Recent advances in imprecise-probabilistic graphical models

Gert de Cooman[1] and **Jasper De Bock** and **Arthur Van Camp**

Abstract. We summarise and provide pointers to recent advances in inference and identification for specific types of probabilistic graphical models using imprecise probabilities. Robust inferences can be made in so-called *credal networks* when the local models attached to their nodes are imprecisely specified as conditional *lower* previsions, by using exact algorithms whose complexity is comparable to that for the precise-probabilistic counterparts.

1 INTRODUCTION

The last twenty years have witnessed a rapid growth of *probabilistic graphical models*, and particular Bayesian nets, in AI. These models combine graphs and probability to address complex multivariate problems in a various domains. Much has been done also on the front of imprecise probability: *credal nets* [3] 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*.

Strong independence is the independence notion used with credal nets in the majority of cases. 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-probabilistic models in each of which X and Y are stochastically independent. For credal nets, strong independence leads to an equivalence: a credal net is mathematically equivalent to a set of Bayesian nets, with the same graph but with different values for the parameters. The net's parameters are not known precisely, and that is why one considers all Bayesian nets that are consistent with the partial specification of the parameters. An important problem here is the complexity of algorithms (usually exponential in the number of nodes) for making inferences.

Recent developments [5, 7, 6, 1, 4, 2, 13] have shown that there is another approach, leading to elegant mathematical formulations and algorithms whose efficiency is much better, and comparable to that of the corresponding precise-probabilistic ones. It uses another way of expressing independence: *epistemic irrelevance* [14]. X is epistemically irrelevant to Y if observing X does not affect our beliefs about Y. When the belief model is a precise probability, both epistemic irrelevance and strong independence reduce to the usual independence notion—if we ignore issues related to events with probability zero. But when the model is an imprecise probability model—a set of probabilities—this is no longer the case. Contrary to strong independence, epistemic irrelevance is not a symmetrical notion: the epistemic irrelevance of X to Y need 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 include those related to strong independence assessments.

In this paper, we give a brief overview of these developments. Due to the limited scope of this contribution, we only hint at the most salient details and provide pointers for further reference. We begin with a very brief introduction to imprecise probability models in Section 2. The main mathematical result is explained in some detail in Section 3: a recursive formula for the joint in a credal *tree* under epistemic irrelevance. Subsequent sections sketch its applications: an algorithm for inferences in credal trees (Section 4), inference in imprecise Markov chains (Section 5), identification of imprecise hidden Markov models (iHMMs, Section 6.1) and an algorithm for state sequence estimation in iHMMs (Section 6.3).

2 IMPRECISE PROBABILITIES

We begin with some basic theory of coherent lower previsions; see [14] for an in-depth study, and [10] for a recent survey. Coherent lower previsions are a special type of imprecise probability model. Roughly speaking, whereas classical probability theory assumes that a subject's uncertainty can be represented by a single probability mass function, the theory of imprecise probabilities effectively works with sets of them, and thereby allows for imprecision as well as indecision to be modelled and represented. Looking at it as a way of robustifying the classical theory is perhaps the easiest way to understand and interpret it; see [14] for different interpretations.

Consider a set $\mathcal{M}$ of probability mass functions, defined on a finite set $\mathcal{X}$ of possible states. With each mass function $p \in \mathcal{M}$, we can associate a *linear prevision* (or expectation operator) P_p, defined on the set $\mathcal{G}(\mathcal{X})$ of all real-valued maps on $\mathcal{X}$. Any $f \in \mathcal{G}(\mathcal{X})$ is also called a *gamble* on $\mathcal{X}$, and $P_p(f) = \sum_{x\in\mathcal{X}} p(x)f(x)$ is the expectation of f, associated with the probability mass function p. We can now define the *lower prevision* $\underline{P}_{\mathcal{M}}$ that corresponds with the set $\mathcal{M}$ as the following *lower envelope* of linear previsions: $\underline{P}_{\mathcal{M}}(f) := \inf\{P_p(f) \colon p \in \mathcal{M}\}$ for all gambles f on $\mathcal{X}$. Similarly, we define the *upper prevision* $\overline{P}_{\mathcal{M}}$ as

$$\overline{P}_{\mathcal{M}}(f) := \sup\{P_p(f) \colon p \in \mathcal{M}\} = -\underline{P}_{\mathcal{M}}(-f) \qquad (1)$$

for all gambles f on $\mathcal{X}$. We will mostly talk about lower previsions, since it follows from the *conjugacy relation* (1) that the two models are mathematically equivalent.

An *event* A is a subset of $\mathcal{X}$: $A \subseteq \mathcal{X}$. With such A, we associate an *indicator* $\mathbb{I}_A$: the gamble that is 1 on A, and 0 outside A. We call $\underline{P}_{\mathcal{M}}(A) := \underline{P}_{\mathcal{M}}(\mathbb{I}_A) = \inf\{\sum_{x\in A} p(x) \colon p \in \mathcal{M}\}$ the *lower probability* of A, and $\overline{P}_{\mathcal{M}}(A) := \overline{P}_{\mathcal{M}}(\mathbb{I}_A)$ its *upper probability*.

The functional $\underline{P}_{\mathcal{M}}$ satisfies the following set of interesting mathematical properties, which define a *coherent lower prevision* [14]:

[1] Ghent University, Belgium, email: gert.decooman@UGent.be

C1. $\underline{P}_{\mathcal{M}}(f) \geq \inf f$ for all $f \in \mathcal{G}(\mathcal{X})$,
C2. $\underline{P}_{\mathcal{M}}(\lambda f) = \lambda \underline{P}_{\mathcal{M}}(f)$ for all $f \in \mathcal{G}(\mathcal{X})$ and all real $\lambda \geq 0$,
C3. $\underline{P}_{\mathcal{M}}(f+g) \geq \underline{P}_{\mathcal{M}}(f) + \underline{P}_{\mathcal{M}}(g)$ for all $f, g \in \mathcal{G}(\mathcal{X})$.

Every set of mass functions $\mathcal{M}$ uniquely defines a coherent lower prevision $\underline{P}_{\mathcal{M}}$, but in general the converse does not hold. However, if we limit ourselves to sets of mass functions $\mathcal{M}$ that are closed and convex—which makes them *credal sets*—they are in a one-to-one correspondence with coherent lower previsions [14]. This implies that we can use the theory of coherent lower previsions as a tool for reasoning with closed convex sets of probability mass functions. From now on, we will no longer explicitly refer to credal sets $\mathcal{M}$, but we will simply talk about coherent lower previsions $\underline{P}$. It is useful to keep in mind that there always is a unique credal set that corresponds to such a coherent lower prevision: $\underline{P} = \underline{P}_{\mathcal{M}}$ for some unique credal set $\mathcal{M}$, given by $\mathcal{M} = \{p\colon (\forall f \in \mathcal{G}(\mathcal{X}))P_p(f) \geq \underline{P}(f)\}$.

Conditional lower and upper previsions, which are extensions of the classical conditional expectation functionals, can be defined in a similar, intuitively obvious way as lower envelopes associated with sets of conditional mass functions. Consider a variable X in $\mathcal{X}$ and a variable Y in $\mathcal{Y}$. A *conditional lower prevision* $\underline{P}(\cdot|X)$ on the set $\mathcal{G}(\mathcal{Y})$ of all gambles on $\mathcal{Y}$ is a two-place real-valued function. For any gamble g on $\mathcal{Y}$, $\underline{P}(g|X)$ is a gamble on $\mathcal{X}$, whose value $\underline{P}(g|x)$ in $x \in \mathcal{X}$ is the lower prevision of g, *conditional on the event* $X = x$. If for any $x \in \mathcal{X}$, the lower prevision $\underline{P}(\cdot|x)$ is coherent—satisfies conditions C1–C3—then we call the conditional lower prevision $\underline{P}(\cdot|X)$ *separately coherent*. It is useful to extend the domain of the conditional lower prevision $\underline{P}(\cdot|x)$ from $\mathcal{G}(\mathcal{Y})$ to $\mathcal{G}(\mathcal{Y} \times \mathcal{X})$ by letting $\underline{P}(f|x) := \underline{P}(f(\cdot, x)|x)$ for all gambles f on $\mathcal{Y} \times \mathcal{X}$.

If we have a number of conditional lower previsions involving a number of variables, each of these must be separately coherent, but they must also satisfy a more stringent *joint coherence* requirement. Explaining this in detail would take us too far, but we refer to [14] for a detailed discussion, with motivation. For our present purposes, it suffices to say that joint coherence is very closely related to making sure that these conditional lower previsions are lower envelopes associated with conditional mass functions that satisfy Bayes's Rule.

3 CONSERVATIVE COHERENT INFERENCE IN IMPRECISE MARKOV TREES

3.1 Basic notions and notation.

Consider a rooted and directed discrete tree with finite width and depth, with set of nodes T. We denote the *root* node by $\square$. For any node s, we denote its *mother node* by $m(s)$; and use the convention $m(\square) = \emptyset$. Also, we denote the set of s's *children* by $C(s)$. If $C(s) = \emptyset$, then we call s a *leaf*. $T^{\Diamond} := \{s \in T\colon C(s) \neq \emptyset\}$ denotes the set of all *non-terminal nodes*.

For nodes s and t, we write $s \sqsubseteq t$ if s *precedes* t: there is a directed segment in the tree from s to t (or $s = t$). $D(s) := \{t \in T\colon s \sqsubset t\}$ denotes the set of *descendants* of s, where $s \sqsubset t$ means that $s \sqsubseteq t$ and $s \neq t$. We also use the notation $\downarrow s := D(s) \cup \{s\}$ for the subtree with root s. Similarly, we let $\downarrow S := \bigcup\{\downarrow s\colon s \in S\}$ 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 $\mathcal{X}_s$. 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 $\mathcal{X}_S := \times_{s \in S}\mathcal{X}_s$. Generic elements of $\mathcal{X}_s$ are denoted by x_s or z_s. Similarly for x_S and z_S in $\mathcal{X}_S$. 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 $\mathcal{X}_S$, for all $\emptyset \subseteq S \subseteq T$. We use the simplifying device of identifying a gamble f_S on $\mathcal{X}_S$ with its *cylindrical extension* to $\mathcal{X}_U$, where $S \subseteq U \subseteq T$. This is the gamble f_U on $\mathcal{X}_U$ defined by $f_U(x_U) := f_S(x_S)$ for all $x_U \in \mathcal{X}_U$.

We consider (conditional) lower previsions as models for a subject's beliefs about the values that variables in the tree may assume. Let $I, O \subseteq T$ be *disjoint* sets of nodes with $O \neq \emptyset$, then we generically[2] denote by $\underline{V}_O(\cdot|X_I)$ a *conditional lower prevision*, defined on the set of gambles $\mathcal{G}(\mathcal{X}_{I \cup O})$. For every gamble f on $\mathcal{X}_{I \cup O}$ and every $x_I \in \mathcal{X}_I$, $\underline{V}_O(f|x_I)$ is the lower prevision (or lower expectation) for/of the gamble f, conditional on the event that $X_I = x_I$.

3.2 Epistemic irrelevance

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.[3]

Consider disjoint subsets C, I, and O of T, with I and O nonempty. When a subject judges X_I to be *epistemically irrelevant to* X_O *conditional on* X_C, he assesses that if he knows the value of X_C, then learning in addition the value of 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 $\mathcal{G}(\mathcal{X}_O)$. 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 $\mathcal{G}(\mathcal{X}_O)$ given by $\underline{V}_O(f|x_{C \cup I}) := \underline{V}_O(f|x_C)$ for all $f \in \mathcal{G}(\mathcal{X}_O)$ and all $x_{C \cup I} \in \mathcal{X}_{C \cup I}$.

3.3 Local and global 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 $\mathcal{G}(\mathcal{X}_s)$: 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 $\mathcal{G}(\mathcal{X}_{\square})$. We use the notation $\underline{Q}_s(\cdot|X_{m(s)})$ for all these local models.

We intend to show 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 $\mathcal{G}(\mathcal{X}_T)$ for all variables in the tree, as well as conditional lower previsions $\underline{P}_{\downarrow S}(\cdot|X_s)$ on $\mathcal{G}(\mathcal{X}_{\downarrow S})$ 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.

[2] Besides the letter V, we will also use the letters P, Q, R and S.
[3] It suffices, in the discussion below, to let $C = \emptyset$.

3.4 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.

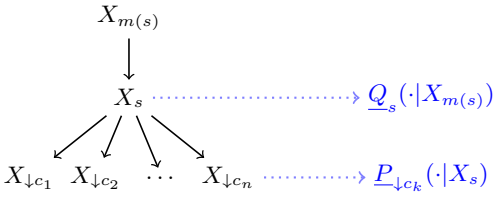
CI. Consider any node s in the tree, any subset S of its set of children $C(s)$, and the set $\overline{S} := \bigcap_{c\in S} \overline{c}$ of their common non-parent non-descendants. Then *conditional on the mother variable X_s, the non-parent non-descendant variables $X_{\overline{S}}$ are assumed 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*. We introduce the term *imprecise Markov tree* (IMT) for it. For global models, CI 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 $z_{\{s\}\cup I} \in \mathcal{X}_{\{s\}\cup I}$ we have:

$$\underline{P}_{\downarrow S}(f|z_{\{s\}\cup I}) := \underline{P}_{\downarrow S}(f(\cdot, z_I)|z_s) \text{ for all } f \text{ in } \mathcal{G}(\mathcal{X}_{\downarrow S\cup I}). \quad (2)$$

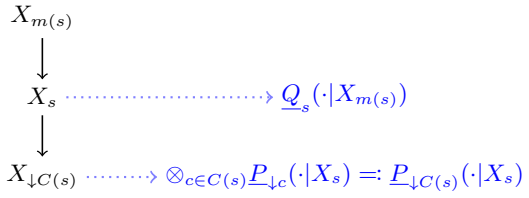
3.5 The most conservative global models

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. 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:

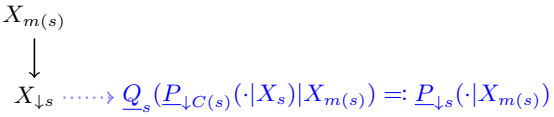


The global models are then also constructed recursively, following the same pattern.

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 $\mathcal{G}(\mathcal{X}_{\{s\}\cup\downarrow c})$. Given that, conditional on X_s, the variables $X_{\downarrow c}$, $c \in C(s)$ are epistemically independent [see Section 3.4, condition CI], this 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, the so-called *conditionally independent natural extension* [8, 14] $\otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s)$, which is a conditional lower prevision $\underline{P}_{\downarrow C(s)}(\cdot|X_s)$ on $\mathcal{G}(\mathcal{X}_{\downarrow s})$:



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 3.4, 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 Equation (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*[4] $\underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_{\{m(s),s\}})|X_{m(s)}) = \underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_s)|X_{m(s)})$; see [11, 14] for 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 $\mathcal{G}(\mathcal{X}_{\downarrow s})$ by backwards recursion:

$$\underline{P}_{\downarrow C(s)}(\cdot|X_s) := \otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s) \quad (3)$$

$$\underline{P}_{\downarrow s}(\cdot|X_{m(s)}) := \underline{Q}_s(\underline{P}_{\downarrow C(s)}(\cdot|X_s)|X_{m(s)}) = \underline{Q}_s(\otimes_{c\in C(s)}\underline{P}_{\downarrow c}(\cdot|X_s)|X_{m(s)}), \quad (4)$$

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

$$\underline{P}_{\downarrow t}(\cdot|X_{m(t)}) := \underline{Q}_t(\cdot|X_{m(t)}) \text{ for all leaves } t, \quad (5)$$

then the recursion relations (3) and (4) 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 $\mathcal{G}(\mathcal{X}_{\downarrow C(s)})$ to the set $\mathcal{G}(\mathcal{X}_{\downarrow S})$:

$$\underline{P}_{\downarrow S}(g|X_s) := \underline{P}_{\downarrow C(s)}(g|X_s) \text{ for all gambles } g \text{ on } \mathcal{X}_{\downarrow S}. \quad (6)$$

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

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

Suppose we have some family of global models

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

associated with the tree. How do we express that such a family is compatible with the assessments encoded in the tree? First of all, 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 $\mathcal{G}(\mathcal{X}_s)$.

Secondly, our models should satisfy the rationality requirement of coherence:

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

Thirdly, our global models should reflect all epistemic irrelevancies encoded in the graphical structure of the tree:

[4] 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.

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 $I \subseteq \overline{S}$ through the epistemic irrelevance requirements $\underline{V}_{\downarrow S}(f|z_{\{s\}\cup I}) := \underline{V}_{\downarrow S}(f(\cdot, z_I)|z_s)$ for all f in $\mathcal{G}(\mathcal{X}_{\downarrow S\cup I})$, then all these models together should be (jointly) coherent with all the available models in the family $\mathcal{T}(\underline{V})$.

The final requirement guarantees that all inferences we make on the basis of our global models are as conservative as possible—are based on no other considerations than what is encoded in the tree:

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

It turns out that the family of models $\mathcal{T}(\underline{P})$ we have been constructing above satisfies all these requirements. We call a real functional Φ on $\mathcal{G}(\mathcal{X})$ *strictly positive* if $\Phi(\mathbb{I}_{\{x\}}) > 0$ for all $x \in \mathcal{X}$.

Theorem 1 *If all local models $\overline{Q}_s(\cdot|X_{m(s)})$ on $\mathcal{G}(\mathcal{X}_s)$, $s \in T$ are strictly positive, then the family of global models $\mathcal{T}(\underline{P})$, obtained through Equations* (3)–(6)*, 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 so-called* regular extension [14]*:*

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

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

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 rest of this paper, where we derive efficient algorithms for doing inferences on such trees using regular extension.

4 THE MEPICTIR ALGORITHM

As a first example of an algorithm capable of making computationally efficient exact inferences in imprecise Markov trees, we introduce the MePiCTIr algorithm [6]. It deals with updating beliefs about the value of a single variable X_t in some *target node* t, after observing the evidence $X_E = x_E$ in a set of *instantiated nodes* E. It calculates the value of $\underline{R}(g|x_E)$ for any given gamble g on $\mathcal{X}_t$, assuming that $\overline{P}(\{x_E\}) > 0$.

The MePiCTIr algorithm solves this problem by cleverly exploiting the tree structure and the recursive nature of the formula for calculating the joint, in a distributed fashion by passing messages up the tree from leaves to root. It has a complexity that is essentially linear in the number of nodes in the tree, which is remarkably efficient, given that it seems that the corresponding inference in credal trees under strong independence is NP-hard.

We now focus on two special cases, which are easier to study due to their simplified structure.

5 IMPRECISE MARKOV CHAINS

The simplest special case is that of an imprecise Markov chain:

$$X_1 \longrightarrow X_2 \longrightarrow X_3 \longrightarrow \cdots \longrightarrow X_{n-1} \longrightarrow X_n$$

with as local models the *marginal* model $\underline{Q}_1$ for X_1 and the conditional so-called *transition* models $\underline{Q}_k(\cdot|X_{k-1})$ for X_k conditional on X_{k-1}, $k = 2, \ldots, n$. All so-called *state variables* X_k assume values in the same *set of states* $\mathcal{X}$. Efficient inference for such models was studied in detail in [7], and their convergence properties in relation to the notion of ergodicity were explored in [9]. We mention one interesting result to illustrate the power of this approach. When all transition models $\underline{Q}_k(\cdot|X_{k-1})$ are the same, the imprecise Markov chain is called *stationary*, and inferences can be summarised using a so-called *lower transition operator* $\underline{\mathbb{T}}\colon \mathcal{G}(\mathcal{X}) \to \mathcal{G}(\mathcal{X})$, defined by

$$(\underline{\mathbb{T}}h)(x) := \underline{Q}(h|x) \text{ for all } h \in \mathcal{G}(\mathcal{X}) \text{ and all } x \in \mathcal{X}.$$

Theorem 1 ensures that the marginal $\underline{P}_n$ for state X_n of the joint model $\underline{P}$ is given by the simple recursion equation

$$\underline{P}_n(h) = \underline{Q}_1(\underline{\mathbb{T}}^{n-1}h) \text{ for all } h \in \mathcal{G}(\mathcal{X}),$$

whose computational complexity is linear in n. If we let $n \to \infty$, there is the following simple convergence result that significantly generalises the classical Perron–Frobenius Theorem. A more refined discussion, yielding a necessary and sufficient condition for convergence, can be found in [9].

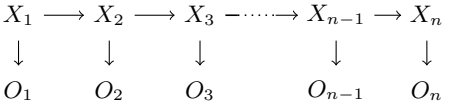
Theorem 2 (Perron–Frobenius Theorem [7]) *Consider a stationary imprecise Markov chain with finite state set $\mathcal{X}$ that is* regular*, meaning that there is some $n > 0$ such that $\max \underline{\mathbb{T}}^n(-\mathbb{I}_{\{x\}}) < 0$ for all $x \in \mathcal{X}$. Then for every marginal model $\underline{Q}_1$, the lower prevision $\underline{P}_n = \underline{Q}_1 \circ \underline{\mathbb{T}}^{n-1}$ for the state at time n converges point-wise to the same lower prevision $\underline{P}_\infty$:*

$$\lim_{n\to\infty} \underline{P}_n(h) = \lim_{n\to\infty} \underline{Q}_1(\underline{\mathbb{T}}^{n-1}h) := \underline{P}_\infty(h) \text{ for all } h \text{ in } \mathcal{G}(\mathcal{X}).$$

Moreover, the limit lower prevision $\underline{P}_\infty$ is the only $\underline{\mathbb{T}}$-invariant lower prevision $\mathcal{G}(\mathcal{X})$, meaning that $\underline{P}_\infty = \underline{P}_\infty \circ \underline{\mathbb{T}}$.

6 IMPRECISE HIDDEN MARKOV MODELS

A second, slightly more advanced special case is that of an imprecise hidden Markov Model (iHMM):



This is a stationary imprecise Markov chain, as defined in Section 5, where the state variables X_k are not directly observable (*hidden*). What we can observe are the so-called *observation variables* O_k, which depend on the corresponding states X_k through the local *emission models* $\underline{S}_k(\cdot|X_k)$ for O_k conditional on X_k, $k = 1, \ldots, n$. We assume for the sake of simplicity that all these O_k assume values in the same finite set $\mathcal{O}$, and that, besides all the local transition models, all the local emission models are the same.

6.1 System identification

One of the main questions in iHMMs is how to learn the local emission and transition models from a sequence of observations $o_{1:n}$. We describe a method [2, 13], based on the Baum–Welch algorithm for

precise hidden Markov models and the imprecise Dirichlet model (IDM, [15]).

The IDM yields imprecise estimates for multinomial probabilities. If $n(A)$ is the number of occurrences of an event A in N experiments, then the lower and upper probability of A according to an IDM are given by $\underline{P}(A) = {}^{n(A)}/_{N+s}$ and $\overline{P}(A) = {}^{n(A)+s}/_{N+s}$, where s is a non-negative hyperparameter. The larger s, the more imprecise the inferences. If $s = 0$, the resulting precise model returns the relative frequency $\underline{P}(A) = \overline{P}(A) = {}^{n(A)}/_{N}$.

We rely on the Baum–Welch algorithm to provide us with suitable quantities to plug into the IDM formulas. Consider states $x, z \in \mathcal{X}$ and observation $o \in \mathcal{O}$. The random variable $N_{x,z} \coloneqq \sum_{k=2}^{n} \mathbb{I}_{\{(x,z)\}}(X_{k-1}, X_k)$ gives the number of transitions from state x to state z. Similarly, $N_x \coloneqq \sum_{k=1}^{n} \mathbb{I}_{\{x\}}(X_k)$ gives the number of times state x is visited, and $N_{x,o} \coloneqq \sum_{k=1}^{n} \mathbb{I}_{\{(x,o)\}}(X_k, o_k)$ the number of emissions of observation o from state x. Since the state sequence $X_{1:n}$ is not known (not observed), the Baum–Welch algorithm uses successive estimates $\hat{n}_{x,z} \coloneqq E(N_{x,z}|o_{1:n})$ for the expected number of transitions conditional on the observations, and similarly for $\hat{n}_x \coloneqq E(N_x|o_{1:n})$ and $\hat{n}_{x,o} \coloneqq E(N_{x,o}|o_{1:n})$. Once the algorithm, and these estimates, have converged to stationary values, they are plugged into the IDM formulas, leading to the following formulas for the estimated local imprecise transition model:

$$\underline{Q}(\{z\}|x) = \frac{\hat{n}_{x,z}}{s + \sum_{z' \in \mathcal{X}} \hat{n}_{x,z'}}, \overline{Q}(\{z\}|x) = \frac{s + \hat{n}_{x,z}}{s + \sum_{z' \in \mathcal{X}} \hat{n}_{x,z'}}$$

and for the estimated local imprecise emission model:

$$\underline{S}(\{o\}|x) = \frac{\hat{n}_{x,o}}{s + \hat{n}_x}, \overline{S}(\{o\}|x) = \frac{s + \hat{n}_{x,o}}{s + \hat{n}_x}.$$

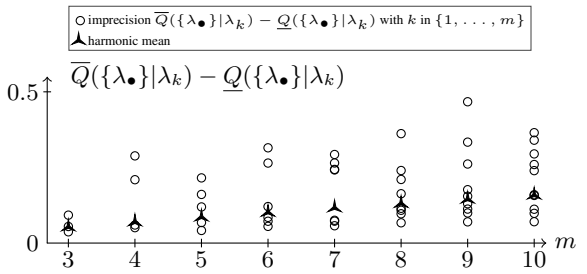
6.2 MePiCTIr

One interesting application of the MePiCTIr algorithm (see Section 4) to iHMMs concerned model tracking [1]. Here we describe a simple application for predicting future major (with magnitude 7 and higher) earthquake rates.

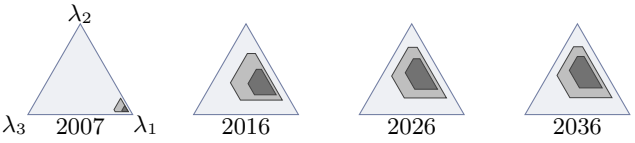
We use a hidden Markov model, where we assume that the earth can be in m different 'seismic' states $\lambda_1, \ldots, \lambda_m$ and that the occurrence of earthquakes in any given year year depends on the seismic state Λ of the Earth in that year. The Earth, being in a seismic state Λ,"emits" a number of earthquakes O governed by a Poisson distribution with parameter Λ: the emission model is assumed to be precise and characterised by the mass function $s(o|\Lambda) = {}^{e^{-\Lambda}\Lambda^o}/_{o!}$.

To learn the transmission and emission models, we have used data of counted annual numbers of major earthquakes over 107 subsequent years, from 1900 to 2006.[5] We have modelled this problem as an iHMM of length 107, in which each observation variable O_i corresponds to one of the 107 yearly earthquake counts. The states correspond to the seismic states Earth can be in. The set of seismic states $\{\lambda_1, \ldots, \lambda_m\}$ defines the space $\mathcal{X}$ of the state variables in the HMM. Since there is only 107 years of data, we believe that a precise local transition model is not justified, so we have done an imprecise estimation. To show how the resulting model imprecision changes with changing number of possible state values m, we have plotted, as a function of m ranging from 3 to 10, the imprecision $\overline{Q}(\{\lambda_\bullet\}|\lambda_k) - \underline{Q}(\{\lambda_\bullet\}|\lambda_k)$ of the transition probability estimates for going from state λ_k to state $\lambda_\bullet$, for $s = 2$ and their harmonic mean H, known to increase with m as $H = {}^{ms}/_{ms+n-1}$.

[5] Freely available from `http://neic.usgs.gov/neis/eqlists`.



With the learned transition model (we choose $m = 3$ for graphical convenience), we have used the MePiCTIr algorithm to predict future earthquake rates, in the years 2007, 2016, 2026 and 2036: we are interested in the imprecise probability model for the state variable Λ in these years, conditional on the observed rates. The figure below shows conservative approximations (the smallest hexagons with vertices parallel with the borders of the simplex) of such updated models, as credal sets in the probability simplex. Dark grey are the estimates corresponding to $s = 2$, light grey the ones for $s = 5$.



The precision of the predictive inferences decreases as we move forward in time. For 2007, we can be fairly confident that seismic rate will be close to λ_1, but for 2036, we can only make very imprecise inferences about the seismic rate. This is a (we believe desirable) property that predictions with precise HMMs do not have.

6.3 The EstiHMM algorithm

Suppose we have observed the output sequence $o_{1:n}$, how do we estimate the state sequence $x_{1:n}$? In precise HMMs, the solution can be calculated efficiently using the well-known Viterbi algorithm. It solves the problem by finding the state sequence with highest posterior probability, after conditioning on the observed outputs. For imprecise HMMs, the solution can be efficiently calculated using the EstiHMM algorithm [4], and allows us to robustify the results obtained through the Viterbi algorithm.

If the local models of the iHMM have been identified, the global model $\underline{P}$ is determined using the recursive construction in Section 3.5. We take into account the observed output sequence $o_{1:n}$ by conditioning the global model on it, using regular extension. By Theorem 1, the resulting conditional model $\underline{P}(\cdot|o_{1:n})$ yields coherent inferences if we assume all local models to be strictly positive.[6]

With imprecise models, solving a decision-making problem does not necessarily lead to a single solution: set-valued results are allowed, containing multiple so-called *optimal* solutions. EstiHMM decides which state sequences are optimal using the criterion of (Walley–Sen) *maximality* [14, 12]: a state sequence $\hat{x}_{1:n}$ is considered to be strictly better than a sequence $x_{1:n}$ if its posterior probability is strictly higher for each conditional mass function $p(\cdot|o_{1:n})$ in the credal set associated with the updated lower prevision $\underline{P}(\cdot|o_{1:n})$. This induces a partial order on the set of all possible sequences. The maximal sequences are those that are undominated under this partial order, meaning that there is no sequence that is strictly better.

[6] This is always the case if the local models are derived using the method proposed in Section 6.1.

Finding all maximal state sequences seems a daunting task: the search space grows exponentially in the length n of the iHMM. However, by exploiting the recursive formulas of Section 3.5, an appropriate version of Bellman's Principle of Optimality can be derived, allowing for an exponential reduction of the search space. By using a number of additional tricks, EstiHMM finds all maximal state sequences in a time essentially linear in the number of such maximal sequences, quadratic in the length of the chain, and cubic in the number of states; a complexity comparable to that of Viterbi's algorithm.

As a first toy application, we used EstiHMM to try and detect mistakes in words. A written word was regarded as a hidden sequence $x_{1:n}$, generating an output sequence $o_{1:n}$ by artificially corrupting the word. This simulates not perfectly reliable observational processes, such as the output of an Optical Character Recognition (OCR) device. As an example, the Italian word QUANTO generated the output OUANTO. The objective was to try and detect such errors by using EstiHMM. We started building an imprecise hidden Markov model by applying IDM estimation to a data set of correct words and their corrupted counterparts. Next, we took a corrupted word, for example OUANTO, and let it play the role of an output sequence, using EstiHMM to try and produce the corresponding hidden sequence (the original correct word QUANTO). For this particular example, EstiHMM returned CUANTO, DUANTO, FUANTO and QUANTO as maximal (undominated) solutions, including the correct one. Applying the Viterbi algorithm to the same problem, using a precise identification, resulted in the single incorrect solution DUANTO. This already illustrates that EstiHMM is able to robustify the results of the Viterbi algorithm. Let us justify this statement by analysing how both algorithms compared in trying to detect errors in a set of 200 words, 63 of which had been corrupted.

	total number	*correct*	*corrupted*
total number	200 (100%)	137 (68.5%)	63 (31.5%)
EstiHMM			
correct solution included	172 (86%)	137	35
correct solution not included	28 (14%)	0	28
Viterbi			
correct solution	157 (78.5%)	132	25
wrong solution	43 (21.5%)	5	38

EstiHMM suggested the original correct word as one of its solutions in 86% of cases. Assuming we are able to detect this correct word (in some way), the percentage of correct words rises from 68.5% to 86% by applying the EstiHMM algorithm, thereby outperforming the Viterbi algorithm by almost 10%. Also, unlike Viterbi's algorithm, EstiHMM did not introduce new errors in already correct words. Since the Viterbi solutions are always contained within EstiHMM's, the difference between both methods is only relevant if EstiHMM returns multiple solutions. We therefore take a closer look at those words for which this was indeed the case.

	total number	*correct*	*corrupted*
total number	45 (100%)	8 (17.8%)	37 (82.2%)
EstiHMM (multiple solutions)			
correct solution included	38 (84.4%)	8	30
correct solution not included	7 (15.6%)	0	7
Viterbi			
correct solution	23 (51.1%)	3	20
wrong solution	22 (48.9%)	5	17

A first conclusion is that EstiHMM's being indecisive serves as a rather strong indication a word contains errors: when EstiHMM returns multiple solutions, the original word was corrupted in 82.2% of cases. A second conclusion, related to the first, is that EstiHMM's being indecisive also indicates that the result returned by the Viterbi algorithm is less reliable: here the percentage of correct words for Viterbi drops to 51.1%, in contrast with the global percentage of 78.5%. EstiHMM, however, still yields the correct word as one of its solutions in 84.4% of cases, which is almost as high as its global percentage of 86%. EstiHMM seems to notice we are dealing with more difficult words and therefore gives us multiple solutions, between which it cannot decide.

We conclude that EstiHMM can be usefully applied to robustify the results of the Viterbi algorithm, and to gain an appreciation of where it is likely to go wrong. If EstiHMM returns multiple solutions between which it cannot decide, this indicates robustness issues for the Viterbi algorithm, which will apparently pick one of them in a fairly arbitrary way, thereby likely increasing the number of errors. EstiHMM's advantage is that it detects such robustness issues, leaving us with the option of resolving the ambiguity by picking the correct word, for instance by using a dictionary or a human expert.

ACKNOWLEDGEMENTS

We would like to acknowledge support from SBO project 060043 of the IWT-Vlaanderen. Jasper De Bock is a Ph.D. Fellow of the Fund for Scientific Research - Flanders (FWO).

REFERENCES

[1] A. Antonucci, A. Benavoli, M. Zaffalon, G. de Cooman, and F. 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)*, pp. 1767–1774, (2009).

[2] A. Antonucci, R. de Rosa, and A. Giusti, 'Action recognition by imprecise hidden Markov models', in *Proceedings of the 2011 International Conference on Image Processing, Computer Vision and Pattern Recognition (IPCV 2011)*, pp. 474–478. CSREA Press, (2011).

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

[4] J. De Bock and G. de Cooman, 'State sequence prediction in imprecise hidden Markov models', in *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, eds., F. Coolen, G. de Cooman, Th. Fetz, and M. Oberguggenberger, pp. 159–168, Innsbruck, (2011). SIPTA.

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

[6] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon, 'Epistemic irrelevance in credal nets: the case of imprecise Markov trees', *International Journal of Approximate Reasoning*, **51**, 1029–1052, (2010).

[7] G. de Cooman, F. Hermans, and E. Quaeghebeur, 'Imprecise Markov chains and their limit behaviour', *Probability in the Engineering and Informational Sciences*, **23**, 597–635, (2009).

[8] G. de Cooman, E. Miranda, and M. Zaffalon, 'Independent natural extension', *Artificial Intelligence*, **175**, 1911–1950, (2011).

[9] F. Hermans and G. de Cooman, 'Characterisation of ergodic upper transition operators', *International Journal of Approximate Reasoning*, **53**, 573–583, (2012).

[10] E. Miranda, 'A survey of the theory of coherent lower previsions', *International Journal of Approximate Reasoning*, **48**, 628–658, (2008).

[11] E. Miranda and G. de Cooman, 'Marginal extension in the theory of coherent lower previsions', *International Journal of Approximate Reasoning*, **46**, 188–225, (2007).

[12] M. C. M. Troffaes, 'Decision making under uncertainty using imprecise probabilities', *International Journal of Approximate Reasoning*, **45**, 17–29, (2007).

[13] A. Van Camp and G. de Cooman, 'A method for learning imprecise hidden Markov models'. Accepted for IPMU 2012.

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

[15] 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.