Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 8068196, 15 pages
https://doi.org/10.1155/2018/8068196

Research Article

A Novel Entropy-Based Decoding Algorithm for a Generalized High-Order Discrete Hidden Markov Model

Jason Chin-Tiong Chan [1] and Hong Choon Ong [2]

[1] *Ted Rogers School of Management, Ryerson University, 350 Victoria St., Toronto, ON, Canada M5B 2K3*
[2] *School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia*

Correspondence should be addressed to Jason Chin-Tiong Chan; chintiongjason.chan@ryerson.ca

Received 15 December 2017; Revised 12 February 2018; Accepted 27 February 2018; Published 2 May 2018

Academic Editor: Steve Su

The optimal state sequence of a generalized High-Order Hidden Markov Model (HHMM) is tracked from a given observational sequence using the classical Viterbi algorithm. This classical algorithm is based on maximum likelihood criterion. We introduce an entropy-based Viterbi algorithm for tracking the optimal state sequence of a HHMM. The entropy of a state sequence is a useful quantity, providing a measure of the uncertainty of a HHMM. There will be no uncertainty if there is only one possible optimal state sequence for HHMM. This entropy-based decoding algorithm can be formulated in an extended or a reduction approach. We extend the entropy-based algorithm for computing the optimal state sequence that was developed from a first-order to a generalized HHMM with a single observational sequence. This extended algorithm performs the computation exponentially with respect to the order of HMM. The computational complexity of this extended algorithm is due to the growth of the model parameters. We introduce an efficient entropy-based decoding algorithm that used reduction approach, namely, entropy-based order-transformation forward algorithm (EOTFA) to compute the optimal state sequence of any generalized HHMM. This EOTFA algorithm involves a transformation of a generalized high-order HMM into an equivalent first-order HMM and an entropy-based decoding algorithm is developed based on the equivalent first-order HMM. This algorithm performs the computation based on the observational sequence and it requires $O(T\widetilde{N}^2)$ calculations, where $\widetilde{N}$ is the number of states in an equivalent first-order model and T is the length of observational sequence.

1. Introduction

State sequence for the Hidden Markov Model (HMM) is invisible but we can track the most likelihood state sequence based on the model parameter and a given observational sequence. The restored state has many applications especially when the hidden state sequence has meaningful interpretations for making predictions. For example, Ciriza et al. [1] have determined the optimal printing rate based on the HMM model parameter and an optimal time-out based on the restored states. The classical Viterbi algorithm is the most common technique for tracking state sequence from a given observational sequence [2]. However, it does not measure the uncertainty present in the solution. Proakis and Salehi [3] proposed a method for measuring the error of a single state but this method is unable to measure the error of the entire state sequence. Hernando et al. [4] proposed a method of using entropy for measuring the uncertainty of the state sequence of a first-order HMM tracked from a single observational sequence with a length of T. The method is based on the forward recursion algorithm integrated with entropy for computing the optimal state sequence. Mann and McCallum [5] developed an algorithm for computing the subsequent constrained entropy of HMM which is similar to the probabilistic model conditional random fields (CRF). Ilic [6] developed an algorithm based on forward-backward recursion over the entropy semiring, namely, the Entropy Semiring Forward-Backward (ESRFB) algorithm for a first-order HMM with a single observational sequence. ESRFB has lower memory requirement as compared with Mann and McCallum's algorithm for subsequent constrained entropy computation.

This paper is organized as follows. In Section 2, we define the generalized HHMM and present the extended entropy-based algorithm for computing the optimal state sequence developed by Hernando et al. [4] from a first-order to a generalized HHMM. In Section 3, we first review the high-order transformation algorithm proposed by Hadar and Messer [7] and then we introduce EOTFA, an entropy-based order-transformation forward algorithm for computing the optimal state sequence for any generalized HHMM. We discuss future research in Section 4 on entropy associated with state sequence of a generalized high-order HMM.

2. Entropy-Based Decoding Algorithm with an Extended Approach

The uncertainty appearing in a HHMM can be quantified by entropy. This concept is applied for quantifying the uncertainty of the state sequence tracked from a single observational sequence and model parameters. The entropy of the state sequence equals 0 if there is only one possible state sequence that could have generated the observation sequence as there is no uncertainty in the solution. The higher this entropy the higher the uncertainty involved in tracking the hidden state sequence. We extend the entropy-based Viterbi algorithm developed by Hernando et al. [4] for computing the optimal state sequence from a first-order HMM to a high-order HMM, that is, kth-order, where $k \geq 2$. The state entropy in HHHM is computed recursively for the reason of reducing the computational complexity from $O(N^{kT})$ which used direct evaluation method to $O(TN^{k+1})$ in a HHMM where N is the number of states, T is the length of observational sequence, and k is the order of the Hidden Markov Model. In terms of memory space, the entropy-based Viterbi algorithm is more efficient which requires $O(N^{k+1})$ as compared to the classical Viterbi algorithm which requires $O(TN^{k+1})$. The memory space for the classical Viterbi algorithm is dependent on the length of the observational sequence due to the involvement of the process of "back tracking" in computing the optimal state sequence.

Before introducing the extended entropy-based Viterbi algorithm, we define a generalized high-order HMM, that is, kth-order HMM, where $k \geq 2$. These are followed by the definition of forward and backward probability variables for a generalized high-order HMM. These variables are required for computing the optimal state sequence in our decoding algorithm.

2.1. Elements of HHMM. HHMM involves two stochastic processes, namely, hidden state process and observation process. The hidden state process cannot be directly observed. However, it can be observed through the observation process. The observational sequence is generated by the observation process incorporated with the hidden state process. For a discrete HHMM, it must satisfy the following conditions.

The hidden state process $\{q_t\}_{t=2-k}^{T}$ is the kth-order Markov chain that satisfies

$$P\left(q_t \mid \{q_l\}_{l<t}\right) = P\left(q_t \mid \{q_l\}_{l=t-k}^{t-1}\right), \tag{1}$$

where q_t denotes the hidden state at time t and $q_t \in S$, where S is the finite set of hidden states.

The observation process $\{o_t\}_{t=1}^{T}$ is incorporated with the hidden state process according to the state probability distribution that satisfies

$$P\left(o_t \mid \{o_l\}_{l<t}, \{q_l\}_{l\leq t}\right) = P\left(o_t \mid \{q_l\}_{l=t-k+1}^{t}\right), \tag{2}$$

where o_t denotes the observation at time t and $o_t \in V$, where V is the finite set of observation symbols.

The elements for the kth-order discrete HMM are as follows:

(i) Number of distinct hidden states, N

(ii) Number of distinct observed symbols, M

(iii) Length of observational sequence, T

(iv) Observational sequence, $O = \{o_t,\ t = 1, 2, \ldots, T\}$

(v) Hidden state sequence, $Q = \{q_t,\ t = 2-k, \ldots, T\}$

(vi) Possible values for each state, $S = \{s_i,\ i = 1, 2, \ldots, N\}$

(vii) Possible symbols per observation, $V = \{v_w,\ w = 1, 2, \ldots, M\}$

(viii) Initial hidden state probability vector, $\pi_{i_1}, \pi_{i_1 i_2}, \ldots, \pi_{i_1\cdots i_k}$

where π_{i_1} is the probability that model will transit from state s_{i_1},

$$\pi_{i_1} = P\left(q_1 = s_{i_1}\right),$$
$$\sum_{i_1=1}^{N} \pi_{i_1} = 1, \tag{3}$$
$$\pi_{i_1} \geq 0,\ 1 \leq i_1 \leq N$$

$\pi_{i_1 i_2}$ is the probability that model will transit from state s_{i_1} and state s_{i_2},

$$\pi_{i_1 i_2} = P\left(q_0 = s_{i_1}, q_1 = s_{i_2}\right),$$
$$\sum_{i_2=1}^{N} \pi_{i_1 i_2} = 1, \tag{4}$$
$$\pi_{i_1 i_2} \geq 0,\ 1 \leq i_1, i_2 \leq N,$$
$$\vdots$$

$\pi_{i_1\cdots i_k}$ is the probability that model will transit from state s_{i_1}, state s_{i_2}, . . ., and state s_{i_k},

$$\pi_{i_1\cdots i_k} = P\left(q_{2-k} = s_{i_1}, q_{3-k} = s_{i_2}, \ldots, q_1 = s_{i_k}\right),$$
$$\sum_{i_k=1}^{N} \pi_{i_1\cdots i_k} = 1, \tag{5}$$
$$\pi_{i_1\cdots i_k} \geq 0,\ 1 \leq i_1, i_2, \ldots, i_k \leq N$$

(ix) State transition probability matrix, $A_1 = \{a_{i_1 i_2}\}, A_2 = \{a_{i_1 i_2 i_3}\}, \ldots, A_k = \{a_{i_1 i_2 \cdots i_{k+1}}\}$,

where A_{j-1} is the j-dimensional state transition probability matrix and $a_{i_1 i_2 \cdots i_j}$, is the probability of a transition to state s_{i_j} given that it has had a transition from state s_{i_1} to state s_{i_2} to $\cdots$ and to state $s_{i_{j-1}}$ where $j = 2, \ldots, k+1$,

$$a_{i_1 \cdots i_j} = P\left(q_t = s_{i_j} \mid q_{t-1} = s_{i_{j-1}}, q_{t-2} = s_{i_{j-2}}, \ldots, q_{t-j+1} = s_{i_1}\right),$$
$$\sum_{i_j=1}^{N} a_{i_1 i_2 \cdots i_j} = 1, \tag{6}$$
$$a_{i_1 i_2 \cdots i_j} \geq 0$$

(x) Emission probability matrix, $B_1 = \{b_{i_1}(v_m)\}, B_2 = \{b_{i_1 i_2}(v_m)\}, \ldots, B_k = \{b_{i_1 \cdots i_k}(v_m)\}$,

where B_1 is the two-dimensional emission probability matrix and $b_{i_1}(v_m)$ is a probability of observing v_m in state s_{i_1},

$$b_{i_1}(v_m) = P\left(o_t = v_m \mid q_t = s_{i_1}\right),$$
$$\sum_{m=1}^{M} b_{i_1}(v_m) = 1, \tag{7}$$
$$b_{i_1}(v_m) \geq 0, \quad 1 \leq i_1 \leq N,$$

where B_j is the $j+1$-dimensional emission probability matrix and $b_{i_1 \cdots i_j}(v_m)$ is a probability of observing v_m in state s_{i_1} at time $t-j+1$, s_{i_2} at time $t-j+2, \ldots$, and s_{i_j} at time t where $j = 2, \ldots, k$,

$$b_{i_1 \cdots i_j}(v_m) = P\left(o_t = v_m \mid q_t = s_{i_j}, q_{t-1} = s_{i_{j-1}}, \ldots, q_{t-j+1} = s_{i_1}\right),$$
$$\sum_{m=1}^{M} b_{i_1 \cdots i_j}(v_m) = 1, \tag{8}$$
$$b_{i_1 \cdots i_j}(v_m) \geq 0, \quad 1 \leq i_1, i_2, \ldots, i_j \leq N$$

For the kth-order discrete HMM, we summarize the parameters by using the components of $\lambda = (\pi_{i_1}, \pi_{i_1 i_2}, \ldots, \pi_{i_1 i_2 \cdots i_k}, A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_k)$.

Note that throughout this paper, we will use the following notations.

(i) $q_{1:t}$ denotes $q_1, q_2, \ldots, q_t$

(ii) $o_{1:t}$ denotes $o_1, o_2, \ldots, o_t$

2.2. Forward and Backward Probability. The entropy-based algorithm proposed by Hernando et al. [4] for computing the optimal state sequence of a first-order HMM is incorporated with forward recursion process. Recently, high-order HMM are widely used in a variety of applications such as speech recognition [8, 9] and longitudinal data analysis [10, 11]. For the HHMM, the Markov assumption has been weakened since the next state not only depends on the current state but also depends on other historical states. The state dependency is subjected to the order of HMM. Hence we have to modify the classical forward and backward probability variables for the HHMM, that is, the kth-order HMM where $k \geq 2$ are shown as follows.

Definition 1. The forward variable $\alpha_t(i_2, i_3, \ldots, i_{k+1})$ in the kth-order HMM is a joint probability of the partial observation sequence $o_1, o_2, \ldots, o_t$ and the hidden state of s_{i_2} at time $t-k+1$, s_{i_3} at time $t-k+2, \ldots,$ $s_{i_{k+1}}$ at time t where $1 \leq t \leq T$. It can be denoted as

$$\alpha_t(i_2, i_3, \ldots, i_{k+1}) = P\left(o_1, o_2, \ldots, o_t, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}} \mid \lambda\right). \tag{9}$$

From (9), $t = 1$ and $1 \leq i_2, i_3, \ldots, i_{k+1} \leq N$, we obtain the initial forward variable as

$$\begin{aligned}\alpha_1(i_2, i_3, \ldots, i_{k+1}) &= P\left(o_1, q_{2-k} = s_{i_2}, q_{3-k} = s_{i_3}, \ldots, q_1 = s_{i_{k+1}} \mid \lambda\right)\\ &= P\left(q_{2-k} = s_{i_2}, q_{3-k} = s_{i_3}, \ldots, q_1 = s_{i_{k+1}}\right)\\ &\quad\cdot P\left(o_1 \mid q_{2-k} = s_{i_2}, q_{3-k} = s_{i_3}, \ldots, q_1 = s_{i_{k+1}}\right)\\ &= \pi_{i_2 i_3 \cdots i_{k+1}} b_{i_2 i_3 \cdots i_{k+1}}(o_1).\end{aligned} \tag{10}$$

From (9), (10), and $1 \leq i_1, i_2, \ldots, i_k, i_{k+1} \leq N$, we obtain the recursive forward variable for $t = 2, \ldots, T$,

$$\begin{aligned}\alpha_t(i_2, i_3, \ldots, i_{k+1}) &= P\left(o_1, o_2, \ldots, o_t, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}} \mid \lambda\right)\\ &= \sum_{i_1=1}^{N} P\left(o_1, o_2, \ldots, o_t, q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}} \mid \lambda\right)\\ &= \sum_{i_1=1}^{N} P\left(o_1, o_2, \ldots, o_{t-1}, q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k} \mid \lambda\right) P\left(q_t = s_{i_{k+1}} \mid q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k}\right)\\ &\quad\times P\left(o_t \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}\right)\\ &= \left[\sum_{i_1=1}^{N} \alpha_{t-1}(i_1, i_2, \ldots, i_k) a_{i_1 i_2 \cdots i_{k+1}}\right] b_{i_2 i_3 \cdots i_{k+1}}(o_t).\end{aligned} \tag{11}$$

Definition 2. The backward probability variable $\beta_t(i_1, i_2, \dots, i_k)$ in the kth-order HMM is a conditional probability of the partial observation sequence $o_{t+1}, o_{t+2}, \dots, o_T$ given the hidden state of s_{i_1} at time $t-k+1$, s_{i_2} at time $t-k+2, \dots$, and s_{i_k} at time t. It can be denoted as

$$\beta_t(i_1, i_2, \dots, i_k) = P\left(o_{t+1}, o_{t+2}, \dots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k}, \lambda\right), \tag{12}$$

where $1 \le t \le T$, $1 \le i_1, i_2, \dots, i_k \le N$.

We obtain the initial backward probability variable as

$$\beta_T(i_1, i_2, \dots, i_k) = 1. \tag{13}$$

From (12) and (13), we obtain the recursive backward probability variable for $t = 1, 2, \dots, T-1$,

$$\begin{aligned}\beta_t(i_1, i_2, \dots, i_k) &= P\left(o_{t+1}, o_{t+2}, \dots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k}, \lambda\right)\\ &= \sum_{i_{k+1}=1}^{N} P\left(o_{t+1}, o_{t+2}, \dots, o_T, q_{t+1} = s_{i_{k+1}} \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k}, \lambda\right)\\ &= \sum_{i_{k+1}=1}^{N} P\left(o_{t+2}, \dots, o_T \mid q_{t-k+2} = s_{i_2}, \dots, q_{t+1} = s_{i_{k+1}}, \lambda\right) P\left(q_{t+1} = s_{i_{k+1}} \mid q_{t-k+1} = s_{i_1}, \dots, q_t = s_{i_k}, \lambda\right) \times P\left(o_{t+1} \mid q_{t-k+2} = s_{i_2}, \dots, q_{t+1} = s_{i_{k+1}}\right)\\ &= \sum_{i_{k+1}=1}^{N} \beta_{t+1}(i_2, i_3, \dots, i_{k+1})\, a_{i_1 i_2 \cdots i_{k+1}} b_{i_2 i_3 \cdots i_{k+1}}(o_{t+1}).\end{aligned} \tag{14}$$

The probability of the observational sequence given the model parameter for the first-order HMM can be represented by using the classical forward probability and backward probability variables [2]. We extend it to HHMM by using our modified forward probability and backward probability variables. The proof is due to Rabiner [2].

Definition 3. Let $\alpha_t(i_1, i_2, \dots, i_k)$ and $\beta_t(i_1, i_2, \dots, i_k)$ be the forward probability variable and backward probability variable, respectively; $P(O \mid \lambda)$ is presented using the forward and backward probability variables as

$$\begin{aligned}P(O \mid \lambda) &= P(o_1, \dots, o_T \mid \lambda)\\ &= \sum_{i_i=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} \alpha_t(i_1, i_2, \dots, i_k)\, \beta_t(i_1, i_2, \dots, i_k).\end{aligned} \tag{15}$$

Proof.

$$\begin{aligned}P(O \mid \lambda) = P(o_1, o_2, \dots, o_T \mid \lambda) &= \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} P\left(o_1, o_2, \dots, o_T, q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k} \mid \lambda\right)\\ &= \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} P\left(o_1, o_2, \dots, o_t, q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k} \mid \lambda\right)\\ &\quad \times P\left(o_{t+1}, o_{t+2}, \dots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \dots, q_t = s_{i_k}, \lambda\right)\\ &= \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} \alpha_t(i_1, i_2, \dots, i_k)\, \beta_t(i_1, i_2, \dots, i_k).\end{aligned} \tag{16}$$

We now normalize both of the forward and backward probability variables. These normalized variables are required as the intermediate variables for the algorithm of state entropy computation. □

Definition 4. The normalized forward probability variable $\hat{\alpha}_t(i_2, i_3, \dots, i_{k+1})$ in the kth-order HMM is defined as the probability of the hidden state of s_{i_2} at time $t-k+1$, s_{i_3} at time $t-k+2, \dots$, $s_{i_{k+1}}$ at time t given the partial observation sequence $o_1, o_2, \dots, o_t$ where $1 \le t \le T$.

$$\hat{\alpha}_t(i_2, i_3, \dots, i_{k+1}) = P\left(q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \dots, q_t = s_{i_{k+1}} \mid o_1, o_2, \dots, o_t\right). \tag{17}$$

From (10), (17), $t = 1$, and $1 \le i_1, i_2, \dots, i_k \le N$, we obtain the initial normalized forward probability variable as

$$\begin{aligned}\hat{\alpha}_1(i_2, i_3, \dots, i_{k+1}) &= P\left(q_{2-k} = s_{i_2}, q_{3-k} = s_{i_3}, \dots, q_1 = s_{i_{k+1}} \mid o_1\right)\end{aligned}$$

$$= \frac{P\left(q_{2-k} = s_{i_2}, q_{3-k} = s_{i_3}, \ldots, q_1 = s_{i_{k+1}}, o_1\right)}{P(o_1)} = \frac{\pi_{i_2 i_3 \cdots i_{k+1}} b_{i_2 i_3 \cdots i_{k+1}}(o_1)}{r_0}, \tag{18}$$

where

$$r_0 = \sum_{j_k=1}^{N} \cdots \sum_{j_1=1}^{N} \pi_{j_1 j_2 \cdots j_k} b_{j_1 j_2 \cdots j_k}(o_1). \tag{19}$$

From (11), (17), (18), and $t = 2, \ldots, T$, $1 \le i_1, i_2, \ldots, i_k, i_{k+1} \le N$, we obtain the recursive normalized forward probability variable as

$$\begin{aligned} &\widehat{\alpha}_t(i_2, i_3, \ldots, i_{k+1}) \\ &= P\left(q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}} \mid o_1, o_2, \ldots, o_t\right) \\ &= \frac{P\left(q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_1, o_2, \ldots, o_t\right)}{P(o_t \mid o_1, o_2, \ldots, o_{t-1})} \\ &= \frac{\left[\sum_{i_1=1}^{N} \widehat{\alpha}_{t-1}(i_1, i_2, \ldots, i_k) a_{i_1 i_2 \cdots i_{k+1}}\right] b_{i_2 i_3 \cdots i_{k+1}}(o_t)}{r_t}, \end{aligned} \tag{20}$$

where

$$r_t = \sum_{j_k=1}^{N} \cdots \sum_{j_1=1}^{N} \sum_{i_1=1}^{N} \alpha_{t-1}(i_1, j_1, \ldots, j_{k-1}) \cdot a_{i_1 j_1 \cdots j_k} b_{j_1 j_2 \cdots j_k}(o_t). \tag{21}$$

Note that the normalization factor r_t ensures that the probabilities sum to one and it also represents the conditional observational probability [2].

Definition 5. The normalized backward probability variable $\widehat{\beta}_t(i_1, i_2, \ldots, i_k)$ in the kth-order HMM is defined as the quotient of a conditional probability of the partial observation sequence $o_{t+1}, o_{t+2}, \ldots, o_T$ given the hidden state of s_{i_1} at time $t-k+1$, s_{i_2} at time $t-k+2, \ldots$, s_{i_k} at time t, and a conditional probability of the partial observation sequence $o_{t+1}, o_{t+2}, \ldots, o_T$ given the entire observation sequence $o_1, o_2, \ldots, o_T$. It can be denoted as

$$\widehat{\beta}_t(i_1, i_2, \ldots, i_k) = \frac{P\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \ldots, q_t = s_{i_k}\right)}{P(o_{t+1}, o_{t+2}, \ldots, o_T \mid o_1, o_2, \ldots, o_T)}, \tag{22}$$

where $1 \le t \le T$, $1 \le i_1, i_2, \ldots, i_k \le N$

From (14) and (22), we obtain the recursive normalized backward probability variable as

$$\begin{aligned} \widehat{\beta}_t(i_1, i_2, \ldots, i_k) &= \frac{P\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \ldots, q_t = s_{i_k}\right)}{P(o_{t+1}, o_{t+2}, \ldots, o_T \mid o_1, o_2, \ldots, o_T)} \\ &= \frac{\sum_{i_{k+1}=1}^{N} P\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid q_{t-k+1} = s_{i_1}, q_{t-k+2} = s_{i_2}, \ldots, q_t = s_{i_k}, q_{t+1} = s_{i_{k+1}}\right)}{P(o_{t+1}, \ldots, o_T \mid o_1, o_2, \ldots, o_T)} \\ &= \frac{\sum_{i_{k+1}=1}^{N} \widehat{\beta}_{t+1}(i_2, i_3, \ldots, i_{k+1}) a_{i_1 i_2 \cdots i_{k+1}} b_{i_2 i_3 \cdots i_{k+1}}(o_{t+1})}{r_{t+1}}, \end{aligned} \tag{23}$$

where

$$r_{t+1} = \sum_{j_k=1}^{N} \cdots \sum_{j_1=1}^{N} \sum_{i_1=1}^{N} \alpha_t(i_1, j_1, \ldots, j_{k-1}) \cdot a_{i_1 j_1 \cdots j_k} b_{j_1 j_2 \cdots j_k}(o_{t+1}). \tag{24}$$

Our extended algorithm includes the normalized forward recursion given by (18) and (20). The extended algorithm for the kth-order HMM requires $O(TN^{k+1})$ calculations if we include either normalized forward recursion given by (18) and (20) or the normalized backward recursion given by (13) and (23). The direct evaluation method, in comparison, requires $O(N^{T+k-1})$ calculations where N is the number of states, T is the length of observational sequence, and k is the order of the Hidden Markov Model.

2.3. The Algorithm by Hernando et al. Hernando et al. [4] are pioneers for using entropy to compute the optimal state sequence of a first-order HMM with a single observational sequence. This algorithm is based on a first-order HMM normalized forward probability,

$$\widehat{\alpha}_t(j) = P\left(q_t = s_j \mid o_1, o_2, \ldots, o_t\right), \tag{25}$$

auxiliary probability,

$$P\left(q_{t-1} = s_i \mid q_t = s_j, o_{1:t}\right), \tag{26}$$

and intermediate entropy,

$$H_t(s_j) = H\left(q_{1:t-1} \mid q_t = s_j, o_{1:t}\right). \tag{27}$$

The entropy-based algorithm for computing the optimal state sequence of a first-order HMM is as follows [4].

(1) Initialization. For $t = 1$ and $1 \le j \le N$,

$$H_1(s_j) = 0, \quad \hat{\alpha}_1(j) = \frac{\pi_j b_j(o_1)}{\sum_{i=1}^{N} \pi_i b_i(o_1)}. \tag{28}$$

(2) Recursion. For $t = 2, \ldots, T-1$, and $1 \le j \le N$,

$$\begin{aligned} \hat{\alpha}_t(j) &= \frac{\sum_{i=1}^{N} \hat{\alpha}_{t-1}(i) a_{ij} b_j(o_t)}{\sum_{k=1}^{N} \sum_{i=1}^{N} \hat{\alpha}_{t-1}(i) a_{ik} b_k(o_t)}, \\ P(q_{t-1} = s_i \mid q_t = s_j, o_{1:t}) &= \frac{a_{ij}\hat{\alpha}_{t-1}(i)}{\sum_{k=1}^{N} \sum_{i=1}^{N} a_{ik}\hat{\alpha}_{t-1}(i)}, \\ H_t(s_j) &= \sum_{i=1}^{N} \left[P(q_{t-1} = s_i \mid q_t = s_j, o_{1:t}) H_{t-1}(s_i) \right] \\ &\quad - \sum_{i=1}^{N} \left[P(q_{t-1} = s_i \mid q_t = s_j, o_{1:t}) \cdot \log_2 P(q_{t-1} = s_i \mid q_t = s_j) \right]. \end{aligned} \tag{29}$$

(3) Termination

$$H_T(q_{1:T} \mid o_{1:T}) = \sum_{i=1}^{N} H_T(s_i) \hat{\alpha}_T(i) - \sum_{i=1}^{N} \hat{\alpha}_T(i) \log_2 \hat{\alpha}_T(i). \tag{30}$$

This algorithm performs the computation linearly with respect to the length of the observation sequence with computational complexity $O(TN^2)$. It requires the memory space of $O(N^2)$ which indicates that the memory space is independent of the observational sequence.

2.4. The Computation of the Optimal State Sequence for a HHMM. The extended classical Viterbi algorithm is commonly used for computing the optimal state sequence for HHMM. This algorithm provides the solution along with its likelihood. This likelihood probability can be determined as follows.

$$P(q_1, q_2, \ldots, q_T \mid o_1, o_2, \ldots, o_T) = \frac{P(q_1, q_2, \ldots, q_T, o_1, o_2, \ldots, o_T)}{P(o_1, o_2, \ldots, o_T)}. \tag{31}$$

This probability can be used as a measure of quality of the solution. The higher the probability of our "solution," the better our "solution." Entropy can also be used for measuring the quality of the state sequence of the kth-order HMM. Hence, state entropy is proposed to be used for obtaining the optimal state sequence of a HHMM.

We define entropy of a discrete random variable as follows [12].

Definition 6. The entropy $H(X)$ of a discrete random variable X with a probability mass function $P(X = x)$ is defined as

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x). \tag{32}$$

When the log has a base of 2, the unit of the entropy is bits. Note that $0 \log 0 = 0$.

From (32), the entropy of the distribution for all possible state sequences is as follows:

$$\begin{aligned} H(q_1, q_2, \ldots, q_T \mid o_1, o_2, \ldots, o_T) &= -\sum_{Q} \left[P(q_1 = s_{i_1}, q_2 = s_{i_2}, \ldots, q_T = s_{i_T} \mid o_1, o_2, \ldots, o_T) \right. \\ &\quad \left. \cdot \log_2 P(q_1 = s_{i_1}, q_2 = s_{i_2}, \ldots, q_T = s_{i_T} \mid o_1, o_2, \ldots, o_T) \right]. \end{aligned} \tag{33}$$

For the first-order HMM, if all N^T possible state sequences are equally likely to generate a single observational sequence with a length of T, then the entropy equals $T \log_2 N$. The entropy is $kT \log_2 N$ in the kth-order HMM if all N^{kT} possible state sequences are equally likely to produce the observational sequence.

For this extended algorithm, we require an intermediate state entropy variable, $H_t(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}})$ that can be computed recursively using the previous variable, $H_{t-1}(s_{i_1}, s_{i_2}, \ldots, s_{i_k})$.

We define the state entropy variable for the kth-order HMM as follows.

Definition 7. The state entropy variable, $H_t(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}})$, in the kth-order HMM is the entropy of all the state sequences that lead to state of s_{i_2} at time $t-k+1$, s_{i_3} at time $t-k+2, \ldots$, and $s_{i_{k+1}}$ at time t, given the observation sequence $o_1, o_2, \ldots, o_t$. It can be denoted as

$$H_t(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}) = H(q_{2-k:t-1} \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}). \tag{34}$$

We analyse the state entropy for the kth-order HMM in detail, shown as follows.

From (34) and $t = 1$, we obtain the initial state entropy variable as

$$H_1(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}) = 0. \tag{35}$$

From (34) and (35) we obtain the recursion on the entropy for $t = 2, \ldots, T$, and $1 \le i_1, i_2, \ldots, i_{k+1} \le N$,

$$H_t(s_{i_2}, s_{i_3}, \ldots, i_{k+1}) = H(q_{2-k:t-1} \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t})$$

$$= H\left(q_{t-k:t-1} \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right) + H\left(q_{2-k:t-2} \mid q_{t-k}, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right), \tag{36}$$

where

$$\begin{aligned}
&H\left(q_{t-k:t-1} \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right) \\
&\quad = -\sum_{i_1=1}^{N} \left[P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right)\right. \\
&\quad \left.\cdot \log_2\left(P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right)\right)\right], \\
&H\left(q_{2-k:t-2} \mid q_{t-k}, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right) \\
&\quad = \sum_{i_1=1}^{N} \left[P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right)\right. \\
&\quad \left.\cdot H\left(q_{2-k:t-2} \mid q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right)\right] \\
&\quad = \sum_{i_1=1}^{N} \left[P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}\right) H_{t-1}\left(s_{i_1}, s_{i_2}, \ldots, s_{i_k}\right)\right].
\end{aligned} \tag{37}$$

The auxiliary probability $P(q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k}, q_t = s_{i_{k+1}}, o_{1:t})$ is required for our extended entropy-based algorithm. It can be computed as follows:

$$\begin{aligned}
&P\left(q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k}, q_t = s_{i_{k+1}}, o_{1:t}\right) \\
&= \frac{P\left(q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{t-k+1}, \ldots, o_t \mid q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k}, o_{1:t-1}\right) P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid o_{1:t-1}\right)}{P\left(q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{t-k+1}, \ldots, o_t \mid o_{1:t-1}\right)} \\
&= \frac{P\left(o_{t-k+1}, \ldots, o_t \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}\right) P\left(q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}} \mid q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k}\right) P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid o_{1:t-1}\right)}{P\left(o_{t-k+1}, \ldots, o_t \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}\right) P\left(q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}} \mid o_{1:t-1}\right)} \\
&= \frac{P\left(q_{t-k} = s_{i_2}, \ldots, q_t = s_{i_{k+1}} \mid q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k}\right) P\left(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid o_{1:t-1}\right)}{\sum_{j_k=1}^{N} \sum_{j_{k-1}=1}^{N} \cdots \sum_{j_1=1}^{N} P\left(q_{t-k+1} = s_{j_2}, \ldots, q_t = s_{j_{k+1}} \mid q_{t-k} = s_{j_1}, \ldots, q_{t-1} = s_{j_k}\right) P\left(q_{t-k} = s_{j_1}, \ldots, q_{t-1} = s_{j_k} \mid o_{1:t-1}\right)} \\
&= \frac{a_{i_1 i_2 \cdots i_k i_{k+1}} \widehat{\alpha}_{t-1}\left(i_1, i_2, \ldots, i_k\right)}{\sum_{j_k=1}^{N} \sum_{j_{k-1}=1}^{N} \cdots \sum_{j_1=1}^{N} a_{j_1 j_2 \cdots j_k i_{k+1}} \widehat{\alpha}_{t-1}\left(j_1, j_2, \ldots, j_k\right)}.
\end{aligned} \tag{38}$$

For the final process of our extended algorithm, we are required to compute the conditional entropy $H(q_{1:T} \mid o_{1:T})$ which can be expanded as follows:

$$\begin{aligned}
&H\left(q_{1:T} \mid o_{1:T}\right) = H\left(q_{1:T-k} \mid q_{T-k+1} = s_{i_1}, q_{T-k+2} = s_{i_2},\right. \\
&\quad \left. q_{T-k+3} = s_{i_3}, \ldots, q_T = s_{i_k}, o_{1:T}\right) + H\left(q_{T-k+1:T} \mid o_{1:T}\right) \\
&\quad = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} H_T\left(s_{i_1}, s_{i_2}, \ldots, s_{i_k}\right) \\
&\quad \cdot \widehat{\alpha}_T\left(i_1, i_2, \ldots, i_k\right) \\
&\quad - \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} \widehat{\alpha}_T\left(i_1, i_2, \ldots, i_k\right) \\
&\quad \cdot \log_2\left(\widehat{\alpha}_T\left(i_1, i_2, \ldots, i_k\right)\right).
\end{aligned} \tag{39}$$

The following basic properties of HMM and entropy are used for proving Lemma 8.

(i) According to the generalized high-order HMM, state $q_{t-k-j+1}$, $j \geq 2$ and q_t are statistically independent given $q_{t-k}, q_{t-k+1}, q_{t-k+2}, \ldots, q_{t-1}$. The same applies to $q_{t-k-j+1}$,

$j \geq 2$ and o_t are statistically independent given $q_{t-k}, q_{t-k+1}, q_{t-k+2}, \ldots, q_{t-1}$.

(ii) According to the basic property of entropy [12],

$$H(X \mid Y = y) = H(X) \quad \text{if } X \text{ and } Y \text{ are independent.} \tag{40}$$

We now introduce the following lemma for the kth-order HMM. The following proof is due to Hernando et al. [4].

Lemma 8. *For the kth-order HMM, the entropy of the state sequence up to time $t-k-1$, given the states from time $t-k$ to time $t-1$ and the observations up to time $t-1$, is conditionally independent of the state and observation at time t*

$$H_{t-1}(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = H(q_{1:t-2} \mid q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}). \tag{41}$$

Proof.

$$\begin{aligned} &H(q_{1:t-2} \mid q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}) = H(q_{1:t-2} \mid q_{t-k} = s_{i_1}, q_{t-k+1} \\ &= s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_{t-1} = s_{i_k}, o_{1:t-1}, q_t = s_{i_{k+1}}, o_t) \\ &= H(q_{1:t-2} \mid q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, q_{t-k+2} \\ &= s_{i_3}, \ldots, q_{t-1} = s_{i_k}, o_{1:t-1}) = H_{t-1}(s_{i_1}, s_{i_2}, \ldots, s_{i_k}). \end{aligned} \tag{42}$$

Our extended entropy-based algorithm for computing the optimal state sequence is based on normalized forward recursion variable, state entropy recursion variable, and auxiliary probability. From (18), (20), (35), (36), (38), and (39), we construct the extended entropy-based decoding algorithm for the kth-order HMM as follows:

(1) Initialization. For $t = 1$ and $1 \leq i_2, i_3, \ldots, i_{k+1} \leq N$,

$$\begin{aligned} &H_1(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}) = 0, \\ &\widehat{\alpha}_1(i_2, i_3, \ldots, i_{k+1}) \\ &\quad = \frac{\pi_{i_2 i_3 \cdots i_{k+1}} b_{i_2 i_3 \cdots i_{k+1}}(o_1)}{\sum_{j_k=1}^{N} \sum_{j_{k-1}=1}^{N} \cdots \sum_{j_1=1}^{N} \pi_{j_1 j_2 \cdots j_k} b_{j_1 j_2 \cdots j_k}(o_1)}. \end{aligned} \tag{43}$$

(2) Recursion. For $t = 2, \ldots, T-1$, and $1 \leq i_1, i_2, \ldots, i_{k+1} \leq N$,

$$\begin{aligned} &\widehat{\alpha}_t(i_2, i_3, \ldots, i_{k+1}) = \frac{\sum_{i_1=1}^{N} \widehat{\alpha}_{t-1}(i_1, i_2, \ldots, i_k) a_{i_1 i_2 \cdots i_k i_{k+1}} b_{i_2 i_3 \cdots i_k i_{k+1}}(o_t)}{\sum_{j_k=1}^{N} \cdots \sum_{j_1=1}^{N} \sum_{i_1=1}^{N} \widehat{\alpha}_{t-1}(i_1, j_1, \ldots, j_{k-1}) a_{i_1 j_1 \cdots j_k} b_{j_2 j_3 \cdots j_{k+1}}(o_t)}, \\ &P(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}) = \frac{a_{i_1 i_2 \cdots i_k i_{k+1}} \widehat{\alpha}_{t-1}(i_1, i_2, \ldots, i_k)}{\sum_{j_k=1}^{N} \sum_{j_{k-1}=1}^{N} \cdots \sum_{j_1=1}^{N} a_{j_1 j_2 \cdots j_k i_{k+1}} \widehat{\alpha}_{t-1}(j_1, j_2, \ldots, j_k)}, \\ &H_t(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}) = \sum_{i_1=1}^{N} \left[P(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}) H_{t-1}(s_{i_1}, s_{i_2}, \ldots, s_{i_k})\right] \\ &\quad - \sum_{i_1=1}^{N} \left[P(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t}) \right. \\ &\quad \left. \cdot \log_2 \left(P(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t})\right)\right]. \end{aligned} \tag{44}$$

(3) Termination

$$\begin{aligned} H(q_{1:T} \mid o_{1:T}) &= \sum_{i_1=1}^{N} \cdots \sum_{i_k=1}^{N} H_T(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) \\ &\quad \cdot \widehat{\alpha}_T(i_1, i_2, \ldots, i_k) \\ &\quad - \sum_{i_1=1}^{N} \cdots \sum_{i_k=1}^{N} \widehat{\alpha}_T(i_1, i_2, \ldots, i_k) \log_2 \widehat{\alpha}_T(i_1, i_2, \ldots, i_k). \end{aligned} \tag{45}$$

This extended algorithm performs the computation of the optimal state sequence linearly with respect to the length of observational sequence which requires $O(TN^{k+1})$ calculation and it has memory space that is independent of the length of observational sequence, $O(N^{k+1})$, since $\widehat{\alpha}_t(i_2, i_3, \ldots, i_{k+1})$, $H_t(s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}})$, $P(q_{t-k} = s_{i_1}, \ldots, q_{t-1} = s_{i_k} \mid q_{t-k+1} = s_{i_2}, \ldots, q_t = s_{i_{k+1}}, o_{1:t})$ should be computed only once in tth iteration and, having been used for the computation of $(t+1)$th, they can be deleted from the space storage. □

2.5. Numerical Illustration for the Second-Order HMM. We consider a second-order HMM for illustrating our extended entropy-based algorithm in computing the optimal state sequence. Let us assume that this second-order HMM has the state space S, which is $S = \{s_1, s_2\}$ and the possible symbols per observation which is $O = \{v_1, v_2, v_3\}$.

The graphical representation of the first-order HMM that is used for the numerical example in this section is given in

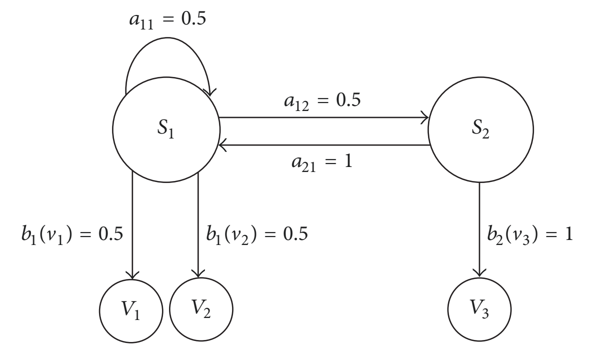

Figure 1: The graphical diagram shows a first-order HMM with 2 states and 3 observational symbols.

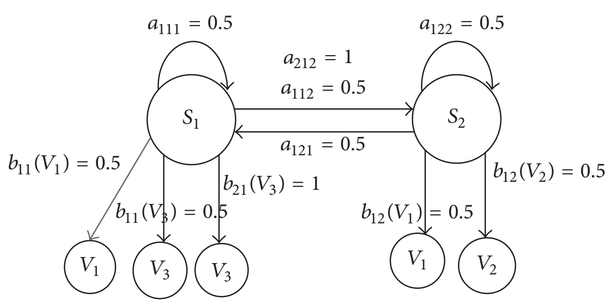

Figure 2: The graphical diagram shows a second-order HMM with 2 states and 3 observational symbols.

Figure 1. The second-order HMM in Figure 2 is developed based on the first-order HMM in Figure 1 which has two states and three observational symbols. A HMM of any order has the parameters of $\lambda = (\pi, A, B)$ where π is the initial state probability vector, A is the state transition probability matrix, and B is the emission probability matrix. Note that the matrices of A and B whose components are indicated as $a_{i_1 i_2}$, $a_{i_1 i_2 i_3}$, $b_{i_2}(o_t = v_m)$ and $b_{i_2 i_3}(o_t = v_m)$ where $1 \le i_1, i_2, i_3 \le 2$ and $1 \le m \le 3$ can be obtained from Figures 1 and 2. However, the initial state probability vector is not shown in the above graphical diagrams.

The initial state probability vectors for the first-order and second-order HMM are shown as follows:

$$\pi_1 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix},$$
$$\pi_2 = \begin{bmatrix} 0.5 & 0 \end{bmatrix}, \qquad (46)$$
$$\pi_3 = \begin{bmatrix} 0.5 & 0 \end{bmatrix}.$$

$\pi_1 = \{\dot{\pi}_{i_2}\}$ is the initial state probability vector for the first-order HMM and $\pi_2 = \{\dot{\pi}_{i_2 1}\}$ and $\pi_3 = \{\dot{\pi}_{i_2 2}\}$ are the initial state probability vectors for the second-order HMM where $\dot{\pi}_{i_2} = P(q_1 = s_{i_2})$, $\dot{\pi}_{i_2 1} = P(q_1 = s_1, q_0 = s_{i_2})$, $\dot{\pi}_{i_2 2} = P(q_1 = s_2, q_0 = s_{i_2})$, and $1 \le i_2 \le 2$.

The state transition probability matrices for the first-order and second-order HMMs are shown as follows:

$$A_1 = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix},$$
$$A_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0 \end{bmatrix}, \qquad (47)$$
$$A_3 = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix}.$$

$A_1 = \{a_{i_1 i_2}\}$ is the state transition probability matrix for the first-order HMM and $A_2 = \{a_{i_1 i_2 1}\}$ and $A_3 = \{a_{i_1 i_2 2}\}$ are the state transition probability matrices for the second-order HMM where $a_{i_1 i_2} = P(q_t = s_{i_2} \mid q_{t-1} = s_{i_1})$, $a_{i_1 i_2 1} = P(q_t = s_1 \mid q_{t-1} = s_{i_2}, q_{t-2} = s_{i_1})$, $a_{i_1 i_2 2} = P(q_t = s_2 \mid q_{t-1} = s_{i_2}, q_{t-2} = s_{i_1})$, and $1 \le i_1, i_2 \le 2$

The emission probability matrices for the first-order and second-order HMMs are shown as follows:

$$B_1 = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix},$$
$$B_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0 \end{bmatrix},$$
$$B_3 = \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix}, \qquad (48)$$
$$B_4 = \begin{bmatrix} 0.5 & 0 \\ 1 & 0 \end{bmatrix}.$$

$B_1 = \{b_{i_2}(o_t = v_m)\}$ is the emission probability matrix for the first-order HMM and $B_2 = \{b_{i_2 i_3}(o_t = v_1)\}$, $B_3 = \{b_{i_2 i_3}(o_t = v_2)\}$, and $B_4 = \{b_{i_2 i_3}(o_t = v_3)\}$ are the emission probability matrices for the second-order HMM where $b_{i_2}(o_t = v_m) = P(o_t = v_m \mid q_t = s_{i_2})$, $b_{i_2 i_3}(o_t = v_1) = P(o_t = v_1 \mid q_t = s_{i_3}, q_{t-1} = s_{i_2})$, $b_{i_2 i_3}(o_t = v_2) = P(o_t = v_2 \mid q_t = s_{i_3}, q_{t-1} = s_{i_2})$, and $b_{i_2 i_3}(o_t = v_3) = P(o_t = v_3 \mid q_t = s_{i_3}, q_{t-1} = s_{i_2})$.

The following is the observational sequence that we used for illustrating our extended algorithm:

$$o_{1:6} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2, o_5 = v_3, o_6 = v_1). \qquad (49)$$

We applied our extended algorithm for computing the optimal state sequence based on state entropy. The computed value of the state entropy is shown in Figure 3.

The total entropy after each time step is displayed at the bottom of Figure 3. For example, after receiving the second observation, that is, $o_{1:2} = (o_1 = v_1, o_2 = v_1)$, it has produced two state sequences which are $q_{1:2} = (q_1 = s_1, q_2 = s_1)$ and $q_{1:2} = (q_1 = s_1, q_2 = s_2)$ as shown by the bold arrows. Each possible state sequence has a probability of 0.5; that is, $\hat{\alpha}_2(1,1) = \hat{\alpha}_2(1,2) = 0.5$, and hence the total entropy is 1 bit.

State \ Obs	$o_1 = v_1$	$o_2 = v_1$	$o_3 = v_3$	$o_4 = v_2$	$o_5 = v_3$	$o_6 = v_1$
s_1	$H_1(1,1) = 0$ $\widehat{\alpha}_1(1,1) = 0.5$ $H_1(2,1) = 0$ $\widehat{\alpha}_1(2,1) = 0$	$H_2(1,1) = 0$ $\widehat{\alpha}_2(1,1) = 0.5$ $H_2(2,1) = 0$ $\widehat{\alpha}_2(2,1) = 0$	$H_3(1,1) = 0.5$ $\widehat{\alpha}_3(1,1) = 0.33$ $H_3(2,1) = 0.5$ $\widehat{\alpha}_3(2,1) = 0.67$	$H_4(1,1) = 0$ $\widehat{\alpha}_4(1,1) = 0$ $H_4(2,1) = 0$ $\widehat{\alpha}_4(2,1) = 0$	$H_5(1,1) = 0$ $\widehat{\alpha}_5(1,1) = 0$ $H_5(2,1) = 0$ $\widehat{\alpha}_5(2,1) = 1$	$H_6(1,1) = 0$ $\widehat{\alpha}_6(1,1) = 0$ $H_6(2,1) = 0$ $\widehat{\alpha}_6(2,1) = 0$
s_2	$H_1(1,2) = 0$ $\widehat{\alpha}_1(1,2) = 0.5$ $H_1(2,2) = 0$ $\widehat{\alpha}_1(2,2) = 0$	$H_2(1,2) = 0$ $\widehat{\alpha}_2(1,2) = 0.5$ $H_2(2,2) = 0$ $\widehat{\alpha}_2(2,2) = 0$	$H_3(1,2) = 0$ $\widehat{\alpha}_3(1,2) = 0$ $H_3(2,2) = 0$ $\widehat{\alpha}_3(2,2) = 0$	$H_4(1,2) = 0$ $\widehat{\alpha}_4(1,2) = 1$ $H_4(2,2) = 0$ $\widehat{\alpha}_4(2,2) = 0$	$H_3(1,2) = 0$ $\widehat{\alpha}_3(1,2) = 0$ $H_3(2,2) = 0$ $\widehat{\alpha}_3(2,2) = 0$	$H_6(1,2) = 0$ $\widehat{\alpha}_6(1,2) = 1$ $H_6(2,2) = 0$ $\widehat{\alpha}_6(2,2) = 0$
Total entropy	1	1	1.41	0	0	0

FIGURE 3: The evolution of the trellis structure of the second-order HMM with the observation sequence $o_{1:6} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2, o_5 = v_3, o_6 = v_1)$.

However, after receiving the fourth observation, that is, $o_{1:4} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2)$, it has produced one state sequence which is $q_{1:4} = (q_1 = s_1, q_2 = s_2, q_3 = s_1, q_4 = s_2)$ as shown by the dashed arrow. This possible state sequence has a probability of 1, that is, $\widehat{\alpha}_4(1,2) = 1$, and hence the total entropy is 0 bit. After receiving the sixth observation, this second-order HMM has produced only one possible optimal state sequence; that is, $q_{1:6} = (q_1 = s_1, q_2 = s_2, q_3 = s_1, q_4 = s_2, q_5 = s_1, q_6 = s_2)$ with the total entropy of 0 which indicates that there is no uncertainty.

3. Entropy-Based Decoding Algorithm with a Reduction Approach

The extended entropy-based Viterbi algorithm in Section 2 has addressed only the issue related to memory space but this algorithm is not able to overcome the computational complexity. In this section, we introduce an efficient entropy-based algorithm that used reduction approach, namely, entropy-based order-transformation forward algorithm (EOTFA) to compute the optimal state sequence based on entropy of any generalized HHMM. This algorithm has addressed issues related to memory space and computational complexity.

3.1. Transforming a High-Order HMM with a Single Observational Sequence. This EOTFA algorithm involves a transformation of a generalized high-order HMM into an equivalent first-order HMM and an algorithm is developed based on the equivalent first-order HMM. This algorithm performs the computation based on the observational sequence and it requires $O(T\widetilde{N}^2)$ calculations, where $\widetilde{N}$ is the number of states in an equivalent first-order model and T is the length of observational sequence.

The transformation of a generalized high-order HMM into an equivalent first-order HMM is based on Hadar and Messer's method [7].

Suppose $\widetilde{Q}_t = (q_t, q_{t-1}, \ldots, q_{t-k+1})$ for $1 \le t \le T$; then the hidden state process $\{\widetilde{Q}_t\}_{t=1}^{T}$ of the kth-order Markov chain satisfies

$$\begin{aligned} &P\left(\widetilde{Q}_t \mid \left\{\widetilde{Q}_l\right\}_{l<t}\right) \\ &\quad = P\left(q_t, q_{t-1}, \ldots, q_{t-k+1} \mid q_{t-1}, q_{t-2}, \ldots, q_{2-k}\right) \\ &\quad = P\left(q_t \mid q_{t-1}, q_{t-2}, \ldots, q_{t-k}\right) \\ &\quad = P\left(q_t, q_{t-1}, \ldots, q_{t-k+1} \mid q_{t-1}, q_{t-2}, \ldots, q_{t-k}\right) \\ &\quad = P\left(\widetilde{Q}_t \mid \widetilde{Q}_{t-1}\right), \end{aligned} \tag{50}$$

where $\widetilde{Q}_t$ takes the value from the set of hidden states $\widetilde{S} = \{s_i, i = 1, 2, \ldots, N\}^k$. Hence, the hidden state process $\{\widetilde{Q}_t\}_{t=1}^{T}$ forms the first-order HMM Markov process.

The observation process $\{o_t\}_{t=1}^{T}$ satisfies

$$\begin{aligned} &P\left(o_t \mid \{o_l\}_{l<t}, \left\{\widetilde{Q}_l\right\}_{l\le t}\right) = P\left(o_t \mid \{o_l\}_{l\le t-1}, \{q_l\}_{l\le t}\right) \\ &\quad = P\left(o_t \mid \{q_l\}_{l\le t}\right) = P\left(o_t \mid \{q_l\}_{l=t-k}^{t}\right) \\ &\quad = P\left(o_t \mid \widetilde{Q}_t\right). \end{aligned} \tag{51}$$

Hence, the hidden state process $\{\widetilde{Q}_t\}_{t=1}^{T}$ and the observation process $\{o_t\}_{t=1}^{T}$ form the first-order HMM.

Remarks 9. (i)

$$\begin{aligned} P\left(\widetilde{Q}_t \mid \widetilde{Q}_{t-1}\right) &= P\left(\widetilde{Q}_t\right. \\ &= \left[q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}\right] \mid \widetilde{Q}_{t-1} \\ &= \left.\left[q_{t-k} = s_{i_1}, q_{t-k+1} = s_{i_2}, \ldots, q_{t-1} = s_{i_k}\right]\right) = P\left(\widetilde{Q}_t\right. \\ &= \left[s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}\right] \mid \widetilde{Q}_{t-1} = \left.\left[s_{i_1}, s_{i_2}, \ldots, s_{i_k}\right]\right) \\ &= P\left(\widetilde{Q}_t = s_{i_2 i_3 \cdots i_{k+1}} \mid \widetilde{Q}_{t-1} = s_{i_1 i_2 \cdots i_k}\right), \end{aligned} \quad (52)$$

where $[s_{i_1}, s_{i_2}, \ldots, s_{i_k}]$ and $[s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}] \in \widetilde{S}$.

(ii)

$$\begin{aligned} P\left(o_t \mid \widetilde{Q}_t\right) &= P\left(o_t \mid \widetilde{Q}_t\right. \\ &= \left.\left[q_{t-k+1} = s_{i_2}, q_{t-k+2} = s_{i_3}, \ldots, q_t = s_{i_{k+1}}\right]\right) \\ &= P\left(o_t \mid \widetilde{Q}_t = \left[s_{i_2}, \ldots, s_{i_{k+1}}\right]\right) = P\left(o_t \mid \widetilde{Q}_t\right. \\ &= \left. s_{i_2 i_3 \cdots i_{k+1}}\right), \end{aligned} \quad (53)$$

where $[s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}] \in \widetilde{S}$.

Note that we assume $s_{i_1}, s_{i_2}, \ldots, s_{i_k} = s_{i_1 i_2 \cdots i_k}$ and $s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}} = s_{i_2 i_3 \cdots i_{k+1}}$.

The elements for the transformation of a high-order into an equivalent first-order discrete HMM are as follows:

(i) Number of distinct hidden states, $\widetilde{N}$

(ii) Number of distinct observed symbols, M

(iii) Length of observational sequence, T

(iv) Observational sequence, $O = \{o_t,\ t = 1, 2, \ldots, T\}$

(v) Hidden state sequence, $\widetilde{Q} = \{\widetilde{Q}_t,\ t = 1, 2, \ldots, T\}$

(vi) Possible values for each state, $\widetilde{S} = \{s_i,\ i = 1, 2, \ldots, N\}^k$

(vii) Possible symbols per observation, $\widetilde{V} = \{v_w,\ w = 1, 2, \ldots, M\}$

(viii) Initial hidden state probability vector, $\widetilde{\pi} = \{\widetilde{\pi}_i\}$, and $\widetilde{\pi}_i$ is the probability that model will transit from state $\widetilde{s}_i = [s_{i_1}, s_{i_2}, \ldots, s_{i_k}] = s_{i_1 i_2 \cdots i_k}$, where

$$\widetilde{\pi}_i = P\left(\widetilde{Q}_1 = \widetilde{s}_i\right), \quad \sum_{i=1}^{\widetilde{N}} \widetilde{\pi}_i = 1, \quad \widetilde{\pi}_i \geq 0 \quad (54)$$

(ix) State transition probability matrix, $\widetilde{A} = \{\widetilde{a}_{ij}\}$ and a_{ij} is the probability of a transition from state $\widetilde{s}_i = [s_{i_1}, s_{i_2}, \ldots, s_{i_k}]$ at time $t-1$ to state $\widetilde{s}_j = [s_{i_2}, s_{i_3}, \ldots, s_{i_{k+1}}]$ at time t where

$$\widetilde{a}_{ij} = P\left(\widetilde{Q}_t = \widetilde{s}_j \mid \widetilde{Q}_{t-1} = \widetilde{s}_i\right), \quad \sum_{j=1}^{\widetilde{N}} \widetilde{a}_{ij} = 1, \quad \widetilde{a}_{ij} \geq 0, \quad (55)$$

where the first $k-1$ entries of $\widetilde{s}_i$ are equal to the last $k-1$ entries of $\widetilde{s}_j$

(x) Emission probability matrix, $\widetilde{B} = \{\widetilde{b}_i(v_m)\}$, and $\widetilde{b}_i(v_m)$ is a probability of observing v_m in state $\widetilde{s}_i = [s_{i_1}, s_{i_2}, \ldots, s_{i_k}]$ at time t:

$$\widetilde{b}_i\left(v_m\right) = P\left(o_t = v_m \mid \widetilde{Q}_t = \widetilde{s}_i\right), \quad \sum_{m=1}^{M} \widetilde{b}_i\left(v_m\right) = 1, \quad \widetilde{b}_i\left(v_m\right) \geq 0. \quad (56)$$

3.2. The Forward and Backward Probabilities Variables for the Transformed Model. In this subsection, we omit the derivations for forward and backward probability variables since the derivations are similar to the derivations in Section 2.2.

The forward recursion variable for the transformed model at time t is as follows:

$$\begin{aligned} \widetilde{\alpha}_t\left(j\right) &= P\left(o_1, o_2, \ldots, o_t, \widetilde{Q}_t = \widetilde{s}_j \mid \lambda\right) \\ &= P\left(o_1, o_2, \ldots, o_t, \widetilde{Q}_t = s_{i_2 i_3 \cdots i_k} \mid \lambda\right) \\ &= \sum_{i=1}^{\widetilde{N}} \widetilde{\alpha}_{t-1}\left(i\right) \widetilde{a}_{ij} \widetilde{b}_j\left(o_t\right). \end{aligned} \quad (57)$$

The backward recursion variable for the transformed model at time t is as follows:

$$\begin{aligned} \widetilde{\beta}_t\left(i\right) &= P\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid \widetilde{Q}_t = \widetilde{s}_i, \lambda\right) \\ &= P\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid \widetilde{Q}_t = s_{i_1 i_2 \cdots i_k}\right) \\ &= \left[\sum_{j=1}^{\widetilde{N}} \widetilde{\beta}_{t+1}\left(j\right) \widetilde{a}_{ij}\right] \widetilde{b}_j\left(o_{t+1}\right). \end{aligned} \quad (58)$$

The normalized forward variable at time t is as follows:

$$\widetilde{\alpha}_t^{\bullet}\left(j\right) = P\left(\widetilde{Q}_t \mid o_{1:t}\right) = \frac{\sum_{i=1}^{\widetilde{N}} \widetilde{\alpha}_{t-1}^{\bullet}\left(i\right) \widetilde{a}_{ij} \widetilde{b}_j\left(o_t\right)}{r_t^{\bullet}}, \quad (59)$$

where $r_t^{\bullet} = \sum_{j=1}^{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} \widetilde{\alpha}_{t-1}(i) \widetilde{a}_{ij} \widetilde{b}_j(o_t)$.

The normalized backward variables at time t is as follows:

$$\widetilde{\beta}_t^{\bullet}\left(i\right) = \frac{P\left(o_{t+1:T} \mid \widetilde{Q}_t\right)}{P\left(o_{t+1:T} \mid o_{o:t}\right)} = \frac{\sum_{j=1}^{\widetilde{N}} \widetilde{\beta}_{t+1}^{\bullet}\left(j\right) \widetilde{a}_{ij} \widetilde{b}_j\left(o_{t+1}\right)}{r_{t+1}^{\bullet}}, \quad (60)$$

where $r_{t+1}^{\bullet} = \sum_{j=1}^{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} \widetilde{\alpha}_t(i) \widetilde{a}_{ij} \widetilde{b}_j(o_{t+1})$.

3.3. The Computation of the Optimal State Sequence for a HHMM. For EOFTA algorithm, we require state entropy variable, $H_t(\tilde{s}_j)$, that can be computed recursively using the previous variable, $H_{t-1}(\tilde{s}_i)$.

We define the state entropy variable as follows.

Definition 10. The state entropy variable, $H_t(\tilde{s}_j)$, in an order-transformation HMM, is the entropy of all the paths that lead to state of $\tilde{s}_j$ at time t, given the observations $o_1, o_2, \ldots, o_t$. It can be denoted as

$$H_t\left(\tilde{s}_j\right) = H\left(\widetilde{Q}_{1:t-1} \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right). \tag{61}$$

From (61) at $t = 1$, we obtain the initial state entropy variable as

$$H_1\left(\tilde{s}_j\right) = 0. \tag{62}$$

From (61) and (62), we obtain the recursion on the entropy for $t = 2, \ldots, T-1$, and $1 \le i, j \le \widetilde{N}$

$$\begin{aligned} H_t\left(\tilde{s}_j\right) &= H\left(\widetilde{Q}_{1:t-1} \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &= H\left(\widetilde{Q}_{1:t-2}, \widetilde{Q}_{t-1} \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &= H\left(\widetilde{Q}_{t-1} \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &\quad + H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1}, \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right), \end{aligned} \tag{63}$$

where

$$\begin{aligned} &H\left(\widetilde{Q}_{t-1} \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &\quad = -\sum_{i=1}^{\widetilde{N}} \left[P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right)\right. \\ &\quad \left.\cdot \log_2\left(P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right)\right)\right], \\ &H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1}, \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &\quad = \sum_{i=1}^{\widetilde{N}} \left[P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right)\right. \\ &\quad \left.\cdot H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1} = \tilde{s}_i, \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right)\right] \\ &\quad = \sum_{i=1}^{\widetilde{N}} \left[P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) H_{t-1}\left(\tilde{s}_i\right)\right]. \end{aligned} \tag{64}$$

The auxiliary probability $P(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t})$ is required for our EOTFA algorithm. It can be computed as follows:

$$\begin{aligned} &P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) = P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_t, o_{1:t-1}\right) \\ &= \frac{P\left(\widetilde{Q}_t = \tilde{s}_j, o_t \mid \widetilde{Q}_{t-1} = \tilde{s}_i, o_{1:t-1}\right) P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid o_{1:t-1}\right)}{P\left(\widetilde{Q}_t = \tilde{s}_j, o_t \mid o_{1:t-1}\right)} \\ &= \frac{P\left(o_t \mid \widetilde{Q}_t = \tilde{s}_j\right) P\left(\widetilde{Q}_t = \tilde{s}_j \mid \widetilde{Q}_{t-1} = \tilde{s}_i\right) P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid o_{1:t-1}\right)}{P\left(o_t \mid \widetilde{Q}_t = \tilde{s}_j\right) P\left(\widetilde{Q}_t = \tilde{s}_j \mid o_{1:t-1}\right)} \\ &= \frac{P\left(\widetilde{Q}_t = \tilde{s}_j \mid q_{t-1} = \tilde{s}_i\right) P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid o_{1:t-1}\right)}{\sum_{k=1}^{\widetilde{N}} P\left(\widetilde{Q}_t = \tilde{s}_j \mid Q_{t-1} = \tilde{s}_k\right) P\left(\widetilde{Q}_{t-1} = \tilde{s}_k \mid o_{1:t-1}\right)} \\ &= \frac{\tilde{a}_{ij}\tilde{\alpha}^{\bullet}_{t-1}(i)}{\sum_{k=1}^{\widetilde{N}} \tilde{a}_{kj}\tilde{\alpha}^{\bullet}_{t-1}(k)}. \end{aligned} \tag{65}$$

For the final process, we compute $H(q_{1:T} \mid o_{1:T})$ which can be expanded as follows:

$$\begin{aligned} H\left(\widetilde{Q}_{1:T} \mid o_{1:T}\right) &= H\left(\widetilde{Q}_{1:T-1} \mid \widetilde{Q}_T = \tilde{s}_j, o_{1:T}\right) \\ &\quad + H\left(\widetilde{Q}_T \mid o_{1:T}\right) \\ &= \sum_{i=1}^{\widetilde{N}} H_T\left(\tilde{s}_i\right) \tilde{\alpha}^{\bullet}_T(i) \\ &\quad - \sum_{i=1}^{\widetilde{N}} \tilde{\alpha}^{\bullet}_T(i) \log_2\left(\tilde{\alpha}^{\bullet}_T(i)\right). \end{aligned} \tag{66}$$

The basic entropy concept in (40) and the following basic properties of HMM are used for proving our Lemma 11. According to the transformation of a high-order into an equivalent first-order HMM, state $\widetilde{Q}_{t-r}$, $r \ge 2$, and $\widetilde{Q}_t$ are statistically independent given $\widetilde{Q}_{t-1}$. The same applies to $\widetilde{Q}_{t-r}$, $r \ge 2$ and o_t are statistically independent given $\widetilde{Q}_{t-1}$.

The following proof is due to Hernando et al. [4].

Lemma 11. *For the transformation of a high-order into an equivalent first-order HMM, the entropy of the state sequence up to time $t-2$, given the states at time $t-1$ and the observations up to time $t-1$, is conditionally independent on the state and observation at time t*

$$H_{t-1}\left(\tilde{s}_i\right) = H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1} = \tilde{s}_i, \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right). \tag{67}$$

Proof.

$$\begin{aligned} &H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1} = \tilde{s}_i, \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ &\quad = H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1} = \tilde{s}_i, o_{1:t-1}, \widetilde{Q}_t = \tilde{s}_j, o_t\right) \\ &\quad = H\left(\widetilde{Q}_{1:t-2} \mid \widetilde{Q}_{t-1} = \tilde{s}_i, o_{1:t-1}\right) = H_{t-1}\left(\tilde{s}_i\right). \end{aligned} \tag{68}$$

Our EOTFA algorithm for computing the optimal state sequence is based on the normalized forward recursion variable, state entropy recursion variable, and auxiliary probability. From (59), (60), (61), (62), (63), and (66), we construct our EOTFA algorithm as follows.

(1) Initialization. For $t = 1$ and $1 \le j \le \widetilde{N}$,

$$\begin{aligned} H_1\left(\tilde{s}_j\right) &= 0, \\ \tilde{\alpha}^{\bullet}_1(j) &= \frac{\tilde{\pi}(j)\tilde{b}_j\left(o_1\right)}{\sum_{i=1}^{\widetilde{N}} \tilde{\pi}(i)\tilde{b}_i\left(o_1\right)}. \end{aligned} \tag{69}$$

(2) Recursion. For $t = 2, \ldots, T$ and $1 \le j \le \widetilde{N}$,

$$\tilde{\alpha}_t^{\bullet}(j) = \frac{\sum_{i=1}^{\widetilde{N}} \tilde{\alpha}_{t-1}^{\bullet}(i)\, \tilde{a}_{ij} \tilde{b}_j(o_t)}{\sum_{k=1}^{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} \tilde{\alpha}_{t-1}^{\bullet}(i)\, \tilde{a}_{ik} \tilde{b}_k(o_t)},$$

$$P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) = \frac{\tilde{a}_{ij} \tilde{\alpha}_{t-1}^{\bullet}(i)}{\sum_{k=1}^{\widetilde{N}} \tilde{a}_{kj} \tilde{\alpha}_{t-1}^{\bullet}(k)},$$

$$\begin{aligned} H_t\left(\tilde{s}_j\right) = \sum_{i=1}^{\widetilde{N}} H_{t-1}\left(\tilde{s}_i\right) P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ - \sum_{i=1}^{\widetilde{N}} P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right) \\ \cdot \log_2\left(P\left(\widetilde{Q}_{t-1} = \tilde{s}_i \mid \widetilde{Q}_t = \tilde{s}_j, o_{1:t}\right)\right)\Big]. \end{aligned} \quad (70)$$

(3) Termination

$$\begin{aligned} H\left(\widetilde{Q}_{1:T} \mid o_{1:T}\right) = \sum_{i=1}^{\widetilde{N}} H_T\left(\tilde{s}_i\right) \tilde{\alpha}_T^{\bullet}(i) \\ - \sum_{i=1}^{\widetilde{N}} \tilde{\alpha}_T^{\bullet}(i) \log_2\left(\tilde{\alpha}_T^{\bullet}(i)\right). \end{aligned} \quad (71)$$

The direct evaluation algorithm, Hernando et al.'s algorithm, and our algorithm perform the computation of state entropy exponentially with respect to the order of HMM. Our algorithm proposes the transformation of a generalized high-order into an equivalent first-order HMM and then compute the state entropy based on the equivalent first-order model; hence our algorithm is the most efficient in which it requires $O(T\widetilde{N}^2)$ calculations as compared to the direct evaluation method which requires $O(N^{T+k-1})$ calculations and the extended algorithm which requires $O(TN^{k+1})$ calculations where N is the number of states in a model, $\widetilde{N}$ is the number of states in an equivalent first-order model, T is the length of observational sequence, and k is the order of HMM. □

3.4. Numerical Illustration for an Equivalent First-Order HMM. We consider the second-order HMM in Section 2.5 for illustrating our EOTFA algorithm in computing the optimal state sequence. According to our proposed novel algorithm, we first transformed the second-order HMM in Section 2.5 into the equivalent first-order HMM by using Hadar and Messer method [7]. The equivalent first-order HMM has the following model parameters $\tilde{\lambda} = (\tilde{\pi}, \widetilde{A}, \widetilde{B})$, where $\tilde{\pi}$ is the initial state probability vector, $\widetilde{A}$ is the state transition probability matrix, and $\widetilde{B}$ is the emission probability matrix.

$$\begin{aligned} \tilde{\pi} &= \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \end{bmatrix}, \\ \widetilde{A} &= \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\ \widetilde{B} &= \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 1 & 0 \end{bmatrix}. \end{aligned} \quad (72)$$

Note that the above state transition probability and the emission probability matrices whose components are indicated as $\tilde{a}_{i_1 i_2}$ and $\tilde{b}_{i_2}(o_t = v_m)$ where $1 \le i_1, i_2 \le 4$ and $1 \le m \le 3$ can be obtained from the graphical diagram in Figure 4.

The state space for the equivalent first-order HMM is $\widetilde{S} = \{\tilde{s}_1, \tilde{s}_2, \tilde{s}_3, \tilde{s}_4\}$, where $\tilde{s}_1 = [s_1, s_1]$, $\tilde{s}_2 = [s_1, s_2]$, $\tilde{s}_3 = [s_2, s_1]$, and $\tilde{s}_4 = [s_2, s_2]$, and the possible symbols per observation are $O = \{v_1, v_2, v_3\}$. Note that $\tilde{\pi}_1 = \{\tilde{\pi}_{i_2}\}$, where $\tilde{\pi}_{i_2} = P(\widetilde{Q}_t = \tilde{s}_{i_2})$, $\widetilde{A} = \{\tilde{a}_{i_1 i_2}\}$, where $\tilde{a}_{i_1 i_2} = P(\widetilde{Q}_t = \tilde{s}_{i_2} \mid \widetilde{Q}_{t-1} = \tilde{s}_{i_1})$, and $\widetilde{B} = \{\tilde{b}_{i_2}(o_t = v_m)\}$, where $\tilde{b}_{i_2}(o_t = v_m) = P(o_t = v_m \mid \widetilde{Q}_t = \tilde{s}_{i_2})$.

The equivalent first-order HMM was developed based on Hadar and Messer's method [7] is shown in Figure 4.

Secondly, the optimal state sequence is computed based on the equivalent first-order HMM by using our proposed algorithm. Finally, the optimal state sequence of the second-order HMM is inferred from the optimal state sequence from the equivalent first-order HMM.

The following is the observational sequence used for illustrating our algorithm:

$$o_{1:6} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2, o_5 = v_3, o_6 = v_1). \quad (73)$$

We applied our EOFTA algorithm for computing the optimal state sequence based on the state entropy. The computed value of state entropy is shown in Figure 5.

The total entropy after each time step for the transformed model, that is, the second-order transformed into the equivalent first-order HMM is displayed at the bottom of Figure 5. For example, this model has produced only one possible state sequence; that is, $\widetilde{Q}_{1:5} = (\widetilde{Q}_1 = \tilde{s}_1, \widetilde{Q}_2 = \tilde{s}_2, \widetilde{Q}_3 = \tilde{s}_3, \widetilde{Q}_4 = \tilde{s}_2, \widetilde{Q}_5 = \tilde{s}_3)$, as shown by the bold arrow with a probability of 1 after receiving the fifth observation. The total entropy is 0 at $t = 5$ which indicates that there is no uncertainty. After receiving the sixth observation, that is, $o_{1:6} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2, o_5 = v_3, o_6 = v_1)$, this equivalent first-order HMM has produced one possible optimal state sequence $\widetilde{Q}_{1:6} = (\widetilde{Q}_1 = \tilde{s}_1, \widetilde{Q}_2 = \tilde{s}_2, \widetilde{Q}_3 = \tilde{s}_3, \widetilde{Q}_4 = \tilde{s}_2, \widetilde{Q}_5 = \tilde{s}_3, \widetilde{Q}_6 = \tilde{s}_2)$ which is similar to $q_{1:6} = (q_1 = s_1, q_2 = s_2, q_3 = s_1, q_4 = s_2, q_5 = s_1, q_6 = s_2)$ that is produced by the second-order HMM in Section 2.5 with a total entropy of 0 which indicates that there is no uncertainty. As a result, the optimal state sequence of the high-order HMM is inferred from the optimal state sequence of the equivalent first-order HMM.

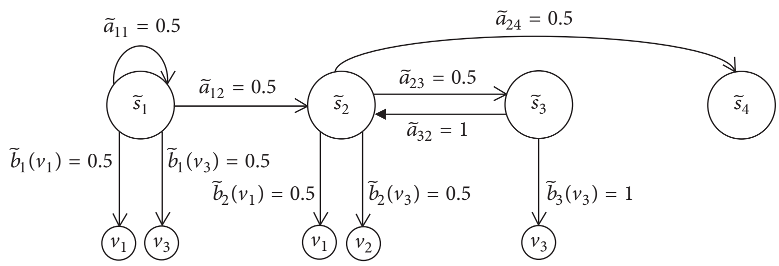

FIGURE 4: The graphical diagram shows an equivalent first-order HMM.

state \ obs	$o_1 = v_1$	$o_2 = v_1$	$o_3 = v_3$	$o_4 = v_2$	$o_5 = v_3$	$o_6 = v_1$
$\tilde{s}_1$	$H_1(1) = 0$ $\hat{\alpha}^{\bullet}_1(1) = 0.5$	$H_2(1) = 0$ $\hat{\alpha}^{\bullet}_2(1) = 0.5$	$H_3(1) = 0.5$ $\hat{\alpha}^{\bullet}_3(1) = 0.33$	$H_4(1) = 0$ $\hat{\alpha}^{\bullet}_4(1) = 0$	$H_5(1) = 0$ $\hat{\alpha}^{\bullet}_5(1) = 0$	$H_6(1) = 0$ $\hat{\alpha}^{\bullet}_6(1) = 0$
$\tilde{s}_2$	$H_1(2) = 0$ $\hat{\alpha}^{\bullet}_1(2) = 0.5$	$H_2(2) = 0$ $\hat{\alpha}^{\bullet}_2(2) = 0.5$	$H_3(2) = 0$ $\hat{\alpha}^{\bullet}_3(2) = 0$	$H_4(2) = 0$ $\hat{\alpha}^{\bullet}_4(2) = 1$	$H_5(2) = 0$ $\hat{\alpha}^{\bullet}_5(2) = 0$	$H_6(2) = 0$ $\hat{\alpha}^{\bullet}_6(2) = 1$
$\tilde{s}_3$	$H_1(3) = 0$ $\hat{\alpha}^{\bullet}_1(3) = 0$	$H_2(3) = 0$ $\hat{\alpha}^{\bullet}_2(3) = 0$	$H_3(3) = 0.5$ $\hat{\alpha}^{\bullet}_3(3) = 0.67$	$H_4(3) = 0$ $\hat{\alpha}^{\bullet}_4(3) = 0$	$H_5(3) = 0$ $\hat{\alpha}^{\bullet}_5(3) = 1$	$H_6(3) = 0$ $\hat{\alpha}^{\bullet}_6(3) = 0$
$\tilde{s}_4$	$H_1(4) = 0$ $\hat{\alpha}^{\bullet}_1(4) = 0$	$H_2(4) = 0$ $\hat{\alpha}^{\bullet}_2(4) = 0$	$H_3(4) = 0$ $\hat{\alpha}^{\bullet}_3(4) = 0$	$H_4(4) = 0$ $\hat{\alpha}^{\bullet}_4(4) = 0$	$H_5(4) = 0$ $\hat{\alpha}^{\bullet}_5(4) = 0$	$H_6(4) = 0$ $\hat{\alpha}^{\bullet}_6(4) = 0$
Total entropy	1	1	1.41	0	0	0

FIGURE 5: The evolution of the trellis structure for a transformation of a second-order into an equivalent first-order HMM with the observation sequence $o_{1:6} = (o_1 = v_1, o_2 = v_1, o_3 = v_3, o_4 = v_2, o_5 = v_3, o_6 = v_1)$.

Our proposed algorithm is based on the equivalent first-order HMM and only requires $O(T\widetilde{N}^2)$ computation and hence we can conclude that our EOTFA algorithm is more efficient.

4. Conclusion and Future Work

We have introduced a novel algorithm for computing the optimal state sequence for HHMM that requires $O(T\widetilde{N}^2)$ calculations and $O(\widetilde{N}^2)$ memory space where $\widetilde{N}$ is the number of states in an equivalent first-order HMM and T is the length of observational sequence. This algorithm is to be running with Viterbi algorithm in tracking the optimal state sequence as well as the entropy of the distribution of the state sequence. We have developed this algorithm for the case of a generalized discrete high-order HMM. This research can be also extended for continuous high-order HMMs and these models are widely used in speech recognition.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] V. Ciriza, L. Donini, J. Durand, and S. Girard, "Optimal time-outs for power management under renewal or hidden Markov processes for requests," Tech. Rep., 2011, http://hal.inria.fr/hal-00412509/en.

[2] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] J. G. Proakis and M. Salehi, *Communications System Engineering*, Prentice-Hall, Upper Saddler River, NJ, USA, 2002.

[4] D. Hernando, V. Crespi, and G. Cybenko, "Efficient computation of the hidden Markov model entropy for a given observation sequence," *Institute of Electrical and Electronics*

Engineers Transactions on Information Theory, vol. 51, no. 7, pp. 2681–2685, 2005.

[5] G. S. Mann and A. McCallum, "Efficient computation of entropy gradient for semi-supervised conditional random fields," in *Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '07)*, pp. 109–112, Association for Computational Linguistics, Morristown, NJ, USA, April 2007.

[6] V. M. Ilic, "Entropy Semiring Forward-backward Algorithm for HMM Entropy Computation," 2011, https://arxiv.org/abs/1108.0347.

[7] U. Hadar and H. Messer, "High-order hidden Markov models—estimation and implementation," in *Proceedings of the 15th IEEE/SP Workshop on Statistical Signal Processing (SSP '09)*, pp. 249–252, Wales, UK, September 2009.

[8] J. A. du Preez, "Efficient training of high-order hidden Markov models using first-order representations," *Computer Speech and Language*, vol. 12, no. 1, pp. 23–39, 1998.

[9] L. M. Lee and J. C. Lee, "A study on high-order hidden Markov Models and applications to speech recognition," in *Proceedings of the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 682–690, 2006.

[10] R. M. Altman, "Mixed hidden MARkov models: an extension of the hidden MARkov model to the longitudinal data setting," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 201–210, 2007.

[11] A. Spagnoli, R. Henderson, R. J. Boys, and J. J. Houwing-Duistermaat, "A hidden Markov model for informative dropout in longitudinal response data with crisis states," *Statistics & Probability Letters*, vol. 81, no. 7, pp. 730–738, 2011.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications and Signal Processing, John Wiley & Sons, New York, NY, USA, 2006.