Training Generalized Hidden Markov Model With Interval Probability Parameters

Yan Wang

Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; PH +1(404)894-4714; FAX +1(404)894-9342; email: yan.wang@me.gatech.edu

ABSTRACT

Recently generalized interval probability was proposed as a new mathematical formalism of imprecise probability. It provides a simplified probabilistic calculus based on its definitions of conditional probability and independence. The Markov property can be described in a form similar to classical probability. In this paper, an expectation-maximization approach is developed to train generalized hidden Markov models with generalized interval probabilities. With the consideration of systematic error in measurement, the training process provides a robust learning mechanism, where data quality requirement is not as restrictive as the traditional hidden Markov model.

INTRODUCTION

Uncertainty in engineering analysis is composed of two components. One is the inherent *randomness* because of fluctuation and perturbation, called *aleatory* uncertainty, and the other is due to *lack of perfect knowledge* about the system, called *epistemic* uncertainty. Epistemic uncertainty has different sources, such as lack of data, conflicting information from multiple sources, conflicting beliefs among experts' opinions, lack of time for introspection, measurement errors, lack of dependency information, etc. Given the very different nature of the two types of uncertainties, it is important to differentiate and treat them separately. Neglecting epistemic uncertainty may lead to decisions that are not robust. Sensitivity analysis is traditionally used in assessing robustness. Conflating epistemic and aleatory uncertainties may increase the cost of risk management. The extra knowledge gained in data clustering or regression analysis can be used to reduce variance.

Aleatory uncertainty is traditionally and predominantly modeled by probability distributions. In contrast, epistemic uncertainty has been modeled in several ways, such as probability, interval, fuzzy set, random set, basic probability assignment, etc. Here interval is used to quantify epistemic uncertainty. An interval is as simple as a pair of numbers, i.e. the lower and upper bounds. The reason to choose intervals is two-fold. First, intervals are natural to human users and simple to use. They have been widely used to represent a range of possible values, estimates of lower and upper bounds for numerical errors, and measurement errors due to the available precision of instruments. Second, intervals can be regarded as the most suitable way to represent the lack of knowledge. Compared to other forms, interval methods require the least assumption. An interval only needs two values for the bounds. In contrast, statistical distributions need assumptions of distribution types, distribution parameters, and the functional mapping from events to real values between 0 and 1. Fuzzy sets need assumptions of not only lower and upper bounds, but also membership functions. Given that the lack of knowledge is the nature of epistemic uncertainty, a representation with the least assumption is the most desirable. Notice that an interval $[L, U]$ only specifies its lower bound L and upper bound U. It does not assume a uniform distribution of values between L and U.

As a generalization of the Markov chain model, a hidden Markov model (HMM) does not assume that the states of Markov chains are directly observable. Rather, probabilistic dependencies exist between the true but hidden states and the observable. The differentiation between the state variables and observable variables in HMMs captures the uncertainty that commonly exists in observation and measurement. HMMs have been successfully used in speech recognition, signature verification, communication and control, bioinformatics, computation vision, network security, and other applications (Khreich et al. (2012)). Yet the traditional HMM does not differentiate the two types of uncertainties, where systematic error in measurement and lack of training data are the major sources of epistemic uncertainties.

Recently, a generalized hidden Markov model (GHMM) (Wang (2011b)) was proposed to extend the HMM with the incorporation of epistemic uncertainty in hidden and observable variables based on a new formalism of imprecise probability, generalized interval probability (Wang (2010)). In this paper, the incremental training of GHMM is studied, where the interval-valued transition and observation probabilities are updated sequentially with newly acquired data. Given the resemblance between GHMM and HMM, the traditional expectation-maximization algorithms can be applied almost directly to the GHMM. It is shown that the robustness of training can be improved when data have measurement errors.

BACKGROUND

Imprecise probability

Imprecise probability $[\underline{p}, \overline{p}]$ combines epistemic uncertainty (as an interval) with aleatory uncertainty (as probability measure), which is regarded as a generalization of traditional probability. Gaining more knowledge can reduce the level of imprecision and indeterminacy, i.e. the interval width. When $\underline{p} = \overline{p}$, the degenerated interval probability becomes a traditional precise one.

Many forms of imprecise probabilities have been developed. For example, the Dempster-Shafer theory (Dempster (1967); Shafer (1976)) characterizes evidence with discrete probability masses associated with a power set of values. The theory of coherent lower previsions (Walley (1991)) models uncertainties with the lower and upper previsions with behavioral interpretations. The possibility theory (Dubois and Prade

(1988)) represents uncertainties with Necessity-Possibility pairs. Probability bound analysis (Ferson et al. (2003)) captures uncertain information with pairs of lower and upper distribution functions or p-boxes. Interval probability (Kuznetsov (1995)) characterizes statistical properties as intervals. F-probability (Weichselberger (2000)) incorporates intervals and represents an interval probability as a set of probabilities which maintain the Kolmogorov properties. A random set (Molchanov (2005)) is a multi-valued mapping from the probability space to the value space. Fuzzy probability (Möller and Beer (2004)) considers probability distributions with fuzzy parameters. A cloud (Neumaier (2000)) combines fuzzy sets, intervals, and probability distributions.

In the applications of interval probability, the interval bounds $\underline{p}$ and $\overline{p}$ can be elicited as the lowest and highest subjective probabilities about a particular event from a domain expert. The expert may hesitate to offer just a precise value of probability. Different experts could have different beliefs. In both cases, the range of probabilities gives the interval bounds. When used in data analysis with frequency interpretation, the interval bounds can be confidence intervals (e.g. the Kolmogorov-Smirnov confidence band) calculated from data to enclose a cumulative distribution function. If extra data are collected, the interval distribution may be reduced to a precise distribution function. For parametric distributions such as exponential and Gaussian, the epistemic uncertainty is represented by interval values of the parameters.

Imprecise probability quantifies aleatory and epistemic uncertainties simultaneously and can be used as an alternative to sensitivity analysis in assessing robustness of probabilistic reasoning. In this paper, the new form of imprecise probability, generalized interval probability, is based on generalized interval. Generalized interval is an algebraic and semantic extension of the classical set-based interval. As a result, the probabilistic calculus in generalized interval probability is greatly simplified.

Generalized interval

The classical set-based interval (Moore (1966)) is defined as $[\![a, b]\!] := \{x \in \mathbb{R} | a \leqslant x \leqslant b\}$. Therefore $[\![a, b]\!]$ is invalid when $a > b$. In contrast, generalized interval (Sainz et al. (2014); Markov (1979); Dimitrova et al. (1992)) does not have such restriction. A generalized interval is defined as a pair of numbers $\mathbf{x} := [\underline{x}, \overline{x}] (\underline{x}, \overline{x} \in \mathbb{R})$. The set of generalized intervals is denoted by $\mathbb{KR} = \{[\underline{x}, \overline{x}] | \underline{x}, \overline{x} \in \mathbb{R}\}$. The set of *proper* intervals is $\mathbb{IR} = \{[\underline{x}, \overline{x}] | \underline{x} \leqslant \overline{x}\}$, and the set of *improper* intervals is $\overline{\mathbb{IR}} = \{[\underline{x}, \overline{x}] | \underline{x} \geqslant \overline{x}\}$. The relationship between proper and improper intervals is established with the operator *dual* as $\mathrm{dual}([\underline{x}, \overline{x}]) := [\overline{x}, \underline{x}]$.

The *inclusion* relationship $\subseteq$ between generalized intervals $\mathbf{x} = [\underline{x}, \overline{x}]$ and $\mathbf{y} = [\underline{y}, \overline{y}]$ is defined as $[\underline{x}, \overline{x}] \subseteq [\underline{y}, \overline{y}] \iff \underline{x} \geqslant \underline{y} \wedge \overline{x} \leqslant \overline{y}$. The *less-than-or-equal-to* relationship $\leqslant$ is defined as $[\underline{x}, \overline{x}] \leqslant [\underline{y}, \overline{y}] \iff \underline{x} \leqslant \underline{y} \wedge \overline{x} \leqslant \overline{y}$. The relationship between generalized interval and classical interval is established with the operator $\triangle$ defined as $[\underline{x}, \overline{x}]^{\triangle} := [\![\min(\underline{x}, \overline{x}), \max(\underline{x}, \overline{x})]\!]$.

The calculation of generalized interval is based on the Kaucher arithmetic (Kaucher (1980)), which is different from the classical interval arithmetic. For the special case of $\mathbf{x} = [\underline{x}, \overline{x}] \geqslant 0$ and $\mathbf{y} = [\underline{y}, \overline{y}] \geqslant 0$, which is applicable to probability values in this paper, the arithmetic is defined as follows. $\mathbf{x} + \mathbf{y} := [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$. $\mathbf{x} - \mathbf{y} := [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$. $\mathbf{x} \times \mathbf{y} := [\underline{x} \times \underline{y}, \overline{x} \times \overline{y}]$. $\mathbf{x}/\mathbf{y} := [\underline{x}/\overline{y}, \overline{x}/\underline{y}] (\mathbf{y} > 0)$. Compared to the semi-group formed by the classical set-based intervals without invertibility, generalized intervals with the operations in Kaucher arithmetic form a group, which satisfies all four conditions of closure, identities, associativity, and invertibility. Invertibility exists in generalized intervals because $\mathbf{x} - \mathrm{dual}\mathbf{x} = 0$ and $\mathbf{x}/\mathrm{dual}\mathbf{x} = 1$. This property significantly simplifies the computational structure. A monotonic interval function as a pair of real-valued functions is defined as $\mathbf{f}(\mathbf{x}) := [f(\underline{x}), f(\overline{x})]$. For instance, $\log(\mathbf{x}) := [\log(\underline{x}), \log(\overline{x})]$. The integral of $\mathbf{f}(x)$ is defined as $\int \mathbf{f}(x)dx := [\int \underline{f}(x)dx, \int \overline{f}(x)dx]$.

Not only generalized interval based on the Kaucher arithmetic simplifies the computational structure, it also provides more semantics than the classical set-based interval. In a functional relation, each generalized interval has an associated logic quantifier, either existential ($\exists$) or universal ($\forall$). The semantics of a generalized interval $\mathbf{x} \in \mathbb{KR}$ is denoted by $(\mathrm{Q}_{\mathbf{x}} x \in \mathbf{x}^{\triangle})$ where $\mathrm{Q} : \mathbb{KR} \mapsto \{\exists, \forall\}$. $\mathbf{x}$ is called *existential* if $\mathrm{Q}_{\mathbf{x}} = \exists$, or *universal* if $\mathrm{Q}_{\mathbf{x}} = \forall$. If a real relation $z = f(x_1, \ldots, x_n)$ is extended to the interval relation $\mathbf{z} = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the interval relation $\mathbf{z}$ is interpretable if there is a semantic relation $(\mathrm{Q}_{\mathbf{x}_1} x_1 \in \mathbf{x}_1^{\triangle}) \cdots (\mathrm{Q}_{\mathbf{x}_n} x_n \in \mathbf{x}_n^{\triangle})(\mathrm{Q}_{\mathbf{z}} z \in \mathbf{z}^{\triangle})(z = f(x_1, \ldots, x_n))$.

Generalized interval provides more semantic power to help verify completeness and soundness of range estimations by logic interpretations. A *complete* range estimation of possible values includes all possible occurrences without underestimation. A *sound* range estimation does not include impossible occurrences without overestimation.

GENERALIZED INTERVAL PROBABILITY

Basic definitions

Given a sample space Ω and a σ-algebra $\mathcal{A}$ of random events over Ω, a *generalized interval probability* $\mathbf{p} \in \mathbb{KR}$ is defined as $\mathbf{p} : \mathcal{A} \to [0,1] \times [0,1]$ which obeys the axioms of Kolmogorov: (1) $\mathbf{p}(\Omega) = [1,1]$; (2) $[0,0] \leqslant \mathbf{p}(E) \leqslant [1,1]$ ($\forall E \in \mathcal{A}$); and (3) for any countable mutually disjoint events $E_i \cap E_j = \varnothing$ ($i \neq j$), $\mathbf{p}(E_i \bigcup E_j) = \mathbf{p}(E_i) + \mathbf{p}(E_j)$.

A generalized interval probability $\mathbf{p} = [\underline{p}, \overline{p}]$ is a generalized interval without the restriction of $\underline{p} \leqslant \overline{p}$. The new definition also implies $\mathbf{p}(\varnothing) = [0,0]$. The probability of *union* is defined by as $\mathbf{p}(A) := \sum_{S \subseteq A} (-\mathrm{dual})^{|A|-|S|} \mathbf{p}(S)$ for $A \subseteq \Omega$.

The assignments of interval-valued probabilities to events are not arbitrary. They should meet certain requirements. In generalized interval probability, the interval probability values should satisfy the *logic coherence constraint* (LCC). That is, for a mutually disjoint event partition $\bigcup_{i=1}^{n} E_i = \Omega$, $\sum_{i=1}^{n} \mathbf{p}(E_i) = 1$. If the sample space is continuous, $\int_{x \in \Omega} \mathbf{p}(x)dx = 1$.

Notice that the LCC is more restrictive than Walley's coherence and avoiding sure loss constraints (Walley (1991)). The LCC ensures that generalized interval probability is logically coherent with precise probability. Suppose that $\mathbf{p}(E_i) \in \mathbb{IR}$ (for $i = 1, \ldots, k$) and $\mathbf{p}(E_i) \in \overline{\mathbb{IR}}$ (for $i = k+1, \ldots, n$). It can be interpreted as

$$\forall p_1 \in \mathbf{p}^{\triangle}(E_1), \ldots, \forall p_k \in \mathbf{p}^{\triangle}(E_k), \exists p_{k+1} \in \mathbf{p}^{\triangle}(E_{k+1}), \ldots, \exists p_n \in \mathbf{p}^{\triangle}(E_n), \sum_{i=1}^{n} p_i = 1$$

For instance, given that $\mathbf{p}(down) = [0.2, 0.3]$, $\mathbf{p}(idle) = [0.3, 0.5]$, and $\mathbf{p}(working) = [0.5, 0.2]$ for a system's working status, we can interpret it as

$$(\forall p_1 \in [\![0.2, 0.3]\!])(\forall p_2 \in [\![0.3, 0.5]\!])(\exists p_3 \in [\![0.2, 0.5]\!])(p_1 + p_2 + p_3 = 1) \qquad (1)$$

With different quantifier assignments, we differentiate non-focal events from focal events based on the respective logic interpretation. An event E is called *focal* if the associated semantics for $\mathbf{p}(E)$ is universal. Otherwise, it is called *non-focal* if the associated semantics is existential. The epistemic uncertainty associated with focal events is 'critical' but 'uncontrollable' to the analyst, whereas the one associated with non-focal events is 'controllable' and 'complementary'. In the above example, the interpretation in Eq.(1) shows that "*down*" and "*idle*" are focal events while "*working*" is non-focal. If the analyst is more interested in "*down*" and "*working*", he/she may assign a different set of interval probability values, for instance, $\mathbf{p}(down) = [0.2, 0.3]$, $\mathbf{p}(working) = [0.2, 0.5]$, and $\mathbf{p}(idle) = [0.6, 0.2]$, with the interpretation

$$(\forall p_1 \in [\![0.2, 0.3]\!])(\forall p_3 \in [\![0.2, 0.5]\!])(\exists p_2 \in [\![0.2, 0.6]\!])(p_1 + p_2 + p_3 = 1)$$

Conditional probability, independence, and generalized interval Bayes' rule

The concepts of conditional probability, independence, and Bayes' rule are essential for the classical probability theory. With independence, we can decompose a complex problem into simpler and manageable components. With Bayes' rule, information can be combined and updated for assessment. Similarly, they are also important for imprecise probabilities.

Different from the definitions in all other forms of imprecise probabilities, the conditional probability in the generalized interval probability theory is defined directly from marginal probability. The *conditional interval probability* $\mathbf{p}(E|C)$ for all $E, C \in \mathcal{A}$ is defined as $\mathbf{p}(E|C) := \mathbf{p}(E \cap C)/\mathrm{dual}\mathbf{p}(C) = \left[\underline{p}(E \cap C)/\underline{p}(C), \overline{p}(E \cap C)/\overline{p}(C)\right]$ when $\mathbf{p}(C) > 0$. Thanks to the unique algebraic properties of generalized intervals, this definition can greatly simplify computation in applications. Only algebraic computation is necessary.

For $A, B, C \in \mathcal{A}$, A is said to be *conditionally independent* with B on C if and only if $\mathbf{p}(A \cap B|C) = \mathbf{p}(A|C)\mathbf{p}(B|C)$. For $A, B \in \mathcal{A}$, A is said to be *independent* with B if and only if $\mathbf{p}(A \cap B) = \mathbf{p}(A)\mathbf{p}(B)$.

The most intuitive meaning of "independence" is that an independence relationship satisfies several graphoid properties. It has been shown that generalized interval probability with the defined independence is graphoid , which has the properties of symmetry, decomposition, composition, contraction, reduction, weak union, redundancy, and intersection (Wang (2011a)).

With the definition of the conditional probability, a *generalized interval Bayes' rule* can be directly derived. For $A \in \mathcal{A}$ and a mutually disjoint event partition $\bigcup_{j=1}^{n} E_j = \Omega$ with $\sum_{j=1}^{n} \mathbf{p}(E_j) = 1$, $\mathbf{p}(E_i|A) = \mathbf{p}(A|E_i)\mathbf{p}(E_i)/\sum_{j=1}^{n} \mathrm{dual}\mathbf{p}(A|E_j)\mathrm{dual}\mathbf{p}(E_j)$.

With the conditional probability and independence defined, a Markov chain model with generalized interval probabilities is used to describe the evolution of a system as the transitions of states which we do not have perfect knowledge about.

Markov chain model with generalized interval probabilities

The imprecise Markov chain based on generalized interval probability can be intuitively kept track of with its resemblance to the classical Markov chain. Given n possible states of a system, a stationary discrete-time imprecise Markov chain is defined by a state transition matrix $\mathbf{A} \in \mathbb{KR}^{n\times n}$ with interval-valued transition probabilities $\mathbf{a}_{ij} = \mathbf{p}(q_{k+1} = i|q_k = j)$, where q_k is the state variable, and $i = 1,\ldots,n$ and $j = 1,\ldots,n$ are its values. Given the probabilistic estimates of the states $\mathbf{\Pi}^{(k)} \in \mathbb{KR}^n$ at time k with elements $\pi_i^{(k)} = \mathbf{p}(q_k = i)$ with $i = 1,\ldots,n$, the probabilistic estimates of states at time $k+1$ is $\mathbf{\Pi}^{(k+1)} = \mathbf{A}\mathbf{\Pi}^{(k)}$. The logic coherence constraint of state probabilities is automatically satisfied during the transition process, stated as follows.

Theorem 1 (Wang (2013)) (*Markov logic coherence constraint*) Given an interval matrix $\mathbf{A}$ and an interval vector $\mathbf{\Pi}^{(k)}$ with their respective elements $\mathbf{a}_{ij}$ ($i = 1,\ldots,n$,$j = 1,\ldots,n$) and $\pi_i^{(k)}$ ($i = 1,\ldots,n$) as generalized interval probabilities, if $\sum_{i=1}^{n} \mathbf{a}_{ij} = [1,1]$ ($\forall j = 1,\ldots,n$) and $\sum_{i=1}^{n} \pi_i^{(k)} = [1,1]$, then the elements of $\mathbf{\Pi}^{(k+1)} = \mathbf{A}\mathbf{\Pi}^{(k)}$ denoted as $\pi_i^{(k+1)}$ ($i = 1,\ldots,n$) also satisfy $\sum_{i=1}^{n} \pi_i^{(k+1)} = [1,1]$.

Notice that the multiplication distributivity of three probability intervals $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ exists, as $(\mathbf{p}_1 + \mathbf{p}_2)\mathbf{p}_3 = \mathbf{p}_1\mathbf{p}_3 + \mathbf{p}_2\mathbf{p}_3$, because $0 \leqslant \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 \leqslant 1$. However, for generalized intervals $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, $(\mathbf{a}+\mathbf{b})\mathbf{c} = \mathbf{ac} + \mathbf{bc}$ does not necessarily hold in general (Gardeñes and Trepat (1980); Markov (1995); Popova (2001)).

GENERALIZED HIDDEN MARKOV MODEL (GHMM)

Given N possible hidden state values $\mathcal{S} = \{1,\ldots,N\}$ and M possible observable symbols $\mathcal{V} = \{v_1,\ldots,v_M\}$, a GHMM with temporal interdependency is defined as $\mathbf{\Lambda} = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$, where $\mathbf{\Pi} = (\pi_i)_{1\times N}$ is a vector of initial interval probability distribution of N possible states with element $\pi_i = \mathbf{p}(q_0 = i)$, q_t denotes the hidden state variable at time t, $\mathbf{A} = (\mathbf{a}_{ij})_{N\times N}$ is the matrix of state transition probabilities that captures the temporal dependencies of hidden state variables with element $\mathbf{a}_{ij} = \mathbf{p}(q_{t+1} = j|q_t = i)$ denoting the interval-valued probability of transition from state i to state j at time t, and $\mathbf{B} = (\mathbf{b}_{jk})_{N\times M}$ is the matrix of observation probabilities with element $\mathbf{b}_{jk} = \mathbf{p}(o_t = v_k|q_t = j)$ denoting the interval-valued probability of observing symbol v_k given the hidden state j and o_t being the observable variable at time t. Given LCC, we have $\sum_{i=1}^{N} \pi_i = [1,1]$, $\sum_{j=1}^{N} \mathbf{a}_{ij} = [1,1]$ for all i, and $\sum_{k=1}^{M} \mathbf{b}_{jk} = [1,1]$ for all j.

The training of a GHMM is to find its parameter $\mathbf{\Lambda}$, given a sequence of T observations $o_{1:T} = \{o_1,\ldots,o_T\}$, such that the maximum likelihood $\mathbf{p}(o_{1:T}|\mathbf{\Lambda})$ can be achieved. In the next section, we will demonstrate that the expected-maximization method developed in HMM can be easily extended to train the GHMM.

Expectation-maximization (EM) for parameter re-estimation

Similar to the traditional incremental learning of HMM parameters, the re-estimation of interval parameters in the GHMM can be performed with the generalized expectation-maximization method (Dempster et al. (1977)) that maximizes the expected log-likelihood $\log \mathbf{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}')$ of the next predicted parameter $\mathbf{\Lambda}'$ within the state space $q_{1:T}$ given the observations $o_{1:T}$ and current parameters $\mathbf{\Lambda}$. If an auxiliary function

$$\mathcal{Q}(\mathbf{\Lambda}'|\mathbf{\Lambda}) = \sum_{q\in\mathcal{S}} \mathbf{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}) \log \mathbf{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}') \tag{2}$$

is defined, the parameter re-estimation is to solve the maximization problem $\max_{\mathbf{\Lambda}'} \mathcal{Q}(\mathbf{\Lambda}'|\mathbf{\Lambda})$, which is more complex than the traditional HMM since $\mathcal{Q}$ is an interval function. Here we define it as two separate maximization problems

$$\max_{\underline{\Lambda}'} \sum_{q\in\mathcal{S}} \overline{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}) \log \underline{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}') \tag{3}$$

and

$$\max_{\overline{\Lambda}'} \sum_{q\in\mathcal{S}} \underline{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}) \log \overline{p}(o_{1:T}, q_{1:T}|\mathbf{\Lambda}') \tag{4}$$

Notice that in Kaucher arithmetic multiplication $\mathbf{x} \times \mathbf{y}$ is defined as $[\overline{x}\underline{y}, \underline{x}\overline{y}]$ when $\underline{x} \geqslant 0, \overline{x} \geqslant 0, \underline{y} < 0, \overline{y} < 0$ (Kaucher (1980)). Similar to the EM training for the HMM, we can define auxiliary functions

$$\gamma_{\tau|t}(i) := \mathbf{p}(q_\tau = i|o_{1:t}, \mathbf{\Lambda}) \tag{5}$$

and

$$\xi_{\tau|t}(i,j) := \mathbf{p}(q_\tau = i, q_{\tau+1} = j|o_{1:t}, \mathbf{\Lambda}) \tag{6}$$

where $\gamma_{1|T}(i) = \pi_i$ and $\gamma_{t|T}(i) = \sum_{j=1}^{N} \xi_{t|T}(i,j)$. The maximization problems in Eqs.(3) and (4) are explicitly expressed in terms of the GHMM parameters as

$$\max_{\underline{\pi}', \underline{a}', \underline{b}'} [\sum_{i=1}^{N} \overline{\gamma}_{1|T}(i) \log \underline{\pi}'_i + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \overline{\xi}_{t|T}(i,j) \log \underline{a}'_{ij} + \sum_{t=1}^{T}\sum_{j=1}^{N} \overline{\gamma}_{t|T}(j) \delta_{o_t=v_m} \log \underline{b}'_{jm}] \tag{7}$$

and

$$\max_{\overline{\pi}', \overline{a}', \overline{b}'} [\sum_{i=1}^{N} \underline{\gamma}_{1|T}(i) \log \overline{\pi}'_i + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \underline{\xi}_{t|T}(i,j) \log \overline{a}'_{ij} + \sum_{t=1}^{T}\sum_{j=1}^{N} \underline{\gamma}_{t|T}(j) \delta_{o_t=v_m} \log \overline{b}'_{jm}] \tag{8}$$

subject to the logic coherent constraints $\sum_{i=1}^{N} \pi_i{}' = [1,1]$, $\sum_{j=1}^{N} \mathbf{a}_{ij}{}' = [1,1]$ for all i, and $\sum_{k=1}^{M} \mathbf{b}_{jk}{}' = [1,1]$ for all j, where $\delta_{true} = 1$ and $\delta_{false} = 0$. The solution is

$$\underline{\pi}'_i = \overline{\gamma}_{1|T}(i), \underline{a}'_{ij} = \frac{\sum_{t=1}^{T-1} \overline{\xi}_{t|T}(i,j)}{\sum_{t=1}^{T-1} \overline{\gamma}_{t|T}(i)}, \underline{b}'_{jm} = \frac{\sum_{t=1}^{T} \overline{\gamma}_{t|T}(j)\delta_{o_t=v_m}}{\sum_{t=1}^{T} \overline{\gamma}_{t|T}(j)} \tag{9}$$

$$\overline{\pi}'_i = \underline{\gamma}_{1|T}(i), \overline{a}'_{ij} = \frac{\sum_{t=1}^{T-1} \underline{\xi}_{t|T}(i,j)}{\sum_{t=1}^{T-1} \underline{\gamma}_{t|T}(i)}, \overline{b}'_{jm} = \frac{\sum_{t=1}^{T} \underline{\gamma}_{t|T}(j)\delta_{o_t=v_m}}{\sum_{t=1}^{T} \underline{\gamma}_{t|T}(j)} \tag{10}$$

The training of the GHMM is based on Eqs.(9) and (10). A forward variable $\alpha_{t|T}(i) = \mathbf{p}(o_{1:t}, q_t = i|\mathbf{\Lambda})$ and a backward variable $\beta_{t|T}(i) = \mathbf{p}(o_{t+1:T}|q_t =$

$i, \mathbf{\Lambda})$ are defined. Given the initial guess of parameters and a sequence of T observed symbols, the prior probabilities $\underline{\pi}(i)$'s and $\overline{\pi}(i)$'s are updated as the expected frequencies of state i's at time $t = 1$, $\overline{\gamma}_{1|T}(i)$ and $\underline{\gamma}_{1|T}(i)$, respectively. This is achieved by computing $\alpha_{t|T}(i)$'s and $\beta_{t|T}(i)$'s, since $\gamma_{t|T}(i) \propto \alpha_{t|T}(i)\beta_{t|T}(i)$. A normalization procedure is needed to make $\gamma_{t|T}(i)$'s interval probability values. Recursively, $\alpha_{1|T}(i) = \pi_i \mathbf{b}_{i,o_1}$ and $\alpha_{t+1|T}(i) = \mathbf{b}_{i,o_{t+1}} \sum_{q_t} \mathbf{a}_{q_t,i}\alpha_{t|T}(q_t)$. $\beta_{T|T}(i) = [1, 1]$ and $\beta_{t|T}(i) = \mathbf{b}_{i,o_{t+1}} \sum_{q_{t+1}} \mathbf{a}_{i,q_{t+1}}\beta_{t+1|T}(q_{t+1})$. $\underline{\pi}'_i$'s and $\overline{\pi}'_i$'s are estimated from the normalized $\overline{\gamma}_{1|T}(i)$'s and $\underline{\gamma}_{1|T}(i)$'s respectively.

Similarly, $\xi_{t|T}(i, j)$'s are computed based on $\xi_{t|T}(i, j) \propto \alpha_{t|T}(i)\mathbf{a}_{ij}\mathbf{b}_{i,o_{t+1}}\beta_{t+1|T}(j)$. Thus $\underline{a}'_{ij}$'s and $\overline{a}'_{ij}$'s are estimated from the normalized $\sum_{t=1}^{T-1} \overline{\xi}_{t|T}(i, j)$'s and $\sum_{t=1}^{T-1} \underline{\xi}_{t|T}(i, j)$'s respectively. The elements of observation matrix $\mathbf{b}'_{jm}$'s are estimated by counting the number of observations for symbols v_m's among all in the sequence then adding the corresponding $\gamma_{t|T}(j)$'s for all t's as $\mathbf{b}'_{jm} \propto \sum_{t=1}^{T} \gamma_{t|T}(j)\delta_{o_t=v_m}$. The final $\underline{b}'_{jm}$ and $\overline{b}'_{jm}$ are the normalized $\sum_{t=1}^{T} \overline{\gamma}_{t|T}(j)\delta_{o_t=v_m}$ and $\sum_{t=1}^{T} \underline{\gamma}_{t|T}(j)\delta_{o_t=v_m}$ respectively.

A numerical example

Here a simple example is used to illustrate the GHMM training process. The number of hidden state is $N = 2$ and the number of observable symbols is $M = 3$. The underlying true transition matrix is assumed to be precise, whereas the true observation matrix is imprecise. Both are randomly generated. These two matrices are used to randomly generate data for training purpose. At the beginning of the training, interval-valued prior probability, transition matrix, and observation matrix are randomly generated. The evolutions of the transition and observation matrices are shown Figures 1 and 2 respectively. In this example, 40 runs of trainings were taken. For each run, 5 sets of observations, each of which has 10 symbols, are randomly generated based on the underlying true matrices. During sampling, either the lower or upper observation matrix is randomly chosen, with equal probability, to generate observations. This is to simulate the observation error. For each run, all 5 sets of data are applied in the training. It stops when either a convergence condition is met or a maximum of 10 iterations is reached. The results are then used as the initial values for the next run.

Figures 1 and 2 show the parameter values at the end of each run, with 40 runs plotted sequentially. The dashed and solid lines correspond to the values of lower and upper bounds for the interval probabilities respectively. The circles indicate the training results when the same data are used to train the HMM with precise probabilities. The plus signs (+) show the training results when the observation data are sampled from the middle values of the imprecise observation matrix, which are the training results when there is no measurement error in observation. The figures show that the plus signs are roughly bounded by the dashed and solid lines. In contrast, the circles drift away from the plus signs, which indicates that the training results can be significantly different from the true ones when measurement error is not considered and the data with measurement errors are used to directly train the HMM. With measurement errors, training the GHMM instead of the HMM can be more robust.

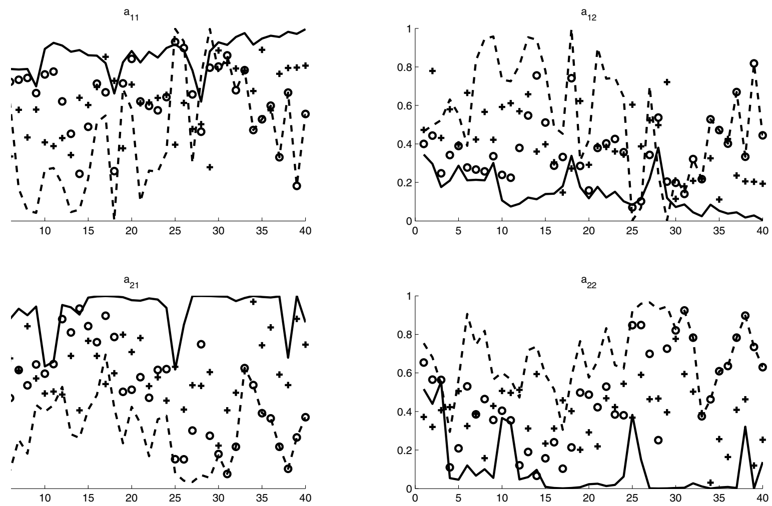

Figure 1. The training of a 2×2 transition matrix A.

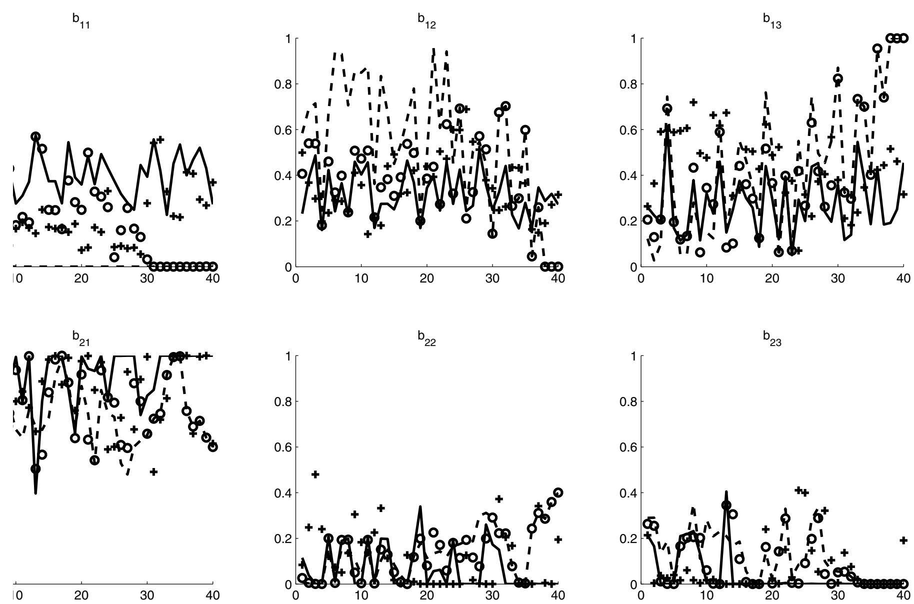

Figure 2. The training of a 2×3 observation matrix B.

CONCLUSION

In this paper, an expectation-maximization approach is developed to train hidden Markov models with generalized interval probability. The unique calculus of generalized interval probability allows for modeling independence and Markov properties under both aleatory and epistemic uncertainty more intuitively than other forms of imprecise probability. It is shown that from imprecise data with measurement errors the training of GHMM's is more robust than the training of HMM's.

ACKNOWLEDGMENTS

This research is being performed using funding received from the U.S. Department of Energy Office of Nuclear Energy's Nuclear Energy University Programs.

REFERENCES

Dempster, A. P. (1967). "Upper and lower probabilities induced by a multivalued mapping." *Annals of Mathematical Statistics*, 38(2), 325–339.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the em algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dimitrova, N. S., Markov, S. M., and Popova, E. D. (1992). "Extended interval arithmetics: new results and applications." *Computer Arithmetic and Enclosure Methods*, L. Atanassova and J. Herzberger, eds., Elsevier, 225–232.

Dubois, D. and Prade, H. (1988). *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum, New York.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., and Sentz, K. (2003). "Constructing probability boxes and dempster-shafer structures." *Report No. SAND2002-4015*, Sandia National Laboratories, Albuquerque, NM.

Gardeñes, E. and Trepat, A. (1980). "Fundamentals of sigla, an interval computing system over the completed set of intervals." *Computing*, 24(2), 161–179.

Kaucher, E. (1980). "Interval analysis in the extended interval space ir." *Computing Supplementa*, Vol. 2, Springer-Verlag, 33–49.

Khreich, W., Granger, E., Miri, A., and Sabourin, R. (2012). "A survey of techniques for incremental learning of hmm parameters." *Information Sciences*, 197, 105–130.

Kuznetsov, V. (1995). "Interval methods for processing statistical characteristics." *Proceedings of the International Workshop on Applications of Interval Computations APIC95*, 23–25.

Markov, S. (1979). "Calculus for interval functions of a real variable." *Computing*, 22(4), 325–337.

Markov, S. (1995). "On directed interval arithmetic and its applications." *Journal of Universal Computer Science*, 1(7), 514–526.

Molchanov, I. (2005). *Theory of Random Sets*. Springer, London.

Möller, B. and Beer, M. (2004). *Fuzzy Randomness: Uncertainty in Civil Engineering and Computational Mechanics*. Springer, Berlin.

Moore, R. E. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

Neumaier, A. (2000). "Clouds, fuzzy sets, and probability intervals." *Reliable Computing*, 10(4), 249–272.

Popova, E. (2001). "Multiplication distributivity of proper and improper intervals." *Reliable computing*, 7(2), 129–140.

Sainz, M. A., Armengol, J., Calm, R., Herrero, P., Jorba, L., and Vehi, J. (2014). *Modal Interval Analysis: New Tools for Numerical Information*. Springer, Heidelberg.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

Wang, Y. (2010). "Imprecise probabilities based on generalised intervals for system reliability assessment." *International Journal of Reliability & Safety*, 4(4), 319–342.

Wang, Y. (2011a). "Independence in generalized interval probability." *Proceedings of 1st International Conference on Vulnerability and Risk Analysis and Management (ICVRAM 2011) and 5th International Symposium on Uncertainty Modeling and Analysis (ISUMA 2011)*, ASCE, 37–44.

Wang, Y. (2011b). "Multiscale uncertainty quantification based on a generalized hidden Markov model." *Journal of Mechanical Design*, 133, 031004.

Wang, Y. (2013). "Generalized Fokker-Planck equation with generalized interval probability." *Mechanical Systems and Signal Processing*, 37(1-2), 92–104.

Weichselberger, K. (2000). "The theory of interval-probability as a unifying concept for uncertainty." *International Journal of Approximate Reasoning*, 24(2–3), 149–170.