State sequence prediction in imprecise hidden Markov models

Jasper De Bock and Gert de Cooman
SYSTeMS, Ghent University, Belgium
{jasper.debock,gert.decooman}@UGent.be

Abstract

We present an efficient exact algorithm for estimating state sequences from outputs (or observations) in imprecise hidden Markov models (iHMM), where both the uncertainty linking one state to the next, and that linking a state to its output, are represented using coherent lower previsions. The notion of independence we associate with the credal network representing the iHMM is that of epistemic irrelevance. We consider as best estimates for state sequences the (Walley–Sen) maximal sequences for the posterior joint state model (conditioned on the observed output sequence), associated with a gain function that is the indicator of the state sequence. This corresponds to (and generalises) finding the state sequence with the highest posterior probability in HMMs with precise transition and output probabilities (pHMMs). We argue that the computational complexity is at worst quadratic in the length of the Markov chain, cubic in the number of states, and essentially linear in the number of maximal state sequences. For binary iHMMs, we investigate experimentally how the number of maximal state sequences depends on the model parameters.

Keywords. Imprecise hidden Markov model, optimal state sequence, maximality, coherent lower prevision, credal network, epistemic irrelevance.

1 Introduction

In a recent paper on inference in credal networks [5], De Cooman et al. developed the so-called MePiCTIr[1] algorithm for coherently updating beliefs about a single node in the tree after instantiating any number of other nodes. The local uncertainty models associated with the nodes of the network are coherent lower previsions [10, 14], and the independence notion used to interpret the graphical structure is that of epistemic irrelevance [2, 14]. This algorithm is quite efficient—it is essentially linear in the number of nodes—but it has a number of limitations. First of all, it only works for very special graphical structures: trees. While this is a serious limitation, there are, nevertheless quite a number of models and applications that involve a tree structure. Amongst these, hidden Markov models (HMMs) are definitely the simplest, and perhaps also the most popular ones. But this brings us to the second limitation: MePiCTIr only allows updating of beliefs about a *single* node. Whereas one of the most important applications for, say, HMMs, involves finding the *sequence* of (hidden) states with the highest posterior probability after observing a sequence of outputs [11]. For HMMs with precise local transition and emission probabilities, there are quite efficient dynamic programming algorithms, such as Viterbi's [11, 13], for performing this task. For imprecise-probabilistic local models, such as coherent lower previsions, we know of no algorithm in the literature for which the computational complexity comes even close to that of Viterbi's.

In this paper, we take the first steps towards remedying this situation. We describe imprecise hidden Markov models as special cases of credal trees (a special case of credal networks) under epistemic irrelevance in Section 2. We show in particular how we can use the ideas underlying the MePiCTIr algorithm (independent natural extension and marginal extension) to construct a most conservative joint model from imprecise local transition and emission models, and derive a number of interesting and useful formulas from that construction. In Section 3 we explain how a sequence of observations leads to (a collection of) so-called maximal state sequences. Finding all of them seems a daunting task at first: it has a search space that grows exponentially in the length of the Markov chain. However, in Section 4 we use the basic formulas found in Section 2 to derive an appropriate version of Bellman's [1] Principle of Optimality, which allows for an exponential reduction of the search space. By using a number of additional tricks, we are able in Section 5 to devise an algorithm for finding all maximal state sequences that is essentially linear in the number of such maximal sequences, quadratic in the length of the chain, and cubic in the number of states. We perceive this complexity to be comparable to that of the Viterbi algorithm, especially after realising that the latter makes the simplifying step of resolving ties more or less arbitrarily in order to produce

[1] MePiCTIr: Message Passing in Credal Trees under Irrelevance.

only a single optimal state sequence. This is something we will not allow our algorithm to do, for reasons that should become clear further on. In Section 6, we consider the special case of binary iHMMs, and investigate experimentally how the number of maximal state sequences depends on the model parameters. We comment on the very interesting structures that emerge, and give an heuristic explanation for them. We show off the algorithm's efficiency in Section 7 by calculating the maximal sequences for a specific iHMM of length 100.

We assume that the reader has a good working knowledge of the theory of coherent lower previsions; see Ref. [14] for an in-depth study, and Ref. [10] for a recent survey.

2 Basic notions

A hidden Markov model can be depicted using the following probabilistic graphical model:

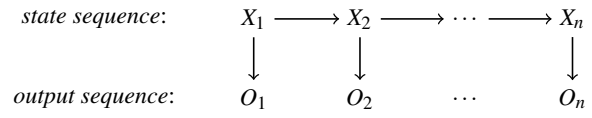


Figure 1: Tree representation of a hidden Markov model

Here n is some natural number. The *state variables* $X_1, \ldots, X_n$ assume values in the respective finite sets $\mathscr{X}_1, \ldots, \mathscr{X}_n$, and the *output variables* $O_1, \ldots, O_n$ assume values in the respective finite sets $\mathscr{O}_1, \ldots, \mathscr{O}_n$. We denote generic values of X_k by x_k, $\hat{x}_k$ or z_k, and generic values of O_k by o_k.

Local uncertainty models. We assume that we have the following local uncertainty models for these variables. For X_1, we have a *marginal* lower prevision $\underline{Q}_1$, defined on the set $\mathscr{G}(\mathscr{X}_1)$ of all real-valued maps (or *gambles*) on $\mathscr{X}_1$. For the subsequent states X_k, with $k \in \{2, \ldots, n\}$, we have a conditional lower prevision $\underline{Q}_k(\cdot|X_{k-1})$ defined on $\mathscr{G}(\mathscr{X}_k)$, called a *transition model*. In order to maintain uniformity of notation, we will also denote the marginal lower prevision $\underline{Q}_1$ as a conditional lower prevision $\underline{Q}_1(\cdot|X_0)$, where X_0 denotes a variable that may only assume a single value, and whose value is therefore certain. For any gamble f_k in $\mathscr{G}(\mathscr{X}_k)$, $\underline{Q}_k(f_k|X_{k-1})$ is interpreted as a gamble on $\mathscr{X}_{k-1}$, whose value $\underline{Q}_k(f_k|z_{k-1})$ in any $z_{k-1} \in \mathscr{X}_{k-1}$ is the lower prevision (or lower expectation) of the gamble $f_k(X_k)$, conditional on $X_{k-1} = z_{k-1}$.

In addition, for each output O_k, with $k \in \{1, \ldots, n\}$, we have a conditional lower prevision $\underline{S}_k(\cdot|X_k)$ defined on $\mathscr{G}(\mathscr{O}_k)$, called an *emission model*. For any gamble g_k in $\mathscr{G}(\mathscr{O}_k)$, $\underline{S}_k(g_k|X_k)$ is interpreted as a gamble on $\mathscr{X}_k$, whose value $\underline{S}_k(g_k|z_k)$ in any $z_k \in \mathscr{X}_k$ is the lower prevision (or lower expectation) of the gamble $g_k(O_k)$, conditional on $X_k = z_k$.

We take all these local (marginal, transition and emission) uncertainty models to be separately coherent; see for instance Ref. [5] for more details about such local uncertainty models and their separate coherence.

Interpretation of the graphical structure. We will assume that the tree in Fig. 1 represents the following irrelevance assessments: *conditional on its mother variable, the non-parent non-descendants of any variable in the tree are epistemically irrelevant to this variable and its descendants.* This is a weaker condition than the one usually associated with credal networks [3], which imposes strong independence rather than epistemic irrelevance. Recent work [5] has shown that using this weaker condition guarantees that an efficient algorithm exists for updating a credal *tree*, that is essentially linear in the number of nodes in the tree.

A joint uncertainty model. By applying the general analysis in Ref. [5] to the special case considered here, we find that the local uncertainty models can always be extended to a point-wise smallest (most conservative or least committal) coherent family of *joint* lower previsions $\underline{P}_k(\cdot|X_{k-1})$ on $\mathscr{G}(\mathscr{X}_{k:n} \times \mathscr{O}_{k:n})$, where $k \in \{1, \ldots, n\}$, $\mathscr{X}_{k:n} := \times_{i=k}^{n} \mathscr{X}_i$ and $\mathscr{O}_{k:n} := \times_{i=k}^{n} \mathscr{O}_i$. Again, for $k = 1$ the joint lower prevision $\underline{P}_1 = \underline{P}_1(\cdot|X_0)$ is effectively an unconditional lower prevision, defined on $\mathscr{G}(\mathscr{X}_{1:n} \times \mathscr{O}_{1:n})$. These joint lower previsions are given by the following recursion equations:

$$\underline{E}_k(\cdot|X_k) := \begin{cases} \underline{S}_n(\cdot|X_n) & k = n \\ \underline{S}_k(\cdot|X_k) \otimes \underline{P}_{k+1}(\cdot|X_k) & k = n-1, \ldots, 1 \end{cases} \tag{1}$$

and

$$\underline{P}_k(\cdot|X_{k-1}) := \underline{Q}_k(\underline{E}_k(\cdot|X_k)|X_{k-1}) \text{ for } k = n, \ldots, 1. \tag{2}$$

Eq. (1) states that, for $k = n-1, \ldots, 1$, the conditional lower prevision $\underline{E}_k(\cdot|X_k)$ on $\mathscr{G}(\mathscr{X}_{k+1:n} \times \mathscr{O}_{k:n})$ is the so-called (conditionally) *independent natural extension* [14, Chapter 9] of the conditional lower previsions $\underline{S}_k(\cdot|X_k)$ and $\underline{P}_{k+1}(\cdot|X_k)$, which was studied in detail in Ref. [6]. For our present purposes, it will suffice to recall from that study that such independent natural extensions are *factorising*, which implies in particular that

$$\begin{aligned} \underline{E}_k(fg|z_k) &= \underline{E}_k(g\underline{E}_k(f|z_k)|z_k) \\ &= \begin{cases} \underline{S}_k(g|z_k)\underline{P}_{k+1}(f|z_k) & \text{if } \underline{P}_{k+1}(f|z_k) \geq 0 \\ \overline{S}_k(g|z_k)\underline{P}_{k+1}(f|z_k) & \text{if } \underline{P}_{k+1}(f|z_k) \leq 0 \end{cases} \\ &= \overline{\underline{S}}_k(g|z_k) \odot \underline{P}_{k+1}(f|z_k), \end{aligned} \tag{3}$$

for all $z_k \in \mathscr{X}_k$, all $f \in \mathscr{G}(\mathscr{X}_{k+1:n} \times \mathscr{O}_{k+1:n})$ and all non-negative $g \in \mathscr{G}(\mathscr{O}_k)$, where $k \in \{1, \ldots, n-1\}$ (we call a gamble non-negative if all its values are). In this expression, we have used the shorthand notation $\overline{a} \odot b := \underline{a}\max\{0, b\} + \overline{a}\min\{0, b\}$.

Interesting lower and upper probabilities. Without too much trouble, we can use Eqs. (1)–(3) to derive the follow-

ing expressions for a number of interesting lower and upper probabilities:

$$\underline{P}_k(\{z_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \underline{Q}_i(\{z_i\}|z_{i-1})$$
$$\overline{P}_k(\{z_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \overline{Q}_i(\{z_i\}|z_{i-1}),$$

and

$$\underline{P}_k(\{z_{k:n}\} \times \{o_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \underline{S}_i(\{o_i\}|z_i)\underline{Q}_i(\{z_i\}|z_{i-1}) \tag{4}$$

$$\overline{P}_k(\{z_{k:n}\} \times \{o_{k:n}\}|z_{k-1}) = \prod_{i=k}^{n} \overline{S}_i(\{o_i\}|z_i)\overline{Q}_i(\{z_i\}|z_{i-1}), \tag{5}$$

for $k = \{1,\ldots,n\}$. We will assume throughout that

$\underline{P}_1(\{z_{1:n}\} \times \{o_{1:n}\}) > 0$ for all $z_{1:n} \in \mathscr{X}_{1:n}$ and $o_{1:n} \in \mathscr{O}_{1:n}$

or equivalently, that all *local lower previsions are positive* [5], in the sense that

$$\underline{Q}_k(\{z_k\}|z_{k-1}) > 0 \text{ and } \underline{S}_k(\{o_k\}|z_k) > 0$$

for all $z_{k-1} \in \mathscr{X}_{k-1}$, $z_k \in \mathscr{X}_k$ and $o_k \in \mathscr{O}_k$, $k \in \{1,\ldots,n\}$. This implies in particular that $\underline{P}_k(\{o_{k:n}\}|z_{k-1}) > 0$ for all $k \in \{1,\ldots,n\}$, $z_{k-1} \in \mathscr{X}_{k-1}$ and $o_{k:n} \in \mathscr{O}_{k:n}$.

We have good reason to believe that our results remain valid, *mutatis mutandis*, on the weaker condition that all local *upper* previsions should be positive, and we intend to deal with this issue in further work.

3 Estimating states from outputs

In a hidden Markov model, the states are not directly observable, but the outputs are, and the general aim is to use the outputs to estimate the states. In the present paper, we concentrate on the following problem: *Suppose we have observed the output sequence $o_{1:n}$, estimate the state sequence $x_{1:n}$.* We will use an essentially Bayesian approach to do so, but need to allow for the fact that we are working with imprecise rather than precise probability models.

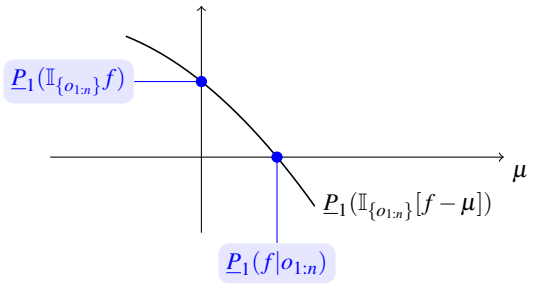
Updating the iHMM. The first step in our approach consists in updating (or conditioning) the joint model $\underline{P}_1$ on the observed outputs $O_{1:n} = o_{1:n}$. Given our positivity assumptions on the local lower prevision, we see that the lower probability $\underline{P}_1(\{o_{1:n}\})$ of the conditioning event $\{o_{1:n}\}$ is strictly positive. This implies [5] that there is only one coherent way to perform this updating, namely using the Generalised Bayes Rule [14], which reduces to Bayes's Rule when all local models are precise. We are thus led to consider the updated lower prevision $\underline{P}_1(\cdot|o_{1:n})$ on $\mathscr{G}(\mathscr{X}_{1:n})$, given by

$$\underline{P}_1(f|o_{1:n}) := \max\left\{\mu \in \mathbb{R} \colon \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu]) \geq 0\right\}, \tag{6}$$

for all gambles f on $\mathscr{X}_{1:n}$. Using the coherence of $\underline{P}_1$, it is not too hard to prove that when $\underline{P}_1(\{o_{1:n}\}) > 0$, $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu])$ constitutes a strictly decreasing and continuous function of μ, which therefore has a unique zero. As a consequence, we have for any $f \in \mathscr{G}(\mathscr{X}_{1:n})$ that

$$\begin{aligned}\underline{P}_1(f|o_{1:n}) \leq 0 &\Leftrightarrow (\forall \mu > 0)\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu]) < 0\\ &\Leftrightarrow \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}f) \leq 0.\end{aligned} \tag{7}$$

In fact, it is not hard to infer from the strictly decreasing and continuous character of $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[f-\mu])$ that $\underline{P}_1(f|o_{1:n})$ and $\underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}f)$ have the same sign. They are either both negative, both positive or both equal to zero; see also the illustration below.



Maximal state sequences. The next step consists in using the posterior model $\underline{P}_1(\cdot|o_{1:n})$ to find best estimates for the state sequence $x_{1:n}$. On the Bayesian approach, this is usually done by solving a decision-making, or optimisation, problem: we associate a gain function $\mathbb{I}_{\{x_{1:n}\}}$ with every candidate state sequence $x_{1:n}$, and select as best estimates those state sequences $\hat{x}_{1:n}$ that maximise the expected gain, resulting in state sequences with maximal posterior probability.

Here we generalise this decision-making approach towards working with imprecise probability models. The criterion we use to decide which estimates are optimal for the given gain functions is that of (Walley–Sen) *maximality* [12, 14]. Maximality has a number of very desirable properties that make sure it works well in optimisation contexts [7, 9], and it is well-justified from a behavioural point of view, as we shall see presently.

We can express a strict preference $\succ$ between two state sequence estimates $\hat{x}_{1:n}$ and $x_{1:n}$ as follows:

$$\hat{x}_{1:n} \succ x_{1:n} \Leftrightarrow \underline{P}_1(\mathbb{I}_{\{\hat{x}_{1:n}\}} - \mathbb{I}_{\{x_{1:n}\}}|o_{1:n}) > 0.$$

On a behavioural interpretation, this expresses that a subject with lower prevision $\underline{P}_1(\cdot|o_{1:n})$ is disposed to pay some strictly positive amount of utility to replace the (gain associated with the) estimate $x_{1:n}$ with the (gain associated with the) estimate $\hat{x}_{1:n}$; see Ref. [14, Section 3.9]. This induces a strict partial order $\succ$ [an irreflexive and transitive binary relation] on the set of state sequences $\mathscr{X}_{1:n}$, and we consider an estimate $\hat{x}_{1:n}$ to be optimal when it is undominated, or maximal, in this strict partial order:

$$
\begin{aligned}
&\hat{x}_{1:n} \in \operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n}) \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) x_{1:n} \not\succ \hat{x}_{1:n} \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) \underline{P}_1(\mathbb{I}_{\{x_{1:n}\}} - \mathbb{I}_{\{\hat{x}_{1:n}\}}|o_{1:n}) \leq 0 \qquad (8) \\
&\quad \Leftrightarrow (\forall x_{1:n} \in \mathscr{X}_{1:n}) \underline{P}_1(\mathbb{I}_{\{o_{1:n}\}}[\mathbb{I}_{\{x_{1:n}\}} - \mathbb{I}_{\{\hat{x}_{1:n}\}}]) \leq 0,
\end{aligned}
$$

where the last equivalence follows from Eq. (7). *In summary then, the aim of this paper is to develop an efficient algorithm for finding the set of maximal estimates* $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$.

Another approach, which we will not consider here, could consist in trying to find the so-called *maximin* state sequences $\bar{x}_{1:n}$, which maximise the posterior lower probability:

$$
\bar{x}_{1:n} \in \operatorname{argmax}_{x_{1:n} \in \mathscr{X}_{1:n}} \underline{P}_1(\{x_{1:n}\}|o_{1:n})
$$

While it is well known that any such maximin sequence is in particular guaranteed to also be a maximal sequence, finding such maximin sequences seems to be a much more complicated affair.[2]

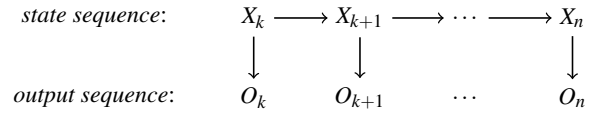
More general optimality operators. We shall see below that in order to find the set of maximal estimates, it is useful to consider a more general collection of 'optimality operators': for any $k \in \{1,\ldots,n\}$ and $z_{k-1} \in \mathscr{X}_{k-1}$, we define the optimality operator

$$
\operatorname{opt}(\cdot|z_{k-1}, o_{k:n}) : \mathscr{P}(\mathscr{X}_{k:n}) \to \mathscr{P}(\mathscr{X}_{k:n})
$$

such that for all $S \in \mathscr{P}(\mathscr{X}_{k:n})$, or in other words $S \subseteq \mathscr{X}_{k:n}$, and all $\hat{x}_{k:n} \in S$:

$$
\begin{aligned}
&\hat{x}_{k:n} \in \operatorname{opt}(S|z_{k-1}, o_{k:n}) \\
&\Leftrightarrow (\forall x_{k:n} \in S) \underline{P}_k(\mathbb{I}_{\{o_{k:n}\}}[\mathbb{I}_{\{x_{k:n}\}} - \mathbb{I}_{\{\hat{x}_{k:n}\}}]|z_{k-1}) \leq 0. \qquad (9)
\end{aligned}
$$

The interpretation of these operators is immediate: consider the following part of the original iHMM:



where we take $\underline{Q}_k(\cdot|z_{k-1})$ as the marginal model for the first state X_k. Then the corresponding joint lower prevision on $\mathscr{G}(\mathscr{X}_{k:n} \times \mathscr{O}_{k:n})$ is precisely $\underline{P}_k(\cdot|z_{k-1})$, and if we have a sequence of outputs $o_{k:n}$, then $\operatorname{opt}(\cdot|z_{k-1}, o_{k:n})$ selects from a set $S \subseteq \mathscr{X}_{k:n}$ those state sequence estimates that are undominated by any other estimate in S. It should be clear that the set $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ we are eventually looking for, can also be written as $\operatorname{opt}(\mathscr{X}_{1:n}|z_0, o_{1:n})$.

[2] Private communication from Cassio de Campos. Of course, once we know all maximal solutions, we could determine which of them are the maximin solutions by comparing their posterior lower probabilities. As far as we can see now, calculating these does not seem a trivial task.

Useful recursion equations. Fix any k in $\{1,\ldots,n\}$. If we look at Eq. (9), we see that it will be useful to derive a manageable expression for $\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1})$, where $\Delta[x_{k:n}, \hat{x}_{k:n}]$ is the gamble on $\mathscr{X}_{k:n} \times \mathscr{O}_{k:n}$ given by:

$$
\Delta[x_{k:n}, \hat{x}_{k:n}] := \mathbb{I}_{\{o_{k:n}\}}[\mathbb{I}_{\{x_{k:n}\}} - \mathbb{I}_{\{\hat{x}_{k:n}\}}].
$$

Using Eqs. (1)–(5) together with a few algebraic manipulations, we can derive the following equations for $\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1})$:

If $k \in \{1,\ldots,n-1\}$ and $\hat{x}_k = x_k$ then, with some fairly obvious abuse of notation:

$$
\begin{aligned}
\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1}) = \overline{\underline{Q}}_k(\{x_k\}|z_{k-1}) \overline{\underline{S}}_k(\{o_k\}|x_k) \qquad (10) \\
\odot \underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|x_k).
\end{aligned}
$$

If $\hat{x}_n = x_n$ then

$$
\underline{P}_n(\Delta[x_n, \hat{x}_n]|z_{n-1}) = 0. \qquad (11)
$$

If $k \in \{1,\ldots,n\}$ and $\hat{x}_k \neq x_k$ then

$$
\begin{aligned}
&\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1}) \\
&\qquad = \underline{Q}_k(\mathbb{I}_{\{x_k\}} \beta(x_{k:n}) - \mathbb{I}_{\{\hat{x}_k\}} \alpha(\hat{x}_{k:n})|z_{k-1}), \qquad (12)
\end{aligned}
$$

where we define, for any $z_{k:n} \in \mathscr{X}_{k:n}$:

$$
\begin{aligned}
\beta(z_{k:n}) &:= \underline{S}_k(\{o_k\}|z_k) \prod_{i=k+1}^{n} \underline{S}_i(\{o_i\}|z_i) \underline{Q}_i(\{z_i\}|z_{i-1}) \\
\alpha(z_{k:n}) &:= \overline{S}_k(\{o_k\}|z_k) \prod_{i=k+1}^{n} \overline{S}_i(\{o_i\}|z_i) \overline{Q}_i(\{z_i\}|z_{i-1}).
\end{aligned}
$$

For any given sequence of states $z_{k:n} \in \mathscr{X}_{k:n}$, the $\alpha(z_{k:n})$ and $\beta(z_{k:n})$ can be found by simple backward recursion:

$$
\alpha(z_{k:n}) = \alpha(z_{k+1:n}) \overline{S}_k(\{o_k\}|z_k) \overline{Q}_{k+1}(\{z_{k+1}\}|z_k) \qquad (13)
$$

$$
\beta(z_{k:n}) = \beta(z_{k+1:n}) \underline{S}_k(\{o_k\}|z_k) \underline{Q}_{k+1}(\{z_{k+1}\}|z_k), \qquad (14)
$$

for $k \in \{1,\ldots,n-1\}$, and starting from:

$$
\begin{aligned}
\alpha(z_{n:n}) &= \alpha(z_n) = \overline{S}_n(\{o_n\}|z_n) \\
\beta(z_{n:n}) &= \beta(z_n) = \underline{S}_n(\{o_n\}|z_n).
\end{aligned}
$$

4 The Principle of Optimality

Determining the state sequences in $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ directly using Eq. (8) clearly has exponential complexity (in the length of the chain). We are now going to take a dynamic programming approach [1] to reducing this complexity by deriving a recursion equation for the optimality operators $\operatorname{opt}(\cdot|z_{k-1}, o_{k:n})$.

Theorem (Principle of Optimality)**.** *For $k \in \{1,\ldots,n-1\}$, all $z_{k-1} \in \mathscr{X}_{k-1}$ and all $\hat{x}_{k:n} \in \mathscr{X}_{k:n}$:*

$$
\begin{aligned}
&\hat{x}_{k:n} \in \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n}) \\
&\qquad \Rightarrow \hat{x}_{k+1:n} \in \operatorname{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k, o_{k+1:n}).
\end{aligned}
$$

Proof. Fix $k \in \{1,\ldots,n-1\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$ and $\hat{x}_{k:n} \in \mathscr{X}_{k:n}$. Assume that $\hat{x}_{k+1:n} \notin \operatorname{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k, o_{k+1:n})$, then we show that $\hat{x}_{k:n} \notin \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n})$. It follows from the assumption that there is some $x_{k+1:n} \in \mathscr{X}_{k+1}$ such that $\underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|\hat{x}_k) > 0$. Now prefix the state sequence $x_{k+1:n}$ with the state $\hat{x}_k$ to form the state sequence $x_{k:n}$, implying that $\hat{x}_k = x_k$. We then infer from Eq. (10) that

$$\begin{aligned}
&\underline{P}_k(\Delta[x_{k:n}, \hat{x}_{k:n}]|z_{k-1}) \\
&= \underline{Q}_k(\{\hat{x}_k\}|z_{k-1})\underline{S}_k(\{o_k\}|\hat{x}_k)\underline{P}_{k+1}(\Delta[x_{k+1:n}, \hat{x}_{k+1:n}]|\hat{x}_k) \\
&> 0,
\end{aligned}$$

which tells us that indeed $\hat{x}_{k:n} \notin \operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n})$. □

As an immediate consequence, we find that

$$\operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n}) \subseteq \bigcup_{z_k \in \mathscr{X}_k} z_k \oplus \operatorname{opt}(\mathscr{X}_{k+1:n}|z_k, o_{k+1:n}), \tag{15}$$

where $\oplus$ denotes concatenation of state sequences. From this we can infer that

$$\begin{aligned}
&\operatorname{opt}(\mathscr{X}_{k:n}|z_{k-1}, o_{k:n}) \\
&= \operatorname{opt}\left(\bigcup_{z_k \in \mathscr{X}_k} z_k \oplus \operatorname{opt}(\mathscr{X}_{k+1:n}|z_k, o_{k+1:n}) \,\middle|\, z_{k-1}, o_{k:n} \right),
\end{aligned} \tag{16}$$

since the optimality operator selecting the maximal elements in a strict partial order is insensitive to the omission of non-optimal elements; see Ref. [7] for a detailed discussion. While Eq. (16) clearly exhibits the reduction in computational complexity that the Principle of Optimality allows for, it is perhaps useful to point out here that we will not use this specific form for it in our algorithm.

5 An algorithm for finding maximal state sequences

Instead, we use Eq. (15) to devise an algorithm for constructing the set $\operatorname{opt}(\mathscr{X}_{1:n}|o_{1:n})$ of maximal state sequences in a recursive manner.

Initial set-up using backward recursion. We begin by defining a few auxiliary notions. First of all, we consider the thresholds:

$$\begin{aligned}
&\theta_k(\hat{x}_k, x_k|z_{k-1}) \\
&\quad := \min\left\{ a \in \mathbb{R} \colon \underline{Q}_k(\mathbb{I}_{\{x_k\}} - a\mathbb{I}_{\{\hat{x}_k\}}|z_{k-1}) \leq 0 \right\}
\end{aligned} \tag{17}$$

for all $k \in \{1,\ldots,n\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$ and $x_k, \hat{x}_k \in \mathscr{X}_k$. Observe that it follows from the positivity assumptions on the $\underline{Q}_k(\cdot|X_{k-1})$ that $\theta_k(\hat{x}_k, x_k|z_{k-1}) > 0$.

Next, we define

$$\alpha_k^{\max}(x_k) := \max_{\substack{z_{k:n} \in \mathscr{Z}_{k:n} \\ z_k = x_k}} \alpha(z_{k:n}) \tag{18}$$

and

$$\beta_k^{\max}(x_k) := \max_{\substack{z_{k:n} \in \mathscr{Z}_{k:n} \\ z_k = x_k}} \beta(z_{k:n}) \tag{19}$$

for all $k \in \{1,\ldots,n\}$ and $x_k \in \mathscr{X}_k$. Using Eq. (13)–(14), these can be calculated efficiently using the following backward recursive (dynamic programming) procedure:

$$\begin{aligned}
&\alpha_k^{\max}(x_k) \\
&= \max_{z_{k+1} \in \mathscr{Z}_{k+1}} \alpha_{k+1}^{\max}(z_{k+1})\overline{S}_k(\{o_k\}|x_k)\overline{Q}_{k+1}(\{z_{k+1}\}|x_k) \\
&= \overline{S}_k(\{o_k\}|x_k) \max_{z_{k+1} \in \mathscr{Z}_{k+1}} \alpha_{k+1}^{\max}(z_{k+1})\overline{Q}_{k+1}(\{z_{k+1}\}|x_k),
\end{aligned} \tag{20}$$

and

$$\begin{aligned}
&\beta_k^{\max}(x_k) \\
&= \max_{z_{k+1} \in \mathscr{Z}_{k+1}} \beta_{k+1}^{\max}(z_{k+1})\underline{S}_k(\{o_k\}|x_k)\underline{Q}_{k+1}(\{z_{k+1}\}|x_k) \\
&= \underline{S}_k(\{o_k\}|x_k) \max_{z_{k+1} \in \mathscr{Z}_{k+1}} \beta_{k+1}^{\max}(z_{k+1})\underline{Q}_{k+1}(\{z_{k+1}\}|x_k),
\end{aligned} \tag{21}$$

for $k \in \{1,\ldots,n-1\}$, starting from

$$\alpha_n^{\max}(x_n) = \alpha(x_n) = \overline{S}_n(\{o_n\}|x_n) \tag{22}$$

and

$$\beta_n^{\max}(x_n) = \beta(x_n) = \underline{S}_n(\{o_n\}|x_n). \tag{23}$$

Finally, we let

$$\alpha_k^{\operatorname{opt}}(\hat{x}_k|z_{k-1}) := \max_{\substack{x_k \in \mathscr{X}_k \\ x_k \neq \hat{x}_k}} \beta_k^{\max}(x_k)\theta_k(\hat{x}_k, x_k|z_{k-1}), \tag{24}$$

for all $k \in \{1,\ldots,n\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$ and $\hat{x}_k \in \mathscr{X}_k$.

Reformulation of the optimality condition. First, we consider $k = n$. For every $z_{n-1} \in \mathscr{Z}_{n-1}$, we determine $\operatorname{opt}(\mathscr{X}_n|z_{n-1}, o_n)$ as the set of those elements $\hat{x}_n$ of $\mathscr{X}_n$ for which

$$(\forall x_n \in \mathscr{X}_n \setminus \{\hat{x}_n\})\underline{Q}_n(\mathbb{I}_{\{x_n\}}\beta(x_n) - \mathbb{I}_{\{\hat{x}_n\}}\alpha(\hat{x}_n)|z_{n-1}) \leq 0,$$

as this condition is equivalent to condition (9) for $k = n$, considering Eqs. (11) and (12). But this condition is also equivalent to

$$(\forall x_n \in \mathscr{X}_n \setminus \{\hat{x}_n\})\frac{\alpha(\hat{x}_n)}{\beta_n^{\max}(x_n)} \geq \theta_n(\hat{x}_n, x_n|z_{n-1}),$$

considering Eqs. (23) and (17). Eq. (24) now tells us that this is equivalent to $\alpha(\hat{x}_n) \geq \alpha_n^{\operatorname{opt}}(\hat{x}_n|z_{n-1})$. In summary,

$$\operatorname{opt}(\mathscr{X}_n|z_{n-1}, o_n) = \left\{ \hat{x}_n \in \mathscr{X}_n \colon \alpha(\hat{x}_n) \geq \alpha_n^{\operatorname{opt}}(\hat{x}_n|z_{n-1}) \right\}. \tag{25}$$

Next, we consider any $k \in \{1,\ldots,n-1\}$. Fix $z_{k-1} \in \mathscr{Z}_{k-1}$, then we must determine $\mathrm{opt}(\mathscr{X}_{k:n}|z_{k-1},o_{k:n})$. We know from the Principle of Optimality (15) that we can limit the candidate optimal sequences $\hat{x}_{k:n}$ to the set

$$\bigcup_{z_k \in \mathscr{Z}_k} z_k \oplus \mathrm{opt}(\mathscr{X}_{k+1:n}|z_k,o_{k+1:n}). \tag{26}$$

Consider any such $\hat{x}_{k:n}$, then we must check for any $x_{k:n} \in \mathscr{X}_{k:n}$ whether $\underline{P}_k(\Delta[x_{k:n},\hat{x}_{k:n}]|z_{k-1}) \leq 0$; see Eq. (9). But if $x_{k:n}$ is such that $x_k = \hat{x}_k$, then it follows from Eq. (10) that $\underline{P}_k(\Delta[x_{k:n},\hat{x}_{k:n}]|z_{k-1}) \leq 0$, because the fact that $\hat{x}_{k+1:n} \in \mathrm{opt}(\mathscr{X}_{k+1:n}|\hat{x}_k,o_{k+1:n})$ also guarantees that $\underline{P}_{k+1}(\Delta[x_{k+1:n},\hat{x}_{k+1:n}]|\hat{x}_k) \leq 0$. So we can limit ourselves to checking the inequality for $x_{k:n}$ for which $x_k \neq \hat{x}_k$.

So fix any $x_k \neq \hat{x}_k$, then we must check whether

$$(\forall x_{k+1:n} \in \mathscr{X}_{k+1:n}) \quad \underline{Q}_k(\mathbb{I}_{\{x_k\}}\beta(x_{k:n}) - \mathbb{I}_{\{\hat{x}_k\}}\alpha(\hat{x}_{k:n})|z_{k-1}) \leq 0;$$

see Eq. (12). Considering Eq. (19), this is equivalent to

$$\underline{Q}_k(\mathbb{I}_{\{x_k\}}\beta_k^{\max}(x_k) - \mathbb{I}_{\{\hat{x}_k\}}\alpha(\hat{x}_{k:n})|z_{k-1}) \leq 0,$$

and therefore also equivalent to

$$\frac{\alpha(\hat{x}_{k:n})}{\beta_k^{\max}(x_k)} \geq \theta_k(\hat{x}_k,x_k|z_{k-1}),$$

considering Eq. (17). Since this inequality must hold for every $x_k \neq \hat{x}_k$, we infer from Eq. (24) that we must have that

$$\alpha(\hat{x}_{k:n}) \geq \alpha_k^{\mathrm{opt}}(\hat{x}_k|z_{k-1}). \tag{27}$$

So we must check this condition for all the candidate sequences $\hat{x}_{k:n}$ in the set (26). We can do this efficiently by using the following backward-forward recursion approach.

Backward-forward recursion. We start by letting k run *backward* from n to 1.

For $k = n$, it is a straightforward matter to determine $\mathrm{opt}(\mathscr{X}_n|z_{n-1},o_n)$ for every $z_{n-1} \in \mathscr{X}_{n-1}$ using Eq. (25).

For each $k < n$, we now show how we can determine $\mathrm{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$ by executing the following *forward running* procedure for every $z_{k-1} \in \mathscr{X}_{k-1}$.

If we combine Eqs. (27) and (18), we see that a necessary condition for $\hat{x}_k$ to be the state at time k in some optimal state sequence in $\mathrm{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$ is that

$$\alpha_k^{\max}(\hat{x}_k) \geq \alpha_k^{\mathrm{opt}}(\hat{x}_k|z_{k-1}), \tag{28}$$

meaning we can eliminate from our search those sequences for which the first state $\hat{x}_k$ does not satisfy this condition. On the other hand, for any $\hat{x}_k$ that satisfies the condition (28), we know from Eq. (18) that there is at least one state sequence with first state $\hat{x}_k$ that satisfies the condition (27).

So now we consider any $\hat{x}_k$ that satisfies the condition (28), and any $\hat{x}_{k+1}$ that is a first state in some optimal sequence in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$. Observe that we can determine whether $\hat{x}_{k+1}$ satisfies this condition, because we have determined $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ in the forward run for $k+1$.

Taking into account the recursion equation (13), we see that the condition (27) is equivalent to

$$\alpha(\hat{x}_{k+1:n}) \geq \alpha^{\mathrm{opt}}(\hat{x}_{k:k+1}|z_{k-1}), \tag{29}$$

where

$$\alpha^{\mathrm{opt}}(\hat{x}_{k:k+1}|z_{k-1}) := \frac{\alpha_k^{\mathrm{opt}}(\hat{x}_k|z_{k-1})}{\overline{S}_k(\{o_k\}|\hat{x}_k)\overline{Q}_{k+1}(\{\hat{x}_{k+1}\}|\hat{x}_k)}.$$

So if we combine Eqs. (29) and (18), we see that a necessary condition for $\hat{x}_{k+1}$ to be a state at time $k+1$ in some optimal sequence starting with $\hat{x}_k$ is that

$$\alpha_{k+1}^{\max}(\hat{x}_{k+1}) \geq \alpha^{\mathrm{opt}}(\hat{x}_{k:k+1}|z_{k-1}), \tag{30}$$

meaning we can eliminate from our search those sequences in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ for which the first state $\hat{x}_{k+1}$ does not satisfy this condition. On the other hand, for any $\hat{x}_{k+1}$ that satisfies the condition (30), we know from Eq. (18) [for $k+1$] that there is at least one state sequence in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ with first state $\hat{x}_{k+1}$ that satisfies the condition (29).

Next, we consider any $\hat{x}_k$ and $\hat{x}_{k+1}$ that satisfy the condition (30) and any $\hat{x}_{k+2}$ for which $\hat{x}_{k+1}$ and $\hat{x}_{k+2}$ are the first two states in some optimal sequence in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$. Taking into account the recursion equation (13), we see that the condition (27) is equivalent to

$$\alpha(\hat{x}_{k+2:n}) \geq \alpha^{\mathrm{opt}}(\hat{x}_{k:k+2}|z_{k-1}), \tag{31}$$

where

$$\alpha^{\mathrm{opt}}(\hat{x}_{k:k+2}|z_{k-1}) := \frac{\alpha^{\mathrm{opt}}(\hat{x}_{k:k+1}|z_{k-1})}{\overline{S}_{k+1}(\{o_{k+1}\}|\hat{x}_{k+1})\overline{Q}_{k+2}(\{\hat{x}_{k+2}\}|\hat{x}_{k+1})}.$$

So if we combine Eqs. (31) and (18), we see that a necessary condition for $\hat{x}_{k+2}$ to be a state at time $k+2$ in some optimal sequence starting with $\hat{x}_{k:k+1}$ is that

$$\alpha_{k+2}^{\max}(\hat{x}_{k+2}) \geq \alpha^{\mathrm{opt}}(\hat{x}_{k:k+2}|z_{k-1}), \tag{32}$$

meaning we can eliminate from our search those sequences in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ for which the second state $\hat{x}_{k+2}$ does not satisfy this condition. On the other hand, for any $\hat{x}_{k+2}$ that satisfies the condition (32), there is at least one state sequence in $\mathrm{opt}(\mathscr{X}_{k+1}|\hat{x}_k,o_{k+1:n})$ with a second state $\hat{x}_{k+2}$ that satisfies the condition (31).

It should be clear that we can go forward in this way until we reach time n, and that in doing so we construct all the sequences $\hat{x}_{k:n}$ in $\mathrm{opt}(\mathscr{X}_k|z_{k-1},o_{k:n})$.

A brief discussion of the algorithm's complexity. We begin with the preparatory calculations of the quantities in Eqs. (17)–(24). For the thresholds $\theta_k(\hat{x}_k, x_k|z_{k-1})$ in Eq. (17), the computational complexity is clearly cubic in the number of states, and linear in the number of nodes. Calculating the $\alpha_k^{\max}(x_k)$ and $\beta_k^{\max}(x_k)$ in Eqs. (20) and (21) is linear in the number of nodes, and quadratic in the number of states. The complexity of finding the $\alpha_k^{\text{opt}}(\hat{x}_k|z_{k-1})$ in Eq. (24) is linear in the number of nodes, and cubic in the number of states.

On the other hand, the computational complexity of the backward-forward loop is clearly quadratic in the number of nodes, quadratic in the number of states, and roughly speaking linear in the number of maximal sequences.[3]

For precise HMMs, the state sequence estimation problem can be solved very efficiently by the Viterbi algorithm [11, 13], whose complexity is linear in the number of nodes, and quadratic in the number of states. However, this algorithm only emits a single optimal (most probable) state sequence, even in cases where there are multiple (equally probable) optimal solutions: this of course simplifies the problem. If we would content ourselves with giving only a single maximal solution, the ensuing algorithm would have a complexity that is similar to Viterbi's. So, to allow for a fair comparison between Viterbi's algorithm and ours, we would need to alter Viterbi's algorithm in such a way that it no longer resolves ties arbitrarily, and emits all (equally probable) optimal state sequences. This new version will remain linear in the number of nodes, and quadratic in the number of states, but emitting the optimal sequences will be linear in the number of them. For the complexity for the most time-consuming part of our algorithm (the backward-forward loop), the only difference is this: Viterbi's approach is linear and ours quadratic in the number of nodes. Where does this difference come from? In iHMMs we have mutually incomparable solutions, whereas in pHMMs the optimal solutions are indifferent, or equally probable. This makes sure that the algorithm for pHMMs requires no forward loops. We believe that this added complexity is a reasonable price to pay for the robustness that working with imprecise-probabilistic models offers.

Additional comments. All that is needed in order to produce the α- and β-functions are assessments for the lower and upper transition and emission mass functions:

$$\underline{Q}_k(\{z_k\}|z_{k-1}), \overline{Q}_k(\{z_k\}|z_{k-1}), \underline{S}_k(\{o_k\}|z_k), \overline{S}_k(\{o_k\}|z_k)$$

for all $k \in \{1,\ldots,n\}$, $z_{k-1} \in \mathscr{Z}_{k-1}$, $z_k \in \mathscr{Z}_k$ and $o_k \in \mathscr{O}_k$. The most conservative coherent models $\underline{Q}_k(\cdot|X_{k-1})$ that correspond to such assessments are 2-monotone [4, 8]. Due to their comonotone additivity, this implies that:

$$\underline{Q}_k(\mathbb{I}_{\{x_k\}} - a\mathbb{I}_{\{\hat{x}_k\}}|z_{k-1}) = \underline{Q}_k(\{x_k\}|z_{k-1}) - a\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})$$

for all $a \geq 0$, and therefore Eq. (17) leads to

$$\theta_k(\hat{x}_k, x_k|z_{k-1}) = \frac{\underline{Q}_k(\{x_k\}|z_{k-1})}{\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})}. \tag{33}$$

The right-hand side is the smallest possible value of the threshold $\theta_k(\hat{x}_k, x_k|z_{k-1})$ corresponding to the assessments $\underline{Q}_k(\{x_k\}|z_{k-1})$ and $\overline{Q}_k(\{\hat{x}_k\}|z_{k-1})$, leading to the most conservative inferences, and therefore the largest possible sets of maximal sequences, that correspond to these assessments.

6 Some experiments

While a linear complexity in the number of maximal sequences is probably the best we can hope for, we also see that we will only be able to find all maximal sequences efficiently provided their number is reasonably small. Should it, say, tend to increase exponentially with the length of the chain, then no algorithm, however cleverly designed, could overcome this hurdle. Because this number of maximal sequences is so important, we study its behaviour in more detail. In order to do so, we take a closer look at how this number of maximal sequences depends on the transition probabilities of the model, and how it evolves when we let the imprecision of the local models grow. We shall see that this number displays very interesting behaviour that can be explained, and even predicted to some extent. To allow for easy visualisation, we limit this discussion to binary iHMMs, where both the state and output variables can assume only two possible values, say 0 and 1.

Describing a binary stationary iHMM. We first consider a binary stationary HMM. The (precise) transition probabilities for going from one state to the next are completely determined by numbers in the unit interval: the probability p to go from state 0 to state 0, and the probability q to go from state 1 to state 0. To further pin down the HMM we also need to specify the (marginal) probability m for the first state to be 0, and the two emission probabilities: the probability r of emitting output 0 from state 0 and the probability s of emitting output 0 from state 1.

In this binary case, all imprecise models can be found by contamination: taking convex mixtures of precise models, with mixture coefficient $1-\varepsilon$, and the vacuous model, with mixture coefficient ε, leading to a so-called linear-vacuous model. To simplify the analysis, we let the emission model remain precise, and use the same mixture coefficient ε for the marginal and the transition models. As ε ranges from zero to one, we then evolve from a precise HMM towards an iHMM with vacuous marginal and transition models (and precise emission models).

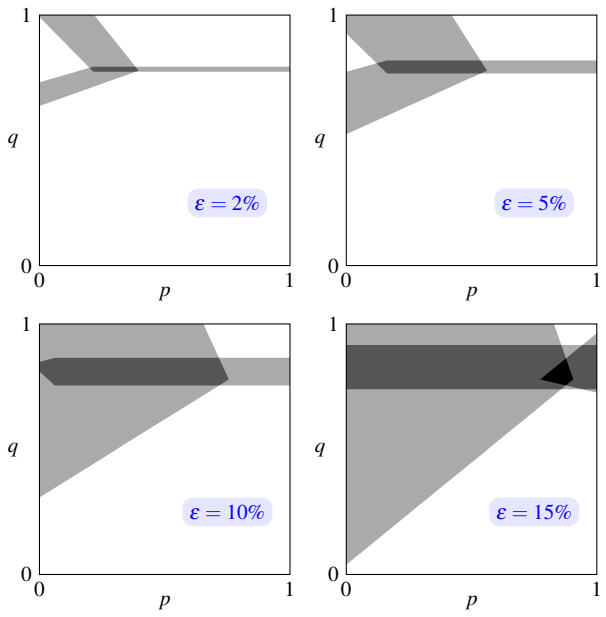
Explaining the basic ideas using a chain of length two. We now examine the behaviour of an iHMM of length two,

[3] Each backward step in the backward-forward loop has a linear complexity in the number of maximal elements at that stage.

with the following (precise) probabilities fixed:[4]

$$m = 0.1,\ r = 0.8 \text{ and } s = 0.3.$$

Fixing an output sequence and a value for ε, we can use our algorithm to calculate the corresponding numbers of maximal state sequences as p and q range over the unit interval. The results can be represented conveniently in the form of a heat plot. The plots below correspond to the output sequence $o_{1:2} = 01$.



The number of maximal state sequences clearly depends on the transition probabilities p and q. In the rather large parts of 'probability space' that are coloured white, we get a single maximal sequence—as we would for HMMs—, but there are contiguous regions where we see a higher number appear. In the present example (binary chain of length two), the highest possible number of maximal sequences is of course four. In the dark grey area, there are three maximal sequences, and two in the light grey regions. The plots show what happens when we let ε increase: the grey areas expand and the number of maximal sequences increases. For $\varepsilon = 15\%$, we even find a small area (coloured black) where all four possible state sequences are maximal: locally, due to the relatively high imprecision of our local models, we cannot give any useful robust estimates of the state sequence producing the output sequence $o_{1:2} = 01$.

For small ε, the areas with more than one maximal state sequence are quite small and seem to resemble strips that narrow down to lines as ε tends to zero. This suggests that we should be able to explain at least qualitatively where these areas come from by looking at compatible precise models: the regions where an iHMM produces different maximal (mutually incomparable) sequences, are widened versions of loci of indifference for precise HMMs.

By a *locus of indifference*, we mean the set of (p,q) that correspond to two given state sequences $x_{1:2}$ and $\hat{x}_{1:2}$ having equal posterior probability:

$$p(x_{1:2}|o_{1:2}) = p(\hat{x}_{1:2}|o_{1:2}),$$

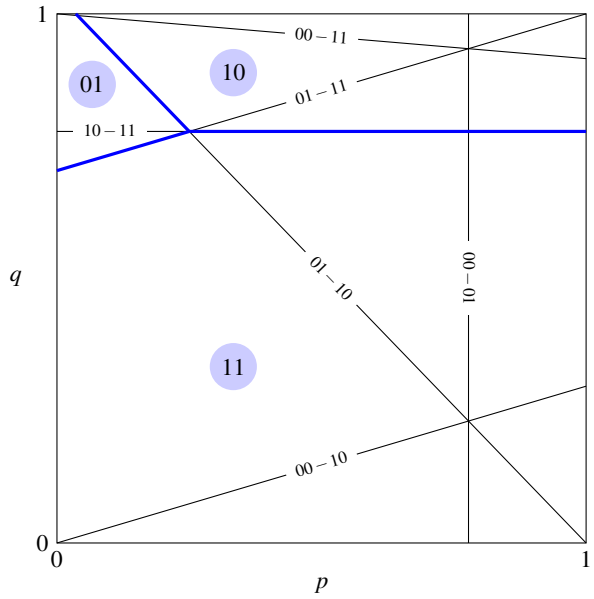
or, provided that $p(o_{1:2}) > 0$,

$$p(x_{1:2}, o_{1:2}) = p(\hat{x}_{1:2}, o_{1:2}).$$

In our example where $o_{1:2} = 01$, we find the following expressions for each of the four possible state sequences:

$$\begin{aligned} p(00,01) &= mr(1-r)p \\ p(01,01) &= mr(1-s)(1-p) \\ p(10,01) &= (1-m)s(1-r)q \\ p(11,01) &= (1-m)s(1-s)(1-q) \end{aligned}$$

By equating any two of these expressions, we express that the corresponding two state sequences have an equal posterior probability. Since the resulting equations are a function of p and q only, each of these six possible combinations defines a locus of indifference. All of them are depicted as lines in the following figure:

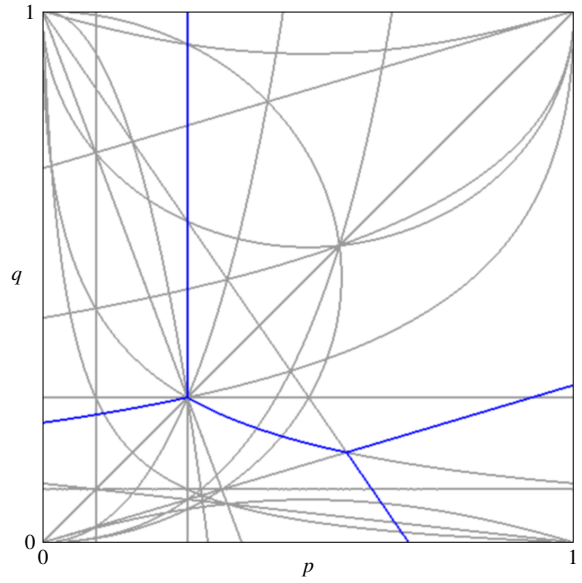


Parts of these loci, depicted in blue (darker and bolder in monochrome versions of this paper) demarcate the three regions where the state sequences 01, 10 and 11 are optimal (have the highest posterior probability).

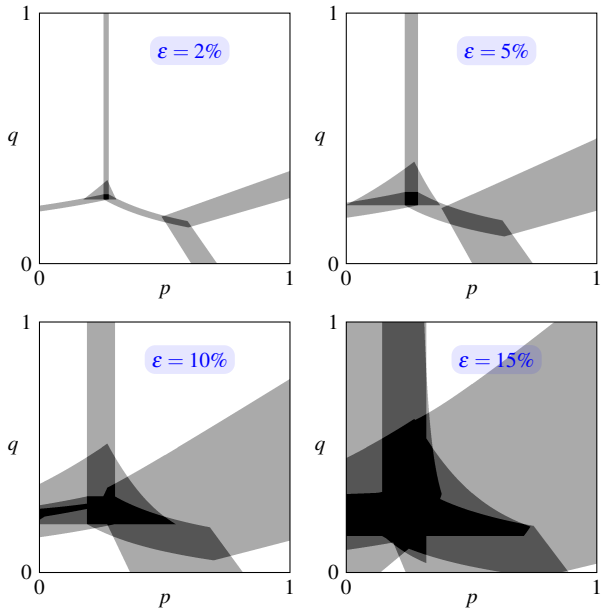
What happens when the transition models become imprecise? Roughly speaking, nearby values of the original p and q enter the picture, effectively turning the loci (lines) of indifference into bands of incomparability: the emergence of regions with two and more maximal sequences can be seen to originate from the loci of indifference; compare the figure for these loci with the heat plots given above.

[4] This choice is of course arbitrary. Different values would yield comparable results.

Extending the argument to a chain of length three. For a chain of length three, we can determine the loci of indifference for precise models in a completely analogous manner. If we use the same marginal model and emission model as in the previous example, the resulting lines of indifference for the output sequence 000 look as follows:



If we compare this with the visualisation below of the number of maximal elements for the same sequence, the resemblance is again quite striking.



7 Showing off the algorithm's power

In order to demonstrate that our algorithm is indeed quite efficient, we let it determine the maximal sequences for a random output sequence of length 100.

We consider the same binary stationary HMM as we presented above, but with the following precise marginal and emission probabilities:

$$m = 0.1,\ r = 0.98 \text{ and } s = 0.01.$$

In practical applications, the probability for an output variable to have the same value as the corresponding hidden state variable is usually quite high, which explains why we have chosen r and s to be close to 1 and to 0, respectively. In contrast with the previous experiments, we do not let the transition probabilities vary, but fix them to the following values:

$$p = 0.6 \text{ and } q = 0.5.$$

The iHMM we use to determine the maximal sequences is then generated by mixing these precise local models with a vacuous one, using the same mixture coefficient ε for the marginal, transition and emission models. On the right, we display the five maximal sequences corresponding to the highlighted output sequence, and $\varepsilon = 2\%$. Since the emission probabilities were chosen to be quite accurate, it is no surprise that the output sequence itself is one of the maximal sequences. In addition, we have indicated in bold face the state values that differ from the outputs in the output sequence. We see that the model represents more indecision about the values of the state variables as we move further away from the end of the sequence. This is a result of a phenomenon called *dilation*, which—as has been noted in another paper [5]—tends to occur when inferences in a credal tree proceed from the leaves towards the root.

As for the efficiency of our algorithm: it took about 0.2 seconds to calculate these 5 maximal sequences. The reason why this could be done so fast is that the algorithm is linear in the number of solutions, which in this case is only 5. If we let ε grow to for example 5%, the number of maximal sequences for the same output sequence is 764 and these can be determined in about 32 seconds. This demonstrates that the complexity is indeed linear in the number of solutions and that the

algorithm can efficiently calculate the maximal sequences even for long output sequences.

8 Conclusions

Interpreting the graphical structure of an imprecise hidden Markov model as a credal network under epistemic irrelevance, leads to an efficient algorithm for finding the maximal state sequences for a given output sequence. Preliminary simulations show that, even for transition models with non-negligible imprecision, the number of maximal elements seems to be reasonably low in fairly large regions of parameter space, with high numbers of maximal elements concentrated in fairly small regions. It remains to be seen whether this observation can be corroborated by a theoretical analysis, and whether increasing the imprecision of the emission models changes this picture appreciably.

It is not clear to us, at this point, whether ideas similar to the ones we discussed above could be used to derive similarly efficient algorithms for imprecise hidden Markov models whose graphical structure is interpreted as a credal network under strong independence [3]. This could be interesting and relevant, as the more stringent independence condition leads to joint models that are less imprecise, and therefore produce fewer maximal state sequences (although they will be contained in our solutions).

Acknowledgements

Jasper De Bock is a Master student of Civil Engineering at Ghent University, and has developed the algorithm described here in the context of his Master's thesis, in close cooperation with Gert de Cooman, who acted as his thesis supervisor.

Research by De Cooman has been supported by SBO project 060043 of the IWT-Vlaanderen. This paper has benefitted from discussions with Marco Zaffalon, Alessandro Antonucci, Alessio Benavoli, Cassio de Campos, Erik Quaeghebeur and Filip Hermans. We are grateful to Marco Zaffalon for providing travel funds allowing us to visit IDSIA and discuss practical applications.

References

[1] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

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

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

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

[5] Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51:1029–1052, 2010.

[6] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design (Proceedings of IPMU 2010, 28 June – 2 July 2010, Dortmund, Germany)*, volume 6178 of *Lecture Notes in Computer Science*, pages 737–746. Springer, Heidelberg, 2010.

[7] Gert de Cooman and Matthias C. M. Troffaes. Dynamic programming for deterministic discrete-time systems with uncertain gain. *International Journal of Approximate Reasoning*, 39:257–278, 2005.

[8] Gert de Cooman, Matthias C. M. Troffaes, and Enrique Miranda. n-Monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347:143–156, 2008.

[9] Nathan Huntley and Matthias C. M. Troffaes. Normal form backward induction for decision trees with coherent lower previsions. *Annals of Operations Research*, 2010. Submitted for publication.

[10] Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658.

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

[12] Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29.

[13] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

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