Contribution of belief functions to hidden markov models with an application to fault diagnosis.

Emmanuel Ramasso

To cite this version:


HAL Id: hal-00414145
https://hal.archives-ouvertes.fr/hal-00414145
Submitted on 8 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CONTRIBUTION OF BELIEF FUNCTIONS TO HIDDEN MARKOV MODELS WITH AN APPLICATION TO FAULT DIAGNOSIS

Emmanuel Ramasso

FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, Automatic Control and Micro-Mechatronic Systems Department, Besançon, France, 25000
emmanuel.ramasso@femto-st.fr

ABSTRACT
Evidence-theoretic propagations of temporal belief functions are proposed to deal with possibly dependent observations and for partially supervised learning of HMM. Solutions are formulated in Transferable Belief Model framework and experiments concern a diagnosis problem.

Keywords: Evidential Hidden Markov Models, State Sequence Recognition, Partially Supervised Learning, System’s Health Assessment

1. INTRODUCTION
Temporal data modelling and analysis are important in many fields [1, 2] and one common solution is to use state sequence representation, learning and inference algorithms. A state generally represents a stationarity or a functioning mode in the data and describes the dynamical system at a given time while transitions represent the system dynamics. In practical applications, states are hidden and only features are observable. An additional modelling step is thus required to relate numerical features to states. In order to manage imperfections, the modelling should cope with uncertainty using [3] probability theory, possibility theory or evidence theory, the latter being more general.

Hidden Markov Model (HMM) [1] is a famous probabilistic and non-deterministic finite state machine for state sequence learning and inference under the concept of hidden states managing both discrete and continuous features. In HMM, the modelling method allows to generate a conditional probability (or a likelihood), denoted \( b_j(O_t) \) for the \( j \)-th state, where \( O_t = [O_1, O_2, \ldots, O_F] \) is the set of \( F \) features. Three problems concerning HMM can be tackled.

What if information on states are represented by possibilities or belief functions in several formalisms? How to express a lack of knowledge without choosing any artificial priors and is it then possible to take it into account in probabilistic HMM learning? What if observations are not independent (conditionally on states)?

In order to tackle these problems, one solution is to marry HMM mechanisms with other representations of uncertainty [3]. Mohamad and Gader [4] introduced the Generalized HMM in which the generalization is narrowed down to possibility measures and thereby their framework is not able to manage belief functions. One advantage of their framework is the possibility to manage dependent observations by using fuzzy operators but the authors used the product thereby assuming statistical independence. Pieczynski et al. [5, 6] pioneered in mixing belief functions with Markov chains leading to promising results. The main idea was to alleviate the problem of prior modelling in Markov chains [7] using Dempster's rule of combination (generalizing Bayesian inference [8]). However, either the prior or the belief state is evidential but not both, thereby underlying probability assumptions are present. Rombaut et al. [9] proposed a generalization of a Petri Net to belief functions based on the Generalized Bayesian Theorem (GBT) [8]. However, it is not robust to noise because links between states at successive instants are given by an evolving and sparse transition matrix depending on sensors measures. Moreover, no classification criterion was proposed. The fourth paper [10] proposed a deterministic state machine called Belief Scheduler based on the Temporal Evidential Filter [11]. An original inference criterion based on conflict is proposed for classification of sequences. The problem is the sensitivity of some threshold settings.

Compared to previous work, evidential forward-backward-Viterbi propagations are extended to belief functions with underlying probability assumptions. The extension is based on special operators developed in the Transferable Belief Model (TBM) [12] in order to consider possibly dependent observations. The problem of partially supervised learning of probabilistic HMM is also tackled using belief functions.

2. PROBABILISTIC HMM IN A NUTSHELL
States are said hidden: only \( F \) observations in \( O_t \) are available at instants \( t \in \{1..T\} \) (\( T \) is the sequence length). The number of states is \( N \) and \( s_i^t \in \Omega_i \) is the \( i \)-th state at \( t \) where \( \Omega_i = \{s_1^i, s_2^i, \ldots, s_N^i\} \) is the frame of discernment. At each \( t \), the likelihood of state \( i \) conditional to the cur-

O_t = [O_1, O_2, \ldots, O_F] \) is the set of \( F \) features. Three problems concerning HMM can be tackled.

What if information on states are represented by possibilities or belief functions in several formalisms? How to express a lack of knowledge without choosing any artificial priors and is it then possible to take it into account in probabilistic HMM learning? What if observations are not independent (conditionally on states)?

In order to tackle these problems, one solution is to marry HMM mechanisms with other representations of uncertainty [3]. Mohamad and Gader [4] introduced the Generalized HMM in which the generalization is narrowed down to possibility measures and thereby their framework is not able to manage belief functions. One advantage of their framework is the possibility to manage dependent observations by using fuzzy operators but the authors used the product thereby assuming statistical independence. Pieczynski et al. [5, 6] pioneered in mixing belief functions with Markov chains leading to promising results. The main idea was to alleviate the problem of prior modelling in Markov chains [7] using Dempster's rule of combination (generalizing Bayesian inference [8]). However, either the prior or the belief state is evidential but not both, thereby underlying probability assumptions are present. Rombaut et al. [9] proposed a generalization of a Petri Net to belief functions based on the Generalized Bayesian Theorem (GBT) [8]. However, it is not robust to noise because links between states at successive instants are given by an evolving and sparse transition matrix depending on sensors measures. Moreover, no classification criterion was proposed. The fourth paper [10] proposed a deterministic state machine called Belief Scheduler based on the Temporal Evidential Filter [11]. An original inference criterion based on conflict is proposed for classification of sequences. The problem is the sensitivity of some threshold settings.

Compared to previous work, evidential forward-backward-Viterbi propagations are extended to belief functions with underlying probability assumptions. The extension is based on special operators developed in the Transferable Belief Model (TBM) [12] in order to consider possibly dependent observations. The problem of partially supervised learning of probabilistic HMM is also tackled using belief functions.

2. PROBABILISTIC HMM IN A NUTSHELL
States are said hidden: only \( F \) observations in \( O_t \) are available at instants \( t \in \{1..T\} \) (\( T \) is the sequence length). The number of states is \( N \) and \( s_i^t \in \Omega_i \) is the \( i \)-th state at \( t \) where \( \Omega_i = \{s_1^i, s_2^i, \ldots, s_N^i\} \) is the frame of discernment. At each \( t \), the likelihood of state \( i \) conditional to the cur-
recent observation is denoted \( b_i(O_i) = P(O_i|s_i) \) and is provided by a modelling technique such as a mixture of Gaussians [1]. The transition from a state \( s_{i-1} \) to a state \( s_i \) is made with a probability \( a_{ij} = P(s_i|s_{i-1}) \) and the transition matrix is \( \kappa = \{a_{ij}\}, i \in \{1, \ldots, N\}, j \in \{1, \ldots, N\} \) with size \( N \times N \) and \( \sum_j a_{ij} = 1 \). Transitions between states start from an initial distribution \( (t = 1) \) denoted \( \pi_i, i \in \{1, \ldots, N\} \) and probabilities on states are then updated along time according to transitions and observations. An HMM \( \lambda = \{\kappa, \mathbb{B}, \Pi\} \) is thus characterized by three elements: the transition matrix \( \kappa \), the observation models \( \mathbb{B} \) (generating likelihoods \( b_i(O_i) \)) and the initial distribution \( \Pi \).

Estimating the likelihood of an observation sequence given the model is computationally feasible using the forward propagation that evaluates the forward variable \( \alpha_t(s_j) = P(O_1, O_2, \ldots, O_t, s_j|\lambda) = \sum_{s_{t-1}} \alpha_{t-1}(s_i) \cdot a_{ij} \cdot b_j(O_t) \). For a classification problem, we need to compute the log-likelihood of observations for each HMM model \( \lambda \) and then to choose the best model by: \( \lambda^* = \text{argmax}_\lambda \log L_p \) with \( \log L_p = -\frac{1}{T} \sum_t \log \left( \sum_{s_t \in \Omega_t} \alpha_t(s_t) \right) \). The backward propagation allows to compute the backward variable \( \beta_t(s_j) = P(O_T, O_{T-1}, \ldots, O_{t+1}|s_j, \lambda) = \sum_{s_{t+1}} \beta_{t+1}(s_j) \cdot a_{ij} \cdot b_j(O_t+1) \). Two particular combinations of these variables are used in HMM learning [1].

The first is the smooth estimate of probabilities \( \gamma_t(s_j) = P(s_j|O_1:T, \lambda) \) with \( \gamma_t(s_j) = \frac{\alpha_t(s_j) \cdot \beta_t(s_j)}{P(O_1:T, \lambda)} \), that is particularly used to both determine the most probable state at a given instant and models parameters. The second combination is \( \xi_t(s_j, s_{j+1}) = P(O_t|s_j, s_{j+1}, \lambda) \) with \( \xi_t(s_j, s_{j+1}) = \alpha_t(s_j) \cdot a_{ij} \cdot b_j(O_t+1) \cdot \beta_{t+1}(s_j) \) and allows to compute the expected probabilities of transition by: \( a_{ij} \propto \sum_t \xi_t(s_j, s_{j+1}) \). Lastly, the Viterbi algorithm determines the best sequence of hidden states. It is actually a forward propagation where the sum-product operator is replaced by a max-product.

3. TRANSFERABLE BELIEF MODEL (TBM)

The belief of an agent on subsets of the frame of discernment \( \Omega_t \) can be represented by a belief mass assignment (BBA) \( m^{\Omega_t} : 2^{\Omega_t} \to [0, 1] \), \( A \to m^{\Omega_t}(A) \). The belief mass satisfies \( \sum_{A \subseteq \Omega_t} m^{\Omega_t}(A) = 1 \) and can be assigned to singleton state (\( |A| = 1 \)) or to subsets (\( |A| > 1 \), without assumption concerning additivity) that is one important difference with probabilities. Conditional BBA can be used to represent state of knowledge, e.g. \( m^{\Omega_t}_{\Omega_{t-1}}(S_i|\cdot) \) is a BBA defined on \( \Omega_t \) conditionally to subset \( S_i \subseteq \Omega_{t-1} \). The mass \( m^{\Omega_t}(\emptyset) \) is the mass of conflict and dividing each belief mass by \( 1 - m^{\Omega_t}(\emptyset) \) while forcing \( m^{\Omega_t}(\emptyset) = 0 \) is called Dempster normalization.

Several functions, which are in one-to-one correspondence [12], can be computed from a BBA which are then used to simplify the computation of combination rules. In the sequel, the function \( w^{\Omega_t} \) (representing the weights of the canonical conjunctive decomposition (WCD) [14]) will be used. They are defined \( \forall \Omega_t \) by:

\[
w^{\Omega_t}(B) = \prod_{B \subseteq C} q^{\Omega_t}(C) |C| - |B| + 1
\]

where the commonality \( q \) is \( q^{\Omega_t}(B) = \sum_{C \supseteq B} m^{\Omega_t}(C) \). We will also use the credibility \( \text{be}^{\Omega_t}(B) = \sum_{C \subseteq B} m^{\Omega_t}(C) \) and the plausibility \( p^{\Omega_t}(B) = \sum_{C \cap B = \emptyset} m^{\Omega_t}(C) \) with:

\[
p^{\Omega_t}(\Omega_t) = 1 - m^{\Omega_t}(\emptyset)
\]

The combinations of BBAs can then be achieved by four main rules [14]: the conjunctive combination (CRC, \( \& \)), the disjunctive combination (DRC), the “cautious” CRC (CCRC, \( \odot \)), and the “bold” DRC. The two last ones are used when sources of belief are not independent. We have [14]:

\[
w^{\Omega_t}_{\odot}(B) = w^{\Omega_t}(B) \odot w^{\Omega_t}(B)
\]

where \( \odot \) is a “generalized” cautious rule (GCR) based on positive t-norms/t-conorms (see [14]-5.1 for details). This offers the possibility to use an infinity number of t-norm/t-conorm operators (\( \wedge \) and \( \vee \)) defined in possibility theory. For example [14] the Frank t-norm can be used:

\[
x_{\tau y} = \begin{cases} x \land y, & \text{if } s = 0 \\ x \cdot y, & \text{if } s = 1 \\ \log_s \left( 1 + \frac{(s^x - 1) \cdot (s^y - 1)}{s - 1} \right), & \text{otherwise} \end{cases}
\]

where \( s \in [0, 1] \) and \( \wedge \) is the minimum. If \( s = 0 \) (resp. \( s = 1 \)), the CCRC (resp. the CRC) is obtained. If the CRC is used and the obtained belief mass is normalized then we obtain Dempster rule [15].

The conditioning process is a special case of conjunctive combination using the CRC between a first belief mass (to be conditioned) with a second one (said categorical) where the latter has the particularity to be not nil only for one element \( A \subseteq \Omega_t, (\wedge A) = \text{the condition} \).

Decision making in TBM is made by choosing the best hypothesis from the pignistic probability distribution [12] obtained \( \forall \Omega_k \subseteq \Omega_t \) by:

\[
\text{BetP}\{m^{\Omega_k}(s_k) = \frac{1}{1 - m^{\Omega_k}(\emptyset)} \sum_{B \subseteq \Omega_t, s_k \in B} \frac{m^{\Omega_k}(B)}{|B|}
\]

4. PARTIALLY-HIDDEN MARKOV MODELS

Before tackling the problem of evidential propagations with dependent observations, let consider an intermediate situation where one uses probabilistic HMM when data are annotated by belief functions. Let \( \Omega_t \) be an instance of the learning set. One can have information about the belief of this observation to belong to one of the K states, i.e.
each observation in the learning set is annotated by a belief function $m_{\Omega}^{[t]}$. This can be useful because one can learn a HMM in a supervised / unsupervised / partially supervised manner. Indeed, in case of supervised learning, the belief function of each data in the learning set is categorical (e.g. for all instances, the whole mass is assigned to one singleton) thus $m_{\Omega}^{[t]}(B) = 1$ for $B \subseteq \Omega_t$ and $\forall B \neq C \subseteq \Omega, m_{\Omega}^{[t]}(C) = 0$. In unsupervised learning, the belief function of each data is vacuous, i.e. $m_{\Omega}^{[t]}(\Omega_t) = 1$ and semi-supervised learning is a mixing of both previous cases. Finally, in partially-supervised learning, belief functions are “general” thus $m_{\Omega}^{[t]}(B) \in [0,1], \forall B \subseteq \Omega_t$. So the problem is: how to take into account partial labels (fourth case) in probabilistic HMM learning? For that, we rely on [16] concerning EM-based parameters learning using partial labels which has shown promising results for static data.

Let consider the forward variable that has a pivotal role in HMM learning: $\alpha_t(s_j) = P(O_1, O_2, \ldots, O_t, s_j | \lambda) = \sum_{s_i \in \Omega_{t-1}} \alpha_{t-1}(s_i) \cdot a_{ij} \cdot b_j(O_t)$. In case a partial knowledge on states at $t$ is available in the form of a belief function on states $m_{\Omega}^{[t]}$, it can be taken into account by the conjunctive combination with $\alpha_t$. Since the probabilistic forward variable has only positive values for singletons, the conjunctive combination leads to:

$$\alpha_t(\bigcirc m_{\Omega}^{[t]})(s_i) = \alpha_t(s_i) \cdot p_{\Omega}^{[t]}(s_i)$$  \hspace{1cm} (6)

and thus the forward variable can be rewritten as:

$$\alpha_t(s_j) = \sum_{\forall s_i \in \Omega_{t-1}} \alpha_{t-1}(s_i) \cdot a_{ij} \cdot b_j(\Omega_t) \cdot p_{\Omega}^{[t]}(s_j)$$  \hspace{1cm} (7)

A similar result can be obtained for the backward variable. Given a learning set annotated by belief functions, only plausibility on singletons need to be stored and then used in the forward and backward recursions for HMM learning.

5. THE CLASSIFICATION PROBLEM

Evidential forward algorithm - The evidential forward propagation is used to assess the forward variable at instant $t$ and requires 1) the BBA $m_{\Omega}^{[t-1]}(\lambda)$ of the forward variable at the previous instant $t-1$ (conditional to the parameters that governs the corresponding EvHMM), 2) the conditional WCD $w_{\Omega}^{[t]}(S_{t}^{t-1})$ defined on $\Omega_t$ conditionally to subsets of preceding states $(S_{t}^{t-1} \subseteq \Omega_{t-1})$ and representing transitions between subsets of states at $t-1$ and $t$, and 3) the WCD $w_{\Omega}^{[t]}(\Omega_t)$ obtained from observations at $t$.

Proposition 1 The computation of the evidential forward variable at $t$ satisfies a “prediction – update” mechanism:

$$w_{\Omega}^{[t]}(\lambda)(S_{t}^{t}) = w_{\Omega}^{[t]}(\Omega_t)(S_{t}^{t}) \oplus \left( \sum_{S_{t}^{t-1} \subseteq \Omega_{t-1}} m_{\Omega}^{[t-1]}(\lambda)(S_{t}^{t-1}) \cdot w_{\Omega}^{[t]}(S_{t}^{t-1})(S_{t}^{t}) \right)$$  \hspace{1cm} (8)

GCR rules endow the evidential forward propagation with a panel of operators that could be chosen according to the application and the data [14]. It has been demonstrated in [14] that these operators are suited for dependent data. Moreover, another advantage is that the prior on states at $t = 1$ can be “vacuous” and this state of knowledge is easily represented by $w_{\Omega}^{[1]}(\lambda)(S_{1}^{1}) = 1, \forall S_{1}^{1} \subseteq \Omega_1$ meaning total ignorance.

A conflict may appear at each instant between prediction and observation meaning both quantities are contradictory. This conflict must be cancelled out because it is absorptive by a conjunctive rule. Actually, the required normalization procedure is the same as in probabilistic HMM [1].

Unlike probabilistic HMM where the normalisation process consists in redistributing uniformly $1 - \sum_{j} \alpha_t(j)$ to each state at $t$, belief function framework provides complex and sound redistribution rules [17]. The parallel between both formalisms is more obvious when seeing the (seldom emphasized) relation between conflict and plausibility (Eq. 2) that represents a bridge between HMM and EvHMM. As shown in many applications (object association [18], image processing [19], change detection and state sequence recognition [11, 10]), conflict is a useful information that is used below for classification of state sequences.

Classification in EvHMM - Given an observation sequence $\Omega_{1:T}$ (length $T$) and a set of $N$ EvHMM $\lambda_1 \ldots N$, how to choose the EvHMM that best fits observations?

**Proposition 2** The lower is the conflict throughout the whole observation sequence, the better is the EvHMM $\lambda$ for explaining these observations. The whole conflict and the best EvHMM are given by:

$$L_c(\lambda_n) = \frac{1}{T} \sum_{t=1}^{T} \log(1 - m_{\Omega}^{[t]}(\lambda_n)(\emptyset))$$  \hspace{1cm} (9a)

$$\lambda^* = \underset{\lambda_n}{\text{argmax}} \ L_c(\lambda_n)$$  \hspace{1cm} (9b)

This criterion reduces to the probabilistic one when using Dempster rule and Bayesian BBAs. Conclusively, the evidential forward algorithm can be used for online filtering of belief functions on states and for the classification of observation sequences in five steps: 1) the initial BBA ($t = 1$) on states can be vacuous, 2) Eq. 8 is applied from $t = 2$ to $T$, 3) the conflict at $t$ is stored and then redistributed, 4) the classification criterion (Eq. 9) is finally computed.

6. TIME-INSTANT STATE RECOGNITION AND TRANSITION ESTIMATION

Evidential backward algorithm - The evidential backward algorithm is similar to the forward one and requires the BBA
Proposition 4 The transition matrix is expected to gradually tend to a probabilistic distribution (Eq. 5) based on evidential HMM. This state is found out by maximizing the pigesimal version of the Viterbi algorithm is proposed ([1], this technique is not always well-suited for state estimation). Therefore, due to the conditioning and propagation processes.

Proposition 3 The belief on states at $t$ in the backward recursion is obtained by:

$$w_{\beta_i}(S^t_i) = \sum_{S^t_{j+1} \subseteq \Omega_{t+1}} m_{\beta_{j+1}}(S^t_{j+1}) \cdot w_{\alpha_i}(S^t_{j+1})(S^t_i)$$

where $m_{\Omega_{t+1}}$ is the conjunctive combination of BBAs at $\Omega_{t+1}$, i.e. $m_{\beta_{j+1}}$ and $m_{\beta_{j+1}}$. The posterior conditional WCDs $w_{\alpha_i}[S^t_{j+1}]$ are obtained from plausibilities by calculating the conjunctive commonalities using the GBT [8]:

$$q_a[m_{\Omega_{t+1}}[S^t_{j+1}]](S^t_i) = \prod_{s_i \in S^t_i} m_{\Omega_{t+1}}[S^t_{j+1}](s^t_i)$$

then used in Eq. 1 to obtain WCDs. Unknown prior $T$ is modelled by a vacuous BBA ($w_{\alpha_i}^{T}[S^t_{j}] = 1, \forall S^t_{j} \subseteq \Omega_T$). Eq. 10 reduces to the probabilistic case if all BBAs are Bayesian and if Dempster rule is used.

Instant state recognition - As in probabilistic HMM, the conjunctive combination of both forward and backward variables (called $\gamma$-variable) is performed when observations are available as a whole.

Proposition 4 The $\gamma$-variable is given by:

$$w_{\gamma_i}(S^t_i) = w_{\alpha_i}(S^t_i) \odot w_{\beta_i}(S^t_i)$$

One has to use the same rule $\odot$ as in the forward and backward variables. The $\gamma$-variable can then be exploited for two functionalities. The first one is offline smoothing of belief functions on states. The second one is the detection of the best state $s^t_i$ at a given instant and for a given EvHMM. This state is found out by maximizing the pigismatic probability (Eq. 5) based on $m_{\gamma_{i,v}}$, $\forall v \in \{1 \ldots T\}$, i.e. $s^t_i = \arg \max_{s_i \in \Omega_T}$ BetP$(m_{\gamma_{i,v}})((S_i))$. As in the probabilistic case [1], this technique is not always well-suited for state sequence recognition because is too “local”. Instead, an evidential version of the Viterbi algorithm is proposed (§ 7).

Transition estimation - One can mimic the computation of the so-called $\xi$-variable [1] that is used in probabilistic HMM for transition estimation in an Expectation-Maximization (EM) learning process. A proposition has been formulated in [13] for independent observations. However, by doing so, one can face a serious problem: loosing the interest of belief functions. Indeed, EM is an iterative procedure while (evidential) transition estimation is based on conjunctive combinations. Therefore, due to the repeated conjunctive combinations at each iteration, the transition matrix is expected to gradually tend to a probabilistic one. Therefore, we rather propose below an estimation formula of transitions that is independent of transitions themselves. It consists in computing the expected belief mass of making a transition from one state to another. For that, it is required to combine conjunctively two observed BBAs defined at two successive instants and then taking the mean of these belief masses as proposed below.

Proposition 5 An estimation of the expected transition from subsets to subsets from observations is given by:

$$m_{\Omega_{t+1}} \times \Omega_{t+1} \propto \sum_{t=1}^{T-1} \left( m_{\beta_{t+1}} \Omega_{t+1} \times \Omega_{t+1} \right)$$

up to a constant $\frac{1}{T}$ and where $m_{\Omega_{t+1}} \times \Omega_{t+1}$ is the vacuous extension [12] of the belief mass $m_{\Omega_{t+1}}[\Omega_{t+1}]$ on the cartesian product defined by: $m_{\Omega_{t+1}}[\Omega_{t+1} \times \Omega_{t+1}](B) = m_{\Omega_{t+1}}(C)$ if $C \times \Omega_{t+1} = B$ and 0 otherwise. If one has prior information on transitions, it can be combined conjunctively with observations in Eq. 13.

The backward recursion requires Eq. 11 to be computed and for that one needs plausibilities conditional to singleton states. They can be obtained from the joint belief mass distribution (Eq. 13) by conditioning it by each singleton state $s^t_i \in \Omega_t$ followed by a marginalization onto $\Omega_{t+1}$, formally $m_{\Omega_t \times \Omega_{t+1}}[s^t_i][\Omega_{t+1}]$ (the two processes are explained in [8, 12]). This solution satisfies the constraint of independence of conditional beliefs (transition) to apply the GBT. Due to the duality GBT-DRC [8], Eq. 11 computes implicitly the beliefs conditional to unions of states by the disjunctive rule. Transitions expressed as a joint belief mass can also be used directly in both forward-backward recursions and this requires to compute the conjunctive combination on a joint space followed by a marginalization onto $\Omega_t$ to obtain the evidential variables at $t$. Eq. 13 has been used to estimate the parameters of the Temporal Evidential Filter [11].

7. STATE SEQUENCE FROM OBSERVATIONS

For a given model, the goal is to determine the best state sequence $s^t_1, s^t_2, \ldots, s^t_T$ given the sequence of observations $\Omega_{1:T}$. In probabilistic HMM, the powerful Viterbi algorithm, based on dynamic programming, achieves this goal by selecting the sequence with maximum likelihood without computing the $N^T$ sequences.

An optimal evidential Viterbi decoding - If we consider sequences of singleton states $s^t_1, s^t_2, \ldots, s^t_T$, then one can apply exactly the same algorithm as in probabilistic HMM except that transitions and observations are represented by the plausibilities on (and conditional to) singletons computed from the related belief masses. The Viterbi metric is a thus propagated plausibility initialized to 1 at the beginning of the sequence reflecting missing prior. The complexity is then the same as in the probabilistic case.

A sub-optimal evidential Viterbi decoding - The previous algorithm, even if optimal in the sense of maximum plausibility / likelihood, reduces belief functions to single values due to the conditioning and propagation processes.
In [13] was proposed an algorithm that draws benefits from belief function modelling and that is able to postpone the decision concerning the best predecessors. For that, at each step, it updates a propagated metric that is a belief function $w_A^t$. This algorithm is a Viterbi-like decoder since decisions are not directly propagated. In the sequel, conditioning on EvHMM parameters $\lambda$ are omitted for simplicity.

**Proposition 6** When observations are not independent, the Viterbi-like metric at $t$ is defined $\forall S^t_j \subseteq \Omega_t$ by:

$$w^{\Omega_t}_a(s^t_j) = w^{\Omega_t}_a(O_t|s^t_j) \otimes \sum_{S^{t-1} \subseteq \Omega_{t-1}} w^{\Omega_{t-1}}_a(S^{t-1} \cap A^{t-1})(s^t_j) \cdot m^{\Omega_{t-1}}_a(s^{t-1}_j) \ (14)$$

where $A^{t-1} = \bigcup_{s'_j \in \Omega_t} \psi'(s'_j)$ is the union of predecessors of states at $t$. The main idea behind conditioning on the set of predecessors $A^{t-1}$ is to make a one-step forward propagation of the belief masses of the most important subsets of $m^{\Omega_{t-1}}_a$. Set $A^{t-1}$ is obtained by first searching which is the predecessor state at $t=1$ of each state at $t$ and this is performed in three steps. First, since the predecessor of a state $s^t_j$ is unknown, it is needed to condition $w^{\Omega_{t-1}}_a$ on all possible candidates $s^{t-1}_j \in \Omega_{t-1}$. Thus, a WCD $w^{\Omega_{t-1}}_a$ is computed by applying Eq. 14 and replacing $w^{\Omega_{t-1}}_a(S^{t-1} \cap A^{t-1})$ by $w^{\Omega_{t-1}}_a(S^{t-1} \cap s^t_{j-1})$. The second step aims at making a decision concerning predecessors for each state $s^t_j$ based on $m^{\Omega_{t-1}}_a$ using the decision criterion (Eq. 5):

$$\mathcal{P}_t(s^{t-1}_j | s^t_j) = \text{BetP}(m^{\Omega_{t-1}}_a)(s^t_j) \ (15)$$

Conditioning on $s^{t-1}_j$ in the first step is equivalent to a conjunctive combination with a categorical BBA (§ 3) and the generated conflict $m^{\Omega_{t-1}}_a(\emptyset)$ quantifies how irrelevant is the hypothesis “the predecessor of $s^t_j$ is $s^{t-1}_j$”. When the two first steps are done for all previous states $s^{t-1}_j \in \Omega_{t-1}$, the most probable predecessor of $s^t_j \in \Omega_t$ is found by:

$$\psi'(s^t_j) = \arg \max_{s^{t-1}_j \in \Omega_{t-1}} [(1 - m^{\Omega_{t-1}}_a(\emptyset)) \cdot \mathcal{P}_t(s^{t-1}_j | s^t_j)] \ (16)$$

where $\psi'(s^t_j)$ stores the best predecessor of $s^t_j$. Note that these three first steps are strictly equivalent to the probabilistic mechanism. Then, values of $\psi'(s^t_j)$ are used to generate $A^{t-1}$ (Eq. 14). They are also used to compute a propagated metric defined at each $t$ by:

$$\mathcal{Q}_t(s^t_j, \lambda) = \mathcal{Q}_{t-1}(\psi'(s^t_j), \lambda) \cdot p^{\Omega_t}_A(s^t_j) \ (17)$$

At $t=1$, $\mathcal{Q}_t(s^t_j, \lambda) = 1$ (reflecting missing prior) and at $t=T$, the best state $s^{*t}_j = \arg \max_{s^t_j \in \Omega_T} \mathcal{Q}_T(s^t_j, \lambda)$ is used for state backtracking by $s^t_j = \psi'(s^{t+1}_j)$. The value of the metric $\mathcal{Q}_T(s^t_j, \lambda)$ at $T$ quantifies the plausibility of the state sequence given the EvHMM and can be used for state sequence classification.

**8. ILLUSTRATION ON FAULT DIAGNOSIS**

The evidential Viterbi-like decoder is illustrated on challenge data of diagnostic and prognostic of machine faults from the Int. Conf. on Prognostics and Health Management [21] (available at https://dashlink.arc.nasa.gov/member/gebel/). Data set consists of multiple multivariate time series with sensor noise (blue dots in Fig. 1). Each time series is from a different engine of the same fleet and each engine starts with different degrees of initial wear and manufacturing variation unknown to the user and considered normal. The engine is operating normally at the start and develops a fault at some point. The fault grows in magnitude until system failure (we have used the file train_FD001.txt for training and test_FD001.txt for testing).

Given a new observation sequence measured on an engine, the goal is to diagnose its current mode and to determine whether the system is faulty. A fault occurs when a sequence of four modes is detected: steady $\rightarrow$ transition $\rightarrow$ up $\rightarrow$ faulty. One detector is built for each mode with an EM run on the training set using mixture of Gaussians with three components. A four states EvHMM is built where the evidential transition matrix is estimated as proposed in this paper and using the training set. The chosen GCR operator $\otimes$ is the Frank $t$-norm (Eq. 4) with parameter $s = 0.9$. To exploit the EvHMM, we first use the GBT to transform the likelihoods generated by each detector into belief function distributions [8, 20]. These beliefs are then used in the Viterbi decoder. To compare EvHMM with HMM, we transform the evidential transition matrix into a pignistic one and use the likelihoods directly.

An illustration of the detection by both systems is given in Fig. 1 (HMM in light magenta and EvHMM in bold red).
The differences in detections come mainly from the lack of data for HMM learning while EvHMM are less sensitive. This lack of sensitivity on the size of the dataset, peculiar to belief functions [16], is important in fault diagnosis to decrease the number of tests on machines to build datasets. Moreover, the evidential Viterbi is better in part because it postpones the decision until the last instant thanks to the conditioned forward propagation. Table 1 presents confusion matrices of detections by HMM and EvHMM to the conditioned forward propagation. Table 1 presents because it postpones the decision until the last instant thanks to the conditioned forward propagation.

<table>
<thead>
<tr>
<th></th>
<th>Steady</th>
<th>Transition</th>
<th>Up</th>
<th>Faulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steady</td>
<td>82</td>
<td>48</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Transition</td>
<td>17</td>
<td>42</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>up</td>
<td>1</td>
<td>10</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>Faulty</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrices for HMM (top, global accuracy: 68%) and EvHMM (bottom, global accuracy: 75%), values are percents.

9. CONCLUSION

Extensions of HMM propagations to belief functions are presented. Contributions of belief functions to HMM are: 1) Possibility to process temporal belief functions, 2) Opportunity to mix several uncertainty formalisms into a flexible one, 3) Partially supervised learning of HMM, 4) Processing possibly-dependent observations by parametrized operators [14], 5) Availability of a variety of combination operators and normalization processes. Higher computational cost is expected but solutions exist to reduce it.

A thorough exploration of EvHMM is required for further applications. Moreover, generalizing the approach developed here to other graphical models as well as the development of a criterion for learning in EvHMM taking into account both models and transitions together are under study.

10. REFERENCES