6th International Symposium on Imprecise Probability: Theories and Applications, Durham, United Kingdom, 2009

Epistemic irrelevance in credal networks: the case of imprecise Markov trees

Gert de Cooman and Filip Hermans
SYSTeMS, Ghent University, Belgium
{gert.decooman,filip.hermans}@ugent.be

Alessandro Antonucci and Marco Zaffalon
IDSIA, Switzerland
{alessandro,zaffalon}@idsia.ch

Abstract

We replace strong independence in credal networks with the weaker notion of epistemic irrelevance. Focusing on directed trees, we show how to combine local credal sets 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 essentially linear in the number of nodes, is formulated entirely in terms of coherent lower previsions. We supply examples of the algorithm's operation, and report an application to on-line character recognition that illustrates the advantages of our model for prediction.

Keywords. Coherence, credal network, epistemic irrelevance, epistemic independence, strong independence, imprecise Markov tree, separation, hidden Markov chain.

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, defense, and environment, to name just a few.

Much has been done also on the front of imprecise probability. *Credal networks* [3] have been and still are the subject of intense research. A credal network 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 [6]). 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 with the *sensitivity analysis* interpretation of credal sets, which regards an imprecise model as arising out of partial ignorance of a precise one. This is a somewhat narrow view, and it does not apply in general.

An alternative and attractive way to express irrelevance that is not committed to the sensitivity analysis interpretation is offered by *epistemic irrelevance* [15]: we say that X is irrelevant to Y if observing X does not affect beliefs about Y. Epistemic irrelevance is defined directly in terms of a subject's beliefs and is therefore very well suited for a behavioural theory of imprecise probability. It is also weaker than strong independence, and it therefore does not lead to overconfident inferences when the sensitivity analysis interpretation is not justified.

At this point the question that 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 only 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 [7] and imprecise Markov chains [8].

How do we address this problem? After giving some preliminary notions and introducing the model in Sec. 2, we discuss in Sec. 3 how to combine marginal models into joint ones reflecting certain irrelevance assessments, in a way that is as conservative as possible. We comment on the graphical separation criteria induced by epistemic irrelevance in Sec. 5. We then go on to develop and justify an inference algorithm for treating the model as an expert system in Sec. 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, that are called *instantiated*, meaning that their value is determined. It is based on message passing, as are the traditional algorithms that have been developed for precise graphical models, and it has some remarkable properties: (i) it works in time essentially linear in the size of the tree; (ii) it natively computes posterior lower and upper

previsions (or expectations) rather than probabilities; (iii) it is an algorithm for credal nets developed for the first time exclusively using the formalism of *coherent lower previsions* [15]; and (iv) it is shown to lead to coherent inferences under mild conditions. We give a step-by-step example of the way inferences can be done in our framework in Sec. 7, where we also comment on the intriguing relationship between the failure of certain classical separation properties in our framework, and dilation [10, 14]. The last part of the paper focuses on numerical simulations. In Sec. 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 [7] that there is no difference between inferences that go forward from instantiated nodes to target under strong independence and epistemic irrelevance. In Sec. 9 we present an application of our algorithm to on-line character recognition. We learn the probabilities from data and compare the predictions of the 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.

Due to lack of space, we must assume the reader has a working knowledge of the basics of Walley's [15] theory of coherent lower previsions. We also refrain from giving proofs of technical results for the same reason, and rather stress motivation, simple justifications and examples.

2 Credal trees under epistemic irrelevance

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 $\square$. Consider any node s, then we denote the set of its parents by $P(s)$. Of course, $P(\square)=\emptyset$, and for $s\neq\square$ we have that $P(s)=\{m(s)\}$ where $m(s)$ is the *mother node* of s. 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*.

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 set 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 $\uparrow s:=A(s)\cup\{s\}$, $\downarrow s:=D(s)\cup\{s\}$, $\uparrow S:=\bigcup\{\uparrow s: s\in S\}$ and $\downarrow S:=\bigcup\{\downarrow s: s\in S\}$ for any subset $S\subseteq T$.

With each node s of the tree, there is associated a variable X_s assuming values in a finite non-empty set $\mathscr{X}_s$. We denote the set of all real-valued maps (*gambles*) on $\mathscr{X}_s$ by $\mathscr{L}(\mathscr{X}_s)$. 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 $\mathscr{L}(\mathscr{X}_S)$. 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.

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 coherent lower prevision on $\mathscr{L}(\mathscr{X}_\square)$. We use the common generic notation $\underline{Q}_s(\cdot|X_{P(s)})$ for all these local models.

Global uncertainty models. In this and the following two sections, we show how all these local models $\underline{Q}_s(\cdot|X_{m(s)})$ can be combined into *global uncertainty models*. If we generically denote by the symbol $\underline{P}_s$ lower previsions on $\mathscr{L}(\mathscr{X}_{\downarrow s})$, representing information about $X_{\downarrow s}$, then this means we want to end up with an unconditional joint lower prevision $\underline{P}:=\underline{P}_\square$ on $\mathscr{L}(\mathscr{X}_T)$ for all variables in the tree, as well conditional lower previsions $\underline{P}_s(\cdot|X_{m(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$ for all non-initial nodes s. Ideally, we want these global (conditional) lower previsions to be coherent with one another, and 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 to be as conservative (small) as possible.

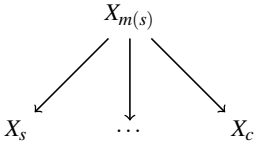
The interpretation of the graphical model. Consider any node s in the tree, and its parent set $P(s)$ [either empty or equal to $\{m(s)\}$]. We also consider the set $\overline{s}:=T\setminus[D(s)\cup P(s)]$ of its non-parent non-descendants. Then *conditional on the parent variables $X_{P(s)}$, the non-parent non-descendant variables $X_{\overline{s}}$ are assumed to be epistemically irrelevant to the variables $X_{\downarrow s}$ associated with s and its descendants.* This interpretation turns the tree into a *credal tree under epistemic irrelevance*, and we shall also use the term *imprecise Markov tree* (IMT) for it.

In terms of the global models, this means that for all $s\in T$, for all $S\subseteq\overline{s}$ and for all $z_{S\cup P(s)}\in\mathscr{X}_{S\cup P(s)}$:

$$\underline{P}_s(\cdot|z_{P(s)})=\underline{P}_s(\cdot|z_{S\cup P(s)}). \tag{1}$$

We discuss the separation properties that accompany this interpretation in some detail in Sec. 5. For now, we focus on one immediate consequence that will help

us go from local to global models in Sec. 4. Consider some non-initial node s. The interpretation of the graphical structure of the tree tells us that for each *sibling* $c \in S(s)$ of s, the variable X_c is epistemically irrelevant to the variable X_s, conditional on $X_{m(s)}$. It even tells us that for any non-empty set $S \subseteq S(s)$ of siblings of s, the variable X_S is epistemically irrelevant to X_s, conditional on $X_{m(s)}$. We conclude that all children of a node are not just epistemically irrelevant to each other: they are even epistemically independent [15, Chapter 9], in some very specific sense.



3 Net-independent natural extension

This leads us to the following small digression. We consider the following problem, the solution of which will help us in our discussion further on. Suppose we have a number of *marginal* lower previsions $\underline{P}_n$ representing beliefs about the values that each of a finite number of (logically independent) variables X_n assume in the respective finite sets $\mathscr{X}_n$, $n \in N$, where N is some finite set.

Net-independent products. We now 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 joint reflects the following structural assessments: for each $o \in N$ and each non-empty $I \subseteq N\setminus\{o\}$, the variables X_I are epistemically irrelevant to the variable X_o. In other words, learning the value of any number of these variables does not affect beliefs about any single other variable amongst them. We then call the variables X_n, $n \in N$ *net-independent*.

Such irrelevance assessments are useful because they allow us to turn marginal into conditional lower previsions. Indeed, for each $o \in N$ and each $I \subseteq N\setminus\{o\}$ we can use the epistemic irrelevance of X_I to X_o to infer from the marginal lower prevision $\underline{P}_o$ a conditional lower prevision $\underline{P}_o(\cdot|X_I)$ on $\mathscr{L}(\mathscr{X}_o)$ given by:

$$\underline{P}_o(h|x_I) := \underline{P}_o(h) \text{ for all gambles } h \text{ on } \mathscr{X}_o.$$

So we can use the assessment of net-independence of the variables X_n, $n \in N$ to infer from the marginals a family of conditional lower previsions:

$$\mathscr{N}(\underline{P}_n, n\in N) := \{\underline{P}_o(\cdot|X_I)\colon o \in N \text{ and } I \subseteq N\setminus\{o\}\}.$$

Definition 1. *A coherent joint 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 $\mathscr{N}(\underline{P}_n, n\in N)$ is called a* net-independent product *of these marginals. If it exists, then the point-wise smallest such net-independent product is called the* net-independent natural extension *of these marginals, and denoted by $\otimes_{n\in N}\underline{P}_n$.*

Conditioning factorising lower previsions. The following notion of factorisation is intimately linked with that of a net-independent product. It will also play a crucial part in our development of an algorithm for treating an imprecise Markov tree as an expert system.

Definition 2. *We call a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$* factorising *if for all $o \in N$ and all non-empty $I \subseteq N\setminus\{o\}$, all $g \in \mathscr{L}(\mathscr{X}_o)$ and all non-negative $f_i \in \mathscr{L}(\mathscr{X}_i)$, $i \in I$, $\underline{P}_N(fg) = \underline{P}_N(f\underline{P}_N(g))$, where $f := \prod_{i\in I} f_i$.*

As an important example, the so-called *strong product* [3] $\times_{n\in N}\underline{P}_n$ of the marginal lower previsions $\underline{P}_n$ is factorising. But for any coherent factorising joint lower prevision $\underline{P}_N$, we see that for any non-empty subset I of N:

$$\underline{P}_N(\times_{i\in I}A_i) = \prod_{i\in I}\underline{P}_N(A_i) \text{ and } \overline{P}_N(\times_{i\in I}A_i) = \prod_{i\in I}\overline{P}_N(A_i), \tag{2}$$

where $A_i \subseteq \mathscr{X}_i$ for all $i \in I$. Let us call any real functional Φ on $\mathscr{L}(\mathscr{X})$ *strictly positive* if $\Phi(I_{\{x\}}) > 0$ for all $x \in \mathscr{X}$. Then the following result is immediate from Eq. (2).

Proposition 1. *A factorising coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ is strictly positive if and only if all its marginals are, and its conjugate upper prevision $\overline{P}_N$ is strictly positive if and only if all its marginals are.*

As a next step, suppose we want to condition a coherent and factorising joint $\underline{P}_N$ on an observation $X_I = x_I$, where I is some proper subset of N. To this end, we calculate the *regular extension* [15, Appendix J]: when $\overline{P}_N(I_{\{x_I\}}) > 0$,

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

where h is any gamble on $\mathscr{X}_O$ and O is any non-empty subset of $N\setminus I$. Otherwise $\underline{R}(\cdot|x_I)$ is vacuous. Then because $\underline{P}_N$ is factorising:

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

so we conclude that, quite interestingly,

$$\underline{R}(h|x_I) = \underline{P}_N(h) \text{ as soon as } \overline{P}_N(\{x_I\}) > 0. \tag{3}$$

Because we are working in a finitary context [$\mathscr{X}_N$ is a finite set], the regular extension $\underline{R}(\cdot|X_I)$ is guaranteed to be coherent with the joint lower prevision $\underline{P}_N$ [15, Sec. J3]. This, together with an interesting recent coherence result by Enrique Miranda [11, Theorem 5], leads us to the following conclusion.

Proposition 2. *Any coherent joint lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ that is factorising and strictly positive,*[1] *is a net-independent product of its marginals.*

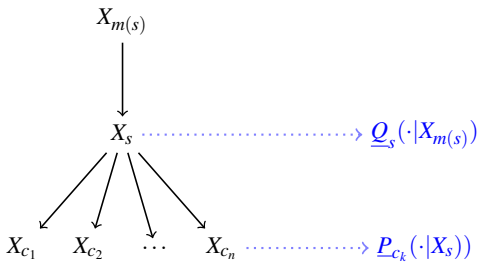
As an immediate consequence, the strong product $\times_{n\in N}\underline{P}_n$ of a collection of strictly positive marginals $\underline{P}_n$, $n \in N$, is also a net-independent product of these marginals, and is therefore coherent with the associated family of conditional lower previsions $\mathscr{N}(\underline{P}_n, n \in N)$. So this family is itself always guaranteed to be coherent, and because all the sets $\mathscr{X}_n$ are finite, we can invoke Walley's Finite Extension Theorem [15, Theorem 8.1.9] to conclude that there always is a point-wise smallest joint lower prevision that is coherent with the family $\mathscr{N}(\underline{P}_n, n \in N)$. This provides the most important step in the proof of the following result. Another crucial step is provided by the fact that, since the strong product is a net-independent product of the marginals $\underline{P}_n$, $n \in N$, it has to dominate the net-independent natural extension: $\times_{n\in N}\underline{P}_n \geq \otimes_{n\in N}\underline{P}_n$.

Proposition 3. *For any collection of strictly positive and coherent marginal lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$, their net-independent natural extension $\otimes_{n\in N}\underline{P}_n$ exists, and it is a factorising and strictly positive coherent lower prevision on $\mathscr{L}(\mathscr{X}_N)$.*

4 Constructing the most conservative joint

We now 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 (1) encoded in the imprecise Markov tree. In the next section, we will use these global models to construct and justify an algorithm for treating the imprecise Markov tree as an expert system.

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 in a recursive manner, following the same pattern. Consider a node s and suppose that, in each of its children $c \in C(s)$, we already have a global conditional lower prevision $\underline{P}_c(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow c})$. We construct a global conditional lower prevision $\underline{P}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$ by backwards recursion:

$$\underline{P}_s(\cdot|X_s) := \otimes_{c\in C(s)}\underline{P}_c(\cdot|X_s) \tag{4}$$

$$\begin{aligned}\underline{P}_s(\cdot|X_{P(s)}) &:= \underline{Q}_s(\underline{P}_s(\cdot|X_s)|X_{P(s)}) \\ &= \underline{Q}_s(\otimes_{c\in C(s)}\underline{P}_c(\cdot|X_s)|X_{P(s)}),\end{aligned} \tag{5}$$

the conditional lower prevision $\underline{P}_s(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow C(s)})$ being the net-independent natural extension of the conditional lower previsions $\underline{P}_c(\cdot|X_s)$ on $\mathscr{L}(\mathscr{X}_{\downarrow c})$, $c \in C(s)$. If we start in leaves t with the 'boundary condition'

$$\underline{P}_t(\cdot|X_{P(t)}) := \underline{Q}_t(\cdot|X_{P(t)}) \text{ for all leaves } t, \tag{6}$$

then the recursion relations (4) and (5) eventually lead to a global model $\underline{P}_s(\cdot|X_{m(s)})$ in all nodes s of the tree, and in particular to a joint model $\underline{P} := \underline{P}_\square$ on $\mathscr{L}(\mathscr{X}_T)$. These are the global (conditional) lower previsions we have been looking for, as the following theorem tells us. Its proof proceeds in a recursive fashion, similar to the construction of the global models. It relies rather heavily on the fact that the net-independent natural extension is factorising, and on the coherence result by Miranda [11, Theorem 5], already mentioned before Proposition 2.

Theorem 4. *If all local models $\underline{Q}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_s)$, $s \in T$ are strictly positive, then the global models $\underline{P}_s(\cdot|X_{P(s)})$ on $\mathscr{L}(\mathscr{X}_{\downarrow s})$, $s \in T$ obtained through Eqs.* (4)–(6)*, constitute the point-wise smallest coherent family of (conditional) lower previsions that (i) extend the local models, and (ii) satisfy the epistemic irrelevance conditions* (1) *encoded in the graphical structure.*

5 Some separation properties

Without going into too much detail, we would like to point out one of the more striking differences between the separation properties in imprecise Markov trees under epistemic irrelevance, and the more usual ones for Bayesian nets [12] and credal nets under strong independence [3].

It is clear from the interpretation of the graphical model described in Sec. 2 that we have the following simple separation results:

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.

[1] We strongly suspect that this proposition, and a number of further results that build on it, such as Proposition 3, can be extended to the case that not $\underline{P}_N$ but $\overline{P}_N$ is strictly positive. We have no proof yet, however.

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:

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

where X_{i_2} *does not necessarily separate* X_t *from* X_{i_1}. We will come across a simple counterexample in Sec. 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 [2], 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 Algorithm for treating the imprecise Markov tree as an expert system

We now consider the case where the imprecise Markov tree is treated as an expert system: 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*.

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 '$X_E = x_E$'. 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* [15, Appendix J]:

$$\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R} \colon \underline{P}(I_{\{x_E\}}[g-\mu]) \geq 0\} \quad (7)$$

for all gambles g on $\mathscr{X}_t$, assuming that $\overline{P}(\{x_E\}) > 0$. Consider the map $\rho_g \colon \mathbb{R} \to \mathbb{R} \colon \mu \mapsto \underline{P}(I_{\{x_E\}}[g-\mu])$. By coherence of $\underline{P}$, $|\rho_g(\mu_1) - \rho_g(\mu_2)| \leq |\mu_1 - \mu_2|\overline{P}(\{x_E\})$, which implies that ρ_g is continuous. 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 (7) is indeed a maximum. $\underline{R}_t(g|x_E)$ is the right-most zero of ρ_g, and it is, again by coherence of $\underline{P}$, guaranteed to lie between $\inf g$ and $\sup g$. If moreover $\underline{P}(\{x_E\}) > 0$, then it is the unique zero. 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 at least check their signs, efficiently.

Calculating the values of ρ_g recursively. Recall that the joint $\underline{P}$ can be constructed recursively from leaves to root. The idea we now use is that calculating $\rho_g(\mu) = \underline{P}(I_{\{x_E\}}[g-\mu])$ becomes easier if we graft the structure of the tree onto the argument $g^\mu := I_{\{x_E\}}[g-\mu]$ as follows. Define $g^\mu_e := I_{\{x_e\}}$ for all $e \in E$, $g^\mu_t := g - \mu$, and $g^\mu_s := 1$ for $s \in T \setminus (E \cup \{t\})$, whence $g^\mu = \prod_{s \in T} g^\mu_s$. 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 $\phi^\mu_\square = g^\mu$, $\phi^\mu_s \geq 0$ if $s \not\sqsupseteq t$, and for any $s \in T$:

$$\phi^\mu_s = g^\mu_s \prod_{c \in C(s)} \phi^\mu_c, \quad (8)$$

where we use the convention that $\prod_{u \in \emptyset} \alpha_u = 1$. Eq. (8) is the argument counterpart of Eq. (5). Also, if $s \not\sqsupseteq t$ then g^μ_s and ϕ^μ_s do not depend on μ, nor on g.

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

$$\underline{\pi}_s := \underline{Q}_s\left(g^\mu_s \prod_{c \in C(s)} \underline{\pi}_c | X_{m(s)}\right) \quad \overline{\pi}_s := \overline{Q}_s\left(g^\mu_s \prod_{c \in C(s)} \overline{\pi}_c | X_{m(s)}\right), \quad (9)$$

summarised by the self-explanatory shorthand notation: $\overline{\pi}_s = \overline{\underline{Q}}_s(g^\mu_s \prod_{c \in C(s)} \overline{\pi}_c | X_{m(s)})$. There are two possibilities:

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

The messages $\underline{\pi}_s$ and $\overline{\pi}_s$ can be seen as tuples of real numbers, with as many components as there are elements in $\mathscr{X}_{m(s)}$: one for each of the possible values of $X_{m(s)}$. 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 values x_E the variables for these instantiated nodes assume. It then follows from Eqs. (5) and (8) and the factorisation property[2] of the local product lower previsions that:

$$\underline{P}_s(\phi^\mu_s | X_{m(s)}) = \underline{\pi}_s \text{ and } \overline{P}_s(\phi^\mu_s | X_{m(s)}) = \overline{\pi}_s. \quad (10)$$

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

$$\pi^\mu_s := \underline{Q}_s(\psi^\mu_s | X_{P(s)}), \quad (11)$$

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

$$\psi^\mu_t := \max\{g-\mu, 0\} \prod_{c \in C(t)} \underline{\pi}_c + \min\{g-\mu, 0\} \prod_{c \in C(t)} \overline{\pi}_c, \quad (12)$$

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

$$\psi^\mu_{m(s)} := g^\mu_{m(s)} \left[\max\{\pi^\mu_s, 0\} \prod_{c \in S(s)} \underline{\pi}_c + \min\{\pi^\mu_s, 0\} \prod_{c \in S(s)} \overline{\pi}_c\right]. \quad (13)$$

[2] This 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 factorising*. For instance, using the strong product and the net-independent natural extension will lead to the same inferences.

The messages π_s^μ are again tuples of real numbers, with one component for each of the possible values of $X_{m(s)}$.[3] They depend on the choice of g or μ, as well as on which nodes are instantiated and on which values x_E the variables for these instantiated nodes assume. It then follows from Eqs. (5) and (8) and the factorisation property that

$$\underline{P}_s(\phi_s^\mu | X_{P(s)}) = \pi_s^\mu, \tag{14}$$

and of course $\rho_g(\mu) = \pi_\square^\mu$. 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. (9) and (11)–(13).

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 t, and let s_t be its successor in the chain $\uparrow t$. If we let $\lambda_g(\mu)$ be the real number

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

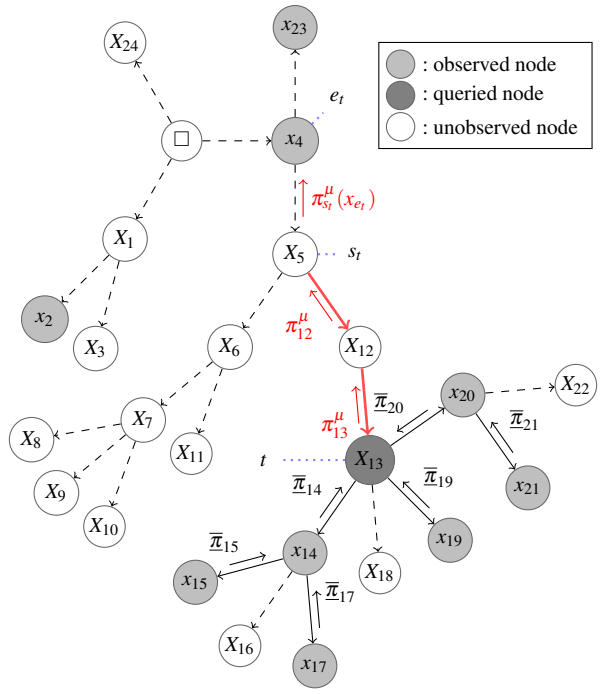
then it follows from Eq. (12) that $\psi_{e_t}^\mu = I_{\{x_{e_t}\}}\lambda_g(\mu)$. If we now continue to use Eqs. (12) and (13) until the root of the tree, we eventually find that

$$\rho_g(\mu) = \begin{cases} \underline{P}(I_{\{x_{e_t}\}})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \geq 0 \\ \overline{P}(I_{\{x_{e_t}\}})\lambda_g(\mu) & \text{if } \lambda_g(\mu) \leq 0. \end{cases}$$

Since we assumed that $\overline{P}(I_{\{x_E\}}) > 0$, it readily follows that $\overline{P}(I_{\{x_{e_t}\}}) > 0$, so we gather from Eq. (7) that $\underline{R}_t(g|x_E) = \max\{\mu \in \mathbb{R}\colon \lambda_g(\mu) \geq 0\}$. In fact, under the assumption that $\overline{P}(I_{\{x_E\}}) > 0$, $\lambda_g(\mu) \geq 0$ can be replaced in this expression by $\pi_{s_t}^\mu(x_{e_t}) \geq 0$. We conclude that in order to do expert system inference of the type described above, we can perform all calculations on the subtree $\downarrow s_t$, 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 Sec. 5.

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 a $\mu \in [\min g, \max g]$; (ii) calculate the value of $\lambda_g(\mu)$ by sending messages from the terminal nodes towards the root; and (iii) look for 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 the separation properties of the tree, and the independence of μ for some of the messages. For a start, as we are only interested in the sign of $\rho_g(\mu)$, which is determined by $\pi_{s_t}^\mu(x_{e_t})$, we only have to take nodes into consideration that strictly follow e_t.

[3] Of course, if s is the root node, then $P(s) = \emptyset$ and π_s^μ is just a single real number, which by Eq. (14) is equal to $\rho_g(\mu)$.



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 element e_t preceding t. For the tree above for instance, where X_{13} (in grey) is the queried node and the light grey nodes $\{X_2, X_4, X_{14}, X_{15}, X_{17}, X_{19}, X_{20}, X_{21}, X_{23}\}$ are instantiated, the trunk consists of $\tilde{T} = \{X_5, X_{12}, X_{13}\}$.

The start of the algorithm can be implemented with the piece of pseudo-code on the left. Here, the queried node t is known in advance and besides the trunk $\tilde{T}$, also the nodes s_t and e_t are computed. We are especially interested in the nodes that constitute the trunk, because only these nodes will send messages to their parents that depend on μ. As a consequence, we can summarise all the μ-independent messages by propagating all messages until they reach the trunk, which means that they have to be calculated only once.

```
s_t := t
T̃ := {t}
while m(s_t) ∉ E do:
    T̃ := T̃ ∪ m(s_t)
    s_t := m(s_t)
end while
e_t := m(s_t)
```

The following piece of pseudocode does the trick. Both $\underline{\pi}_c$ and $\overline{\pi}_c$ can be calculated in the recursive manner outlined in Eq. (10), 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,

```
for n ∈ T̃ do:
    for c ∈ C(n) do:
        if c ∉ T̃ then:
            calculate π̄_c
        end if
    end for
    Π̄_n := ∏_{c∈C(s)\T̃} π̄_c
end for
```

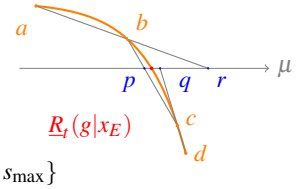
which is equivalent to deleting the node from the tree. We could envisage removing *barren nodes* (all of whose descendants are uninstantiated, such as $X_6, \ldots, X_{11}, X_{16}, X_{18}, X_{22}$ in the example tree above) from the tree beforehand, but we believe the computational overhead created by the search for them will void the gain.

At this point we can calculate $\pi^{\mu}_{s_t}(e_t)$. If we assume that t, s_t, g, $\underline{\Pi}_n$ and $\overline{\Pi}_n$ for $n \in \tilde{T}$ are stored as global variables, the following function will do the job. Now that we have the code 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 considering the computational complexity of the getJoint function, and using that $\pi^{\mu}_{s_t}(e_t)$ is concave, we can speed up the calculation of the maximal root drastically as shown in the figure below.

function getJoint(μ)
  $s := t$
  while $s \neq s_t$ **do:**
    calculate ψ^{μ}_s
    $\pi^{\mu}_s := \underline{Q}_s(\psi^{\mu}_s | X_{m(s)})$
    $s := m(s)$
  end while
  calculate $\psi^{\mu}_{s_t}$
  $\pi^{\mu}_{s_t}(e_t) := \underline{Q}_{s_t}(\psi^{\mu}_{s_t} | x_{e_t})$
  return $\pi^{\mu}_{s_t}(e_t)$

If a, b, c, and d are distributed in such a way that $\rho_g(a) \geq \rho_g(b) \geq 0 \geq \rho_g(c) \geq \rho_g(d)$, then the root of ρ_g is in the interval $[s_{\min}, s_{\max}] := [p, \min\{p, r\}]$.

function concaveRoot($a, b, c, d, s_{\min}, s_{\max}$)
  $\mu := \frac{1}{2}(s_{\min} + s_{\max})$
  $f(\mu) :=$ getJoint(μ)
  if $f(\mu) > 0$ **then:**
    $a := b$
    $b := (\mu, f(\mu))$
    $s_{\max} = \min\{b_x - \frac{b_x - a_x}{b_y - a_y} b_y, s_{\max}\}$
  else
    $d := c$
    $c := (\mu, f(\mu))$
    $s_{\max} = \min\{d_x - \frac{d_x - c_x}{d_y - c_y} d_y, s_{\max}\}$
  end if
  $s_{\min} = b_x - \frac{b_x - c_x}{b_y - c_y} b_y$
  if $s_{\max} - s_{\min} <$ tolerance **then:**
    return $s_{\min}$
  else
    return concaveRoot($a, b, c, d, s_{\min}, s_{\max}$)
  end if



Here, $s_{\min}$ is prefered over $s_{\max}$ as return value to stay on the conservative (small) side. If $b_y - a_y = 0$, then we define $\min\{b_x - \frac{b_x - a_x}{b_y - a_y} b_y, s_{\max}\}$ to be equal to $s_{\max}$ and similarly for $d_y - c_y = 0$. Keeping this in mind, we can finalise our algorithm by invoking a call to the following function.

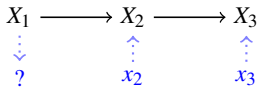
function getLowerPrevisionGivenEvidence(g)
  $a := (\min(g),$ getJoint(a_x))
  $d_x := (\max(g),$ getJoint(a_d))
  return concaveRoot(a, a, d, d, a_x, d_x)

The complexity of our algorithm is something that should be investigated further. But we can say something taking into account that for a fixed μ each node makes a single local computation and then propagates the result to the mother node: this implies that, with μ fixed, the algorithm is linear in the number of nodes. The iterations on μ create some additional complexity, but the number of iterations is usually small: a quick graphical investigation shows that the computational complexity of our root-finding algorithm must be lower than for the secant and bisection algorithms. We even have some experimental evidence that our root finder can outperform the Newton-Raphson method. Therefore, we can reasonably take the number of iterations to be a small constant for all practical applications, and conclude that the complexity of the algorithm is essentially 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 expert system 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 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 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 coherence that it suffices to calculate $\underline{r}$ and $\overline{r}$, because

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

We let $g^{\mu} = [I_{\{a\}} - \mu] I_{\{x_2\}} I_{\{x_3\}}$, 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 $\overline{\pi}_3 = \overline{Q}_3(\{x_3\}|X_2)$ to the instantiated node 2, who transforms them into the messages

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

These 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 $\psi_1^\mu = q(x_2|X_1)\overline{q}\langle I_{\{a\}} - \mu\rangle$. 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

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

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,

$$\overline{r} = \underline{r} = \frac{q(a)q(x_2|a)}{q(a)q(x_2|a) + q(b)q(x_2|b)} =: p(a|x_2) \qquad (15)$$

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

$$\overline{r} - \underline{r} = \frac{\alpha\beta(\overline{q}^2 - \underline{q}^2)}{(\alpha^2+\beta^2)\underline{q}\overline{q} + \alpha\beta(\underline{q}^2 + \overline{q}^2)}$$
$$\overline{r} - p(a|x_2) = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\overline{q}+\beta\underline{q}}$$
$$p(a|x_2) - \underline{r} = \frac{\alpha\beta}{\alpha+\beta}\frac{\overline{q}-\underline{q}}{\alpha\underline{q}+\beta\overline{q}}.$$

As soon as $\overline{q} > \underline{q}$, X_2 no longer separates X_3 from X_1, and we witness *dilation* [10, 14] because of the additional observation of X_3!

8 Numerical comparison

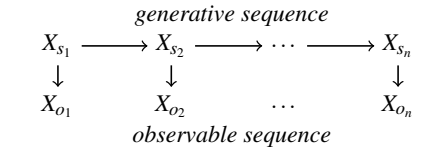
Strong independence implies epistemic irrelevance, but the converse does not generally hold. This implies that inferred probability intervals for epistemic irrelevance will generally include the ones for strong independence [3]. Here, we report on results of a number of numerical tests involving updating the tree. As noted in Sec. 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 (lower and upper) probabilities for the root variable of a *chain* when the leaf node variable is instantiated.

We have used the algorithm in Sec. 6 to compute posterior probability intervals in the irrelevance case, while the procedure in [5] is employed in the strong independence case. Inferred intervals for the former turn out to be clearly wider, and a mean square difference of about .2 is observed when considering 100 chains with three or four ternary variables and credal sets with three randomly generated extreme points. For longer chains, the updating with strong independence is too slow and no comparison can be made. Yet, similar results are observed in binary chains, for which the *2U algorithm* [9] can be used for efficient update in the strong independence case. In summary, there is a non-negligible difference between inferences based on the two notions of 'independence'.

9 An application: imprecise HMMs

Hidden Markov models (HMMs, [13]) are popular tools for modelling generative sequences, characterised by an underlying process generating an observable sequence. They have applications in many areas of signal processing, and more specifically in speech and text processing.

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.



A local uncertainty model should be defined for each variable. In the more usual case of precise probabilistic assessments, this corresponds to linear 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, [16]), 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:

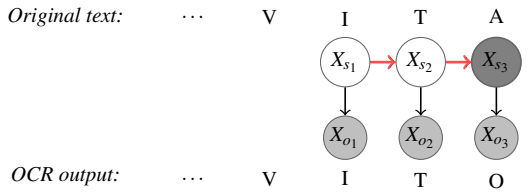
$$\underline{Q}_{s_1}(\{x_1\}) = \frac{n_{x_1}^{s_1}}{s + \sum_{x\in\mathscr{X}} n_x^{s_1}} \quad \overline{Q}_{s_1}(\{x_1\}) = \frac{s + n_{x_1}^{s_1}}{s + \sum_{x\in\mathscr{X}} n_x^{s_1}}, \qquad (16)$$

where $n_{x_1}^{s_1}$ counts the units in the sample for which $X_{s_1} = x_1$, and s is a hyperparameter 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 Sec. 2.

Generally speaking, the algorithm described in Sec. 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. We consider *maximality*: we order the states by $x_{s_n} > z_{s_n}$ iff $\underline{P}(I_{\{x_{s_n}\}} - I_{\{z_{s_n}\}}|x_{o_1}, \ldots, x_{o_n}) > 0$, and we look for the *undominated* or *maximal* states under this order. This may produce *indeterminate* predictions: the set of the undominated states can have more than one element.

Online character recognition by imprecise HMMs. As a very first application of the imprecise HMM, we have considered a *character recognition* problem. 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 (16), 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 correlation there might be between a character and its nth predecessor (with $n \geq 2$). 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 [1] in using *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).

Accuracy	93.96%	(7275/7743)
Accuracy (if imprecise indeterminate)	64.97%	(243/374)
Determinacy	95.17%	(7369/7743)
Set-accuracy	93.58%	(350/374)
Single accuracy	95.43%	(7032/7369)
Indeterminate output size	2.97	over 21

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 (with the modification suggested in [17]). 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 descriptor values for a large set 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 three). Interestingly, the instances for which the imprecise probability model returns more than a single 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 credal trees using Walley's epistemic irrelevance and have developed 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. This leads to computing lower and upper conditional previsions (expectations) with a complexity that is essentially linear in the number of nodes in the tree.

It has been unclear until recently whether an algorithm with the features described above was at all feasible. Epistemic irrelevance is most easily formulated using coherent lower previsions, which have never been used before in the context of credal networks. Moreover, epistemic irrelevance is not as "well-behaved" as strong independence is with respect to the graphoid axioms for propagation of

probability in graphical models [4]. Our results are therefore very encouraging, and they have the potential to open up new avenues of research in credal nets. This is important because strong independence is not always the most suitable notion of independence in an imprecise probability context, and epistemic irrelevance has wider scope, as well as a natural behavioural interpretation.

There is one more issue we would like to clarify at this point. While our algorithm clearly is fully functional as soon as all observations have positive upper probability, we have only proved that it produces coherent inferences when their lower probability is positive; see Theorem 4. At the time of writing this, we have strong indications that our coherence results can be extended to include observations with zero lower but positive upper probability.

Avenues for future research seem to be many. It would be important to extend the algorithm at least to so-called *polytrees*, which are substantially more expressive graphs than trees are. It would be interesting also to study in more detail the separation properties induced by epistemic irrelevance on a graph. 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 in our experiments.

Acknowledgements

Research by De Cooman and Hermans has been supported by Flemish BOF-project 01107505. 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 benefitted from discussions with Serafín Moral and Fabio Cozman.

References

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

[2] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

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

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

[5] C. P. de Campos and F. 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.

[6] C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.

[7] G. de Cooman and F. Hermans. Imprecise probability trees: Bridging two theories of imprecise probability. *Artificial Intelligence*, 172(11):1400–1427, 2008.

[8] G. de Cooman, F. Hermans, and E. Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 2009. Accepted for publication.

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

[10] T. Herron, T. Seidenfeld, and L. Wasserman. Divisive conditioning: further results on dilation. *Philosophy of Science*, 64:411–444, 1997.

[11] E. Miranda. Updating coherent lower previsions on finite spaces. *Fuzzy Sets and Systems*, 2009. In press.

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

[13] L. Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[14] T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21:1139–54, 1993.

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

[16] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[17] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, Maastricht, 2000.