An Efficient Forward–Backward Algorithm for an Explicit-Duration Hidden Markov Model

Shun-Zheng Yu and Hisashi Kobayashi, *Fellow, IEEE*

***Abstract*—Existing algorithms for estimating the model parameters of an explicit-duration hidden Markov model (HMM) usually require computations as large as $O((MD^2 + M^2)T)$ or $O(M^2DT)$, where M is the number of states; D is the maximum possible interval between state transitions; and T is the period of observations used to estimate the model parameters. Because of such computational requirements, these algorithms are not practical when we wish to construct an HMM model with large state space and large explicit state duration and process a large amount of measurement data to obtain high accuracy. We propose a new forward–backward algorithm whose computational complexity is only $O((MD+M^2)T)$, a reduction by almost a factor of D when $D > M$ and whose memory requirement is $O(MT)$. As an application example, we discuss an HMM characterization of access traffic observed at a large-scale Web site: we formulate the Web access pattern in terms of an HMM with explicit duration and estimate the model parameters using our algorithm.**

***Index Terms*—Explicit-duration HMM, hidden Markov model (HMM), hidden semi-Markov model, traffic characterization, variable-duration HMM.**

I. Introduction

THE HIDDEN Markov model (HMM) has been successfully applied to a number of scientific and engineering problems [4]. Most studies found in the literature, however, implicitly assume that the duration of any system state is constant (i.e., a unit time in a discrete-time model) or exponentially (i.e., geometrically) distributed. This simplifying assumption is made because efficient computation algorithms have been well developed to deal with such HMMs. There are a few studies that discuss more general situations, where the duration of any state is explicitly assumed to be nonexponential. Such an HMM is often referred to as an explicit-duration HMM, HMM with variable duration, HMM with explicit duration, and hidden semi-Markov model [1]–[3].

To the best of our knowledge, Ferguson [1] (see also [4]) is the first to investigate estimation algorithms for the explicit-duration HMM. However, Ferguson's algorithm is computationally too expensive to be of practical use in many applications, since its computational complexity is $O((MD^2 + M^2)T)$. The product term, i.e., the joint probability distribution of a sequence of observations, required in Ferguson's algorithm can be calculated more efficiently by a recursive method as suggested in [2] and further refined in [3]. In [2] and [3], the product terms require only $O(D)$ computations; but at every (discrete) time t, the previous D observations must be retrieved, and the recursive steps must be performed. Therefore, the total number of recursive steps required in [2] and [3] increases by a factor of D compared with Ferguson's algorithm.

An alternative approach is to transform a given explicit-duration HMM into an equivalent "super HMM" [1], [5], [6]. In this case, however, the number of superstates becomes MD, i.e., D times as large as the original state size M. The number of nonzero entries of the transition probability matrix for the superstates largely determines the computational complexity, which turns out to be $O(M^2DT)$. Therefore, these refined algorithms are still too computationally intensive in some applications.

In Section II, we discuss our new forward–backward algorithm, and we show that its computational complexity is $O((MD+M^2)T)$ and that its memory requirement is $O(MT)$. In Section III, we discuss an application of the new algorithm to Web workload characterization. We use actual Web access data obtained at a large-scale Web site and represent the Web access traffic pattern as a probabilistic function of an underlying explicit-duration HMM and then estimate the model parameters from the data.

II. New Forward–Backward Algorithm

Consider a semi-Markov chain of M states, denoted $s_1, s_2, \ldots, s_M$, with the probability of transition from state s_m to state s_n being denoted $a_{mn}(m, n = 1, 2, \ldots, M)$. The initial state probability distribution is given by $\{\pi_m\}$. Let q_t denote the state of the semi-Markov chain at time t, where $t = 1, 2, \ldots, T$ and let o_t stand for the observable output at t. The observable and the state are related through the conditional probability distribution $b_m(v_k) = \Pr[o_t = v_k \mid q_t = s_m]$, where $\{v_k\}$ is a set of K distinct values that may be assumed by the observation o_t. In the sequel, however, we shall often write $b_m(o_t)$ instead to simplify the notation. We assume the "conditional independence" of outputs given the state in the sense that $\Pr[\mathbf{o}_a^b \mid s_m] = \prod_{t=a}^b b_m(o_t)$, where $\mathbf{o}_a^b = \{o_t; a \le t \le b\}$. We also assume that the duration of a given state is a discrete random variable, taking value d with probability $p_m(d)$, where $d \in \{1, 2, \ldots, D\}$. The integer D is the maximum duration possible in any state, or equivalently, the maximum interval between any two consecutive state transitions.

Let τ_t denote the remaining (or residual) time of the current state q_t. Then, if the pair process (q_t, τ_t) takes on value (s_m, d),

Manuscript received January 11, 2002; revised July 6, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Geert Leus.

S.-Z. Yu is with the Department of Electrical and Communication Engineering, Zhongshan University, Guangzhou 510275, China (e-mail: syu@zsu.edu.cn).

H. Kobayashi is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: hisashi@ee.princeton.edu).

Digital Object Identifier 10.1109/LSP.2002.806705

the semi-Markov chain will remain in the current state s_m until time $t+d-1$ and transits to another state at time $t+d$, where $d \geq 1$. For brevity of notation, let λ stand for the complete set of model parameters: $\lambda = (\{a_{mn}\}, \{\pi_m\}, \{b_m(v_k)\}, \{p_m(d)\})$. We initially estimate these model parameters, but after we collect observations $\{o_t\}$, we reestimate the parameters and update the model accordingly. This process is referred to as *parameter reestimation*. Thus, we first evaluate the various probabilities *conditioned on* λ. For brevity of notation, however, we drop this conditioning on λ.

We define the *forward variable* by

$$\alpha_t(m,d) \overset{\text{def}}{=} \Pr\left[\mathbf{o}_1^t, (q_t, \tau_t) = (s_m, d)\right]. \tag{1}$$

A transition into state $(q_t, \tau_t) = (s_m, d)$ takes place either from $(q_{t-1}, \tau_{t-1}) = (s_m, d+1)$ or from $(q_{t-1}, \tau_{t-1}) = (s_n, 1)$ for some $n \neq m$. Therefore, we readily obtain the following forward recursion formula:

$$\begin{aligned} \alpha_t(m,d) &= \alpha_{t-1}(m,d+1) b_m(o_t) \\ &+ \left(\sum_{n \neq m} \alpha_{t-1}(n,1) a_{nm}\right) \cdot b_m(o_t) p_m(d), \qquad d \geq 1 \end{aligned} \tag{2}$$

for a given state s_m and time $t > 1$, with the initial condition

$$\alpha_1(m,d) = \pi_m b_m(o_1) p_m(d). \tag{3}$$

We define the *backward variable* by

$$\beta_t(m,d) \overset{\text{def}}{=} \Pr\left[\mathbf{o}_{t+1}^T \,|\, (q_t, \tau_t) = (s_m, d)\right]. \tag{4}$$

By examining the possible states that follow $(q_t, \tau_t) = (s_m, d)$, we see that when $d > 1$ the next state must be $(q_{t+1}, \tau_{t+1}) = (s_m, d-1)$, and when $d = 1$ it must be $(q_{t+1}, \tau_{t+1}) = (s_n, d')$ for some $n \neq m$ and $d' \geq 1$. We thus have the following backward recursion formula:

$$\beta_t(m,d) = b_m(o_{t+1}) \beta_{t+1}(m, d-1), \qquad \text{for } d > 1 \tag{5}$$

and

$$\beta_t(m,1) = \sum_{n \neq m} a_{mn} b_n(o_{t+1}) \left(\sum_{d \geq 1} p_n(d) \beta_{t+1}(n,d)\right) \tag{6}$$

for a given states s_m and time $t < T$, with the initial condition (in the backward recursive steps)

$$\beta_T(m,d) = 1, \qquad d \geq 1. \tag{7}$$

To evaluate the forward variable $\alpha_t(m,d)$ of (2), we first compute the sum $\sum_{n \neq m} \alpha_{t-1}(n,1) a_{nm}$ for all m (in M^2 steps) and then $\alpha_t(m,d)$ for all m and d (in $O(MD)$ steps). Therefore, updating the forward variables at every t requires $O(MD+M^2)$ steps. Similarly, in the backward formulae (5) and (6), we first compute $\sum_{d \geq 1} p_n(d) \beta_{t+1}(n,d)$ for all n (in $O(MD)$ steps) and then $\beta_t(m,1)$ for all m (in $O(M^2)$ steps). Therefore, evaluation of $\beta_t(m,d)$ at each t also requires $O(MD+M^2)$ steps. Hence, the total number of computation steps for evaluating the forward and backward variables is $O((MD+M^2)T)$, where T is the total number of observations.

III. State Estimation and Parameter Reestimation

Now we discuss how to estimate the state q_t from the observation sequence $\mathbf{o}_1^T$ and reestimate the model parameters λ. First, we find that the joint probability of observing $\mathbf{o}_1^T$ and a transition from s_m to another state s_n $(n \neq m)$ at time t given λ can be expressed in terms of the assumed model parameters and the forward and backward variables defined above

$$\begin{aligned} \zeta_t(m,n) &\overset{\text{def}}{=} \Pr\left[\mathbf{o}_1^T, q_{t-1} = s_m, q_t = s_n\right] \\ &= \alpha_{t-1}(m,1) a_{mn} b_n(o_t) \cdot \left(\sum_{d \geq 1} p_n(d) \beta_t(n,d)\right) \end{aligned} \tag{8}$$

for $m \neq n$. Next we find that the joint probability of observing $\mathbf{o}_1^T$ and a transition to state s_m at time t and remaining in state s_m for d time units can also be expressed in terms of the variables and parameters defined earlier

$$\begin{aligned} \eta_t(m,d) &\overset{\text{def}}{=} \Pr\left[\mathbf{o}_1^T, q_{t-1} \neq s_m, q_t = s_m, \tau_t = d\right] \\ &= \left(\sum_{n \neq m} \alpha_{t-1}(n,1) a_{nm}\right) b_m(o_t) p_m(d) \beta_t(m,d). \end{aligned} \tag{9}$$

In order to estimate the state q_t from the observation sequence $\mathbf{o}_1^T$, let us consider the joint probability of $\mathbf{o}_1^T$ and $q_t = s_m$

$$\gamma_t(m) \overset{\text{def}}{=} \Pr\left[\mathbf{o}_1^T, q_t = s_m\right]. \tag{10}$$

Then using the following identity

$$\begin{aligned} &\Pr[\mathbf{o}_1^T, q_t = s_m, q_{t+1} = s_m] \\ &\quad = \Pr[\mathbf{o}_1^T, q_t = s_m] - \Pr[\mathbf{o}_1^T, q_t = s_m, q_{t+1} \neq s_m] \\ &\quad = \Pr[\mathbf{o}_1^T, q_{t+1} = s_m] \\ &\qquad - \Pr[\mathbf{o}_1^T, q_t \neq s_m, q_{t+1} = s_m] \end{aligned} \tag{11}$$

and the definitions of (8) and (10), we obtain the following backward recursion formula for $\gamma_t(m)$:

$$\gamma_t(m) = \gamma_{t+1}(m) + \sum_{n \neq m} \left(\zeta_{t+1}(m,n) - \zeta_{t+1}(n,m)\right) \tag{12}$$

with the initial condition

$$\gamma_T(m) = \sum_{d \geq 1} \alpha_T(m,d). \tag{13}$$

Now we can readily derive various estimation and reestimation formulae of our interest.

- The maximum *a posteriori* (MAP) estimate of state q_t is

$$\begin{aligned} \hat{q}_t &\overset{\text{def}}{=} \arg \max_{1 \leq m \leq M} \Pr\left[q_t = s_m | \mathbf{o}_1^T\right] \\ &= \arg \max_{1 \leq m \leq M} \gamma_t(m), \qquad \text{for } t = T, T-1, \ldots, 1. \end{aligned} \tag{14}$$

- The maximum-likelihood reestimate of the initial state probability π_m is $\hat{\pi}_m = \gamma_1(m)/G_1$, where G_1 is the normalization constant obtained by summing $\gamma_1(m)$ over $m = 1, \ldots, M$.
- The maximum-likelihood reestimate of the transition probability a_{mn} is $\hat{a}_{mn} = \sum_{t=1}^T \zeta_t(m,n)/G(m)$, for

$n \neq m$, where $G(m)$ is the normalization constant. Note that in the explicit-duration HMM formulation, a transition from a state back to itself cannot occur. Thus, $a_{mm} = 0$ and $\hat{a}_{mm} = 0$ for all m.

- The maximum-likelihood reestimate of the state duration probability $p_m(d)$ is $\hat{p}_m(d) = \sum_{t=1}^{T} \eta_t(m,d)/H(m)$ where $H(m) = \sum_{d=1}^{D} \sum_{t=1}^{T} \eta_t(m,d)$.
- The reestimate of the conditional probability distribution of observing v_k under a given state is $\hat{b}_m(v_k) = \sum_{t=1}^{T} \gamma_t(m)\delta(o_t - v_k)/V(m)$, where $\{v_k\}$ is the set of values that an observation can take on, and $\delta(o_t - v_k) = 1$, if $o_t = v_k$ and 0 if $o_t \neq v_k$. Here, $V(m)$ is the normalization constant, i.e., $V(m) = \sum_k \sum_{t=1}^{T} \gamma_t(m)\delta(o_t - v_k)$.

In the state estimation and model parameter reestimation formulae given above, we find that the estimation algorithm can be combined with the backward algorithm. Therefore, the backward variables $\beta_t(m,d)$ and the probabilities $\zeta_t(m,n)$, $\eta_t(m,d)$, and $\gamma_t(m)$ do not have to be stored for later use. Among the forward variables, only $\alpha_t(m,1)$ and $\sum_{n \neq m} \alpha_{t-1}(n,1)a_{nm}$ (for all m and t) need to be stored, since they are used in (8) and (9), respectively. Hence, the storage requirement is $O(MT)$.

Because the sums $\sum_{n \neq m} \alpha_{t-1}(n,1)a_{nm}$ (for all m) and $\sum_{d \geq 1} p_n(d)\beta_{t+1}(n,d)$ (for all n) are obtained during the computation of the forward–backward variables, the number of computation steps required for evaluating $\zeta_t(m,n)$ of (8) and $\eta_t(m,d)$ of (9) is linearly proportional to the number of parameters. Evaluation of $\gamma_t(m)$ in (10) requires M additions, but no multiplications. Hence, the computational complexity of the reestimation algorithm is $O(|\lambda|T)$, where $|\lambda| = M^2 + M + MK + MD$ is the total number of model parameters. Recall that the integer K stands for the number of distinct values that an observation o_t can take on, i.e., the cardinality of the set $\{v_k\}$.

IV. Application to Web Workload Characterization

In this section, we apply the explicit-duration HMM formulation to characterize the user request patterns to a Web server. We use the example discussed in [10] to illustrate this application. The empirical data we use were extracted from actual traffic data collected at a large-scale Web site.

Measurements of real workload often indicate that a significant amount of variability is present in the traffic observed over a wide range of time scales, exhibiting self-similar or long-range-dependent characteristics [7]. Such characteristics can have a significant impact on the performance of networks and systems [8], [9]. Therefore, better understanding of the nature of the Web workload is critical to the proper design and implementation of Web servers. A major advantage of using an explicit-duration HMM is its efficiency in estimating the model parameters to account for an observed sequence. Furthermore, the estimated parameters can capture various statistical properties of the workload, including long-range and short-range dependence. Thus, the model can be applied to generate synthetic workload patterns to be used for system performance evaluation and capacity planning [10]. They can also be used together with, for example, matrix-analytic methods to obtain analytically tractable solutions to queueing-theoretic models of Web server performance.

TABLE I
PARAMETERS OVER HOURS OF THE WORKLOAD

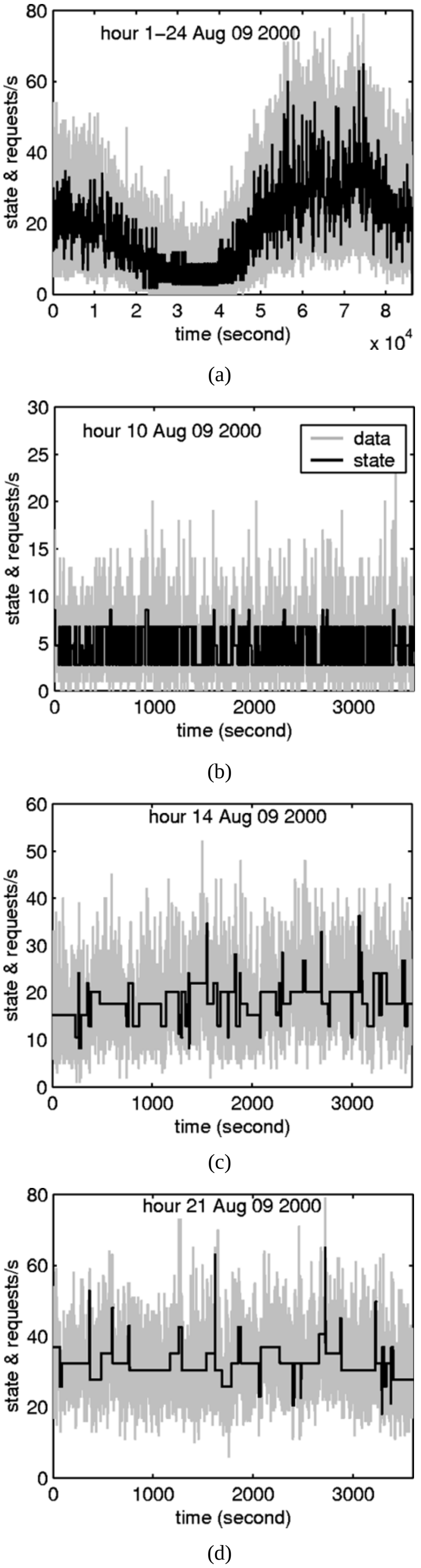
Hour	1	**10**	**11**	14	15	**20**	**21**	24
		valley		rising		peak		
Mean arrvls/s	21.2	**4.6**	**4.8**	18.3	22.6	**30.2**	**32.2**	21.2
Maximum arrvls/s	51	**27**	**25**	52	55	**71**	**74**	56
No. of states entered	16	**4**	**4**	15	17	**19**	**20**	13
Maximum duration	344	**65**	**143**	278	318	**388**	**405**	327

In the Web application example, the observation sequence $\{o_t\}$ represents the number of user requests arriving at the Web site in the tth second, and the maximum observed value was $\max\{v_k\} = 74$. The total number of observations is $T = 86\,400$, where the time unit is 1 s. Hence, the observation period is over one day. We characterize the request arrivals as a discrete-time random process modulated by an underlying (hidden state) semi-Markov process.

We consider the case in which the probability distributions of the model parameters are assumed to be general. Based on empirical analysis of the workload data, we find that a rule of thumb for an appropriate choice of M, the number of Markov states, should be approximately $0.8E[o_t]$. We chose the value $M = 30$ in this study. Similarly, we set D to be sufficiently large to cover the maximum duration of any state. In this study, we used $D = 500$, which is close to the average HTTP session length from our measurement data.

Given these model assumptions, we now apply our forward–backward algorithm to estimate the model parameters λ, where the initial values of λ are simply assumed to be uniformly distributed or randomly selected. In our experiments, the final results are found to be robust in terms of their convergence to the estimated value of model parameters: the final value is independent of the initially selected values and is reached in about 20 iterations. We summarize the results in Table I. From this table, we can see that during the "peak" hours (i.e., Hours 20 and 21) as many as 19 or 20 hidden states are actually used to modulate the rate of the Web access traffic, and the remaining 10 to 11 states are never visited. During the "valley" hours (i.e., Hours 10 and 11) only four hidden states are visited. The maximum duration D goes up to 405 (seconds) during the peak time (Hour 21), whereas during the slack time (Hour 10) it only reaches 65 (seconds). This suggests that the state duration distributions in the peak period and those in the slack period may be significantly different in their tails.

Fig. 1 plots the number of requests per second together with the estimated hidden states for 1) the whole day, 2) the valley hour (Hour 10), 3) the sharpest rising hour (Hour 14), and 4) the peak hour (Hour 21). From Fig. 1(a), we can see that the hidden states modulate the arrival rates over the day. When the arrival rate is low, as shown in Fig. 1(b), there are frequent transitions among the hidden states. In this case, the model is essentially a two-state Markov chain with transitions between states 1 and 3. When the arrival rate has a deterministic trend, as shown in Fig. 1(c), the transition probabilities will have a corresponding trend. For example, a given state exhibits more frequent transitions to states with higher indices than to those states with lower



(a)

(b)

(c)

(d)

Fig. 1. Observed data (i.e., number of requests per second) and the hidden states of the modulate. (a) Whole day. (b) Valley hour. (c) Sharpest rising hour. (d) Peak hour.

indices, while the deterministic trend is increasing. When the arrival rate reaches its peak, as shown in Fig. 1(d), the process will stay in the same state for a long period of time with a mean duration of 87.8 s and only 41 state transitions occurring during the period of 3600 s.

V. Conclusion

Our forward–backward algorithm for an explicit-duration HMM requires only $O((MD + M^2)T)$ computations and $O(MT)$ memory capacity to evaluate all the forward and backward variables, where M is the number of Markov states; D is the maximum duration between successive state transitions; and T is the period of the observation data. Once the forward and backward variables are obtained, various performance measures of interest such as the MAP estimate of a state sequence and the maximum-likelihood estimate of the state probability distribution can be obtained with little computation and memory cost.

Reestimation of the set of model parameters λ from the observation sequence $\mathbf{o}_1^T$ can be done with additional computations of only $O((M^2 + M + MD + MK)T)$. The algorithm has been successfully applied to estimate the parameters of a hidden semi-Markov model that is designed to characterize the access traffic pattern observed at a large-scale Web site. Such a model requires a large number of states M, a long maximum state duration D, and a long observation period T for an accurate characterization, and a practical solution is made possible by our efficient algorithm.

Acknowledgment

The first author would like to thank Z. Liu, M. Sequillante, C. Xia, and L. Zhang (IBM Research Center, Yorktown Heights, NY) for the opportunity to work with them on the Web traffic analysis reported in [10], a part of which is reproduced here as an illustrative example. The authors also thank B. L. Mark (George Mason University, Fairfax, VA) for his careful reading and valuable suggestions to improve this manuscript. This work was performed while the author was with Princeton University.

References

[1] J. D. Ferguson, "Variable duration models for speech," in *Symp. Application of Hidden Markov Models to Text and Speech*, Oct. 1980, pp. 143–179.

[2] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, vol. 1, no. 1, pp. 29–45, 1986.

[3] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMM's," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 213–217, May 1995.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[5] V. Krishnamurthy, J. B. Moore, and S. H. Chung, "Hidden fractal model signal processing," *Signal Processing*, vol. 24, no. 2, pp. 177–192, Aug. 1991.

[6] P. Ramesh and J. G. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition," in *Proc. ICASSP*, 1992, pp. 381–384.

[7] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.

[8] T. Tuan and K. Park, "Multiple time scale congestion control for self-similar network traffic," in *Perf. Eval.*, 1999, vol. 36, pp. 359–386.

[9] K. Park, G. T. Kim, and M. E. Crovella, "On the effect of traffic self-similarity on network performance," in *Proc. SPIE Int. Conf. Performance and Control of Network Systems*, Nov. 1997, pp. 296–310.

[10] S. Z. Yu, Z. Liu, M. Squillante, C. Xia, and L. Zhang, "A hidden semi-Markov model for web workload self-similarity," in *Proc. IPCCC*, Phoenix, AZ, Apr. 3–5, 2002, pp. 65–72.