with probability zero.

natural extension has a very important *strong factorisation* property, which has a crucial part in our algorithm for updating credal trees under epistemic irrelevance.

In Section 4, we turn to the problem of constructing the most conservative global model based only on the local models in the tree and our Markov condition. We show that this task can be achieved by a recursive construction that proceeds from the leaves to the root of the tree using two operations: the *independent natural extension* discussed in Section 3, and the *marginal extension*, defined and studied in [24, 17]. We also show that all uncertainty models we consider, the local ones as well as the global ones that we create, satisfy a consistency criterion that generalises (and is based on the same ideas as) the usual consistency criterion in Bayesian nets: they are (separately and jointly) *coherent* [24, 15, 16, 27] (see in particular [18, Section 8.1]). This is an important rationality requirement.

We briefly comment on some of the graphical separation criteria induced by epistemic irrelevance in Section 5. We then go on to develop and justify an algorithm for making inferences on credal trees under epistemic irrelevance in Section 6. The algorithm is used to *update* the tree: it computes posterior beliefs about a *target* variable in the tree conditional on the observation of other variables, which are called *instantiated*, meaning that their value is determined. It can in particular be used for treating the model as an expert system.

Our algorithm is based on message passing, as are the traditional algorithms that have been developed for precise graphical models. It has some remarkable properties: (i) it works in time linear in the number of nodes in the tree; (ii) it natively computes posterior lower and upper *previsions* (or expectations) rather than probabilities; (iii) it is the first algorithm developed for credal nets that exclusively uses the formalism of coherent lower previsions; and (iv) it is shown that, under very mild conditions, using the tree for updating beliefs cannot lead to inferences that are inconsistent with the local models we have started from, nor with one another.

We give a step-by-step example of the way inferences can be done using our algorithm in Section 7. We also comment there on the intriguing relationship between the failure of certain classical separation properties in our framework, and dilation [14, 22].

The last part of the paper focuses on numerical simulations. In Section 8 we empirically measure the amount of imprecision introduced by using epistemic irrelevance rather than strong independence in a credal tree, when propagating inferences backwards (towards the root) from instantiated nodes to the target node. Indeed, it can be shown [9] that there is *no difference* between inferences that go forward from instantiated nodes to the target node under strong independence and epistemic irrelevance. In Section 9 we present an application of our algorithm to on-line character recognition. We learn the probabilities from data and compare the predictions of our approach with those of its precise probability counterpart. The results are encouraging: they show that the tree can be used for real applications, and that the imprecision it originates is justified.

In order to keep this paper reasonably short, we have to assume the reader has a good working knowledge of the basics of Peter Walley's [24] theory of coherent lower previsions. This is needed in particular for the most important proofs, collected in Appendix A. For a fairly detailed discussion of the coherence notions and results needed in the context of this paper, we refer to recent work by Enrique Miranda [15, 16].

2. Credal trees under epistemic irrelevance

2.1. **Basic notions and notation.** We consider a rooted and directed discrete tree with finite width and depth. We call T the set of its nodes s, and we denote the *root*, or initial, node by $\square$. For any node s, we denote its *mother node* by $m(s)$. Of course, $\square$ has no mother node, and we use the convention $m(\square) = \emptyset$. Also, for each node s, we denote the set of its *children* by $C(s)$, and the set of its *siblings* by $S(s)$. Clearly, $S(\square) = \emptyset$, and if $s \neq \square$ then $S(s) = C(m(s)) \setminus \{s\}$. If $C(s) = \emptyset$, then we call s a *leaf*, or *terminal node*. We denote by $T^{\Diamond} := \{s \in T : C(s) \neq \emptyset\}$ the set of all non-terminal nodes.

For nodes s and t, we write $s \sqsubseteq t$ if *s precedes t*, i.e., if there is a directed segment in the tree from s to t. The relation $\sqsubseteq$ is a special partial order on the set T. $A(s) := \{t \in T : t \sqsubset s\}$ denotes the chain of *ancestors* of s, and $D(s) := \{t \in T : s \sqsubset t\}$ its set of *descendants*. Here $s \sqsubset t$ means that $s \sqsubseteq t$ and $s \neq t$. We also use the notation $\uparrow s := A(s) \cup \{s\}$ for the chain (segment) connecting $\square$ and s, and $\downarrow s := D(s) \cup \{s\}$ for the sub-tree with root s. Similarly, we let $\uparrow S := \bigcup\{\uparrow s : s \in S\}$ and $\downarrow S := \bigcup\{\downarrow s : s \in S\}$ for any subset $S \subseteq T$. For any node s, its set of non-parent non-descendants is given by $\overline{s} := T \setminus (\{m(s)\} \cup \downarrow s)$.

With each node s of the tree, there is associated a variable X_s assuming values in a non-empty finite set $\mathscr{X}_s$. We denote by $\mathscr{L}(\mathscr{X}_s)$ the set of all real-valued maps (also called *gambles*) on $\mathscr{X}_s$. We extend this notation to more complicated situations as follows. If S is any subset of T, then we denote by X_S the tuple of variables whose components are the X_s for all $s \in S$. This new joint variable assumes values in the finite set $\mathscr{X}_S := \times_{s \in S} \mathscr{X}_s$, and the corresponding set of gambles is denoted by $\mathscr{L}(\mathscr{X}_S)$. [4] Generic elements of $\mathscr{X}_s$ are denoted by x_s or z_s. Similarly for x_S and z_S in $\mathscr{X}_S$. Also, if we mention a tuple z_S, then for any $t \in S$, the corresponding element in the tuple will be denoted by z_t. We assume all variables in the tree to be logically independent, meaning that the variable X_S may assume *all* values in $\mathscr{X}_S$, for all $\emptyset \subseteq S \subseteq T$.

We will frequently use the simplifying device of identifying a gamble f_S on $\mathscr{X}_S$ with its *cylindrical extension* to $\mathscr{X}_U$, where $S \subseteq U \subseteq T$. This is the gamble f_U on $\mathscr{X}_U$ defined by $f_U(x_U) := f_S(x_S)$ for all $x_U \in \mathscr{X}_U$. To give an example, if $\mathscr{K} \subseteq \mathscr{L}(\mathscr{X}_T)$, this trick allows us to consider $\mathscr{K} \cap \mathscr{L}(\mathscr{X}_S)$ as the set of those gambles in $\mathscr{K}$ that depend only on the variable X_S. As another example, this device allows us to identify the gambles $\mathbb{I}_{\{x_S\}}$ and $\mathbb{I}_{\{x_S\} \times \mathscr{X}_{T \setminus S}}$, and therefore also the events $\{x_S\}$ and $\{x_S\} \times \mathscr{X}_{T \setminus S}$. More generally, for any event $A \subseteq \mathscr{X}_S$, we can identify the gambles $\mathbb{I}_A$ and $\mathbb{I}_{A \times \mathscr{X}_{T \setminus S}}$, and therefore also the events A and $A \times \mathscr{X}_{T \setminus S}$. In the same spirit, a lower prevision on all gambles in $\mathscr{L}(\mathscr{X}_S)$ can be identified with a lower prevision defined on the set of corresponding gambles on $\mathscr{X}_T$, a subset of $\mathscr{L}(\mathscr{X}_T)$.

Throughout the paper, we consider (conditional) lower previsions as models for a subject's beliefs about the values that certain variables in the tree may assume. We use a systematic notation for such (conditional) lower previsions. Let $I, O \subseteq T$ be *disjoint* sets of nodes with $O \neq \emptyset$, then we generically[5] denote by $\underline{V}_O(\cdot|X_I)$ a *conditional lower prevision*, defined on the set of gambles $\mathscr{L}(\mathscr{X}_{I \cup O})$.[6] For every gamble f on $\mathscr{X}_{I \cup O}$ and every $x_I \in \mathscr{X}_I$, $\underline{V}_O(f|x_I)$ is the lower prevision (or lower expectation, or a subject's supremum buying price) for/of the gamble f, conditional on the event that $X_I = x_I$. We interpret $\underline{V}_O(f|X_I)$ as a real-valued map (gamble) on $\mathscr{X}_I$ that assumes the value $\underline{V}_O(f|x_I)$ in the element x_I of $\mathscr{X}_I$. The conjugate *conditional upper prevision* $\overline{V}_O(\cdot|X_I)$ is defined on $\mathscr{L}(\mathscr{X}_{I \cup O})$ by $\overline{V}_O(f|X_I) := -\underline{V}_O(-f|X_I)$ for all gambles f on $\mathscr{X}_{I \cup O}$.

We will always implicitly assume that all conditional models $\underline{V}_O(\cdot|X_I)$ we use are *separately coherent*, meaning that:

SC1. $\underline{V}_O(f|x_I) \geq \min_{z_O \in \mathscr{X}_O} f(x_I, z_O)$ for all $f \in \mathscr{L}(\mathscr{X}_{I \cup O})$ and all $x_I \in \mathscr{X}_I$ [accepting partial gains];

SC2. $\underline{V}_O(f_1 + f_2|x_I) \geq \underline{V}_O(f_1|x_I) + \underline{V}_O(f_2|x_I)$ for all $f_1, f_2 \in \mathscr{L}(\mathscr{X}_{I \cup O})$ and all $x_I \in \mathscr{X}_I$ [super-additivity];

---

[4] For any subset S of T, $\mathscr{X}_S$ is defined formally as the set of all maps x_S of S to $\bigcup_{s \in S} \mathscr{X}_s$, such that $x_S(s) = x_s \in \mathscr{X}_s$ for all $s \in S$. So when $S = \emptyset$, the empty product $\mathscr{X}_\emptyset$ is defined as the set of all maps from $\emptyset$ to $\emptyset$, which is a singleton. The corresponding variable $X_\emptyset$ can then only assume this single value, so there is no uncertainty about it. $\mathscr{L}(\mathscr{X}_\emptyset)$ can be identified with the set $\mathbb{R}$ of real numbers.

[5] Besides the letter V, we will also use the letters P, Q and R.

[6] In keeping with the observation in footnote 4, we also allow $I = \emptyset$, which means conditioning on the variable $X_I = X_\emptyset$, which can only assume one single value. This means that $\underline{V}_O(\cdot|X_\emptyset) =: \underline{V}_O$ effectively becomes an unconditional lower prevision on $\mathscr{L}(\mathscr{X}_{O \cup \emptyset}) = \mathscr{L}(\mathscr{X}_O)$. This a very useful device that allows us to use the same generic notation for both conditional and unconditional lower previsions.

SC3. $\underline{V}_O(\lambda f|x_I) = \lambda \underline{V}_O(f|x_I)$ for all $f \in \mathscr{L}(\mathscr{X}_{I\cup O})$, all non-negative real λ and all $x_I \in \mathscr{X}_I$ [non-negative homogeneity].

By combining SC1–SC3, it follows that for all $f \in \mathscr{L}(\mathscr{X}_{I\cup O})$, $x_I \in \mathscr{X}_I$ and $z_O \in \mathscr{X}_O$:

$$\min_{z_O\in\mathscr{X}_O} f(x_I,z_O) \leq \underline{V}_O(f|x_I) \leq \overline{V}_O(f|x_I) \leq \max_{z_O\in\mathscr{X}_O} f(x_I,z_O).$$

If we let f be the indicator $\mathbb{I}_{\{z_I\}}$ of the set $\{z_I\}$ in these inequalities, they reduce to the following, intuitively obvious, property:[7]

SC4. $\underline{V}_O(\{z_I\}\times\mathscr{X}_O|x_I) = \overline{V}_O(\{z_I\}\times\mathscr{X}_O|x_I) = \mathbb{I}_{\{x_I\}}(z_I)$ for all $x_I, z_I \in \mathscr{X}_I$.

From SC1, SC2 and SC4 we can also derive that, with obvious notations:

$$\underline{V}_O(f|x_I) = \underline{V}_O(\mathbb{I}_{\{x_I\}}f|x_I) = \underline{V}_O(f(x_I,\cdot)|x_I) \text{ for all gambles } f \text{ on } \mathscr{X}_{I\cup O} \text{ and } x_I \in \mathscr{X}_I, \quad (1)$$

where $f(x_I,\cdot)$ is a partial map defined on $\mathscr{X}_O$. This implies that $\underline{V}_O(\cdot|X_I)$ is completely determined by its behaviour on (cylindrical extensions of maps in) $\mathscr{L}(\mathscr{X}_O)$.

Hereafter, we will frequently introduce conditional lower previsions of the type $\underline{V}_O(\cdot|X_I)$ as if they are defined on $\mathscr{L}(\mathscr{X}_O)$, simply because that is a very natural thing to do: such a conditional lower prevision is usually interpreted as representing beliefs about the variable X_O, conditional on values of the variable X_I. But the reader should keep in mind that, by the separate coherence property (1), $\underline{V}_O(\cdot|X_I)$ can (and should) always be uniquely extended to the larger domain $\mathscr{L}(\mathscr{X}_{I\cup O})$.

As soon as we consider a number of such conditional lower previsions $\underline{V}_{O_k}(\cdot|X_{I_k})$, $k = 1,\ldots,n$, they should satisfy more stringent consistency criteria than that each of them should be separately coherent: they should also be consistent with one another in the sense of Walley's *(joint) coherence* [24, Section 7.1.4(b)]. For more details about this much more involved type of coherence, we refer also to [15, 16].

Finally, let us introduce one of the most important concepts for this paper, that of epistemic irrelevance. We describe the case of conditional irrelevance, as the unconditional version of epistemic irrelevance can easily be recovered as a special case.[8]

Consider three disjoint subsets C, I, and O of N, where both I and O are non-empty. When a subject judges X_I to be *epistemically irrelevant to X_O conditional on X_C*, he assumes that if he knows the value of X_C, then learning in addition which value X_I assumes in $\mathscr{X}_I$ will not affect his beliefs about X_O. More formally, assume that a subject has a separately coherent conditional lower prevision $\underline{V}_O(\cdot|X_C)$ on $\mathscr{L}(\mathscr{X}_O)$. If he assesses X_I to be epistemically irrelevant to X_O conditional on X_C, this implies that he can infer from his model $\underline{V}_O(\cdot|X_C)$ a conditional model $\underline{V}_O(\cdot|X_{C\cup I})$ on $\mathscr{L}(\mathscr{X}_O)$ given by

$$\underline{V}_O(f|x_{C\cup I}) := \underline{V}_O(f|x_C) \text{ for all } f \in \mathscr{L}(\mathscr{X}_O) \text{ and all } x_{C\cup I} \in \mathscr{X}_{C\cup I}.$$

2.2. **Local uncertainty models.** We now add a *local uncertainty model* to each of the nodes s. If s is not the root node, i.e. has a mother $m(s)$, then this local model is a (separately coherent) conditional lower prevision $\underline{Q}_s(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_s)$: for each possible value $z_{m(s)}$ of the variable $X_{m(s)}$ associated with its mother $m(s)$, we have a coherent lower prevision $\underline{Q}_s(\cdot|z_{m(s)})$ for the value of X_s, conditional on $X_{m(s)} = z_{m(s)}$. In the root, we have an unconditional local uncertainty model $\underline{Q}_\square$ for the value of $X_\square$. $\underline{Q}_\square$ is a (separately) coherent lower prevision on $\mathscr{L}(\mathscr{X}_\square)$. We use the common generic notation $\underline{Q}_s(\cdot|X_{m(s)})$ for all these local models.[9]

---

[7] For any event $A \subseteq \mathscr{X}_{I\cup O}$, we denote $\underline{V}_O(\mathbb{I}_A|x_I)$ also as $\underline{V}_O(A|x_I)$ and call this real number the (conditional) lower *probability* of A. Similarly $\overline{V}_O(A|x_I) := \overline{V}_O(\mathbb{I}_A|x_I)$ is the (conditional) upper *probability* of A.

[8] It suffices, in the discussion below, to let $C = \emptyset$. As we indicated in footnote 4, this makes sure the variable X_C has only one possible value, so conditioning on that variable amounts to not conditioning at all.

[9] We can do this because $X_{m(\square)} = X_\emptyset$ has only one possible value, so conditioning on that variable amounts to not conditioning at all.

2.3. **Global uncertainty models.** We intend to show in Section 4 how all these local models $\underline{Q}_s(\cdot|X_{m(s)})$ can be combined into *global uncertainty models*. We generically denote such global models using the letter P. More specifically, we want to end up with an unconditional joint lower prevision $\underline{P} := \underline{P}_{\downarrow\square} = \underline{P}_T$ on $\mathscr{L}(\mathscr{X}_T)$ for all variables in the tree, as well as conditional lower previsions $\underline{P}_{\downarrow S}(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow S})$ for all non-terminal nodes s and all non-empty $S \subseteq C(s)$.

Ideally, we want these global (conditional) lower previsions (i) to be compatible with the local assessments $\underline{Q}_s(\cdot|X_{m(s)})$, $s \in T$, (ii) to be coherent with one another, and (iii) to reflect the conditional irrelevancies (or Markov-type conditions) that we want the graphical structure of the tree to encode. In addition, we want them (iv) to be as conservative (small) as possible.

In this list, the only item that needs more explanation concerns the Markov-type conditions that the tree structure encodes. This is what we turn to now.

2.4. **The interpretation of the graphical model.** In classical Bayesian nets, the graphical structure is taken to represent the following assessments: for any node s, conditional on its parent variables, its non-parent non-descendant variables are epistemically irrelevant to it (and therefore also independent).

In the present context, we assume that the tree structure embodies the following conditional irrelevance assessment, which turns out to be equivalent with the conditional independence assessment above in the special case of a Bayesian tree.

CI. Consider any node s in the tree, any subset S of its set of children $C(s)$, and the set $\overline{S} := \bigcap_{c\in S}\overline{c}$ of their common non-parent non-descendants. Then *conditional on the mother variable X_s, the non-parent non-descendant variables $X_{\overline{S}}$ are assumed to be epistemically irrelevant to the variables $X_{\downarrow S}$ associated with the children in S and their descendants.*

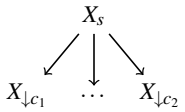
This interpretation turns the tree into a *credal tree under epistemic irrelevance*, and we also introduce the term *imprecise Markov tree* (IMT) for it. For the global models we are considering here, CI has the following consequences. It implies that for all $s \in T^{\Diamond}$, all non-empty $S \subseteq C(s)$ and all $I \subseteq \overline{S}$, we can infer from $\underline{P}_{\downarrow S}(\cdot|X_s)$ a model $\underline{P}_{\downarrow S}(\cdot|X_{\{s\}\cup I})$, where for all $z_{\{s\}\cup I} \in \mathscr{X}_{\{s\}\cup I}$, with obvious notations:[10]

$$\underline{P}_{\downarrow S}(f|z_{\{s\}\cup I}) := \underline{P}_{\downarrow S}(f(\cdot,z_I)|z_s) \text{ for all gambles } f \text{ in } \mathscr{L}(\mathscr{X}_{\downarrow S\cup I}), \tag{2}$$

where $f(\cdot,z_I)$ denotes a partial map of f defined on $\mathscr{X}_{\downarrow S}$.

We discuss some of the separation properties that accompany this interpretation in Section 5. For now, we focus on two immediate consequences that will help us go from local to global models in Section 4.

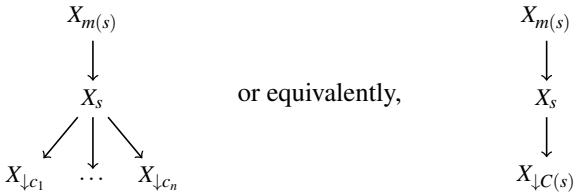
First, consider some node s. Then CI tells us that for any two children $c_1, c_2 \in C(s)$ of s, the variable $X_{\downarrow c_1}$ is epistemically irrelevant to the variable $X_{\downarrow c_2}$, conditional on X_s.



It even tells us that for any two disjoint non-empty sets $S_1 \subseteq C(s)$ and $S_2 \subseteq C(s)$ of children of s, the variable $X_{\downarrow S_1}$ is epistemically irrelevant to $X_{\downarrow S_2}$, conditional on X_s. We conclude that, conditional on a node, all its children c (and the variables associated with their sub-trees $\downarrow c$) are *epistemically independent* [24, Chapter 9], in the specific sense to be discussed in the next section.

Next, consider some non-terminal node s different from $\square$, and its mother variable $X_{m(s)}$. We infer from CI that this mother variable $X_{m(s)}$ is epistemically irrelevant to the variable $X_{\downarrow C(s)}$ conditional on X_s:

[10] For leaves s, the corresponding irrelevance condition is trivial, as the set $C(s)$ of children of s is empty.



3. Independent natural extension

Let us make a small digression on epistemic independence, which will help us in our discussion further on. The material in this section is based on work that some of us have published elsewhere [11], and we refer to that paper for more details and proofs for the results mentioned in this section.

3.1. **Independent products.**

Suppose we have a number of (separately) coherent marginal lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$ representing beliefs about the values that each of a finite number of (logically independent) variables X_n assume in the respective non-empty finite sets $\mathscr{X}_n$, $n \in N$, where N is some non-empty finite set.

We want to construct a joint lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$, where $\mathscr{X}_N = \times_{n\in N}\mathscr{X}_n$, that coincides with the marginals $\underline{P}_n$ on their respective domains $\mathscr{L}(\mathscr{X}_n)$, and such that this $\underline{P}_N$ reflects the following structural assessments: for any disjoint proper subsets O and I of N, the variables X_I are epistemically irrelevant to the variables X_O. In other words, learning the value of any number of these variables would not affect beliefs about the remaining variables. We then call the variables X_n, $n \in N$, *epistemically independent*.

Generally speaking, such irrelevance assessments are useful because they allow us to turn unconditional into conditional lower previsions. In particular, for any disjoint proper subsets O and I of N, we can use the epistemic irrelevance assessment of X_I to X_O to infer from the joint lower prevision $\underline{P}_N$ a conditional lower prevision $\underline{P}_O(\cdot|X_I)$ on $\mathscr{L}(\mathscr{X}_{O\cup I})$ given by:

$$\underline{P}_O(h|z_I) \coloneqq \underline{P}_N(h(\cdot,z_I)) \text{ for all gambles } h \text{ on } \mathscr{X}_{O\cup I} \text{ and all } z_I \in \mathscr{X}_I.$$

So we can use the symmetrised assessment of epistemic independence of the variables X_n, $n \in N$ to infer from $\underline{P}_N$ the following family of conditional lower previsions:

$$\mathscr{I}(\underline{P}_N) \coloneqq \{\underline{P}_O(\cdot|X_I)\colon O \text{ and } I \text{ disjoint proper subsets of } N\}.$$

This idea leads to the definition of an independent product, which generalises the existing notion for (precise) probability models.

Definition 1. *A (separately) coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ that coincides with the marginal lower previsions $\underline{P}_n$ on their domains $\mathscr{L}(\mathscr{X}_n)$, $n \in N$ and that is coherent with the family of conditional lower previsions $\mathscr{I}(\underline{P}_N)$ is called an* independent product[11] *of these marginals $\underline{P}_n$.*

It turns out that there always is a point-wise smallest independent product:

Proposition 1. *Any collection of (separately) coherent lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$, has a point-wise smallest independent product. We call it their* independent natural extension *and denote it by $\otimes_{n\in N}\underline{P}_n$. Moreover, $\otimes_{n\in N}\underline{P}_n$ is a strongly factorising coherent lower prevision on $\mathscr{L}(\mathscr{X}_N)$.*

Strong factorisation is strongly linked with independent products, and will play a crucial part in our development of an algorithm for updating an imprecise Markov tree in Section 6. It is defined as follows:

[11] In [11], we distinguish between many-to-many and many-to-one independent products. It is not necessary to make this distinction here, but whenever we use the term 'independent product' in the present paper, we implicitly refer to the more stringent many-to-many version introduced there.

Definition 2. *We call a (separately) coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$* strongly factorising *if for all disjoint proper subsets O and I of N, all $g \in \mathscr{L}(\mathscr{X}_O)$ and all non-negative $f \in \mathscr{L}(\mathscr{X}_I)$, $\underline{P}_N(fg) = \underline{P}_N(f\underline{P}_N(g))$.*

As another important example, the so-called *strong product* $\times_{n\in N}\underline{P}_n$ [4] of marginal lower previsions $\underline{P}_n$ is strongly factorising.[12]

As a consequence of the separate coherence of the joint lower prevision $\underline{P}_N$, the right-hand side of the equality in this definition can be rewritten as:

$$\underline{P}_N(f\underline{P}_N(g)) = \begin{cases} \underline{P}_N(f)\underline{P}_N(g) & \text{if } \underline{P}_N(g) \geq 0 \\ \overline{P}_N(f)\underline{P}_N(g) & \text{if } \underline{P}_N(g) \leq 0, \end{cases}$$

which explains where the term 'factorising' comes from. In particular, for any (separately) coherent strongly factorising joint lower prevision $\underline{P}_N$, we see that for any partition $N_1, \ldots, N_m$ of N:

$$\underline{P}_N(\times_{k=1}^m A_k) = \prod_{k=1}^m \underline{P}_N(A_k) \text{ and } \overline{P}_N(\times_{k=1}^m A_k) = \prod_{k=1}^m \overline{P}_N(A_k), \tag{3}$$

where $A_k \subseteq \mathscr{X}_{I_k}$ for $k = 1, \ldots, m$.

The independent natural extension has very interesting and non-trivial *marginalisation and associativity properties*. Consider any non-empty subset R of N, then the independent natural extension $\otimes_{r\in R}\underline{P}_r$ of the marginals $\underline{P}_r$, $r \in R$ coincides with the restriction of $\otimes_{n\in N}\underline{P}_n$ to the set of gambles $\mathscr{L}(\mathscr{X}_R)$:

$$(\otimes_{r\in R}\underline{P}_r)(g) = (\otimes_{n\in N}\underline{P}_n)(g) \text{ for all gambles } g \text{ on } \mathscr{X}_R. \tag{4}$$

Moreover, for any partition N_1 and N_2 of N, we have that

$$\otimes_{n\in N}\underline{P}_n = \left(\otimes_{n_1\in N_1}\underline{P}_{n_1}\right) \otimes \left(\otimes_{n_2\in N_2}\underline{P}_{n_2}\right), \tag{5}$$

so $\otimes_{n\in N}\underline{P}_n$ is the independent natural extension of its $\mathscr{X}_{N_1}$-marginal $\otimes_{n_1\in N_1}\underline{P}_{n_1}$ and its $\mathscr{X}_{N_2}$-marginal $\otimes_{n_2\in N_2}\underline{P}_{n_2}$.

3.2. **Regular extension.** As a next step, suppose we want to condition a separately coherent and strongly factorising joint $\underline{P}_N$ on observations of the type $X_I = z_I$, where I is some proper subset of N. In other words, we want to find conditional lower previsions $\underline{P}_O(\cdot|X_I)$ on $\mathscr{L}(\mathscr{X}_{I\cup O})$ that are (jointly) coherent with the joint lower prevision $\underline{P}_N$. To this end, we calculate the so-called *regular extension* as follows. Consider z_I in $\mathscr{X}_I$. When $\overline{P}_N(\{z_I\}) > 0$,

$$\underline{R}(h|z_I) := \max\{\mu \in \mathbb{R} \colon \underline{P}_N(\mathbb{I}_{\{z_I\}}[h-\mu]) \geq 0\},$$

where O is any non-empty subset of $N \setminus I$ and h is any gamble on $\mathscr{X}_{I\cup O}$. When $\overline{P}_N(\{z_I\}) = 0$, $\underline{R}(\cdot|x_I)$ is *vacuous*, meaning that $\underline{R}(h|z_I) = \min_{x_O\in\mathscr{X}_O} h(z_I, x_O)$ for all gambles h on $\mathscr{X}_{I\cup O}$.

Generally speaking, coherence only determines $\underline{P}_O(\cdot|z_I)$ uniquely if $\underline{P}_N(\{z_I\}) > 0$, and in that case regular extension yields this uniquely coherent conditional lower prevision: $\underline{P}_O(\cdot|z_I) = \underline{R}(\cdot|z_I)$. When $\underline{P}_N(\{z_I\}) = 0$, regular extension is still coherent, and it even still characterises the coherent $\underline{P}_O(\cdot|z_I)$, because these all lie between the vacuous lower prevision and $\underline{R}(\cdot|z_I)$. For more details about this regular extension, we refer to [24, Appendix J] and [16, Section 4].

If the joint $\underline{P}_N$ is strongly factorising, we get:

$$\begin{aligned} \underline{P}_N(\mathbb{I}_{\{x_I\}}[h-\mu]) &= \underline{P}_N(\mathbb{I}_{\{x_I\}}\underline{P}_N(h(x_I,\cdot)-\mu)) \\ &= \begin{cases} \underline{P}_N(\{x_I\})[\underline{P}_N(h(x_I,\cdot))-\mu] & \text{if } \underline{P}_N(h(x_I,\cdot)) \geq \mu \\ \overline{P}_N(\{x_I\})[\underline{P}_N(h(x_I,\cdot))-\mu] & \text{if } \underline{P}_N(h(x_I,\cdot)) \leq \mu, \end{cases} \end{aligned}$$

so we conclude that, quite interestingly,

$$\underline{R}(h|x_I) = \underline{P}_N(h(x_I,\cdot)) \text{ as soon as } \overline{P}_N(\{x_I\}) > 0. \tag{6}$$

[12] This type of independent product comes to the fore in a study of credal nets under strong independence.

In other words, the conditional lower previsions found by regular extension satisfy all epistemic irrelevance conditions present in an assessment of epistemic independence. We shall have occasion to use this idea several times in the course of this paper, especially in the proofs.

3.3. **Conditionally independent products.** To end this section, we generalise the notion of an independent product to that of a conditionally independent product. In this case we have a number of 'marginal' conditional lower previsions $\underline{P}_n(\cdot|Y)$ on $\mathscr{L}(\mathscr{X}_n)$ representing beliefs (conditional on a variable Y in a finite set $\mathscr{Y}$) about the values that each of a finite number of (logically independent) variables X_n assume in the respective non-empty finite sets $\mathscr{X}_n$, $n \in N$.

We want to construct a conditional lower prevision $\underline{P}_N(\cdot|Y)$ on $\mathscr{L}(\mathscr{X}_N)$, where $\mathscr{X}_N = \times_{n\in N}\mathscr{X}_n$, that coincides with the marginal conditional lower previsions $\underline{P}_n(\cdot|Y)$ on their respective domains $\mathscr{L}(\mathscr{X}_n)$, and such that this $\underline{P}_N(\cdot|Y)$ reflects the following structural assessments: for any disjoint proper subsets O and I of N, the variables X_I are epistemically irrelevant to the variables X_O, *conditional on* Y. In other words, if the value of Y was known, then learning the value of any number of these variables would not affect beliefs about the remaining variables. We then call the variables X_n, $n \in N$ *epistemically independent, conditional on* Y.

Generally speaking, such conditional irrelevance assessments are useful because they allow us to turn lower previsions conditional on Y alone into other, more involved conditional lower previsions. In particular, for any disjoint proper subsets O and I of N, we can use the epistemic irrelevance assessment of X_I to X_O conditional on Y to infer from the joint lower prevision $\underline{P}_N(\cdot|Y)$ a conditional lower prevision $\underline{P}_O(\cdot|X_I,Y)$ on $\mathscr{L}(\mathscr{X}_{O\cup I})$ [or equivalently on $\mathscr{L}(\mathscr{X}_{O\cup I}\times\mathscr{Y})$] given by:

$$\underline{P}_O(h|z_I,y) := \underline{P}_N(h(\cdot,z_I)|y) \text{ for all gambles } h \text{ on } \mathscr{X}_{O\cup I} \text{ and all } z_I \in \mathscr{X}_I.$$

So we can use the symmetrised assessment of epistemic independence of the variables X_n, $n \in N$ conditional on Y to infer from the $\underline{P}_N(\cdot|Y)$ the following family of conditional lower previsions:

$$\mathscr{I}(\underline{P}_N(\cdot|Y)) := \{\underline{P}_O(\cdot|X_I,Y)\colon O \text{ and } I \text{ disjoint proper subsets of } N\}.$$

This idea leads to the definition of a conditionally independent product.

Definition 3. *A (separately) coherent conditional lower prevision $\underline{P}_N(\cdot|Y)$ on $\mathscr{L}(\mathscr{X}_N)$ that coincides with the 'marginal' conditional lower previsions $\underline{P}_n(\cdot|Y)$ on their domains $\mathscr{L}(\mathscr{X}_n)$, $n \in N$ and that is coherent with the family of conditional lower previsions $\mathscr{I}(\underline{P}_N(\cdot|Y))$ is called a* conditionally independent product *of these marginals $\underline{P}_n(\cdot|Y)$.*

It turns out that there always is a point-wise smallest conditionally independent product:

Proposition 2. *Any collection of (separately) coherent conditional lower previsions $\underline{P}_n(\cdot|Y)$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$, has a point-wise smallest conditionally independent product. We call it their* conditionally independent natural extension *and denote it by $\otimes_{n\in N}\underline{P}_n(\cdot|Y)$.*

The notation we use for the conditionally independent natural extension is appropriately suggestive: for each y in $\mathscr{Y}$, $\otimes_{n\in N}\underline{P}_n(\cdot|y)$ is indeed the independent natural extension of the marginal lower previsions $\underline{P}_n(\cdot|y)$. This implies that each $\otimes_{n\in N}\underline{P}_n(\cdot|y)$ is a strongly factorising coherent lower prevision on $\mathscr{L}(\mathscr{X}_N)$.

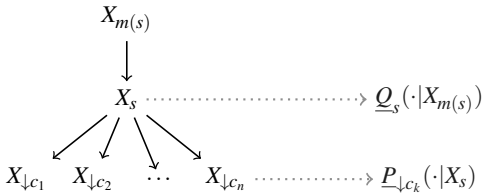
We are now ready to go back to our discussion of imprecise Markov trees.

4. Constructing the most conservative joint

Let us show how to construct specific global models for the variables in the tree, and argue that these are the most conservative coherent models that extend the local models and express all conditional irrelevancies (2), encoded in the imprecise Markov tree. In Section 6,

we will use these global models to construct and justify an algorithm for updating the imprecise Markov tree.

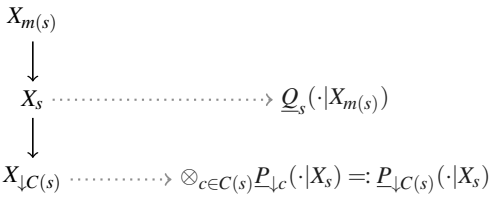
The crucial step lies in the recognition that any tree can be constructed recursively from the leaves up to the root, by using basic building blocks of the following type:



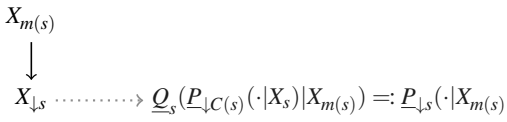
The global models are then also constructed recursively, following the same pattern. In what follows, we first derive the recursion equations for these global models in a heuristic manner. The real justification for using the global models thus derived is then given in Theorem 5.

Consider a node s and suppose that, in each of its children $c \in C(s)$, we already have a global conditional lower prevision $\underline{P}_{\downarrow c}(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow c})$ [or equivalently, on $\mathscr{L}(\mathscr{X}_{\{s\}\cup\downarrow c})$].

Given that, conditional on X_s, the variables $X_{\downarrow c}$, $c \in C(s)$ are epistemically independent [see Section 2.4, condition CI], the discussion in Section 3 leads us to combine the 'marginals' $\underline{P}_{\downarrow c}(\cdot|X_s)$, $c \in C(s)$ into their point-wise smallest conditionally independent product (conditionally independent natural extension) $\otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s)$, which is a conditional lower prevision $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow C(s)})$ [or equivalently, on $\mathscr{L}(\mathscr{X}_{\downarrow s})$]:



Next, we need to combine the conditional models $\underline{Q}_s(\cdot|X_{m(s)})$ and $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ into a global conditional model about $X_{\downarrow s}$. Given that, conditional on X_s, the variable $X_{m(s)}$ is epistemically irrelevant to the variable $X_{\downarrow C(s)}$ [see Section 2.4, condition CI], we expect $\underline{P}_{\downarrow C(s)}(\cdot|X_{\{m(s),s\}})$ and $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ to coincide [this is a special instance of Eq. (2)]. The most conservative (point-wise smallest) coherent way of combining the conditional lower previsions $\underline{P}_{\downarrow C(s)}(\cdot|X_{\{m(s),s\}})$ and $\underline{Q}_s(\cdot|X_{m(s)})$ consists in taking their *marginal extension*[13] $\underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_{\{m(s),s\}})|X_{m(s)}) = \underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_s)|X_{m(s)})$; see [17, 24] for more details. Graphically:



Summarising, and also accounting for the case $s = \square$, we can construct a global conditional lower prevision $\underline{P}_{\downarrow s}(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$ by backwards recursion:

$$\underline{P}_{\downarrow C(s)}(\cdot|X_s) := \otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s) \tag{7}$$

$$\underline{P}_{\downarrow s}(\cdot|X_{m(s)}) := \underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_s)|X_{m(s)}) = \underline{Q}_s(\otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s)|X_{m(s)}), \tag{8}$$

for all $s \in T^{\Diamond}$. If we start with the 'boundary conditions'

$$\underline{P}_{\downarrow t}(\cdot|X_{m(t)}) := \underline{Q}_t(\cdot|X_{m(t)}) \text{ for all leaves } t, \tag{9}$$

[13]Marginal extension is, in the special case of precise probability models, also known as the law of total probability, or the law or iterated expectations.

then the recursion relations (7) and (8) eventually lead to the global joint model $\underline{P}_\square = \underline{P}_{\downarrow\square}(\cdot|X_{m(\square)})$, and to the global conditional models $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ for all non-terminal nodes s. For any subset $S \subseteq C(s)$, the global conditional model $\underline{P}_{\downarrow S}(\cdot|X_s)$ can then be defined simply as the restriction of the model $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow C(s)})$ to the set $\mathscr{L}(\mathscr{X}_{\downarrow S})$:

$$\underline{P}_{\downarrow S}(g|X_s) := \underline{P}_{\downarrow C(s)}(g|X_s) \text{ for all gambles } g \text{ on } \mathscr{X}_{\downarrow S}. \tag{10}$$

It follows from the discussion in Section 3 that, alternatively [see Eq. (4)],

$$\underline{P}_{\downarrow S}(\cdot|X_s) = \otimes_{c\in S}\underline{P}_{\downarrow c}(\cdot|X_s). \tag{11}$$

For easy reference, we will in what follows refer to this collection of global models as the *family of global models* $\mathscr{T}(\underline{P})$, so

$$\mathscr{T}(\underline{P}) := \{\underline{P}\} \cup \{\underline{P}_{\downarrow S}(\cdot|X_s)\colon s \in T^\Diamond \text{ and non-empty } S \subseteq C(s)\}.$$

We end this section by discussing a number of interesting properties for the family of global models $\mathscr{T}(\underline{P})$ we can derive in this way. Let us call any real functional Φ on $\mathscr{L}(\mathscr{X})$ *strictly positive* if $\Phi(\mathbb{I}_{\{x\}}) > 0$ for all $x \in \mathscr{X}$.

Proposition 3. *If all the local models $\overline{Q}_s(\cdot|X_{m(s)})$, $s \in T$ are strictly positive, then so are all the global models in $\mathscr{T}(\underline{P})$.*

Proposition 4. *Consider any non-empty subset E of T and any $x_E \in \mathscr{X}_E$. If $\overline{P}(\{x_E\}) > 0$ then also $\overline{P}_{\downarrow c}(\{x_{E\cap\downarrow c}\}|x_e) > 0$ for all $e \in E$ and all $c \in C(e)$.*[14]

Before we formulate the most important result in this section (and arguably, in this paper), we provide some motivation. Suppose we have some family of global models

$$\mathscr{T}(\underline{V}) := \{\underline{V}\} \cup \{\underline{V}_{\downarrow S}(\cdot|X_s)\colon s \in T^\Diamond \text{ and non-empty } S \subseteq C(s)\}.$$

associated with the tree. How do we express that such a family is compatible with the assessments encoded in the tree?

First of all, we require that our global models should extend the local models:

T1. For each $s \in T$, $\underline{Q}_s(\cdot|X_{m(s)})$ is the restriction of $\underline{V}_{\downarrow s}(\cdot|X_{m(s)})$ to $\mathscr{L}(\mathscr{X}_s)$.

The second requirement is that our models should satisfy the rationality requirement of coherence:

T2. The (conditional) lower previsions in $\mathscr{T}(\underline{V})$ are jointly coherent.

The third requirement requires more explanation: the global models should reflect all epistemic irrelevancies encoded in the graphical structure of the tree. Naively, we would want condition (2) to be satisfied. The problem is that only the right-hand side in Eq. (2), involving the model $\underline{V}_{\downarrow S}(\cdot|X_s)$ is directly available to us. To get to the left-hand side involving the model $\underline{V}_{\downarrow S}(\cdot|X_{\{s\}\cup I})$, one naive approach would be to 'condition the joint model $\underline{V} = \underline{V}_T$ on the variable $X_{\{s\}\cup I}$'. But we have seen in Section 3.1 that given a joint model, coherence in general only determines the conditional models uniquely, provided that the *lower probability* of the conditioning event is non-zero. This is a fairly strong condition, and in what follows we would generally prefer to work with the much weaker condition that the *upper probability* of the conditioning event is non-zero.[15] Since in that case the left-hand side of Eq. (2) need not be uniquely determined from the joint $\underline{V}$ by coherence, this approach becomes unfeasible.

Nevertheless, as soon as we realise that all we can reasonably require from our models is that they should be coherent, the right approach readily suggests itself:[16] we should require

[14] Observe that this holds trivially also if $E \cap \downarrow c = \emptyset$, because then $\mathscr{X}_{E\cap\downarrow c} = \mathscr{X}_\emptyset$ is a singleton [see footnote 4] whose upper probability should be 1 by separate coherence.

[15] As the results in [11] suggest, it might be possible to go even further, and prove a counterpart to Theorem 5 with no positivity restrictions on the local models. We leave this as an avenue for future research, however.

[16] This is also the approach implicit in Definition 1, as well as the one used in [11]. It coincides with the usual, naive approach as soon as all the relevant conditional models are uniquely determined from the joint by coherence.

that if we use the available models $\underline{V}_{\downarrow S}(\cdot|X_s)$ to *define* the models $\underline{V}_{\downarrow S}(\cdot|X_{\{s\}\cup I})$ through the epistemic irrelevance condition (2), then the result should still be coherent:

T3. If we define the conditional lower previsions $\underline{V}_{\downarrow S}(\cdot|X_{\{s\}\cup I})$, $s \in T^{\Diamond}$, $S \subseteq C(s)$ and $R \subseteq \overline{S}$ through the epistemic irrelevance requirements

$$\underline{V}_{\downarrow S}(f|z_{\{s\}\cup R}) := \underline{V}_{\downarrow S}(f(\cdot, z_R)|z_s) \text{ for all gambles } f \text{ in } \mathscr{L}(\mathscr{X}_{\downarrow S\cup R}),$$

then all these models together should be (jointly) coherent with all the available models in the family $\mathscr{T}(\underline{V})$.

And there is a final requirement, which guarantees that all inferences we make on the basis of our global models are as conservative as possible, and are therefore based on no other considerations than what is encoded in the tree:

T4. The models in the family $\mathscr{T}(\underline{V})$ are dominated (point-wise) by the corresponding models in all other families satisfying requirements T1–T3.

It turns out that the family of models $\mathscr{T}(\underline{P})$ we have been constructing above satisfy all these requirements.

Theorem 5. *If all local models $\overline{Q}_s(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_s)$, $s \in T$ are strictly positive, then the family of global models $\mathscr{T}(\underline{P})$, obtained through Eqs.* (7)–(10)*, constitutes the point-wise smallest family of (conditional) lower previsions that satisfy* T1–T3*. It is therefore the unique family to also satisfy* T4*. Finally, consider any non-empty set of nodes $E \subseteq T$ and the corresponding conditional lower prevision derived by applying regular extension:*[17]

$$\underline{R}(f|x_E) := \max\{\mu \in \mathbb{R} \colon \underline{P}_{\downarrow T}(\mathbb{I}_{\{x_E\}}[f-\mu]) \geq 0\} \text{ for all } f \in \mathscr{L}(\mathscr{X}_T) \text{ and all } x_E \in \mathscr{X}_E.$$

Then the conditional lower prevision $\underline{R}(\cdot|X_E)$ is (jointly) coherent with the global models in the family $\mathscr{T}(\underline{P})$.

The last statement of this theorem guarantees that if we use regular extension to *update the tree* given evidence $X_E = x_E$, i.e., derive conditional models $\underline{R}(\cdot|x_E)$ from the joint model $\underline{P} = \underline{P}_{\downarrow T}$, such inferences will always be coherent. This is of particular relevance for the discussion in Section 6, where we derive an efficient algorithm for updating the tree using regular extension. It implies in particular that our algorithm produces coherent inferences.

5. Some separation properties

Without going into too much detail, we would like to point out some of the more striking differences between the separation properties in imprecise Markov trees under epistemic irrelevance, and the more usual ones that are valid for Bayesian nets [20], which, by the way, are also inherited from Bayesian nets by credal nets under strong independence [4].

It is clear from the interpretation of the graphical model described in Section 2.4 that we have the following simple separation results:

$$X_{i_1} \longrightarrow X_{i_2} \longrightarrow X_t \qquad X_{i_1} \longleftarrow X_{i_2} \longrightarrow X_t$$

where in both cases, X_{i_2} *separates* X_t *from* X_{i_1}: when the value of X_{i_2} is known, additional information about the value of X_{i_1} does not affect beliefs about the value of X_t. In this figure, between i_1 and i_2, and between i_2 and t, there may be other nodes, but the arrows along the path segment through these nodes should all point in the indicated directions. The underlying idea is that t is a (descendant of some) child c of i_2, and conditional on the mother i_2 of c, the non-parent non-descendant i_1 of c is epistemically irrelevant to c and all of its descendants.

On the other hand, and in contradistinction with what we are used to in Bayesian nets, we will not generally have separation in the following configuration:

[17] If we look at the proof of this result in the Appendix, it is not hard to see that similar statements can be made about the (joint) coherence of the regular extensions $\underline{R}(\cdot|X_{E_k})$ for any finite collection E_k, $k = 1, \ldots, n$ of sets of nodes.

$$X_{i_1} \longleftarrow X_{i_2} \longleftarrow X_t$$

where X_{i_2} *does not necessarily separate* X_t *from* X_{i_1}. We will come across a simple counterexample in Section 7. Where does this difference with the case of Bayesian nets originate? It is clear from the reasoning above that X_{i_2} separates X_{i_1} from X_t: conditional on X_{i_2}, X_t is epistemically irrelevant to X_{i_1}. For precise probability models, irrelevance generally implies symmetrical independence, and therefore this will generally imply that conditional on X_{i_2}, X_{i_1} is epistemically irrelevant to X_t as well. But for imprecise probability models no such symmetry is guaranteed [3], and we therefore cannot infer that, generally speaking, X_{i_2} will separate X_{i_1} from X_t. *As a general rule, we can only infer separation if the arrows point from the 'separating' variable* X_{i_2} *towards the 'target' variable* X_t.

6. A FAST ALGORITHM FOR UPDATING IN AN IMPRECISE MARKOV TREE

We now consider the case where we are interested in making inferences about the value of the variable X_t in some *target node* t, when we know the values x_E of the variables X_E in a set $E \subseteq T \setminus \{t\}$ of *evidence nodes*.

6.1. **The formulation of the problem.** If we assume that the values of the remaining variables are *missing at random*, then we can do this by conditioning the joint $\underline{P}$ obtained above on the available evidence '$X_E = x_E$'; see for instance [12, 29].

We will address this problem by updating the lower prevision $\underline{P}$ to the lower prevision $\underline{R}_t(\cdot|x_E)$ on $\mathscr{L}(\mathscr{X}_t)$ using *regular extension* [24, Appendix J]:

$$\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R} \colon \underline{P}(\mathbb{I}_{\{x_E\}}[g-\mu]) \geq 0\} \tag{12}$$

for all gambles g on $\mathscr{X}_t$, *assuming that* $\overline{P}(\{x_E\}) > 0$. Theorem 5 guarantees that such inferences are coherent. Sufficient conditions on the local models for this positivity assumption to hold are given in Proposition 3.

Consider the map

$$\rho_g \colon \mathbb{R} \to \mathbb{R} \colon \mu \mapsto \underline{P}(\mathbb{I}_{\{x_E\}}[g-\mu]).$$

We can infer from the separate coherence of $\underline{P}$ that $|\rho_g(\mu_1) - \rho_g(\mu_2)| \leq |\mu_1 - \mu_2|\overline{P}(\{x_E\})$ for all $\mu_1, \mu_2 \in \mathbb{R}$, which implies that ρ_g is (Lipschitz) continuous. Separate coherence of $\underline{P}$ also guarantees that ρ_g is concave and non-increasing. Hence $\{\mu \in \mathbb{R} \colon \rho_g(\mu) \geq 0\} = (-\infty, \underline{R}_t(g|x_E)]$, which shows that the supremum that we should have *a priori* used in (12) is indeed a maximum. $\underline{R}_t(g|x_E)$ is the right-most zero of ρ_g, and it is, again by separate coherence of $\underline{P}$, guaranteed to lie between the smallest value $\min g$ and the largest value $\max g$ of g. If moreover $\underline{P}(\{x_E\}) > 0$, then separate coherence of $\underline{P}$ implies that $\underline{R}_t(g|x_E)$ is the unique zero of ρ_g. If on the other hand $\underline{P}(\{x_E\}) = 0$, then $(-\infty, \underline{R}_t(g|x_E)]$ is the set of all zeros of ρ_g. It appears that any algorithm for calculating $\underline{R}_t(g|x_E)$ will benefit from being able to calculate the values of ρ_g, or even more simply check their signs, efficiently.

6.2. **Calculating the values of ρ_g recursively.** We now recall from Section 4 that the joint $\underline{P}$ can be constructed recursively from leaves to root. The idea we now use is that calculating $\rho_g(\mu) = \underline{P}(\mathbb{I}_{\{x_E\}}[g-\mu])$ becomes easier if we graft the structure of the tree onto the argument $g^\mu := \mathbb{I}_{\{x_E\}}[g-\mu]$ as follows. Define

$$g^\mu_s := \begin{cases} \mathbb{I}_{\{x_s\}} & \text{if } s \in E \\ g - \mu & \text{if } s = t \\ 1 & \text{if } s \in T \setminus (E \cup \{t\}), \end{cases}$$

then $g^\mu_s \in \mathscr{L}(\mathscr{X}_s)$ and $g^\mu = \prod_{s \in T} g^\mu_s$. Also define, for any $s \in T$, the gamble ϕ^μ_s on $\mathscr{X}_{\downarrow s}$ by $\phi^\mu_s := \prod_{u \in \downarrow s} g^\mu_u$. Then

$$\phi^\mu_\square = g^\mu \text{ and } \phi^\mu_s \geq 0 \text{ if } s \not\sqsubseteq t,$$

and

$$\phi_s^\mu = g_s^\mu \prod_{c\in C(s)} \phi_c^\mu \text{ for all } s \in T, \tag{13}$$

where we use the convention that any product over an empty set of indices equals one. Eq. (13) is the argument counterpart of Eq. (8). Also, if $s \not\sqsubseteq t$ then g_s^μ and ϕ_s^μ do not depend on μ, nor on g. Indeed, in that case

$$\phi_s^\mu = \mathbb{I}_{\{x_{E\cap\downarrow s}\}}. \tag{14}$$

First, let us consider the nodes $s \not\sqsubseteq t$. We define the *messages* $\underline{\pi}_s$ and $\overline{\pi}_s$ recursively by

$$\underline{\pi}_s := \underline{Q}_s\left(g_s^\mu \prod_{c\in C(s)} \underline{\pi}_c \middle| X_{m(s)}\right) \text{ and } \overline{\pi}_s := \overline{Q}_s\left(g_s^\mu \prod_{c\in C(s)} \overline{\pi}_c \middle| X_{m(s)}\right). \tag{15}$$

We summarise such a pair by the notation: $\underline{\overline{\pi}}_s := \underline{\overline{Q}}_s(g_s^\mu \prod_{c\in C(s)} \underline{\overline{\pi}}_c | X_{m(s)}) := (\underline{\pi}_s, \overline{\pi}_s)$. Then there are two possibilities:

$$\overline{\pi}_s = \begin{cases} \overline{Q}_s(\{x_s\}|X_{m(s)}) \prod_{c\in C(s)} \overline{\pi}_c(x_s) & \text{if } s \in E \\ \overline{Q}_s\left(\prod_{c\in C(s)} \overline{\pi}_c \middle| X_{m(s)}\right) & \text{if } s \notin E. \end{cases}$$

The messages $\underline{\pi}_s$ and $\overline{\pi}_s$ are gambles on $\mathscr{X}_{m(s)}$, and can therefore be seen as tuples of real numbers, with as many components $\underline{\overline{\pi}}_s(x_{m(s)})$ as there are elements $x_{m(s)}$ in $\mathscr{X}_{m(s)}$. They are all non-negative. As their notation suggests, they do not depend on the choice of g or μ, but only (at most) on which nodes are *instantiated*, i.e., belong to E, and on which value x_E the variable X_E for these instantiated nodes assumes.

It then follows from Eqs. (8) and (13) and the strong factorisation property[18] that

$$\underline{P}_{\downarrow s}(\phi_s^\mu | X_{m(s)}) = \underline{\pi}_s \text{ and } \overline{P}_{\downarrow s}(\phi_s^\mu | X_{m(s)}) = \overline{\pi}_s. \tag{16}$$

Next, we turn to nodes $s \sqsubseteq t$. Define the messages π_s^μ by

$$\pi_s^\mu := \underline{Q}_s(\psi_s^\mu | X_{m(s)}), \tag{17}$$

where the gambles ψ_s^μ on $\mathscr{X}_s$ are given by the recursion relations:

$$\psi_t^\mu := \max\{g-\mu, 0\} \prod_{c\in C(t)} \underline{\pi}_c + \min\{g-\mu, 0\} \prod_{c\in C(t)} \overline{\pi}_c, \tag{18}$$

and for each $\square \neq s \sqsubseteq t$, so $m(s)$ exists,

$$\psi_{m(s)}^\mu := \left[\max\{\pi_s^\mu, 0\} \prod_{c\in S(s)} \underline{\pi}_c + \min\{\pi_s^\mu, 0\} \prod_{c\in S(s)} \overline{\pi}_c\right] g_{m(s)}^\mu. \tag{19}$$

The messages π_s^μ are again tuples of real numbers, with one component $\pi_s^\mu(x_{m(s)})$ for each of the possible values $x_{m(s)}$ of $X_{m(s)}$.[19] They do depend on the choice of g or μ, as well as on which nodes are instantiated and on which value x_E the variable X_E for these instantiated nodes assumes.

It then follows from Eqs. (8) and (13) and the strong factorisation property of the local independent products that

$$\underline{P}_{\downarrow s}(\phi_s^\mu | X_{m(s)}) = \pi_s^\mu \text{ and of course } \rho_g(\mu) = \pi_\square^\mu. \tag{20}$$

[18] This, together with the course of reasoning leading to Eq. (20), shows that the results of updating the tree (and the algorithm we are deriving) in this way will be exactly the same *for any way* of forming a product of the local models for the children of s, *provided only that this product is strongly factorising*. For instance, replacing the conditionally independent natural extension with the strong product in Eq. (7) will lead to exactly the same inferences. Of course, this should not be taken to mean that our algorithm also works for updating credal trees under strong independence.

[19] If s is the root node, then $m(s) = \emptyset$ and π_s^μ is a single real number, which by Eq. (20) is equal to $\rho_g(\mu)$. See also footnote 4.

We conclude that we can find the value of $\rho_g(\mu)$ by a backwards recursion method consisting in passing messages up to the root of the tree, and in transforming them in each node using the local uncertainty models; see Eqs. (15) and (17)–(19).

There is a further simplification, because we are not necessarily interested in the actual value of $\rho_g(\mu)$, but rather in its sign. It arises whenever there are instantiated nodes above the target node: $E \cap A(t) \neq \emptyset$. Let in that case e_t be the greatest element of the chain $E \cap A(t)$, i.e., the instantiated node closest to and preceding the target node t, and let s_t be its successor in the chain $\uparrow t$; see for instance Fig. 1. If we let

$$\lambda_g(\mu) := \max\{\pi^\mu_{s_t}(x_{e_t}),0\}\prod_{c\in S(s_t)}\underline{\pi}_c(x_{e_t}) + \min\{\pi^\mu_{s_t}(x_{e_t}),0\}\prod_{c\in S(s_t)}\overline{\pi}_c(x_{e_t}),$$

then it follows from Eq. (19) [with $s = s_t$ and $m(s) = e_t$] that $\psi^\mu_{e_t} = \mathbb{I}_{\{x_{e_t}\}}\lambda_g(\mu)$. If we now continue to use Eqs. (18) and (19) until we reach the root of the tree, we eventually find that[20]

$$\rho_g(\mu) = \begin{cases} \underline{P}(\{x_E\})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \geq 0 \\ \overline{P}(\{x_E\})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \leq 0. \end{cases} \tag{21}$$

Since we assumed from the outset that $\overline{P}(\mathbb{I}_{\{x_E\}}) > 0$, we gather from Eq. (12) that $\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R}\colon \lambda_g(\mu) \geq 0\}$. Moreover, by combining Eqs. (14) and (16) with Proposition 4, we find that $\overline{\pi}_c(x_{e_t}) = \overline{P}_{\downarrow c}(\{x_{E\cap\downarrow c}\}|x_{e_t}) > 0$ for all $c \in S(s_t)$, and therefore $\lambda_g(\mu) \geq 0 \Leftrightarrow \pi^\mu_{s_t}(x_{e_t}) \geq 0$. Hence $\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R}\colon \pi^\mu_{s_t}(x_{e_t}) \geq 0\}$.

We conclude that in order to update the tree in the situation described above, we can perform all calculations on the sub-tree $\downarrow s_t$, where the new root s_t has local model $\underline{Q}_{s_t}(\cdot|x_{e_t})$. This is also borne out by the discussion of the separation properties in Section 5.

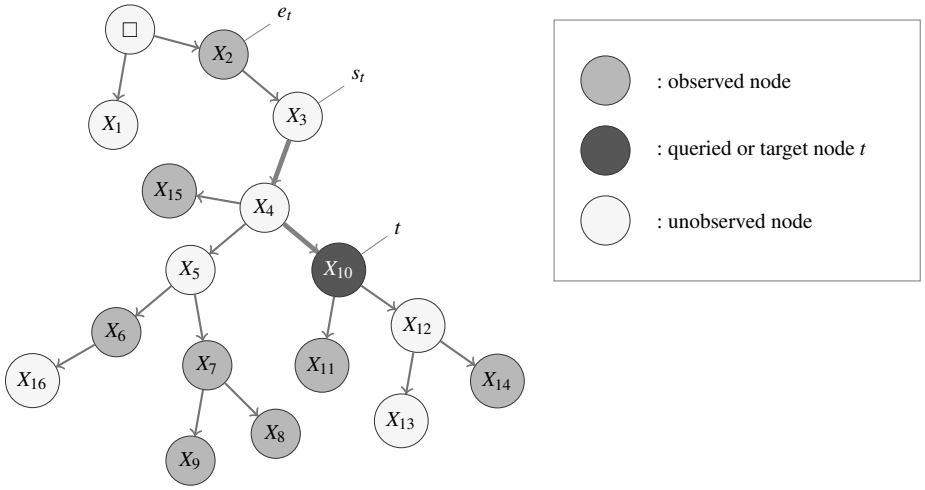


FIGURE 1. Example imprecise Markov tree. The target node is $t = 10$, $e_t = 2$ is the 'greatest' observed ancestor of t and $s_t = 3$ is the child of e_t that precedes t. The bolder arrows represent the trunk $\tilde{T} = \{3,4,10\}$ of the tree.

6.3. **An algorithm.** We now convert these observations into a workable algorithm.

Using regular extension and message passing, we are able to compute $\underline{R}_t(g|x_E)$: we (i) choose any $\mu \in [\min g, \max g]$; (ii) calculate the value of $\lambda_g(\mu)$ by sending messages from the terminal nodes towards the root; and (iii) repeat this in some clever way to find

[20] Actually, we easily derive that $\rho_g(\mu) = a\max\{\lambda_g(\mu),0\} + b\min\{\lambda_g(\mu),0\}$, where a and b are real constants that do not depend on g and μ. Letting $g := \mu \pm 1$ then allows us to identify the constants a and b.

the maximal μ that will make this $\lambda_g(\mu)$ zero. But we have seen above that this naive approach can be sped up by exploiting (a) the separation properties of the tree, and (b) the independence of μ (and g) for some of the messages, namely those associated with nodes that do not precede the target node t.

For a start, as we are only interested in the sign of $\rho_g(\mu)$ [or equivalently, that of $\lambda_g(\mu)$], which we have seen is determined by the sign of $\pi^{\mu}_{s_t}(x_{e_t})$, we only have to take into consideration nodes that strictly follow e_t.

The next thing a smarter implementation of the algorithm can do, is determine the *trunk* $\tilde{T}$ of the tree: those nodes that precede the queried node t and strictly follow the greatest observed node e_t preceding t. We can define the trunk more formally as follows: $\tilde{T} := \uparrow t \cap \downarrow C(e_t)$. For the tree in Fig. 1 for instance, where the darker X_{10} is the queried variable and the lighter nodes $\{2,6,7,8,9,11,14,15\}$ are instantiated, the trunk is given by $\tilde{T} = \{3,4,10\}$, and indicated by bolder arrows.

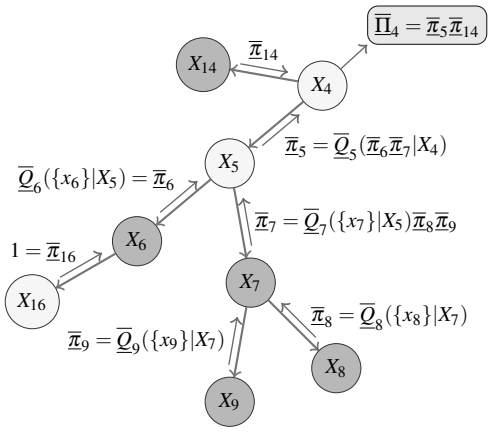


FIGURE 2. Calculation of $\overline{\Pi}_4$, which is a summary of the μ-independent messages in the trunk node 4.

We have a special interest in the nodes that constitute the trunk, because only they will send messages to their mother nodes that actually depend on μ. As a consequence, all other nodes (all descendants of the trunk that are not in the trunk themselves) send messages that have to be calculated only once. This implies that we can summarise all the μ-independent messages by propagating all of them until they reach the trunk. The μ-independent messages $\overline{\pi}_s$ that arrive in a trunk node s can be represented more succinctly by their point-wise products $\overline{\Pi}_s := \prod_{c \in C(s) \setminus \tilde{T}} \overline{\pi}_c$, because Eqs. (18) and (19) only depend on them through on these products.

This means that for every trunk node $s \in \tilde{T}$, we have to find the lower (upper) messages of every child c of s that is not in the trunk itself. Both $\underline{\pi}_c$ and $\overline{\pi}_c$ can be calculated recursively using Eq. (16), where the recursion starts at the leaves and moves up to (but stops right before) the trunk. In the leaves, the local lower and upper previsions of the indicator of the evidence are sent upwards if the leaf is instantiated; if not the constant 1 is sent up, which is equivalent to deleting the node from the tree. We could envisage removing *barren nodes* (all of whose descendants are uninstantiated, such as X_1, X_{13}, X_{16} in the example tree above) from the tree beforehand, but we believe the computational overhead created by the search for them will void the gain.

The only recursion that is still left to do, is the calculation of the μ-dependent messages π^{μ}_s along the trunk. As demonstrated in Fig. 3, we can calculate $\pi^{\mu}_{s_t}(e_t)$ using the following

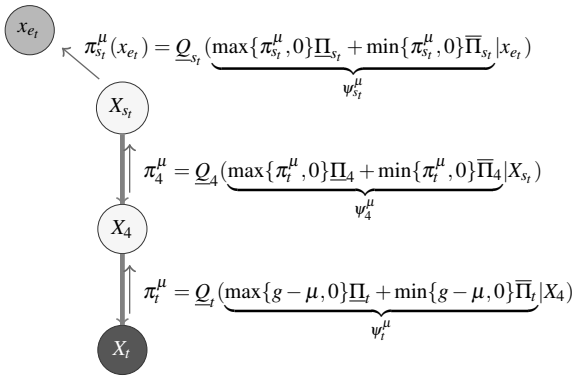
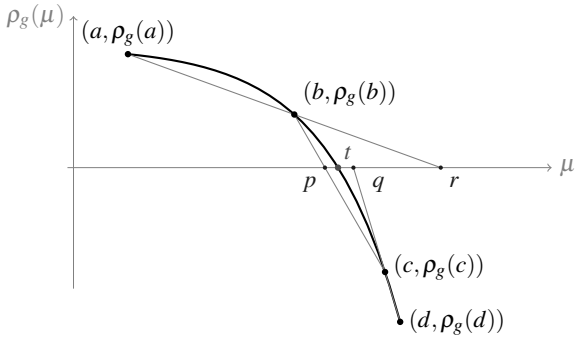


FIGURE 3. Calculation of $\pi^{\mu}_{s_t}(x_{e_t})$, whose sign is the same as that of $\underline{P}(\mathbb{I}_{\{x_E\}}[g-\mu])$.



$p \coloneqq \rho_g(c)b - \rho_g(b)c / \rho_g(c) - \rho_g(b)$;
$m \coloneqq c$; $t \coloneqq p+m/2$;
while $m - p > tol$ **and** $\rho_g(t) \neq 0.0$
if $\rho_g(t) > 0$
$a \coloneqq b$; $b \coloneqq t$;
$s \coloneqq \rho_g(a)b - \rho_g(b)a / \rho_g(a) - \rho_g(b)$;
else
$d \coloneqq c$; $c \coloneqq t$;
$s \coloneqq \rho_g(c)d - \rho_g(d)c / \rho_g(c) - \rho_g(d)$;
$p \coloneqq \rho_g(c)b - \rho_g(b)c / \rho_g(c) - \rho_g(b)$;
$m \coloneqq \min\{m,s\}$; $t \coloneqq p+m/2$;

FIGURE 4. The root of a concave and non-increasing function ρ_g whose values $\rho_g(a) > \rho_g(b) > 0 > \rho_g(c) > \rho_g(d)$ are known, will always be in the interval $[p,m]$ with $m \coloneqq \min\{q,r\}$. Here p,q and r are the intersections with the horizontal axis of the straight lines through $(b,\rho_g(b))$ and $(c,\rho_g(c))$, $(c,\rho_g(c))$ and $(d,\rho_g(d))$, and $(a,\rho_g(a))$ and $(b,\rho_g(b))$, respectively. The next function evaluation of ρ_g will be in t which bisects the error interval $[p,m]$. If $\rho_g(t) > 0$, then a becomes b and b becomes t, otherwise d becomes c and c becomes t and a new interval $[p,m]$ and matching t can be calculated. We stop iterating as soon as the error interval $[p,m]$ is smaller than a given tolerance *tol*, or $\rho_g(t)$ is exactly zero.

recursion formula:

$$\pi^{\mu}_{s} \coloneqq \begin{cases} \underline{Q}_s(\max\{g-\mu,0\}\underline{\Pi}_s + \min\{g-\mu,0\}\overline{\Pi}_s|X_{m(s)}) & s = t, \\ \underline{Q}_s(\max\{\pi^{\mu}_{c_t},0\}\underline{\Pi}_s + \min\{\pi^{\mu}_{c_t},0\}\overline{\Pi}_s|X_{m(s)}) & s \in \tilde{T} \setminus \{t\} \text{ and } C(s) \cap \tilde{T} = \{c_t\}. \end{cases}$$

These formulas are reformulations of Eqs. (17)–(19), where the influence of the $\overline{\Pi}$ has been made explicit.

Since we now know how to calculate $\pi^{\mu}_{s_t}(e_t)$, we can tackle the final problem: find the maximal μ for which $\pi^{\mu}_{s_t}(e_t) = 0$. In principle, a secant root-finding method could be used, but using the concavity and non-increasing character of $\pi^{\mu}_{s_t}(e_t)$ as a function of μ, we can speed up the calculation of the maximal root drastically as shown in Fig. 4.

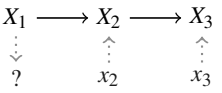
Let us briefly discuss the complexity of our algorithm. Consider for a start that for a fixed μ each node makes a single local computation and then propagates the result to its mother node: this implies that, with μ fixed, the algorithm is linear in the number of nodes.

Iterating on μ then amounts to multiplying such a linear complexity with the number of iterations. This number depends on the function g, as the iterations are made to compute the root of a function that is known to belong to the real interval $[\min g, \max g]$. If we assume that the bisection algorithm is employed to find the root—for the sake of simplicity—and let $r := \max g - \min g$ be the range of the function, then the number of iterations is bounded by $\log_2 \frac{r}{tol} + 1$, where *tol* is some fixed tolerance. In other words, the number of iterations is linear in the number b of bits needed to represent r in base 2. This means that the overall complexity of the algorithm is $O(b \cdot |T|)$, taking into account that the computational complexity of our root-finding algorithm must be lower than for the bisection (and actually also for the secant) algorithm. Since b will be a small number[21] in most cases (e.g. when the focus is on probabilities), we simply refer to the complexity of our algorithm as linear in the number of nodes.

7. A SIMPLE EXAMPLE INVOLVING DILATION

We present a very simple example that allows us to (i) follow the inference method discussed above in a step-by-step fashion; (ii) see that there are separation properties for credal nets under strong independence that fail for credal trees under epistemic irrelevance; and (iii) see that in that case we will typically observe dilation.

Consider the following imprecise Markov chain:



To make things as simple as possible, we suppose that $\mathscr{X}_1 = \{a,b\}$ and that $\underline{Q}_1$ is a linear (or precise, or expectation-like) model Q_1 with mass function q. We also assume that $\underline{Q}_2(\cdot|X_1)$ is a linear model $Q_2(\cdot|X_1)$ with conditional mass function $q(\cdot|X_1)$. We make no such restrictions on the local model $\underline{Q}_3(\cdot|X_2)$. We also use the following simplifying notational device: if we have three real numbers $\underline{\kappa}$, $\overline{\kappa}$ and γ, we let

$$\overline{\kappa}\langle\gamma\rangle := \underline{\kappa}\max\{\gamma,0\} + \overline{\kappa}\min\{\gamma,0\}.$$

We observe $X_2 = x_2$ and $X_3 = x_3$, and want to make inferences about the target variable X_1: for any $g \in \mathscr{L}(\mathscr{X}_1)$, we want to know $\underline{R}_1(g|x_{\{2,3\}})$. Letting $\underline{r} := \underline{R}_1(\{a\}|x_{\{2,3\}})$ and $\overline{r} := \overline{R}_1(\{a\}|x_{\{2,3\}})$, we infer from the separate coherence of $\underline{R}_1(\cdot|x_{\{2,3\}})$ that it suffices to calculate $\underline{r}$ and $\overline{r}$, because

$$\underline{R}_1(g|x_{\{2,3\}}) = g(b) + \overline{r}\langle g(a) - g(b)\rangle.$$

We let $g^\mu = [\mathbb{I}_{\{a\}} - \mu]\mathbb{I}_{\{x_2\}}\mathbb{I}_{\{x_3\}}$, and apply the approach of the previous section. We see that the trunk $\tilde{T} = \{1\}$, and the instantiated leaf node 3 sends up the messages $\overline{\pi}_3 = \overline{\underline{Q}}_3(\{x_3\}|X_2)$ to the instantiated node 2, which transforms them into the messages

$$\overline{\pi}_2 = \overline{\underline{Q}}_2(\{x_2\}|X_1)\overline{\pi}_3(x_2) =: q(x_2|X_1)\overline{q},$$

where we let $q(x_2|X_1) := \overline{\underline{Q}}_2(\{x_2\}|X_1)$ and $\overline{q} := \overline{\pi}_3(x_2)$. These messages are sent up to the (target) root node $t = 1$, which transforms them into the message $\pi_1^\mu = Q_1(\psi_1^\mu)$ with $\psi_1^\mu = q(x_2|X_1)\overline{q}\langle\mathbb{I}_{\{a\}} - \mu\rangle$. If we also use that $0 \leq \mu \leq 1$, this leads to

$$\underline{P}_1(g^\mu) = \pi_1^\mu = q(a)q(x_2|a)\underline{q}[1-\mu] + q(b)q(x_2|b)\overline{q}[-\mu],$$

so we find after applying regular extension that

$$\underline{r} = \underline{R}_1(\{a\}|x_{\{2,3\}}) = \frac{q(a)q(x_2|a)\underline{q}}{q(a)q(x_2|a)\underline{q} + q(b)q(x_2|b)\overline{q}}$$

[21] It could be argued that b should be bounded given the finiteness of a computer's way to represent numbers.

$$\overline{r} = \overline{R}_1(\{a\}|x_{\{2,3\}}) = \frac{q(a)q(x_2|a)\overline{q}}{q(a)q(x_2|a)\overline{q} + q(b)q(x_2|b)\underline{q}}.$$

When $\underline{q} = \overline{q}$, which happens for instance if the local model for X_3 is precise, then we see that, with obvious notations,

$$\overline{r} = \underline{r} = \frac{q(a)q(x_2|a)}{q(a)q(x_2|a) + q(b)q(x_2|b)} =: p(a|x_2) \tag{22}$$

and therefore X_2 indeed separates X_3 from X_1. But in general, letting $\alpha := q(a)q(x_2|a)$ and $\beta := q(b)q(x_2|b)$, we get

$$\overline{r} - p(a|x_2) = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\overline{q}+\beta\underline{q}} \text{ and } p(a|x_2) - \underline{r} = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\underline{q}+\beta\overline{q}}.$$

As soon as $\overline{q} > \underline{q}$, X_2 no longer separates X_3 from X_1, and we witness *dilation* [14, 22] because of the additional observation of X_3!

8. Numerical comparison with strong independence

Strong independence implies epistemic irrelevance, and hence inferred (lower-upper) probability intervals for imprecise Markov trees with epistemic irrelevance will include those obtained assuming strong independence. This suggests that our algorithm could be used also as a tool to make conservative (also called outer) approximations of the computations made in a credal tree under strong independence. This could be an important application of our algorithm since at the moment it is unclear whether or not updating probabilities in a tree is a polynomial task under strong independence. If it were not, addressing the problem would definitely benefit from the availability of fast approximations.

With this idea in mind, here we make a preliminary empirical exploration about the quality of the approximation. As noted in Section 5, the two models have different separation properties: this is particularly important when evidence is back-propagated from leaves to root. For this reason, we compare posterior probability intervals for the root variable of a *chain* where only the leaf node is instantiated.

Fig. 5 reports the results of this comparison for chains with binary nodes, randomly generated local models, and variable length (from 5 up to 100 nodes). The algorithm in Section 6 has been used to compute the posterior probability intervals in the chains under epistemic irrelevance, while the *2U algorithm* [13] was used for updating in the chains under strong independence. The inferred probability intervals for the former turn out to be clearly wider, and the mean difference between the two intervals is about 0.3 irrespective of the length of the chain, at least for chains with more than ten nodes.

For non-binary nodes there are no efficient algorithms known for updating chains with strong independence. We used the procedure in [6] to update chains with less than seven ternary nodes and credal sets with three randomly generated extreme points in the strong independence case. A similar difference between the posterior intervals was observed also in these cases. For longer chains, updating for the chain under strong independence is too slow and no comparison can be made. In summary, there is a non-negligible difference between inferences based on the two notions of 'independence'. This means that the epistemic approximations to the strong case could be quite crude in practise. However, their being outer (that is, safe) approximations together with their light complexity could still make of them very useful tools, whenever the strong independence approach is deemed necessary or appropriate.

9. An application

The tree topology of the graphs considered in this paper is expressive enough to model useful and interesting problems. These problems can then be solved efficiently by means of the algorithm described in the previous sections. We make this point clearer with an

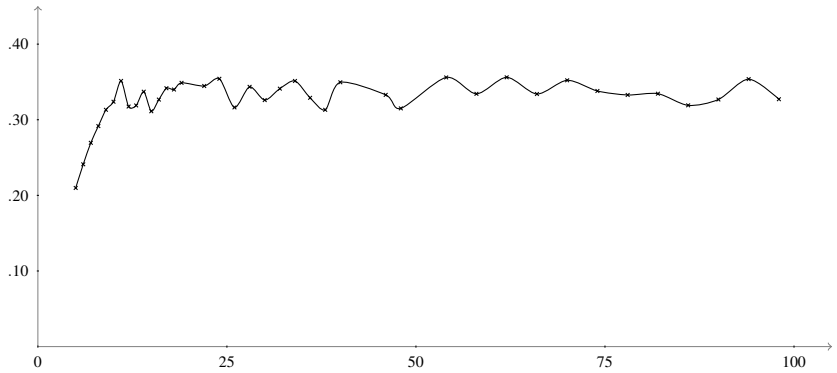


FIGURE 5. Numerical evaluation of the difference between the sizes of the posterior intervals for inferences on credal chains over binary variables with epistemic irrelevance and strong independence. The plot reports the mean difference (in ordinate) as a function of the number of nodes in the chains (in abscissa). Means are estimated over 200 Monte Carlo runs.

example application about character recognition. This is also an opportunity to illustrate the differences between the traditional, precise-probability, approach to the problem and the imprecise-probability one. Most notably, these differences arise because the imprecise-probability methods come with the inherent ability to suspend judgement when the information available is deemed insufficient to reliably recognise a character, whereas the precise-probability ones do not.

9.1. **Imprecise hidden Markov models.** Hidden Markov models (HMMs) [21] are popular tools for modelling a sequence of hidden variables that generate a related sequence of observable variables. These are respectively referred to as the *generative* and *observable* sequences. HMMs have applications in many areas of signal processing, and more specifically in speech and text recognition.

Both the generative and the observable sequence are described by sets of variables over the same domain $\mathscr{X}$, denoted respectively by $X_{s_1}, \ldots, X_{s_n}$ and $X_{o_1}, \ldots, X_{o_n}$. The independence assumptions between these variables, which characterise HMMs, are those corresponding to the tree structure below. Informally, this topology states that every element of the generative sequence depends only on its predecessor, while each observation depends only on the corresponding element of the generative sequence.

$$\begin{array}{lccccccc} \textit{generative sequence}: & X_{s_1} & \longrightarrow & X_{s_2} & \longrightarrow & \cdots & \longrightarrow & X_{s_n} \\ & \downarrow & & \downarrow & & & & \downarrow \\ \textit{observable sequence}: & X_{o_1} & & X_{o_2} & & \cdots & & X_{o_n} \end{array}$$

A local uncertainty model should be defined for each variable. In the case of precise probabilistic assessments, this corresponds to linear (precise, or expectation-like) versions of the local models $\underline{Q}_{s_1}$, $\underline{Q}_{s_{k+1}}(\cdot|X_{s_k})$ and $\underline{Q}_{o_k}(\cdot|X_{s_k})$, $k = 1, \ldots, n$, where the conditional models are assumed to be *stationary*, i.e., independent of k. These model, respectively, beliefs about the first state in the generative sequence, the transitions between adjacent states, and the observation process.

Bayesian techniques for learning from multinomial data are usually employed for identifying these models. But, especially if only few data are available, other methods leading to imprecise assessments, such as the *imprecise Dirichlet model* (IDM, [25]), might offer a more realistic model of the local uncertainty. For example, for the unconditional local

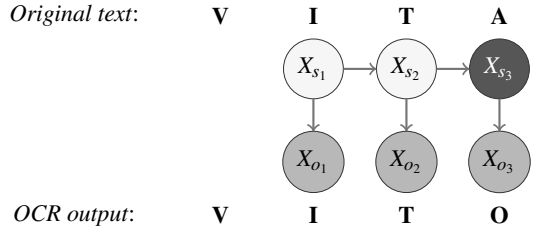
model $\underline{Q}_{s_1}$, applying the IDM leads to the following simple identification:

$$\underline{Q}_{s_1}(\{x_1\}) = \frac{n_{x_1}^{s_1}}{s + \sum_{x \in \mathscr{X}} n_x^{s_1}} \quad \overline{Q}_{s_1}(\{x_1\}) = \frac{s + n_{x_1}^{s_1}}{s + \sum_{x \in \mathscr{X}} n_x^{s_1}}, \tag{23}$$

where $n_{x_1}^{s_1}$ counts the units in the sample for which $X_{s_1} = x_1$, and s is a (positive real) hyper-parameter that expresses the degree of caution in the inferences. For the conditional local models, we can proceed similarly. This leads to the identification of an *imprecise HMM*, a special credal tree under epistemic irrelevance, like the ones introduced in Section 2.

Generally speaking, the algorithm described in Section 6 can be used for computing inferences with such imprecise HMMs. Below, we address the more specific problem of *on-line recognition*, which consists in the identification of the most likely value of X_{s_n}, given the evidence for the whole observational sequence $X_{o_1} = x_{o_1}, \ldots, X_{o_n} = x_{o_n}$. For precise local models, this problem requires the computation of the state $\tilde{x}_{s_n} := \operatorname{argmax}_{x_{s_n} \in \mathscr{X}} P(\{x_{s_n}\} | x_{o_1}, \ldots, x_{o_n})$ that is most probable after the observation. For imprecise local models different criteria can be adopted; see [23] for an overview. We consider *maximality*: we order the states by $x_{s_n} > z_{s_n}$ if and only if $\underline{P}(\mathbb{I}_{\{x_{s_n}\}} - \mathbb{I}_{\{z_{s_n}\}} | x_{o_1}, \ldots, x_{o_n}) > 0$, and we look for the *undominated* or *maximal* states under this order. This may produce *indeterminate* predictions: the set of undominated states may have more than one element.

9.2. **On-line character recognition.** As a very first application of the imprecise HMM, we have considered a *character recognition* problem.[22] A written text was regarded as a generative sequence, while the observable sequence was obtained by artificially corrupting the text. This is a model for a not perfectly reliable observation process, such as the output of an OCR device. The local models were identified using the IDM, as in (23), by counting the occurrences of single characters and the 'transitions' from one character to another in the generative sequence, and by matchings between the elements of the two sequences. By modelling text as a generative sequence, we obviously ignore any dependence there might be between a character and its n-th predecessor, for any $n \geq 2$. A better, albeit still not completely realistic, model would resort to using n-grams (i.e., clusters of n characters with $n \geq 2$) instead of monograms. Such models might lead to higher accuracy, but they need larger data sets for their quantification, because of the exponentially larger number of possible transitions for which probabilities have to be estimated. The figure below depicts how on-line recognition through HMM might apply to this setup.



The performance of the precise model can be characterised by its *accuracy* (the percentage of correct predictions) alone. The imprecise HMM requires more indicators. We follow [2] in using the following:

determinacy: percentage of determinate predictions,
set-accuracy: percentage of indeterminate predictions containing the right state,
single accuracy: percentage of correct predictions computed considering only determinate predictions, and
indeterminate output size: average number of states returned when the prediction is indeterminate over number of possible states.

[22] For a more involved application, related to aircraft trajectory model tracking, see [1].

Precise HMM		
Accuracy	93.96%	(7275/7743)
Accuracy (if imprecise indeterminate)	64.97%	(243/374)
Imprecise HMM		
Determinacy	95.17%	(7369/7743)
Set-accuracy	93.58%	(350/374)
Single accuracy	95.43%	(7032/7369)
Indeterminate output size	2.97 out of 21 classes	(1112/374)

TABLE 1. Precise vs. imprecise HMMs. Test results obtained by twofold cross-validation on the first two chants of Dante's *Divina Commedia* and $n = 2$. Quantification is achieved by IDM with $s = 2$ and Perks' prior modified as suggested in [28, Section 5.2]. The single-character output by the precise model is then guaranteed to be included in the set of characters the imprecise HMM identifies.

The recognition using our algorithm is fast: it never takes more than one second for each character. Table 1 reports descriptive values for a large set (7743) of simulations, and a comparison with precise model performance. Imprecise HMMs guarantee quite accurate predictions. In contrast with the precise model, there are 'indeterminate' instances for which they do not output a single state. Yet, this happens rarely, and even then we witness a remarkable reduction in the number of undominated states (from the 21 letters of the Italian alphabet to less than 3). Interestingly, the instances for which the imprecise probability model returns more than one state appear to be 'difficult' for the precise probability model: the accuracy of the precise models displays a strong decrease if we focus only on these instances, while the imprecise models here display basically the same performance as for other instances, by returning about three characters instead of a single one.

10. CONCLUSIONS

We have defined imprecise-probability (or credal) trees using Walley's notion of epistemic irrelevance. Credal trees generalise tree-shaped Bayesian nets in two ways: by allowing the parameters of the tree to be imprecisely specified, and moreover by replacing the notion of stochastic independence with that of epistemic irrelevance. Our focusing on epistemic irrelevance is the most original aspect of this work, as this notion has received limited attention so far in the context of credal nets.

We have focused in particular on developing an efficient exact algorithm for updating beliefs on the tree. Like the algorithms developed for precise graphical models, our algorithm works in a distributed fashion by passing messages along the tree. It computes lower and upper conditional previsions (expectations) with a complexity that is linear in the number of nodes in the tree. This is remarkable because until now it was unclear whether an algorithm with the features described above was at all feasible: in fact, epistemic irrelevance is most easily formulated using coherent lower previsions, which have never before been used as such in practical applications of credal nets. Moreover, it is at this point not clear that epistemic irrelevance is as 'well-behaved' as strong independence is with respect to the graphoid axioms for propagation of probability in graphical models [5, 19].[23] Our results therefore appear very encouraging, and seem to have the potential to open up new avenues of research in credal nets.

[23]Unlike credal nets based on strong independence, a credal net based on epistemic irrelevance cannot generally be seen as equivalent with a set of Bayesian nets *with the same graphical structure*: if it were, then all separation properties of Bayesian nets would simply be inherited, and we have seen in Section 7 that such is not the case.

On a more theoretical side, we have also shown that our credal trees satisfy the important rationality requirement of coherence. This has been established under the assumption that the *upper* probability of any possible observation in the tree is positive, which is a very mild requirement. The same assumption also allowed us to show that all inferences made by updating the tree will be coherent with each other as well as with the local uncertainty models in the nodes of the tree.

On the applied side, we have presented an application of the credal tree model to the problem of character recognition, where the parameters of the model are inferred from data. The empirical results are positive, especially because they show that our credal trees are able to make more reliable predictions than their precise-probability counterparts.

Where to go from here? There are many possible avenues for future research.

It would be very useful to be able to extend the algorithm at least to so-called *polytrees*, which are substantially more expressive graphs than trees are. This could be a difficult task to achieve. In fact, updating credal nets based on strong independence is an NP-hard task when the graph is more general than a tree [7]. Similar problems might affect the algorithms for credal nets based on irrelevance.

For applications, it would be very important to develop statistical methods specialised for credal nets under irrelevance that avoid introducing excessive imprecision in the process of inferring probabilities from data. This could be achieved, for instance, by using a single global IDM over the variables of the tree rather than many local ones, as we did in our experiments.

Another research direction could be concerned with trying to strengthen the conclusions that epistemic trees lead to. There might be cases where our Markov condition based on epistemic irrelevance is too weak as a structural assessment. We have discussed situations where this type of Markov condition systematically leads to a dilation of uncertainty when updating beliefs with observations, and indicated that this dilation is related to the (lack of) certain separation properties induced by epistemic irrelevance on a graph. Dilation might not be desirable in some applications, and we could be called upon to strengthen the model in order to rule out such behaviour. One way to address the issue of dilation—but not necessarily the easiest—could consist in adding additional irrelevance statements to the model, other than those derived from the Markov condition. An easier avenue could be based on designing assumptions that together with the Markov condition lead to some stronger separation properties, while not necessarily requiring them to match the common ones used in Bayesian nets.

ACKNOWLEDGEMENTS

Research by De Cooman and Hermans was supported by Flemish BOF project 01107505 and SBO project 060043 of the IWT-Vlaanderen. Research by Antonucci and Zaffalon has been partially supported by the Swiss NSF grants n. 200020-116674/1 and n. 200020-121785/1. This paper has benefited from discussions with Serafín Moral, Fabio G. Cozman and Cassio P. de Campos, and from the generous comments provided by two anonymous referees.

REFERENCES

[1] Alessandro Antonucci, Alessio Benavoli, Marco Zaffalon, Gert de Cooman, and Filip Hermans. Multiple model tracking by imprecise Markov trees. In *Proceedings of the 12th International Conference on Information Fusion (Seattle, WA, USA, July 6–9, 2009)*, pages 1767–1774, 2009.

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

[3] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

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

[5] Fabio G. Cozman and Peter Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1–2):173–195, 2005.

[6] Cassio P. de Campos and Fabio G. Cozman. Inference in credal networks using multilinear programming. In *Proceedings of the Second Starting AI Researcher Symposium*, pages 50–61, Valencia, 2004. IOS Press.
[7] Cassio P. de Campos and Fabio G. Cozman. The inferential complexity of Bayesian and credal networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1313–1318, Edinburgh, 2005.
[8] Cassio P. de Campos and Fabio G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.
[9] Gert de Cooman and Filip Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008.
[10] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, January 2009. arXiv:0801.0980.
[11] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. 2009. In preparation.
[12] Gert de Cooman and Marco Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125, November 2004.
[13] Enrico Fagiuoli and Marco Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106:77–107, 1998.
[14] Timothy Herron, Teddy Seidenfeld, and Larry Wasserman. Divisive conditioning: further results on dilation. *Philosophy of Science*, 64:411–444, 1997.
[15] Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, January 2008.
[16] Enrique Miranda. Updating coherent lower previsions on finite spaces. *Fuzzy Sets and Systems*, 160(9):1286–1307, January 2009.
[17] Enrique Miranda and Gert de Cooman. Marginal extension in the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 46(1):188–225, September 2007.
[18] Enrique Miranda and Marco Zaffalon. Coherence graphs. *Artificial Intelligence*, 173:104–144, 2009.
[19] Serafín Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45:197–214, 2005.
[20] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
[21] Lawrence R. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
[22] Teddy Seidenfeld and Larry Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21:1139–54, 1993.
[23] Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, January 2007.
[24] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
[25] Peter Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.
[26] Peter M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Revised journal version: [27].
[27] Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007. Revised journal version of [26].
[28] Marco Zaffalon. Statistical inference of the naive credal classifier. In Gert de Cooman, Terrence. L. Fine, and Teddy Seidenfeld, editors, *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, Maastricht, 2000.
[29] Marco Zaffalon and Enrique Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821, Jan 2009.

APPENDIX A. PROOFS OF IMPORTANT RESULTS

In this Appendix, we justify formulas (16) and (20), and give proofs for Propositions 3 and 4, and Theorem 5.

Proof of Eqs. (16) *and* (20). Let us define the gambles

$$\varpi_s^\mu := \underline{P}_{\downarrow s}(\phi_s^\mu | X_{m(s)}) \in \mathscr{L}(\mathscr{X}_{m(s)}), \quad s \sqsubset t$$

and, with obvious notations,

$$\overline{\varpi}_s := \overline{P}_{\downarrow s}(\phi_s^\mu | X_{m(s)}) \in \mathscr{L}(\mathscr{X}_{m(s)}), \quad s \not\sqsubseteq t.$$

Let the chain $\uparrow t$ be given by $\{t_1, \ldots, t_r\}$, where $t_1 := \square$, $t_r := t$ and $t_k := m(t_{k+1})$ for $k = 1, \ldots, r-1$. If we apply the recursion equation (8) in $s = t_1$ and take into account the separate

coherence and the strong factorisation of the conditionally independent natural extension $\underline{P}_{\downarrow C(t_1)}(\cdot|X_{t_1})$, we see that

$$\varpi^\mu_{t_1} = \varpi^\mu_\square = \rho_g(\mu) = \underline{P}(\phi^\mu_{t_1}) = \underline{Q}_{t_1}(\psi^\mu_{t_1}), \tag{24}$$

where [provided that $t_1 \neq t$ and therefore $r > 1$]

$$\begin{aligned}
\psi^\mu_{t_1} &:= \underline{P}_{\downarrow C(t_1)}(\phi^\mu_{t_1}|X_{t_1}) = \underline{P}_{\downarrow C(t_1)}\left(g^\mu_{t_1}\prod_{c\in C(t_1)}\phi^\mu_c\middle|X_{t_1}\right) = g^\mu_{t_1}\underline{P}_{\downarrow C(t_1)}\left(\prod_{c\in C(t_1)}\phi^\mu_c\middle|X_{t_1}\right)\\
&= g^\mu_{t_1}\underline{P}_{\downarrow C(t_1)}\left(\underline{P}_{\downarrow t_2}(\phi^\mu_{t_2}|X_{t_1})\prod_{c\in C(t_1)\setminus\{t_2\}}\phi^\mu_c\middle|X_{t_1}\right) = g^\mu_{t_1}\underline{P}_{\downarrow C(t_1)}\left(\varpi^\mu_{t_2}\prod_{c\in C(t_1)\setminus\{t_2\}}\phi^\mu_c\middle|X_{t_1}\right)\\
&= \left[\max\{\varpi^\mu_{t_2},0\}\underline{P}_{\downarrow C(t_1)}\left(\prod_{c\in C(t_1)\setminus\{t_2\}}\phi^\mu_c\middle|X_{t_1}\right) + \min\{\varpi^\mu_{t_2},0\}\overline{P}_{\downarrow C(t_1)}\left(\prod_{c\in C(t_1)\setminus\{t_2\}}\phi^\mu_c\middle|X_{t_1}\right)\right]g^\mu_{t_1}\\
&= \left[\max\{\varpi^\mu_{t_2},0\}\prod_{c\in C(t_1)\setminus\{t_2\}}\underline{\varpi}_c + \min\{\varpi^\mu_{t_2},0\}\prod_{c\in C(t_1)\setminus\{t_2\}}\overline{\varpi}_c\right]g^\mu_{t_1}.
\end{aligned} \tag{25}$$

Similarly, we find that

$$\varpi^\mu_{t_2} = \underline{P}_{\downarrow t_2}(\phi^\mu_{t_2}|X_{t_1}) = \underline{Q}_{t_2}(\psi^\mu_{t_2}|X_{t_1}), \tag{26}$$

where [provided that $t_2 \neq t$ and therefore $r > 2$] in a completely similar way as above

$$\psi^\mu_{t_2} := \underline{P}_{\downarrow C(t_2)}(\phi^\mu_{t_2}|X_{t_2}) = \left[\max\{\varpi^\mu_{t_3},0\}\prod_{c\in C(t_2)\setminus\{t_3\}}\underline{\varpi}_c + \min\{\varpi^\mu_{t_3},0\}\prod_{c\in C(t_2)\setminus\{t_3\}}\overline{\varpi}_c\right]g^\mu_{t_2}. \tag{27}$$

We can go on in this way until we come to $t_r = t$:

$$\varpi^\mu_{t_r} = \underline{P}_{\downarrow t_r}(\phi^\mu_{t_r}|X_{t_{r-1}}) = \underline{Q}_{t_r}(\psi^\mu_{t_r}|X_{t_{r-1}}), \tag{28}$$

where, using the separate coherence and the strong factorisation of the conditionally independent natural extension $\underline{P}_{\downarrow C(t_r)}(\cdot|X_{t_r})$,

$$\begin{aligned}
\psi^\mu_t = \psi^\mu_{t_r} &:= \underline{P}_{\downarrow C(t_r)}(\phi^\mu_{t_r}|X_{t_r}) = \underline{P}_{\downarrow C(t_r)}\left(g^\mu_{t_r}\prod_{c\in C(t_r)}\phi^\mu_c\middle|X_{t_r}\right) = \underline{P}_{\downarrow C(t_r)}\left([g-\mu]\prod_{c\in C(t_r)}\phi^\mu_c\middle|X_{t_r}\right)\\
&= \max\{g-\mu,0\}\underline{P}_{\downarrow C(t_r)}\left(\prod_{c\in C(t_r)}\phi^\mu_c\middle|X_{t_r}\right) + \min\{g-\mu,0\}\overline{P}_{\downarrow C(t_r)}\left(\prod_{c\in C(t_r)}\phi^\mu_c\middle|X_{t_r}\right)\\
&= \max\{g-\mu,0\}\prod_{c\in C(t)}\underline{\varpi}_c + \min\{g-\mu,0\}\prod_{c\in C(t)}\overline{\varpi}_c.
\end{aligned} \tag{29}$$

Clearly, if we can prove that $\overline{\varpi}_s = \overline{\pi}_s$ for all $s \not\sqsubseteq t$, it will follow from the considerations above that also $\varpi^\mu_s = \pi^\mu_s$ for all $s \sqsubseteq t$, and then the proof is complete. This is what we now set out to do. Consider any $s \not\sqsubseteq t$. Then applying the recursion equation (8) and taking into account the separate coherence and the strong factorisation of the conditionally independent natural extension $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$, we see that, provided s is not a terminal node, and with obvious notations,

$$\overline{\varpi}_s = \overline{P}_{\downarrow s}(\phi^\mu_s|X_{m(s)}) = \overline{Q}_s(\underline{\Psi}_s|X_{m(s)}), \tag{30}$$

where

$$\begin{aligned}
\underline{\Psi}_s &:= \overline{P}_{\downarrow C(s)}(\phi^\mu_s|X_s) = \overline{P}_{\downarrow C(s)}\left(g^\mu_s\prod_{c\in C(s)}\phi^\mu_c\middle|X_s\right) = g^\mu_s\overline{P}_{\downarrow C(s)}\left(\prod_{c\in C(s)}\phi^\mu_c\middle|X_s\right)\\
&= g^\mu_s\prod_{c\in C(s)}\overline{P}_{\downarrow c}(\phi^\mu_c|X_s) = g^\mu_s\prod_{c\in C(s)}\overline{\varpi}_c.
\end{aligned} \tag{31}$$

If on the other hand s is a terminal node, then we can use Eq. (9) to find that

$$\overline{\varpi}_s = \overline{P}_{\downarrow s}(\phi^\mu_s|X_{m(s)}) = \overline{Q}_s(\phi^\mu_s|X_{m(s)}) = \overline{Q}_s(g^\mu_s|X_{m(s)}) = \overline{\pi}_s, \tag{32}$$

where the last equality follows from Eq. (15). Now combine Eqs. (30)–(32) and use recursion to complete the proof. □

Proof of Proposition 3. Fix any x_T in $\mathscr{X}_T$. We need to prove that $\overline{P}(\{x_T\}) > 0$ and that $\overline{P}_{\downarrow S}(\{x_{\downarrow S}\}|z_s) > 0$ for all $s \in T^\Diamond$, non-empty $S \subseteq C(s)$ and $z_s \in \mathscr{X}_s$. Our argumentation is similar to a special case of the one in Section 6.2. We use the notation established there, but with in particular $g := \mu + 1$, $t := \square$ and $E := T$. This implies that $g^\mu = \mathbb{I}_{\{x_T\}}$, $g_s^\mu = \mathbb{I}_{\{x_s\}}$ and $\phi_s^\mu = \mathbb{I}_{\{x_{\downarrow s}\}}$. In accordance with Eq. (15), we define the messages $\overline{\pi}_s \in \mathscr{L}(\mathscr{X}_{m(s)})$ and $\overline{\lambda}_s \in \mathscr{L}(\mathscr{X}_s)$ recursively by:

$$\overline{\lambda}_s := \prod_{c \in C(s)} \overline{\pi}_c \text{ and } \overline{\pi}_s := \overline{Q}_s(\mathbb{I}_{\{x_s\}}\overline{\lambda}_s|X_{m(s)}) = \overline{\lambda}_s(x_s)\overline{Q}_s(\{x_s\}|X_{m(s)}), \quad s \in T \tag{33}$$

with, as before by convention $\overline{\lambda}_s := 1$ in all leaves s. The last equality follows from the separate coherence of the local models $\overline{Q}_s(\cdot|X_{m(s)})$ and the fact that all messages $\overline{\pi}_s$ and $\overline{\lambda}_s$ are non-negative. It is clear from the recursion equations (7) and (8) [see also Eq. (16), the proof is completely similar to that of Eqs. (16) and (20) given above] that $\overline{P}_{\downarrow s}(\{x_{\downarrow s}\}|X_{m(s)}) = \overline{\pi}_s$, for all $s \in T$, and that $\overline{P}_{\downarrow C(s)}(\{x_{\downarrow C(s)}\}|X_s) = \overline{\lambda}_s$ for all $s \in T^\Diamond$. Similarly, it follows from Eq. (11), conjugacy and the strong factorisation property of the conditionally independent natural extension that $\overline{P}_{\downarrow S}(\{x_{\downarrow S}\}|X_s) = \prod_{c \in S} \overline{\pi}_c$ for all $s \in T^\Diamond$ and all non-empty $S \subseteq C(s)$. So we have to prove that all values (all components) of all messages $\overline{\pi}_s$, $s \in T$ are (strictly) positive. This follows at once from the recursion equation (33) and the assumed strict positivity of the local models $\overline{Q}_s(\cdot|X_{m(s)})$. □

Proof of Proposition 4. Our argumentation is similar to a special case of the one in Section 6.2. We also use notation similar to that established there, with in particular $g := \mu + 1$ and $t := \square$. In accordance with Eq. (15), we define the messages $\overline{\pi}_s \in \mathscr{L}(\mathscr{X}_{m(s)})$ and $\overline{\lambda}_s \in \mathscr{L}(\mathscr{X}_s)$ recursively by:

$$\overline{\lambda}_s := \prod_{c \in C(s)} \overline{\pi}_c, \quad s \in T \tag{34}$$

and

$$\overline{\pi}_s := \begin{cases} \overline{\lambda}_s(x_s)\overline{Q}_s(\{x_s\}|X_{m(s)}) & \text{if } s \in E \\ \overline{Q}_s(\overline{\lambda}_s|X_{m(s)}) & \text{if } s \in T \setminus E. \end{cases} \tag{35}$$

with, as before by convention $\overline{\lambda}_s := 1$ in all leaves s. All these messages are non-negative by construction. It is clear from the recursion equations (7) and (8) [see also Eq. (16), the proof is completely similar to that of Eqs. (16) and (20) given above] that $\overline{P}_{\downarrow s}(\{x_{E \cap \downarrow s}\}|X_{m(s)}) = \overline{\pi}_s$ for all $s \in T$. Now it follows from the recursion equations (34) and (35) and the assumption $\overline{P}(\{x_E\}) = \overline{\pi}_\square > 0$ that $\overline{\lambda}_e(x_e) > 0$ for all $e \in E$. Again applying Eq. (34), we find that indeed $\overline{\pi}_c(x_e) > 0$ for all $c \in C(e)$. □

Our proof of Theorem 5 relies heavily on a very convenient coherence result proved by Enrique Miranda [16, Theorem 6], which we relate here to make the paper more self-contained. We use the notations established in the context of Section 3.

Theorem 6. *Let $\underline{P}$ be a (separately) coherent lower prevision on $\mathscr{L}(\mathscr{X}_N)$, and consider m disjoint pairs of subsets O_k and I_k of N, $k = 1, \ldots, m$. Assume that $\overline{P}(\{x_{I_k}\}) > 0$ for all $x_{I_k} \in \mathscr{X}_{I_k}$, $k = 1, \ldots, m$ and use regular extension to define the conditional lower previsions $\underline{R}(\cdot|X_{I_k})$ on $\mathscr{L}(\mathscr{X}_{O_k})$, for $k = 1, \ldots, m$. Then the (conditional) lower previsions $\underline{P}$, $\underline{R}(\cdot|X_{I_1})$, $\ldots$, $\underline{R}(\cdot|X_{I_m})$ are (jointly) coherent.*

Proof of Theorem 5. We begin by showing that the family of models $\mathscr{T}(\underline{P})$ satisfies requirements T1–T3.

To prove T1, consider any $s \in T$, and any $f \in \mathscr{L}(\mathscr{X}_s)$, then it follows from separate coherence that $\underline{P}_{\downarrow C(s)}(f|X_s) = f$, and therefore we infer from the recursion equation (8) that indeed $\underline{P}_{\downarrow s}(f|X_{m(s)}) = \underline{Q}_s(\underline{P}_{\downarrow C(s)}(f|X_s)|X_{m(s)}) = \underline{Q}_s(f|X_{m(s)})$.

Next, we turn to the proof of T2 and T3. Consider any $s \in T^{\Diamond}$, $S \subseteq C(s)$ and $R \subseteq \overline{S}$. Let $x_{\{s\}\cup R} \in \mathscr{X}_{\{s\}\cup R}$ and $f \in \mathscr{L}(\mathscr{X}_{\downarrow S\cup\{s\}\cup R})$. We calculate the following regular extension of the joint:

$$\underline{R}(f|x_{\{s\}\cup R}) = \max\{\mu \in \mathbb{R} \colon \underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) \geq 0\}. \tag{36}$$

Consider that

$$\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu] = \mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_R\}}[g-\mu],$$

where $g := f(\cdot, x_s, x_R) \in \mathscr{L}(\mathscr{X}_{\downarrow S})$. Let t_2 be the unique child of $t_1 := \square$ such that $s \in \downarrow t_2$ [assuming of course that $s \neq t_1 = \square$]. By using separate coherence, recursion equations (7), (8) and (10), and the strong factorisation property [see Proposition 1] of the conditionally independent natural extension, in a way similar to the argumentation in Section 6.2, we see that

$$\begin{aligned}
\underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) &= \underline{Q}_{t_1}(\underline{P}_{\downarrow C(t_1)}(\mathbb{I}_{\{x_{R\setminus\downarrow t_2}\}}\underline{P}_{\downarrow C(t_1)}(\mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_{R\cap\downarrow t_2}\}}[g-\mu]|X_{t_1})|X_{t_1})) \\
&= \underline{Q}_{t_1}(\underline{P}_{\downarrow C(t_1)}(h_2\underline{P}_{\downarrow t_2}(\mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_{R\cap\downarrow t_2}\}}[g-\mu]|X_{t_1})|X_{t_1})) \\
&= \underline{P}_{\downarrow t_1}(h_2\underline{P}_{\downarrow t_2}(\mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_{R\cap\downarrow t_2}\}}[g-\mu]|X_{t_1})),
\end{aligned}$$

where $h_2 := \mathbb{I}_{\{x_{R\setminus\downarrow t_2}\}} \geq 0$. Similarly, let t_3 be the unique child of t_2 such that $s \in \downarrow t_3$ [assuming of course that $s \neq t_2$]. Then we see in the same way as above that

$$\underline{P}_{\downarrow t_2}(\mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_{R\cap\downarrow t_2}\}}[g-\mu]|X_{t_1}) = \underline{P}_{\downarrow t_2}(h_3\underline{P}_{\downarrow t_3}(\mathbb{I}_{\{x_s\}}\mathbb{I}_{\{x_{R\cap\downarrow t_3}\}}[g-\mu]|X_{t_2})|X_{t_1}),$$

where $h_3 := \mathbb{I}_{\{x_{R\setminus\downarrow t_3}\}} \geq 0$. Continuing in this way, we eventually come to the conclusion that

$$\underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) = \underline{G}(\underline{P}_{\downarrow C(s)}(h\mathbb{I}_{\{x_s\}}[g-\mu]|X_s)), \tag{37}$$

where $h := \mathbb{I}_{\{x_{R\cap\downarrow C(s)}\}} = \mathbb{I}_{\{x_{R\cap(\downarrow C(s)\setminus\downarrow S)}\}}$, and where the real functional $\underline{G}$ on $\mathscr{L}(\mathscr{X}_s)$ is essentially constructed as follows. Consider the segment $t_1t_2\ldots t_{r-1}t_r$ connecting $\square$ and s, i.e. $t_r := s, t_{r-1} := m(t_r), \ldots, t_k := m(t_{k+1}), \ldots, t_1 := m(t_2) = \square$. Then there are non-negative h_k on $\mathscr{X}_{\downarrow t_k}$ such that for all $f \in \mathscr{L}(\mathscr{X}_s)$, $\underline{G}(f) = f_1$, where $f_r := f$ and $f_k := \underline{P}_{\downarrow t_{k+1}}(h_{k+1}f_{k+1}|X_{t_k})$, $k = 1, \ldots, r-1$. [If $r = 1$, or in other words $s = t_1 = \square$, just let $\underline{G} := \underline{P}_{\downarrow\square}$.] In other words, the functional $\underline{G}$ results from recursively multiplying with non-negative maps and applying global conditional lower previsions. As such, $\underline{G}$ is non-negatively homogeneous and superadditive [because the successive multiplication and composition preserves these properties]. In addition, it does not depend on g nor μ. If we use the separate coherence of $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$, the strong factorisation, associativity and marginalisation properties of the conditionally independent natural extension $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ [see Proposition 1, Eqs. (4) and (5), and the recursion equations (7) and (10)], and the separate coherence of the conditional lower prevision $\underline{P}_{\downarrow S}(\cdot|x_s)$, we get:

$$\begin{aligned}
\underline{P}_{\downarrow C(s)}(h\mathbb{I}_{\{x_s\}}[g-\mu]|X_s) &= \mathbb{I}_{\{x_s\}}\underline{P}_{\downarrow C(s)}(h[g-\mu]|x_s) \\
&= \mathbb{I}_{\{x_s\}}\begin{cases}\underline{P}_{\downarrow C(s)}(h|x_s)[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \geq \mu \\ \overline{P}_{\downarrow C(s)}(h|x_s)[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \leq \mu.\end{cases}
\end{aligned} \tag{38}$$

Combining Eqs. (37) and (38), and invoking the non-negative homogeneity of the real functional $\underline{G}$, this leads to:

$$\underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) = \begin{cases}\underline{G}(\mathbb{I}_{\{x_s\}})\underline{P}_{\downarrow C(s)}(h|x_s)[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \geq \mu \\ \overline{G}(\mathbb{I}_{\{x_s\}})\overline{P}_{\downarrow C(s)}(h|x_s)[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \leq \mu,\end{cases}$$

where we let $\overline{G}(\mathbb{I}_{\{x_s\}}) := -\underline{G}(-\mathbb{I}_{\{x_s\}})$. By letting $f = \mu \pm 1$ [and therefore also $g = \mu \pm 1$] in this expression, we derive in particular that

$$\underline{P}(\{x_{\{s\}\cup R}\}) = \underline{G}(\mathbb{I}_{\{x_s\}})\underline{P}_{\downarrow C(s)}(h|x_s) \text{ and } \overline{P}(\{x_{\{s\}\cup R}\}) = \overline{G}(\mathbb{I}_{\{x_s\}})\overline{P}_{\downarrow C(s)}(h|x_s),$$

and therefore also

$$\underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) = \begin{cases} \underline{P}(\{x_{\{s\}\cup R}\})[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \geq \mu \\ \overline{P}(\{x_{\{s\}\cup R}\})[\underline{P}_{\downarrow S}(g|x_s)-\mu] & \text{if } \underline{P}_{\downarrow S}(g|x_s) \leq \mu. \end{cases}$$

Since we have assumed that all local models $\overline{Q}_s(\cdot|X_{m(s)})$ are strictly positive, we gather from Proposition 3 that $\overline{P}(\{x_{\{s\}\cup R}\}) > 0$, and therefore

$$\underline{P}(\mathbb{I}_{\{x_{\{s\}\cup R}\}}[f-\mu]) \geq 0 \Leftrightarrow \underline{P}_{\downarrow S}(g|x_s) \geq \mu.$$

This allows us to infer from Eq. (36) that

$$\underline{R}(f|x_{\{s\}\cup R}) = \underline{P}_{\downarrow S}(f(\cdot,x_s,x_R)|x_s) \text{ for all } f \in \mathscr{L}(\mathscr{X}_{\downarrow S\cup\{s\}\cup R}) \text{ and } x_{\{s\}\cup R} \in \mathscr{X}_{\{s\}\cup R}. \quad (39)$$

If we now combine Eq. (39) with Theorem 6, we see that both T2 and T3 hold.

To complete the proof, we consider T4. Consider any family of models $\mathscr{T}(\underline{V})$ that satisfies conditions T1–T3. Then we want to show that

$$\underline{V}_{\downarrow S}(\cdot|X_t) \geq \underline{P}_{\downarrow S}(\cdot|X_t) \text{ for all } t \in T^{\Diamond} \text{ and all non-empty } S \subseteq C(s) \quad (40)$$

and

$$\underline{V} \geq \underline{P}. \quad (41)$$

The proof proceeds in a recursive (inductive) fashion. Since the $\underline{V}_{\downarrow t}(\cdot|X_{m(t)})$ satisfy T1, we infer in particular that

$$\underline{V}_{\downarrow t}(\cdot|X_{m(t)}) = \underline{P}_{\downarrow t}(\cdot|X_{m(t)}) = \underline{Q}_t(\cdot|X_{m(t)}) \text{ for all terminal nodes } t.$$

It is therefore clearly sufficient to prove the following statement for all non-terminal nodes $t \in T^{\Diamond}$:

$$(\forall c \in C(t))(\underline{V}_{\downarrow c}(\cdot|X_t) \geq \underline{P}_{\downarrow c}(\cdot|X_t)) \Rightarrow \begin{cases} \underline{V}_{\downarrow S}(\cdot|X_t) \geq \underline{P}_{\downarrow S}(\cdot|X_t) \text{ for all non-empty } S \subseteq C(t) \\ \underline{V}_{\downarrow t}(\cdot|X_{m(t)}) \geq \underline{P}_{\downarrow t}(\cdot|X_{m(t)}). \end{cases} \quad (42)$$

This is what we now set out to do. Fix any non-terminal node $t \in T^{\Diamond}$ and any non-empty $S \subseteq C(t)$, and assume that $\underline{V}_{\downarrow c}(\cdot|X_t) \geq \underline{P}_{\downarrow c}(\cdot|X_t)$ for all $c \in C(t)$.

First of all, define for any disjoint proper subsets I and O of S, the conditional lower previsions $\underline{V}_{\downarrow O}(\cdot|X_{\{t\}\cup\downarrow I})$ through:

$$\underline{V}_{\downarrow O}(f|x_{\{t\}\cup\downarrow I}) = \underline{V}_{\downarrow O}(f(\cdot,x_{\downarrow I})|x_t) \text{ for all } f \in \mathscr{L}(\mathscr{X}_{\downarrow O\cup\downarrow I}) \text{ and all } x_{\{t\}\cup\downarrow I} \in \mathscr{X}_{\{t\}\cup\downarrow I}.$$

Then we infer from T3 [with $S = O$, $s = t$ and $R = \downarrow I \subseteq \overline{O}$] that all these conditional lower previsions are in particular (jointly) coherent with the conditional lower prevision $\underline{V}_{\downarrow S}(\cdot|X_t)$. If we recall Definition 3 [with $N = \{\downarrow c\colon c \in S\}$ and $Y = X_t$], we conclude that $\underline{V}_{\downarrow C(t)}(\cdot|X_t)$ is a conditionally independent product of the 'marginals' $\underline{V}_{\downarrow c}(\cdot|X_t)$, $c \in S$, which therefore dominates the smallest such independent product:

$$\underline{V}_{\downarrow S}(\cdot|X_t) \geq \otimes_{c\in S}\underline{V}_{\downarrow c}(\cdot|X_t)$$

and therefore, using the assumption, we infer from this inequality that

$$\underline{V}_{\downarrow S}(\cdot|X_t) \geq \otimes_{c\in S}\underline{V}_{\downarrow c}(\cdot|X_t) \geq \otimes_{c\in S}\underline{P}_{\downarrow c}(\cdot|X_t) = \underline{P}_{\downarrow S}(\cdot|X_t), \quad (43)$$

where we have also used, successively, the monotonicity property of the conditionally independent natural extension [see [11] for a proof] and the recursion equations (7) and (11).

Next, define the conditional lower prevision $\underline{V}_{\downarrow C(t)}(\cdot|X_{\{m(t),t\}})$ on $\mathscr{L}(\mathscr{X}_{\{m(t)\}\cup\downarrow t})$ through:

$$\underline{V}_{\downarrow C(t)}(f|x_{\{m(t),t\}}) := \underline{V}_{\downarrow C(t)}(f(x_{m(t)},x_t,\cdot)|x_t) \quad \text{for all } f \in \mathscr{L}(\mathscr{X}_{\{m(t)\}\cup\downarrow t}) \text{ and all } x_{\{m(t),t\}} \in \mathscr{X}_{\{m(t),t\}}. \quad (44)$$

Then we infer from T3 [with $s=t$, $S=C(t)$ and $R=\{m(t)\}\subseteq\overline{C(t)}$] that this conditional lower prevision $\underline{V}_{\downarrow C(t)}(\cdot|X_{\{m(t),t\}})$ is in particular (jointly) coherent with the conditional lower prevision $\underline{V}_{\downarrow t}(\cdot|X_{m(t)})$ defined on $\mathscr{L}(\mathscr{X}_{\{m(t)\}\cup\downarrow t})$. We then see that for all $g\in\mathscr{L}(\mathscr{X}_{\downarrow t})$:

$$\begin{aligned}\underline{V}_{\downarrow t}(g|X_{m(t)}) &\geq \underline{V}_{\downarrow t}(\underline{V}_{\downarrow C(t)}(g|X_{\{m(t),t\}})|X_{m(t)})\\ &= \underline{V}_{\downarrow t}(\underline{V}_{\downarrow C(t)}(g|X_t)|X_{m(t)}) = \underline{Q}_t(\underline{V}_{\downarrow C(t)}(g|X_t)|X_{m(t)})\\ &\geq \underline{Q}_t(\underline{P}_{\downarrow C(t)}(g|X_t)|X_{m(t)})\\ &= \underline{P}_{\downarrow t}(\cdot|X_{m(t)}).\end{aligned}$$

The first equality follows from Eq. (44), the second one holds because the global models $\underline{V}_{\downarrow t}(\cdot|X_{m(t)})$ satisfy T1, and the third one follows from recursion equation (8). The first inequality follows if we apply Walley's Marginal Extension Theorem[24] [24, Theorem 6.7.2] in the formulation of [17, Theorem 4]. The second inequality follows from the inequality (43) and the non-decreasing character of $\underline{Q}_t(\cdot|X_{m(t)})$, which follows from separate coherence. This completes our proof that T4 is also satisfied.

The last part of the proof follows at once from Eqs. (39) [with $R=\emptyset$], and Theorem 6. □

E-mail address: {gert.decooman,filip.hermans}@UGent.be

GHENT UNIVERSITY, SYSTEMS RESEARCH GROUP, TECHNOLOGIEPARK 914, 9052 ZWIJNAARDE, BELGIUM.

E-mail address: {alessandro,zaffalon}@idsia.ch

IDSIA, GALLERIA 2, 6928 MANNO (LUGANO), SWITZERLAND

[24]Recall that this is a coherence result that generalises the so-called Law of Iterated Expectations to coherent lower previsions.