Hidden Markov Models With Imprecisely Specified Parameters

Denis Deratani Mauá
Escola Politécnica
Universidade de São Paulo
Email: denis.maua@usp.br

Cassio Polpo de Campos and Alessandro Antonucci
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Manno, Switzerland
Email: {cassio,alessandro}@idsia.ch

***Abstract*—Hidden Markov models (HMMs) are widely used models for sequential data. As with other probabilistic graphical models, they require the specification of precise probability values, which can be too restrictive for some domains, especially when data are scarce or costly to acquire. We present a generalized version of HMMs, whose quantification can be done by sets of, instead of single, probability distributions. Our models have the ability to suspend judgment when there is not enough statistical evidence, and can serve as a sensitivity analysis tool for standard non-stationary HMMs. Efficient inference algorithms are developed to address standard HMM usage such as the computation of likelihoods and most probable explanations. Experiments with real data show that the use of imprecise probabilities leads to more reliable inferences without compromising efficiency.**

I. Introduction

Sequential data naturally arises in a wide range of domains such as speech [1] and text processing [2], activity recognition [3] and computational biology [4]. Hidden Markov models (HMMs) are widely used generative probabilistic models of sequential data that assume observations to be produced from a chain of hidden (i.e., unobserved) states [1]. An HMM comprises a *prior probability distribution*, which specifies the probability that the process originates in a given state, a *transition probability distribution*, which specifies the probability that the process will transit from a given state to another, and an *emission probability distribution*, which specifies the probability of observing a symbol conditional on a state.

In many real world applications, the transitions between consecutive hidden states and the relation between a hidden variable and the corresponding observation are affected by severe uncertainty. This is the case, for instance, when data are scarce [5], observations are missing not-at-random [6], and information is conflicting. In such cases, the use of probability distributions to represent uncertainty might be inadequate and lead to overly confident inferences [7].

Credal sets [8] are closed convex sets of probability distributions that for a more general representation of uncertainty, including the situations just described. For instance, complete ignorance about a variable is represented as the credal set of all probability distributions in that variable, instead of the more common representation as a uniform probability distribution. The *imprecise Dirichlet model* (IDM, [9]) learns credal sets from categorical data in a situation of near prior ignorance, providing a more flexible (although less informative) model of the underlying distribution than the more common Multinomial-Dirichlet model.

This paper develops the *imprecise hidden Markov model* (iHMM), which allows the specification of a time- and state-discrete HMM with prior, transition and emission credal sets in lieu of probability distributions. iHMMs provide a sound way to handle severe uncertainty, with two direct benefits. First, they allow us to suspend judgment when there is not sufficient evidence in the training set to make a reliable decision [10]. Second, they provide an efficient tool for performing sensitivity analysis [11] in standard non-stationary HMMs, allowing parameters to vary jointly, and in time. To our knowledge, we present the first polynomial-time algorithm to perform such an analysis.

In the rest of the paper, we review the related work (Sec. II), present the iHMM, and describe algorithms to deal with common uses such as comparing models according to the data likelihood (Sec. IV), and finding the most likely hidden state sequence for a given sequence of observations (Sec. V). Experiments with a part-of-speech tagging and a activity recognition tasks (Sec. VI) provide evidence that iHMMs are indeed capable of making reliable decisions, and evaluating the sensitivity of HMMs to the learning sample size. Conclusions and future work are described in Sec. VII.

II. Related Work

Bayesian networks [12] are probabilistic graphical models where conditional independences are represented by a graph whose nodes are identified with random variables. HMMs are part of a special class of tractable Bayesian networks, one for which efficient inference algorithms are available. As with HMMs, Bayesian networks require uncertainty to be represented by conditional probability distributions. Credal networks [13] extend Bayesian networks to allow uncertainty to be modeled as credal sets. The iHMMs we develop here are special cases of tree-shaped credal networks.

Drawing inferences with credal networks is a notoriously hard problem. Exact marginal inference is NP-hard even for polytree-shaped networks [14]. A few exceptions to that appear in the literature. [15] developed an efficient algorithm for the special case of updating polytree-shaped credal networks with binary variables. [16] described a method to compute joint queries in tree-shaped networks when there is no evidence.

The algorithmic techniques discussed in the paragraph above deal with the interpretation of imprecision in the parameters known as strong independence. Strong independence, which we adopt in this work, assumes the existence of an ideal probability distribution which we cannot characterize for lack of resources. Epistemic irrelevance (or its symmetrical counterpart epistemic independence) on the other hand makes no such claim, and allows for the possibility that there might not be any single probability distribution capable of representing a certain piece of (uncertain) knowledge. [17] presented an efficient algorithm for single-query marginal inferences in tree-shaped credal networks under epistemic irrelevance. As an HMM is a tree-shaped credal network, filtering (i.e., estimating the marginal probability of the future state given a sequence of observations) can be performed in polynomial-time if epistemic irrelevance by their algorithm. Recently, [18] showed that filtering on iHMM provides the same results whether one adopts strong independence or epistemic irrelevance. Hence, filtering is also polynomial-time computable in iHMMs under strong independence.

For other marginal inference tasks, efficiency can be achieved at the expense of accuracy. [19] developed an approximate method based on linear programming relaxations that was shown to outperform other approximate methods for marginal inference.

The use of credal sets in modeling sequential data is not new. [20] investigated Markov chains with interval-valued transition probabilities. [21] used credal sets for sensitivity analysis in Markov chains. [22] and [23] defined imprecise Markov chains, and analyzed some basic asymptotic behaviors such as regularity and ergodicity. [24] studied imprecise Markov chains with absorbing states. [25] investigated the use of iHMMs under epistemic irrelevance for tracking tasks. [7] defined an iHMM over continuous variables aimed at performing robust filtering. An imprecise version of the Baum-Welch procedure [1], used to estimate the parameters of an HMM when the state sequence is not observable, was developed by [26], and tested on a activity recognition task. [27] extended the learning of iHMMs from data to the case of epistemic irrelevance. [28] designed an algorithm that computes the maximal joint state sequences of an iHMM. A state sequence is maximal if there is no other state sequence with greater probability under any distribution induced by the model. [29] designed a method for comparing two iHMMs according to their asymptotic data likelihood, and also applied it on activity recognition.

III. Imprecise Hidden Markov Models

Let Q_t denote a family of discrete variables parametrized by t and taking values in a finite set of states $\mathcal{Q} = \{z_1, \dots, z_N\}$. The parameter t is called *time (step)*. For convenience, we identify each state z_i with its subscript i. Similarly, let O_t denote a family of discrete variables taking values in a set of possible outcomes $\mathcal{O} = \{y_1, \dots, y_M\}$.

Definition 1. *A hidden Markov model (HMM) is a tuple $\lambda = (a_2^1, \dots, a_T^N, b_1^1, \dots, b_T^N, \pi)$, where a_t^i,$i = 1, \dots, N$,$t = 2, \dots, T$, and b_t^i, $i = 1, \dots, N$,$t = 1, \dots, T$, are transition and emission probability distributions, respectively, and π is the prior probability distribution. The model is said to be* stationary *if for any i, t and t' we have that $a_t^i = a_{t'}^i$ and $b_t^i = b_{t'}^i$.*

An HMM defines a joint probability distribution

$$p(q_{1:T}, o_{1:T}|\lambda) = \pi(q_1) b_1^{q_1}(o_1) \prod_{t=1}^{T} a_t^{q_{t-1}}(q_t) b_t^{q_t}(o_t) \qquad (1)$$

over the set of *hidden* variables $Q_1, \dots, Q_T$ and *manifest* variables $O_1, \dots, O_T$ such that the sequence $Q_{1:T} := (Q_1, \dots, Q_T)$ is a Markov chain (i.e., Q_{t+1} is conditionally independent of $Q_1, \dots, Q_{t-1}$ given Q_t, for each t), and the probability of the variables in $O_{1:T} := (O_1, \dots, O_T)$ are only affected by the corresponding hidden variables (i.e., O_t is conditionally independent of all other variables given Q_t).

For any set of random variables $X_{1:r} = \{X_1, \dots, X_r\}$, $K_{X_{1:r}}$ denotes a credal set over $X_{1:r}$, that is, a closed convex set of probability distributions $p(X_{1:r})$. A conditional credal set K_X^w of conditional probability distributions $p(X|W = w)$ is obtained by element-wise application of Bayes' rule over $K_{X,W}$ for $W = w$. The lower and upper probability for an event $X = x$ are defined, respectively, by $\underline{p}(x|K_X) := \min_{p \in K_X} p(x)$ and $\overline{p}(x|K_X) := \max_{p \in K_X} p(x)$.

Definition 2. *An imprecise hidden Markov model (iHMM) is a tuple $\Lambda = (A_2^1, \dots, A_T^N, B_1^1, \dots, B_T^N, \Pi)$, where $A_t^i := K_{Q_t}^i$, $i = 2, \dots, N$,$t = 1, \dots, T$, and $B_t^i := K_{O_t}^i$, $i = 1, \dots, N$,$t = 1, \dots, T$, are credal sets of transition and emission distributions, respectively, and Π is the credal set of prior distributions. An iHMM is said to be* homogenous *if for all i: (i) transition credal sets $A_2^i, \dots, A_T^i$ are equal, and (ii) emission credal sets $B_1^i, \dots, B_T^i$ are equal.*

The above notation emphasizes the analogies with standard HMMs. In fact, an iHMM can be seen as a set of precise HMMs, one for each combination of probability distributions $a_2^1 \in A_2^1, \dots, a_T^N \in A_T^N, b_1^1 \in B_1^1, \dots, b_T^N \in B_T^N, \pi \in \Pi$. Yet another view of an iHMM is as the set of joint probability distributions that factorizes as in (1).

Example. *Consider the following matrices containing lower and upper probabilities.*

$$\underline{A} = \begin{pmatrix} 0.3 & 0.4 \\ 0.5 & 0.4 \end{pmatrix} \qquad \overline{A} = \begin{pmatrix} 0.6 & 0.7 \\ 0.6 & 0.5 \end{pmatrix}$$

$$\underline{B} = \begin{pmatrix} 0.1 & 0.3 \\ 0.2 & 0.3 \end{pmatrix} \qquad \overline{B} = \begin{pmatrix} 0.7 & 0.9 \\ 0.7 & 0.8 \end{pmatrix}$$

$$\underline{\Pi} = (0.3 \quad 0.6) \qquad \overline{\Pi} = (0.4 \quad 0.7)$$

Let Π be the set of prior probability distributions π such that $\pi(j) \in \left[\underline{\Pi}_j, \overline{\Pi}_j\right]$, A_2^i be the set of transition probability distributions a_2^i such that $a_2^i(j) \in \left[\underline{A}_{ij}, \overline{A}_{ij}\right]$, and B_t^i be the set of emission probability distributions b_t^i such that $b_t^i(j) \in \left[\underline{B}_{ij}, \overline{B}_{ij}\right]$, $i = 1, 2$, $j = 1, 2$, and $t = 1, 2$. The tuple $(A_2^1, A_2^2, B_1^1, B_1^2, B_2^1 B_2^2, \Pi)$ specifies an homogenous iHMM.

Given an iHMM $\Lambda = (A_2^1, \dots, A_T^N, B_1^1, \dots, B_T^N, \Pi)$, we write $A_t := \times_{i=1}^N A_t^i$ to denote the Cartesian product of transition credal sets at time t (analogously for B_t), $A_{t_0:t_f} := \times_{t=t_0}^{t_f} A_t$ to denote the Cartesian product of credal sets from time t_0 to time t_f (analogously for $B_{t_0:t_f}$), and $A_1^i := \Pi$ for any i.

IV. Likelihood

HMMs are commonly used to classify sequential data by choosing the model that best fits a sequence of observations according to the likelihood. Formally, we can define the task as follows. Given a finite set Λ of (precise) HMMs, and a sequence of observations $o_{1:T} \in \times_{t=1}^{T} \mathcal{O}$, determine $\lambda^* = \operatorname{argmax}_{\lambda \in \Lambda} p(o_{1:T}|\lambda)$. In order to generalize to the credal setting, we use the following notion of dominance.

Definition 3. *Given two iHMMs Λ_1 and Λ_2 and an observation sequence $o_{1:T}$, we say that Λ_1* dominates *Λ_2 for $o_{1:T}$, denoted $\Lambda_1 \succ \Lambda_2$, if and only if*

$$\underline{p}(o_{1:T}|\Lambda_1) - \overline{p}(o_{1:T}|\Lambda_2) > 0 \; . \tag{2}$$

Dominance suggests that iHMMs can be used as *credal classifiers* [16] for reliable/robust sequence classification in the same way as HMMs are used for classifying sequential data. A class label associated to model Λ_1 is preferred as a classification of $o_{1:T}$ over a class label associated to model Λ_2 if and only if $\Lambda_1 \succ \Lambda_2$. Given a finite set $\Lambda_1, \ldots, \Lambda_K$ of iHMMs, credal classification outputs the set of undominated models

$$\Lambda^* = \{\Lambda_k | \nexists j \neq k \text{ such that } \Lambda_j \succ \Lambda_k\} \; . \tag{3}$$

The elements of Λ^* are determined by computing the upper and lower likelihood for each model Λ_k, $k = 1, \ldots, K$. In the rest of this section, we present an algorithm to compute such probabilities that is inspired on the algorithm of [16] for computing joint queries with no evidence in credal trees. For the sake of brevity, we present only the derivation for the lower bound. The case for upper likelihoods is analogous.

Let the *lower backward* variable be given by

$$\beta_t(i) = \underline{p}(o_{t:T}|Q_{t-1} = i, \Lambda) \; . \tag{4}$$

The credal sets $K^i_{O_{t:T}}$, $i = 1, \ldots, N$, are said to be *compatible* if it is possible to simultaneously consider probability distributions that attain lower probabilities $\underline{p}(o_{t:T}|Q_{t-1} = i, \Lambda)$ for all i. The variables $\beta_t(i)$ can be computed recursively, according to the following inductive procedure.

1. Basis: For $t = T + 1$,

$$\beta_{T+1}(i) = 1, \qquad \text{for all } i \tag{5}$$

Clearly, $\{\beta_{T+1}(i)\}_{i=1}^{N}$ are compatible.

2. Induction: From $t = T$ to $t = 2$,

$$\begin{aligned} \beta_t(i) &= \min_{\substack{a_t^i \in A_t^i \\ b_t \in B_t}} \sum_{j=1}^{N} a_t^i(j) b_t^j(o_t) \beta_{t+1}(j) && (6) \\ &= \min_{\substack{a_t^i \in A_t^i \\ b_t \in B_t}} \sum_{j=1}^{N} a_t^i(j) b_t^j(o_t) \underline{p}(o_{t+1:T}|j, \Lambda) && (7) \\ &= \underline{p}(o_{t:T}|Q_{t-1} = i, \Lambda) \; , && (8) \end{aligned}$$

where we have assumed by inductive hypothesis in (6)–(7) that (4) holds at time $t + 1$ for all j. In the passage (7)–(8), the lower bound can be obtained by assuming compatibility at time $t + 1$ and noting the separate specification of transition and emission credal sets, followed by simple marginalization over Q_t.

3. Termination: For $t = 1$,

$$\begin{aligned} \beta_1(i) &= \min_{\substack{\pi \in \Pi \\ b_t \in B_t}} \sum_{j=1}^{N} \pi(j) b_1^j(o_t) \beta_2(j) && (9) \\ &= \underline{p}(o_{1:T}|\Lambda) \; . && (10) \end{aligned}$$

Let U denote the worst case time spent to compute one iteration of the recursion in (6). The algorithm solves $O(NT)$ recursions of $\beta_t(i)$, taking a total time of $O(TNU)$. The time to solve each recursion depends on how credal sets are specified.

Consider the case where credal sets are specified by probability intervals. A (conditional) credal set K_X^z is said to be specified by interval-valued probabilities if it can be described by a set of inequalities in the form

$$\begin{aligned} 0 \leq \ell(x) \leq p(x|z) \leq u(x) \leq 1, \forall x \in \mathcal{X} && (11) \\ \sum_x p(x|z) = 1 \; . && (12) \end{aligned}$$

Let us define the auxiliary variable $v_t^j = \min_{b_t^j \in B_t^j} b_t^j(o_t) \beta_{t+1}(j)$. Since credal sets are separately specified, variables v_t^j for $j = 1, \ldots, N$ can be optimized independently, and recursion (6) can be solved by $\min_{a_t^i \in A_t^i} \sum_{j=1}^{N} a_t^i(j) v_t^j$. Because each A_t^i is defined through intervals and the values v_t^j are known, this problem reduces to a continuous knapsack problem, which can be solved in $U = O(N)$ time [16].

If the credal sets are specified by finite sets of distributions, the optimization in (6) can be solved by enumeration of all possible solutions [15], which can be done in $O(E)$ time, where E is the maximum cardinality of a set in Λ.

V. Most Probable Explanation

An important use of (precise) HMMs is to infer the most probable configuration of the hidden states for a given sequence of observations, that is,

$$q_{1:T}^* = \operatorname*{argmax}_{q_{1:T}} \frac{p(q_{1:T}, o_{1:T})}{\sum_{i_{1:T}} p(i_{1:T}, o_{1:T})} \; . \tag{13}$$

The Viterbi algorithm [30] can be used to efficiently solve (13) for the optimal sequence based on the fact that the denominator on the right-hand side is constant with respect to the choice of $q_{1:T}$ and can thus be excluded from the computation. This approach can be generalized to iHMMs by adopting the following criteria.

Definition 4. *Given an iHMM Λ, the* joint maximin *and* joint maximax explanations *for a sequence $o_{1:T}$ are given by, respectively,*

$$\begin{aligned} &\operatorname*{argmax}_{q_{1:T}} \underline{p}(q_{1:T}, o_{1:T}|\Lambda) \; , && (14) \\ &\operatorname*{argmax}_{q_{1:T}} \overline{p}(q_{1:T}, o_{1:T}|\Lambda) \; . && (15) \end{aligned}$$

The maximin and maximax explanations describe extreme scenarios where the distributions are chosen, respectively, in a pessimistic and optimistic way. We can use this fact to analyze the sensitivity of precise HMMs to fluctuations of the parameters over time and within the bounds defined by the local credal sets. In this sense, explanations provided by HMMs comprised in an iHMM can be regarded as reliable if both maximin and maximax explanations coincide.

Note that, unlike the case for precise HMMs, the explanations obtained using (14) and (15) may differ from the explanations provided by extreme scenarios of the posterior probabilities, since in the equations

$$\operatorname*{argmax}_{q_{1:T}} \min_{p \in K_{Q_{1:T},O_{1:T}}} \frac{p(q_{1:T}, o_{1:T})}{\sum_{i_{1:T}} p(i_{1:T}, o_{1:T})} , \tag{16}$$

$$\operatorname*{argmax}_{q_{1:T}} \max_{p \in K_{Q_{1:T},O_{1:T}}} \frac{p(q_{1:T}, o_{1:T})}{\sum_{i_{1:T}} p(i_{1:T}, o_{1:T})} . \tag{17}$$

the denominators vary with the choice of state sequence and hence cannot be excluded. Empirical results with artificial data (omitted from this paper due to lack of space) show that, at least for small chains, the explanations obtained using joint probability do not differ significantly from those obtained by the posterior probabilities. Additionally, the joint seems to provide explanations at least as reliable as the posterior.

In what follows, we present an algorithm for computing joint maximin explanations based in (14). A similar algorithm based in (15) can be obtained by analogy.

Let $\delta_t(i)$ and $\phi_t(i)$ be defined, respectively, as

$$\delta_t(i) = \max_{q_{t:T}} \underline{p}(q_{t:T}, o_{t:T} | Q_{t-1} = i, \Lambda) , \tag{18}$$

$$\phi_t(i) = \operatorname*{argmax}_{q_{t:T}} \underline{p}(q_{t:T}, o_{t:T} | Q_{t-1} = i) . \tag{19}$$

We can solve variables $\delta_t(i)$ and $\phi_t(i)$ recursively by a Viterbi-like algorithm as follows.

1. Basis: For $t = T + 1$,

$$\delta_{T+1}(i) = 1, \quad \text{for all } i , \tag{20}$$

$$\phi_{T+1}(i) \text{ arbitrary for all } i . \tag{21}$$

2. Induction: From $t = T$ to $t = 2$,

$$\delta_t(i) = \max_j \min_{\substack{a_t^i \in A_t^i \\ b_t \in B_t}} a_t^i(j) b_t^j(o_t) \delta_{t+1}(j) \tag{22}$$

$$= \max_j \Big[\underline{p}(Q_t = j, o_t | Q_{t-1} = i, \lambda) \times \max_{q_{t+1:T}} \underline{p}(q_{t+1:T}, o_{t+1:T} | j, \lambda) \Big] \tag{23}$$

$$= \max_{q_{t:T}} \underline{p}(q_{t:T}, o_{t:T} | Q_{t-1} = i, \Lambda) . \tag{24}$$

The passage from (22)–(23) was obtained by assuming the inductive hypothesis of (18) true at time step $t+1$ and solving the minimization locally due to the separate specification. Equation (24) was reached by moving the inner maximization out, as it is independent with respect to the choice of $p(Q_t = j, o_t | Q_{t-1} = i, \Lambda)$, and then applying probability laws. $\phi_t(i)$ is chosen as the argument of (22).

3. Termination: For $t = 1$,

$$\delta_1(i) = \max_j \min_{\substack{\pi \in \Pi \\ b_t \in B_t}} \pi(j) b_q^j(o_q) \delta_2(j) \tag{25}$$

$$= \max_{q_{1:T}} \underline{p}(q_{1:T}, o_{1:T} | Q_{t-1} = i, \Lambda). \tag{26}$$

The most probable explanation can be obtained by backtracking the partial decisions stored in $\phi_t(i)$,

$$q_t^* = \phi_t(q_{t-1}^*), \quad \text{for } t = 2, \ldots, T ,$$

where q_1^* is the solution of (25).

The algorithm runs in $O(TN^2)$ time, since the minimizations can be solved by choosing lower probabilities, which are already available for interval-base probability sets or can be computed in advance for credal sets specified by probability distributions.

VI. Experiments

In this section, we describe the results of experiments with real data that provide evidence of the efficiency and applicability of the algorithms described here.[1]

A. Human Action Recognition

As a purely demonstrative application of the likelihood algorithm described in Sec. IV, we consider an action recognition task. Given a video sequence, the goal is to determine which action, among a given number of possible alternatives, is showed there. The frames of the sequence can be easily regarded as the observations of a generative sequence, and HMMs are therefore a natural choice for the modeling. The Weizmann human action data set [3] is composed by a number of sequences, each one tagged by its action. Nine different actions (see second column of Table I) are considered.

We adopt a simple approach to action recognition based on the likelihood algorithm for iHMMs. Nine "representative" video sequences, one for each action, are selected from the benchmark. Then, for each selected sequence, an iHMM is obtained using local IDMs [9] with fractional counts estimated from data using the standard Baum-Welch algorithm [1] in the place of data counts. We denote these iHMMs by $\Lambda_1, \ldots, \Lambda_9$. The use of iHMMs here are justified by the fact that, likewise single probability distributions estimated from data, fractional count estimates from the Baum-Welch algorithm are prone to inaccuracies introduced by the scarceness of data and other factors [26].[2] For each frame, the first five features, obtained among 50 selected features processed with principal components analysis, are treated as a single manifest variable. No data are available about the hidden variable, whose number of states (also called *canonical poses*) is an input for the Baum-Welch algorithm (we set $N = 5$).

As an illustrative example, we consider a random sequence not used as a representative to be classified based on the models in Λ^*. According to the dominance criterion (Def. 3),

[1] The inference algorithms are available upon request to the first author.

[2] Note that we assume MAR as a valid hypothesis since all states are missing in a presumably non selective way. In fact, iHMMs can also deal with other learning approaches that do not make assumptions about the missingness process [6].

TABLE I. Comparison between the decisions made by iHMMs using lower and upper likelihood and HMM using likelihood for a single test sequence that shows a “walk” action. Bold face is used to outline the actions not rejected by the iHMMs and that maximizing the likelihood for the HMM.

j	action	$\underline{p}(o_{1:T}\|\lambda_j)$	$\overline{p}(o_{1:T}\|\lambda_j)$	$p(o_{1:T}\|\lambda_j)$
1	bend	0.8152	0.8607	0.8605
2	jack	**0.8721**	**0.9207**	0.9206
3	jump	0.8425	0.8895	0.8893
4	pjump	**0.8670**	**0.9153**	0.9151
5	side	0.8139	0.8594	0.8591
6	run	**0.9083**	**0.9588**	**0.9587**
7	**walk**	**0.8622**	**0.9103**	0.9101
8	wave1	0.8561	0.9039	0.9037
9	wave2	0.8390	0.8858	0.8856

the result of the classification is a set of non rejected actions, that is, a set of possible explanations comprising the true one. Table I shows the results of a comparison with precise HMMs (with parameters learned with Baum-Welch) for a single video sequence in the test set. Notably, the precise approach returns a wrong answer, while the imprecise approach returns a set of four actions, including both the correct one and the wrong one returned by the HMM.

B. Part-of-Speech Tagging

We evaluated the ability of iHMMs to discriminate between reliable and non reliable explanations by comparing joint maximin and maximax explanations in a part-of-speech (PoS) tagging task [2].

We performed experiments with reduced versions of two common data sets used for PoS that are freely available on the *nltk* package distribution.[3] The reduced versions of the Brown and Penn data sets contain, respectively, 38 and 31 distinct syntactic tags and ~ 5500 and ~ 3500 distinct words, and allowed us to exploit the performance of HMMs with small training samples, where the impact of single probability distributions learned from data is greater. The interval-based credal sets were learned using local IDMs (with hyperparameter $s = 1$). Tags not occurring in the training data were omitted from the model (instead of having vacuous credal sets). Words appearing less than 4 times in the training data were collapsed into a single term and used for estimating the probability (or probability interval) of unseen words during classification. Coherently, the precise HMMs were learned with maximum likelihood smoothed by Perks' priors (with $s = 1$), and same preprocessing steps. In order to assess the difficulty of the task we also performed tests with a simple unigram tagger. Fig. 1 reports the results of numerical simulations for 5-fold cross validation and increasing size of the training set for the two data sets.

We evaluated the ability of discriminating reliable and non-reliable explanations by comparing the average accuracy of the predictions (PoS tags) of the precise HMM in the full test set and in two partitions of test set. The first, denominated *match set* consisted only of tokens for which the maximin and maximax criteria provided the same explanation. The second, denominated *mismatch set*, consisted of instances where maximin and maximax disagreed. According to our rationale, the accuracy of the precise should be much greater on the match set than on the mismatch set, as the former consisted of reliable predictions of the precise HMM. The results show that the predictions in the match set are considerably superior to the predictions in the mismatch set in both data set. The rate of agreement, represented in the graph as the distance between the full set and the match set accuracies, was low in both data sets, indicating the unreliability of explanations generated by precise HMMs learned from very few data.

VII. Conclusion

We introduce imprecise hidden Markov models as an extension of standard HMMs that allows proper handling of the imprecision in the parameters that arise in many domains. We presented algorithms for computing likelihood and most probable explanation queries in iHMMs that are comparable in terms of complexity to the corresponding algorithms available for HMMs.

Experiments with real data showed that iHMMs do work as “cautious” classifiers that make decisions only when there is enough statistical evidence to support them. In addition, iHMMs can serve as valuable tools to perform analysis of the sensitivity of precise HMMs to variations of the parameters. More complete experiments with the classification of sequences are necessary.

The imprecision in the numerical parameters of the model translates to indeterminacy when using the models to make decisions as in the applications we show. We have adopted here interval dominance to compare models based on likelihood and maximin and maximax to evaluate explanations. The literature counts with other criteria that are worth evaluating. Implementing such criteria will require developing efficient algorithms. We leave that as future work.

There is also interest in investigating in the future the use of iHMMs in tasks such as filtering, for which efficient algorithms exist, and to extend the model to cases for which fast algorithms could still be achieved.

References

[1] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[4] D. Nur, D. Allingham, J. Rousseau, K. Mengersen, and R. McVinish, “Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis,” *Computational Statistics and Data Analysis*, vol. 53, pp. 1873–1882, 2009.

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

[6] M. Zaffalon and E. Miranda, “Conservative inference rule for uncertain reasoning under incompleteness,” *Journal of Artificial Intelligence Research*, vol. 34, pp. 757–821, 2009.

[7] A. Benavoli, M. Zaffalon, and E. Miranda, “Reliable hidden Markov model filtering through coherent lower previsions,” in *Proceedings of the 12th International Conference on Information Fusion (FUSION)*, 2009, pp. 1743–1750.

[8] I. Levi, *The Enterprise of Knowledge*. MIT Press, 1980.

[3] http://www.nltk.org

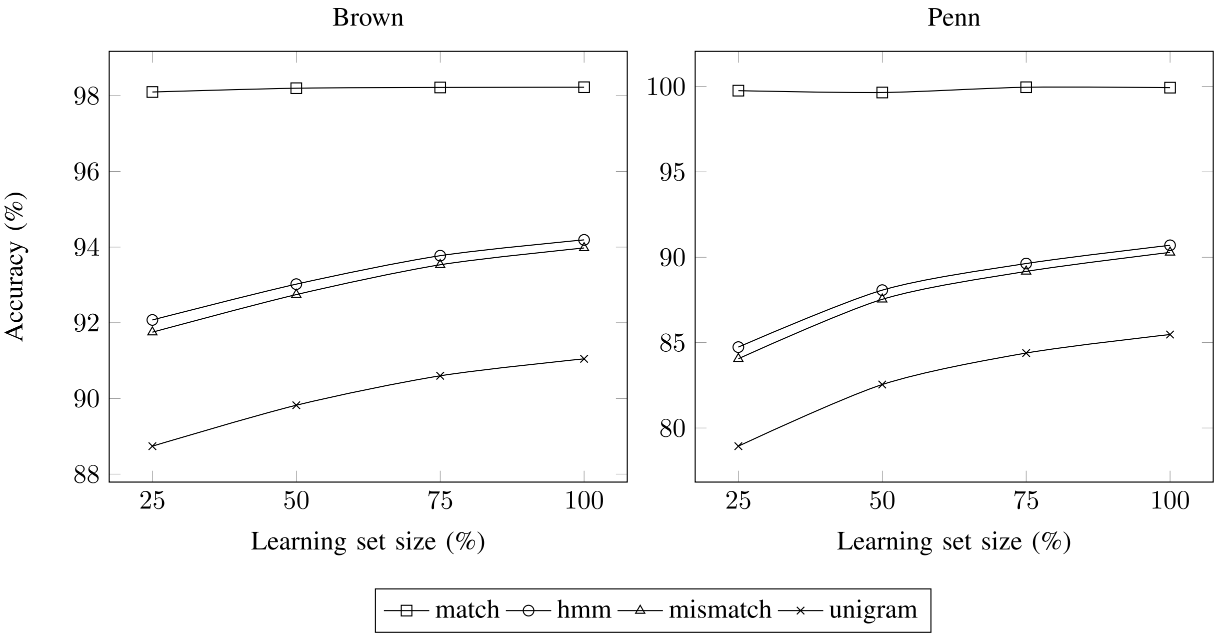

Fig. 1. Results of the PoS tagging experiments. Left and right plots show explanation accuracy for different criteria in the Brown and Penn data sets, respectively.

[9] P. Walley, "Inferences from multinomial data: Learning about a bag of marbles," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996.

[10] M. Zaffalon, G. Corani, and D. D. Mauá, "Evaluating credal classifiers by utility-discounted predictive accuracy," *International Journal of Approximate Reasoning*, vol. 53, no. 8, pp. 1282–1301, 2012.

[11] H. Chuan and A. Darwiche, "Sensitivity analysis in Bayesian networks: From single to multiple parameters," in *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004, pp. 67–75.

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

[13] F. Cozman, "Credal networks," *Artificial Intelligence*, vol. 120, pp. 199–233, 2000.

[14] C. P. de Campos and F. Cozman, "The inferential complexity of Bayesian and credal networks," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 1313–1318.

[15] E. Fagiuoli and M. Zaffalon, "2U: An exact interval propagation algorithm for polytrees with binary variables," *Artificial Intelligence*, vol. 106, no. 1, pp. 77–107, 1998.

[16] M. Zaffalon and E. Fagiuoli, "Tree-based credal networks for classification," *Reliable Computing*, vol. 9, no. 6, pp. 487–509, 2003.

[17] 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*, vol. 51, no. 9, pp. 1029–1052, 2010.

[18] D. D. Mauá, C. P. de Campos, A. Benavoli, and A. Antonucci, "On the complexity of strong and epistemic credal networks," in *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013, pp. 391–400.

[19] A. Antonucci, C. P. de Campos, D. Huber, and M. Zaffalon, "Approximating credal network inferences by linear programming," in *Proceedings of the 12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, L. C. van der Gaag, Ed., vol. 7958. Springer, 2013, pp. 13–25.

[20] I. O. Kozine and L. V. Utkin, "Interval-valued finite Markov chains," *Reliable Computing*, vol. 8, no. 2, pp. 97–113, 2002.

[21] G. de Cooman, F. Hermans, and E. Quaeghebeur, "Sensitivity analysis for finite Markov chains in discrete time," in *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI)*, 2008, pp. 129–136.

[22] D. Škulj, "Discrete time Markov chains with interval probabilities," *International Journal of Approximate Reasoning*, vol. 50, no. 8, pp. 1314–1329, 2009.

[23] D. Škulj and R. Hable, "Coefficients of ergodicity for imprecise Markov chains," in *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, 2009, pp. 377–386.

[24] R. Crossman, P. Coolen-Schrijner, D. Škulj, and F. Coolen, "Imprecise Markov chains with an absorbing state," in *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, 2009, pp. 119–128.

[25] 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 (FUSION)*, 2009.

[26] 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.

[27] A. Camp and G. Cooman, "A new method for learning imprecise hidden Markov models," in *Advances in Computational Intelligence*, 2012, vol. 299, pp. 460–469.

[28] J. de Bock and G. de Cooman, "State sequence prediction in imprecise hidden markov models," in *Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, 2011, pp. 159–168.

[29] A. Antonucci, R. de Rosa, A. Giusti, and F. Cuzzolin, "Temporal data classification by imprecise dynamical models," in *Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, F. Cozman, T. Denoeux, S. Destercke, and T. Seidenfeld, Eds., 2013, pp. 13–22.

[30] G. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.