Asymptotic Smoothing Errors for Hidden Markov Models

Louis Shue, Brian D. O. Anderson, *Fellow, IEEE*, and Franky De Bruyne

***Abstract*—In this paper, the asymptotic smoothing error for hidden Markov models (HMMs) is investigated using hypothesis testing ideas. A family of HMMs is studied parametrised by a positive constant ϵ, which is a measure of the frequency of change. Thus, when $\epsilon \rightarrow 0$, the HMM becomes increasingly slower moving. We show that the smoothing error is $O(\epsilon)$. These theoretical predictions are confirmed by a series of simulations.**

***Index Terms*—Asymptotic error, hidden Markov model, smoothing.**

I. Introduction

THE CLASS of stochastic models known as hidden Markov models (HMMs) has received much attention in recent years, with many applications in speech recognition and telecommunications [1], [2]. An HMM consists of an underlying Markov chain, whose states are observed indirectly through a series of noise-corrupted measurements. In many applications, it is of importance to estimate these hidden Markov states accurately, and filtering and fixed-lag smoothing are two techniques for doing so. Suppose we are interested in estimating the Markov state at time k. The filtering problem consists of estimating from the measurements collected up to time k the Markov state at time k. Fixed-lag smoothing, when used to estimate the state at time k, requires measurements up to and including time $k + \Delta$, where $\Delta > 0$ is known as the smoothing lag.

In [3], filters and fixed-lag smoothers for discrete-time and discrete-state HMMs were investigated, where certain stability properties for HMM filters and smoothers were demonstrated. Specifically, it was shown that the benefit of fixed-lag smoothing decreases at an exponential rate with the smoothing lag used. Similar results have also been postulated in [4] for the case of a noisy random telegraph signal, which may be crudely considered a two-state HMM. However, the relative merits of smoothing over filtering were not addressed fully.

In particular, [4] postulated also that at high signal-to-noise ratio (SNR), smoothing would offer more improvement over filtering than at low SNR. This paper investigates that idea for a particular class of models. The class is characterized by a parameter $\epsilon > 0$ and is such that with ϵ small, the filtering error is proportional to $\epsilon \log(1/\epsilon)$, whereas the smoothing error is proportional to ϵ. This means that the smaller ϵ is, the better smoothing performs relative to filtering.

The parameter ϵ will appear in the models of this paper in the probability transition matrix for the state but not the probability transition matrix linking the state to the output. By letting ϵ tend to zero, the Markov process becomes increasingly slower. Very roughly, this is equivalent to a situation where the signal energy is increasingly more located near frequency zero, and the maximum over frequency of the SNR always becomes large.

General comparisons of the performance of Wiener and Kalman filters and smoothers have already been presented in [5] and [6], where [5] discusses the absolute benefit to be gained from smoothing over filtering as a function of SNR, whereas [6] considered the relative performance gains in the high SNR case. In [4], there was a limited investigation using only simulation of the relative performance gains for varying SNR in an HMM. However, theory was lacking, in part because of the unavailability of an error formula. To the best of our knowledge, it was not until [7] and [8] that an asymptotic formula (in the limit $\epsilon \rightarrow 0$) for the optimal filtering error for finite state-space Markov chains observed in independent noise was presented. Motivated by the existence of the filtering formula in this asymptotic case, we were led to seek a smoothing formula with similar validity. Indeed, in this paper, we will take a similar approach to the smoothing problem in that we will derive lower and upper bounds independently and show that asymptotically, as $\epsilon \rightarrow 0$, the two bounds are of the same order of magnitude. The smoother is constructed by combining the filtered estimates from two asymptotically optimal filters (see also [3]).

Another distinction of the present paper from [3] is that we will consider smoothing as being approximated as a hypothesis testing problem. This is in contrast with [3], where smoothed estimates were obtained from equations governing the time evolution of conditional probabilities. In most cases, the hypothesis testing problem amounts to the consideration of whether or a not a change of state has occurred in a given time interval. In this regard, we are not concerned with using a single optimal smoothing lag[1] but let the lag also vary according to the quantity ϵ. This is consistent with the scheme used in [7] because the interval length is chosen such

Manuscript received November 3, 1998; revised July 30, 2000. The associate editor coordinating the review of this paper and approving it for publication was Dr. Jean Jacques Fuchs.

L. Shue is with the Centre for Signal Processing, Nanyang Technological University, Singapore (e-mail: elouis@ntu.edu.sg).

B. D. O. Anderson is with the Department of Systems Engineering, Australian National University, Canberra, Australia (e-mail: Brian.Anderson@anu.edu.au).

F. De Bruyne is with the Advanced Process Control Group, Siemens Brussels, Huizingen, Belgium (e-mail: Franky.De-Bruyne@siemens.be).

Publisher Item Identifier S 1053-587X(00)10150-3.

[1]It was shown in [3] that due to the exponential-forgetting property exhibited by HMM filters and smoothers, there exists a finite smoothing lag for which practically all the benefit derivable from smoothing over filtering is achieved. However, since this "optimal" smoothing lag is essentially a time constant of the underlying system, as $\epsilon \rightarrow 0$ the "optimal" lag will also tend to ∞ using the arguments of this paper.

that the probability of error in detection of a change in state is small, and we are primarily interested in the asymptotic smoothing error when $\epsilon \to 0$.

Our main result (Theorem V.1) follows. Denoting the estimate of X_k for a smoothing lag of $T(\epsilon)$ as $\hat{X}_{k\,|\,k+T(\epsilon)}$, then as $\epsilon \to 0$

$$\Pr\left(X_k \neq \hat{X}_{k\,|\,k+T(\epsilon)}\right) = \left(\sum_i \pi_i \sum_{j \neq i} \gamma(i,j)\lambda_{ji}\right) \epsilon(1+o(1)) \tag{1}$$

where λ_{ji} denotes the rate of transition from state i to state j (see Section II), $\gamma(i,j)$ depends only on the conditional distributions of the output at time k given state i at time k and given the state j at time k, and $\Pi = (\pi_i)$ denotes the stationary distribution. The quantity $\gamma(i,j)$ is related to the probability of error in a certain hypothesis testing problem. This formula should be compared with the formula from [7], which reads (as $\epsilon \to 0$)

$$\Pr(X_k \neq \hat{X}_{k\,|\,k}) = \left(\sum_i \pi_i \sum_{j \neq i} \frac{\lambda_{ji}}{K(j,i)}\right) \epsilon \log\left(\frac{1}{\epsilon}\right)(1+o(1)) \tag{2}$$

where $\hat{X}_{k\,|\,k}$ denotes the filtered estimate of X_k, and $K(j,i)$ denotes Kullback–Leibler distance between the conditional distributions linking state to output when the states are i and j, respectively (see Section II). The increasingly superior performance of smoothing at low ϵ is evident. Note, however, that our methods are insufficiently precise to allow analytical computation of $\gamma(i,j)$.

This paper is organized as follows. Section II describes the signal model followed by a brief discussion of the mechanisms of filtering and smoothing in Section III. In Section IV, we will outline the method used for deriving the smoothing error bounds; the technical details can be found in the Appendixes. The overall smoothing error is presented in Section V, and some simulation results appear in Section VI.

II. Signal Model

Consider a first-order, discrete-time, and discrete-state Markov process X_k, the subscript k denoting time. For simplicity, we will define the states to be the values $1, 2, \ldots, N$. At each time instant k, a corresponding discrete-valued signal Y_k is observed in the range $1, 2, \ldots, M$. Here, we restrict ourselves to M, N, both being finite. We will adopt the convention that a lower-case x_k denotes the actual state value and likewise for y_k.

Denoting the state probability vector for X_k as $\Pi_k = (\Pr(X_k = i))$, we will adopt as in [7] the parametrization for the transition probability matrix $A = (a_{ij})$ with

$$a_{ij} = \Pr(X_{k+1} = i \,|\, X_k = j) = \begin{cases} \epsilon\lambda_{ij}, & i \neq j \\ 1 - \epsilon\lambda_{jj}, & i = j \end{cases} \tag{3}$$

where $\lambda_{jj} = \sum_{i \neq j} \lambda_{ij}$. Here, λ_{ij} has the interpretation of a transition rate, from state j to state i; ϵ ultimately determines how often such transitions take place.

The observations Y_k denote a sequence of random variables such that given the sequence X_k, Y_k are independent. The probability of obtaining a particular measurement y_k is determined by the observation matrix $C = (c_{mn})$, where

$$c_{mn} = \Pr(Y_k = m \,|\, X_k = n). \tag{4}$$

Unless otherwise stated, $a_{ij} > 0$ and $c_{mn} > 0$, $\forall i, j, n \in \{1, 2, \ldots, N\}$, $\forall m \in \{1, 2, \ldots, M\}$. In subsequent discussions, we will denote the ith column of C as $C_{\bullet i}$.

Remark II.1: In the present definition, A and C are both column-stochastic matrices.

We will also make the following assumptions concerning C. For any i, j, the Kullback–Leibler (KL) divergence

$$K(i,j) = \sum_k c_{ki} \log \frac{c_{ki}}{c_{kj}} \stackrel{\text{def}}{=} \mathbb{E}_i \log \frac{c_{\bullet i}}{c_{\bullet j}} \tag{5}$$

exists, and $\min_{i,j} K(i,j) > 0$, where $\mathbb{E}_i$ denotes expectation with respect to $C_{\bullet i}$. Note that the non-negativity of (5) is guaranteed due to the stochasticity of C, but strict positivity is required to preclude the case of C having two identical columns. If two columns in C are identical, then observations arising from the two different states corresponding to these columns are statistically indistinguishable.

Remark II.2: In [7], there was an additional explicit requirement on the KL divergence, namely, that $\mathbb{E}_i |\log(c_{\bullet i}/c_{\bullet j})|^\alpha < \infty$ for some $\alpha > 2$. This is automatically assured in the present discussion because we have restricted our investigations to HMMs with finite state and observation states.

III. Filtering and Smoothing

In this section, we will recall some general concepts of filtering and smoothing: first from the conventional point of view and then in the framework of hypothesis testing.

A. Evolution of Probabilities

Suppose we wish to estimate X_k. In the conventional framework of filtering, we can proceed by deriving the conditional probabilities $\Pr(X_k = i \,|\, Y_0, Y_1, \ldots, Y_k)$ at time k for each of the candidate states $i \in \{1, 2, \ldots, N\}$. The equations for evolution of such conditional probabilities can be found in, for example, [1]. The *maximum a posteriori* (MAP) state estimate at time k, whichi s denoted as $\hat{X}_{k\,|\,k}$, is then

$$\hat{X}_{k\,|\,k} = \arg\max_i \Pr(X_k = i \,|\, Y_0, Y_1, \ldots, Y_k) \tag{6}$$

where the notation $\hat{X}_{k\,|\,k}$ indicates an estimate of X_k using measurements from time 0 to time k. In estimating X_k, fixed-lag smoothing uses more measurements than filtering. The corresponding MAP state estimate for a smoothing lag of $\Delta > 0$ is

$$\hat{X}_{k\,|\,k+\Delta} = \arg\max_i \Pr(X_k = i \,|\, Y_0, Y_1, \ldots, Y_{k+\Delta}). \tag{7}$$

As shown in [3], the fixed-lag smoothed conditional probability can also be expressed as (8), shown at the bottom of the

page, where Θ is a normalizing constant. As can be seen, the numerator of (8)[2] consists of two terms: a forward filter and a reverse-time one-step ahead predictor. The reverse-time one-step ahead predictor operates on a backward Markov process associated with the Markov chain in the original HMM; the backward process can be constructed from the forward Markov state process by procedures set out in [9]. In this fashion, a fixed-lag HMM smoother can be considered as consisting of two HMM filters (strictly, one filter and one predictor) operating in combination.

B. Hypothesis Testing Perspective

In this section, we will recast the task of filtering and smoothing of HMMs in the form of (binary) hypothesis testing problems. We proceed by first reviewing some concepts in conventional hypothesis testing.

Let $\{v_1, v_2, \ldots, v_N\}$ be a sequence of independent and identically distributed (i.i.d.) random variables. The binary hypothesis testing problem consists of deciding, based on this sequence, whether the distribution generating the sequence is one of the two distributions [$p(\cdot)$ or $q(\cdot)$] when there is no prior knowledge of which distribution generated the sequence. Mathematically, the problem is to correctly discriminate between these two distributions by considering the following hypotheses for the composition of the observations:

H_0: i.i.d. data with probability density $q(v)$.
H_1: i.i.d. data with probability density $p(v)$.

For the purposes of the present argument, we will also assume that H_1 is the desired hypothesis, which may mean the presence of a target, rather than just noise as modeled by $q(\cdot)$, for example.

Remark III.1: We will comment briefly on some terminology used in hypothesis testing. Let us assume first that H_1 has been designated as the desired response, as opposed to H_0, which is sometimes referred to as the null hypothesis. A false alarm is then a situation where we decide (as a result of a hypothesis test) that H_1 is true when, in fact, H_0 is true; conversely, a miss occurs when we say that H_0 is true when H_1 is the correct hypothesis. Last, a detection occurs when we conclude that H_1 is true, and it is also the true hypothesis.

Given that the prior probabilities of the hypotheses being true are unknown, the Neyman–Pearson formulation [10] minimizes the false alarm rate, i.e., $\Pr(\text{decide } H_1 \,|\, H_0 \text{ is true})$, while maintaining a fixed detection rate, i.e., $\Pr(\text{decide } H_1 \,|\, H_1 \text{ is true})$. From the Neyman–Pearson lemma, the optimal test to minimize the false alarm rate is the log-likelihood ratio test

$$\frac{1}{N}\sum_{i=1}^{N} \log \frac{p(v_i)}{q(v_i)} \begin{cases} > u_N & \text{decide } H_1 \\ \leq u_N & \text{decide } H_0 \end{cases}$$

where the threshold u_N is chosen so that a prescribed fixed detection probability is maintained. Consequently, we can rewrite the detection probability as

$$\Pr(\text{decide } H_1 \,|\, H_1 \text{ is true}) = \Pr_1\left(\frac{1}{N}\sum_{i=1}^{N} \log \frac{p(v_i)}{q(v_i)} > u_N\right)$$

where the notation $\Pr_i(Z)$ denotes the probability of event Z, given that hypothesis $i \in \{0, 1\}$ is true.

Remark III.2: For the rest of the paper, we will use the term error to mean the Bayes probability of error, which takes into account both false alarm and miss probabilities, which are defined in the binary hypothesis case as

$$P_N \stackrel{\text{def}}{=} \alpha_N \Pr(H_0) + \beta_N \Pr(H_1) \tag{9}$$

where $\alpha_N = \Pr(\text{Declare } H_1 \,|\, H_0 \text{ true})$ is the probability of a false alarm, and $\beta_N = \Pr(\text{Declare } H_0 \,|\, H_1 \text{ true})$ is the probability of a miss. The best Bayes exponential error rate is achieved by a Neyman–Pearson test with zero threshold (see [10] and [11]).

In [7], the asymptotic filtering error of HMMs was investigated via a series of hypothesis testing problems. In order to obtain a lower bound to the filtering error, the authors considered a binary hypothesis testing problem ([7, Lemma 1]), whereby the hypothesis H_1 in the previous notation consists of not one but multiple events. The two hypotheses are that the chain either remains at the same known state i throughout the entire interval $[1, T]$ or changes to a different known state j at some time τ for $i \neq j \in \{1, 2, \ldots, N\}$. Formally, this is written as

H_0: $\{y_1, y_2, \ldots, y_T\}$ is a sequence of i.i.d. random variables, each of law $C_{\bullet i} = \{\Pr(Y_k = l \,|\, X_k = i)\}$, $k \in \{1, 2, \ldots, T\}$ and $l \in \{1, 2, \ldots, M\}$.

H_1: $\{y_1, y_2, \ldots, y_{\tau-1}\}$ is a sequence of i.i.d. random variables, each of law $C_{\bullet i} = \{\Pr(Y_k = l \,|\, X_k = i)\}$, $k \in \{1, 2, \ldots, \tau - 1\}$, and $\{y_\tau, y_{\tau+1}, \ldots, y_T\}$ is a sequence of i.i.d. random variables each of law $C_{\bullet j} = \{\Pr(Y_k = l \,|\, X_k = j)\}$, $k \in \{\tau, \tau + 1, \ldots, T\}$, $l \in \{1, 2, \ldots, M\}$, τ is uniformly distributed in $\{1, 2, \ldots, T\}$.

However, we emphasize that this hypothesis testing problem is only an approximation to the real filtering task since in order to formulate the hypothesis testing problem, the two terminating states between which the transitions take place are postulated *a*

[2]Since (8) contains the factor $\Pr(X_k = i)$ in the denominator, it appears that even when the forward filter and the reverse-time one-step ahead predictor agree in their estimates of X_k, the resulting MAP smoothed estimate may still differ from that of either filter. However, it can be shown that provided errors in the forward and backward estimates of X_k are small, a MAP estimate obtained by combining the forward and backward estimates has a small probability of error when the two estimates, in fact, agree.

$$\Pr(X_k = i \,|\, Y_0, Y_1, \ldots, Y_{k+\Delta}) = \frac{\Pr(X_k = i \,|\, Y_0, \ldots, Y_k) \Pr(X_k = i \,|\, Y_{k+1}, \ldots, Y_{k+\Delta})}{\Theta \Pr(X_k = i)} \tag{8}$$

priori. Furthermore, it has been implicitly assumed that at most exactly a single transition can take place in $[1,T]$, which is not likely to hold true in a Markov chain over any finite interval. Nevertheless, this is partly justified by using a suitably chosen T to ensure the probability of multiple jumps is small—at most $O(\epsilon^2 \log(1/\epsilon))$—as $\epsilon \to 0$. Since there can be only two outcomes from the hypothesis test just mentioned, a decision of H_1 is roughly equivalent to obtaining a filtered estimate of $\hat{X}_T = j$, i.e., $N = 2$. To obtain a more realistic filtered estimate, the same test is repeated over all $i \neq j \in \{1, 2, \ldots, N\}$.

The data length T, for the purposes of obtaining asymptotic error bounds, is defined as

$$T = \frac{1+\delta}{\Lambda^*} \log \frac{1}{\epsilon} \tag{10}$$

with $\delta > 0$ arbitrary and

$$\Lambda^* = \sup_{\theta} \left[-\log \mathbb{E}_i \left(\frac{c_{y_k j}}{c_{y_k i}} \right)^{\theta} \right] \tag{11}$$

where i, j are the two Markov states postulated in the hypothesis testing problem, and $\mathbb{E}_i(\cdot)$ denotes expectation with respect to the conditional density $C_{\bullet i}$. The technical reasons for this definition of T can be found in Appendix A (see also [7]). Suffice it to say here that T is chosen such that for a related hypothesis testing problem of choosing between two simple hypotheses where the state is constant over $[1,T]$ and is either i or j, the probability of a false alarm is $O(\epsilon^{1+\delta})$. It was shown in [7] that for this choice of T, the asymptotic error as $\epsilon \to 0$ for the hypothesis testing problem involving H_0 and H_1 (not the related problem) is lower bounded by the probability of a single transition over the interval $[1,T]$ or $O(\epsilon T)$.

For an HMM smoother, measurements are available both before and after the time of interest, which we will denote as k. Using as motivation the MAP smoothed conditional probability density (8), we will construct a suboptimal smoother by simply combining two filtered estimates obtained by a forward filter over $T_f = [k-T, k]$ and from a backward predictor on $T_b = [k+1, k+T]$ (see Fig. 1), with T defined in (10).

As an outline, a lower bound on the smoothing error is obtained as follows. We postulate that we know exactly the states at time $k-T$ and at time $k+T$, and we further postulate that we know there is at most one jump in the interval $T_f \cup T_b$. Whether there is zero or one jump is then self-evident from knowledge of the states at times $k-T$ and $k+T$. If there is no jump, the smoothed estimate is obtained with zero error. If there is one jump, the determination of the smoothed estimate is equivalent to determining whether the jump occurs before or after k. In Section IV-A, we will investigate this latter problem. Full details of the lower bound calculations are provided in Section IV-B.

As an outline, the upper bound is determined as follows. Using measurements over T_f but no internal state information or further measurements, we determine a filtered estimate (call it $\hat{X}_{fk}$) of X_k. Similarly, using measurements over T_b, we determine a filtered estimate (call it $\hat{X}_{bk}$) of X_k. If these estimates are the same, we set the smoothed estimate (call it $\hat{X}^{\text{sub}}_{k\,|\,k+T}$) equal to the common value. Otherwise, we adopt a supplementary procedure to determine whether $\hat{X}_{fk}$ or $\hat{X}_{bk}$ is more reliable and then set $\hat{X}^{\text{sub}}_{k\,|\,k+T}$ accordingly. This supplementary procedure, details of which are introduced in Section IV–A, involves introducing an auxiliary hypothesis testing problem like that used in the lower bound calculations. The details of the upper bound calculations are provided in Section IV-C.

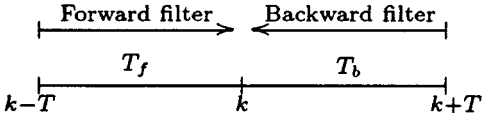

Fig. 1. Smoother as combination of ttwo filters.

IV. General Approach

In this section, we will outline the techniques used in bounding the smoothing error probabilities from below and above for general distributions. The technical details can be found in the Appendixes.

A. Detection of Location of Jump

As we will see, the determination of bounds on the smoothing error rate is strongly connected to an associated problem of partially localizing the time at which changes in the distributions of the measurements occur. Let us consider observations of a process $\{y_{k-T}, \ldots, y_k, \ldots, y_{k+T}\}$ that are derived from two distributions $p(\cdot)$ and $q(\cdot)$. The samples are independent, and we assume that there is a single change from $p(\cdot)$ to $q(\cdot)$ at some time $t_c \in [k-T, k+T]$ so that y_i is distributed with density $p(\cdot)$ for $i \in [k-T, t_c - 1]$ and with density $q(\cdot)$ for $i \in [t_c, k+T]$. We further postulate that t_c is uniformly distributed on $[k-T+1, k+T]$ or for all $s \in [k-T+1, k+T]$

$$\Pr(t_c = s) = \frac{1}{2T}.$$

To obtain bounds to the smoothing error, we will study a hypothesis testing problem. The hypotheses in question—both actually multiple hypotheses—are

H_0: $k - T < t_c \leq k$;
H_1: $k + 1 \leq t_c \leq k + T$.

Our immediate task is to indicate how t_c can be estimated by some $\hat{t}_c$ and then to obtain error bounds on the declaration that H_0 or H_1 occurred when, in reality, the opposite has occurred. This result is summarized in Theorem IV.1.

Remark IV.1: For a Markov chain, the probability distribution of the actual time of jump from state i to state j, given that $X_{k-T} = i$ and $X_{k+T} = j$, can be expressed as (12), shown at the bottom of the next page. However, as will be shown in the proof of Theorem 5.1, as $\epsilon \to 0$, the denominator of (12) is $2T(1+o(1))$, whereas the numerator is $1+o(1)$; hence, the uniform distribution is a good approximation for the time of jump of a Markov chain under such limiting conditions.

Definition IV.1: Denoting the events for different time of jump as $E^{t:i\to j}_{[k-T,k+T]} \stackrel{\text{def}}{=} \{X_{k-T} = \cdots = X_{t-1} = i, X_t = \cdots = X_{k+T} = j\}$ and $E^{i}_{[k-T,k+T]} \stackrel{\text{def}}{=} \{X_{k-T} = \cdots =$

$X_{k+T} = i\}$, where $t \in [k-T, k+T]$, let us define the log-likelihood ratio function

$$S(t) = \log \frac{\Pr\left(y_{k-T}, \ldots, y_{k+T} \,\middle|\, E^{t:i\to j}_{[k-T,k+T]}\right)}{\Pr\left(y_{k-T}, \ldots, y_{k+T} \,\middle|\, E^{i}_{[k-T,k+T]}\right)}.$$

Theorem IV.1: Consider a set of independent observations $\{y_{k-T}, y_{k-T+1}, \ldots, y_{k+T}\}$, where for $\ell = k-T, \ldots, t_c - 1$, y_ℓ is of law $C_{\bullet i} = (\Pr(Y_\ell \,|\, X_\ell = i))$ and for $\ell = t_c, \ldots, k+T$, y_ℓ is of a distinct law $C_{\bullet j} = (\Pr(Y_\ell \,|\, X_\ell = j)$, with $K(i,j)$ and $K(j,i) > 0$ for $i \neq j \in \{1, 2, \ldots, N\}$. Assume that t_c is uniformly distributed in $[k-T, k+T]$; then, the two hypotheses for the unknown time of change t_c are

H_0: $\quad k - T < t_c \leq k$;

H_1: $\quad k + 1 \leq t_c \leq k + T$.

By choosing the estimator $\hat{t}_c$ to maximize the log-likelihood ratio function

$$\begin{aligned} S(t) &= \log \frac{\Pr\left(y_{k-T}, \ldots, y_{k+T} \,\middle|\, E^{t:i\to j}_{[k-T,k+T]}\right)}{\Pr\left(y_{k-T}, \ldots, y_{k+T} \,\middle|\, E^{i}_{[k-T,k+T]}\right)} \\ &= \log \frac{c_{y_{k-T}i} \cdots c_{y_{t-1}i} c_{y_t j} \cdots c_{y_{k+T}j}}{c_{y_{k-T}i} c_{y_{k-T+1}i} \cdots c_{y_{k+T}i}} \\ &= \sum_{\ell=t}^{k+T} \log \frac{c_{y_\ell j}}{c_{y_\ell i}} \end{aligned}$$

the asymptotic error in localizing t_c is

$$\begin{aligned} &\Pr(\text{Declare incorrect hypothesis}, t_c \in T_f \cup T_b) \\ &\quad = \frac{\gamma(i,j)}{2T} \end{aligned}$$

(using the notation of Fig. 1) for T sufficiently large and some $\gamma(i,j) > 0$, depending on the conditional distributions $C_{\bullet i}$ and $C_{\bullet j}$ but independent of T and A.

Proof: Upper Bound: A necessary condition for overestimating t_c as, say $t_c + J \leq k + T$, is that $S(t_c + J) > S(t_c)$. Hence

$$\Pr(\hat{t}_c = t_c + J) \leq \Pr(S(t_c + J) > S(t_c)). \tag{13}$$

Now, by appealing to large deviations theory (see also Appendix B), we note that

$$\Pr(\hat{t}_c = t_c + J) \leq \exp\left[-J I_{C_{\bullet j}}(0)\right] \tag{14a}$$

$$\Pr(\hat{t}_c = t_c - J) \leq \exp\left[-J I_{C_{\bullet i}}(0)\right] \tag{14b}$$

where with $w_\ell = \log(c_{y_\ell j}/c_{y_\ell i})$, $I_{C_{\bullet j}}(0) = \inf_{w_\ell < 0} I_{C_{\bullet j}}(w_\ell)$, and $I_{C_{\bullet i}}(0) = \inf_{w_\ell > 0} I_{C_{\bullet i}}(w_\ell)$. Note that the latter two quantities are positive.

Suppose now that $t_c = s$ for some $s \in T_b$, i.e., H_1 is the true hypothesis. H_0 will be declared if $\hat{t}_c \leq k$, and the probability of such an occurrence can be bounded as

$$\begin{aligned} &\Pr(\hat{t}_c \in T_f \,|\, t_c = s) \\ &\quad = \sum_{i \in T_f} \Pr(\hat{t}_c = i \,|\, t_c = s) \\ &\quad \leq \sum_{\ell=s-k}^{s-(k-T+1)} \Pr(S(t_c - \ell) > S(t_c)) \\ &\quad \leq \sum_{\ell=s-k}^{s-(k-T+1)} \exp[-\ell I_{C_{\bullet i}}(0)]. \end{aligned} \tag{15}$$

The probability of declaring H_0 when H_1 is the true hypothesis is therefore bounded by

$$\begin{aligned} &\Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true}) \\ &\quad = \sum_{s \in T_b} \Pr(\hat{t}_c \in T_f \,|\, t_c = s) \Pr(t_c = s) \\ &\quad = \sum_{s \in T_b} \sum_{\ell=s-k}^{s-(k-T+1)} \Pr(\hat{t}_c = t_c - \ell \,|\, t_c = s) \Pr(t_c = s) \\ &\quad \leq \frac{1}{2T\left(1 - \exp[-I_{C_{\bullet i}}(0)]\right)^2}. \end{aligned} \tag{16}$$

Similarly, if $t_c \in T_f$, then the probability of concluding $\hat{t}_c \in T_b$ is bounded by

$$\begin{aligned} \Pr(\hat{t}_c \in T_b] \,|\, t_c = s) &= \sum_{i \in T_b} \Pr(\hat{t}_c = i \,|\, t_c = s) \\ &\leq \sum_{\ell=k+1-s}^{k+T-s} \Pr(S(t_c + \ell) > S(t_c)) \\ &\leq \sum_{\ell=k+1-s}^{k+T-s} \exp\left[-\ell I_{C_{\bullet j}}(0)\right]. \end{aligned} \tag{17}$$

Consequently, the estimation error is overbounded as follows:

$$\begin{aligned} &\Pr(\text{Declare incorrect hypothesis}, t_c \in T_f \cup T_b) \\ &\quad = \Pr(\text{Declare } H_1 \,|\, H_0 \text{ is true}) \Pr(H_0 \text{ is true}) \\ &\qquad + \Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true}) \Pr(H_1 \text{ is true}) \\ &\quad \leq \frac{1}{2T(1-x)^2} = \frac{\beta(i,j)}{2T} \end{aligned} \tag{18}$$

where $x = \exp[-\min(I_{C_{\bullet j}}(0), I_{C_{\bullet i}}(0))]$ is independent of T, and we have also used the fact that $\Pr(H_0 \text{ is true}) = \Pr(H_1 \text{ is true}) = 1/2$.

$$\begin{aligned} &\Pr(t_c = s \,|\, X_{k-T} = i, X_{k+T} = j, \text{one jump in } [k-T, k+T]) \\ &\quad = \frac{\Pr(X_{k-T} = \ldots X_{s-1} = i, X_s = \ldots X_{k+T} = j)}{\sum_{t_c=k-T+1}^{k+T} \Pr(X_{k-T} \ldots X_{t_c-1} = i, X_{t_c} \ldots X_{k+T} = j)} = \frac{a_{ii}^{s-1} a_{ji} a_{jj}^{2T-s}}{\sum_{u=1}^{2T} a_{ii}^{u-1} a_{ji} a_{jj}^{2T-u}} \end{aligned} \tag{12}$$

Lower Bound: By picking out just one term in the following summation, we have

$$\begin{aligned}&\Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true})\\ &\quad=\sum_{s\in T_b}\Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true}, t_c=s)\Pr(t_c=s)\\ &\quad\geq \Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true}, t_c=k+1)\\ &\qquad\times\Pr(t_c=k+1)\\ &\quad=\frac{1}{2T}\Pr(\hat{t}_c<t_c \,|\, t_c=k+1)=\frac{\alpha(i,j)}{2T}\end{aligned}$$

because $\Pr(\hat{t}_c < t_c \,|\, t_c = k+1)$ is asymptotically independent of T as $T \to \infty$ (see Appendix C). Since $\alpha(i,j), \beta(i,j) > 0$ and are independent of T, there exists $\gamma(i,j) \in [\alpha(i,j), \beta(i,j)]$ such that

$$\Pr(\text{Declare } H_0 \,|\, H_1 \text{ is true}) = \frac{\gamma(i,j)}{2T}.$$

Like $\alpha(i,j)$ and $\beta(i,j)$, $\gamma(i,j)$ depends on $C_{\bullet i}$ and $C_{\bullet j}$. Hence, as $T \to \infty$

$$\begin{aligned}&\Pr(\text{Declare incorrect hypothesis}, t_c \in T_f \cup T_b)\\ &\quad=\frac{\gamma(i,j)}{2T}.\end{aligned}$$

■

B. Lower Bound on Smoothing Error

We will proceed by first observing that MAP estimates minimize the errors, and in general, the quality of state estimates depends on the abundance (or lack) of prior information available.

Let X be a random variable that takes one of N possible values $\{1, 2, \ldots, N\}$. Let A denote an event, and let p_A be a conditional probability vector with $p_A(i) = \Pr(X = i \,|\, A)$. Define the MAP estimate of X based on A as

$$\hat{X}_A = \arg\max_{\ell} \Pr(X = \ell \,|\, A). \tag{19}$$

It is well known that this choice for $\hat{X}_A$ minimizes the error rate $\Pr(X \neq \hat{X}_A \,|\, A)$ over all estimators for X that use A. Suppose a particular estimate using A sets $\hat{X}_A = j$; then, an error will arise if $X = 1, 2, \ldots, j-1, j+1, \ldots, N$. That is, given A, the conditional probability of estimation error with this estimator is $\sum_{\substack{i=1 \\ i\neq j}}^{N} p_A(i) = 1 - p_A(j)$, and as a function of j, this is minimized by maximizing $p_A(j)$, which is the MAP estimator in (19).

Suppose that in addition to A, an event B is also observed. This information is used to construct a MAP estimate $\hat{X}_{AB}$ of the process X. Intuitively, we would expect the overall error rate for the estimator $\hat{X}_{AB}$ to be smaller than that for $\hat{X}_A$ since $\hat{X}_A$ is a suboptimal estimator (the suboptimality arising from the fact that it does not utilize all available information) when both events A and B are observed. That is

$$\Pr(X \neq \hat{X}_{AB} \,|\, A, B) \leq \Pr(X \neq \hat{X}_A \,|\, A, B)$$

and consequently

$$\Pr(X \neq \hat{X}_{AB}) \leq \Pr(X \neq \hat{X}_A).$$

In the subsequent discussions, event B corresponds to the additional knowledge of the states at times $k-T$ and $k+T$.

Denote $\hat{X}_{k\,|\,k+\Delta}$ as the MAP smoothed estimate of X_k. Now, suppose that the states X_{k-T} and X_{k+T} are known precisely, where $T \leq \Delta$. Using the notation of Fig. 1, label the events that the state has no jump, exactly one jump, and i jumps in $T_f \cup T_b$, respectively, as $0J$, $1J$, and iJ[3]. Denote also by $\hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}$ a MAP smoothed estimate of X_k obtained by using the Y_i as before, together with the knowledge of the states X_{k-T} and X_{k+T}. Then, we have

$$\begin{aligned}&\hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\\ &\quad=\arg\max_j \Pr(X_k = j \,|\, \ldots, Y_{k-1}, \ldots\\ &\qquad Y_{k+\Delta}, X_{k-T}, X_{k+T})\\ &\quad=\arg\max_j \Pr(X_k = j \,|\, Y_{k-T+1}, \ldots\\ &\qquad Y_{k+T-1}, X_{k-T}, X_{k+T})\end{aligned}$$

using the Markov property. Since more information is used in computing $\hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}$ than $\hat{X}_{k\,|\,k+\Delta}$, we can obtain a lower bound to the error rate for $\hat{X}_{k\,|\,k+\Delta}$ as follows:

$$\begin{aligned}&\Pr(X_k \neq \hat{X}_{k\,|\,k+\Delta})\\ &\quad\geq \Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\Big)\\ &\quad\geq \Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}, 1J\Big)\\ &\quad=\sum_i\sum_{j\neq i}\Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\\ &\qquad X_{k-T}=i, X_{k+T}=j, 1J\Big)\\ &\quad=\sum_i\sum_{j\neq i}\Big\{\Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\;\Big|\; X_{k-T}=i\\ &\qquad X_{k+T}=j, 1J\Big)\Pr(X_{k-T}=i, X_{k+T}=j, 1J)\Big\}. \quad (20)\end{aligned}$$

Last, denote $t_c \in T_f \cup T_b$ as the true time of (single) jump and $\hat{t}_c$ as the MAP estimate of t_c. Then, we have

$$\begin{aligned}&\Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\;\Big|\; X_{k-T}=i, X_{k+T}=j, 1J\Big)\\ &\quad=\Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}, t_c \in T_f \,|\, X_{k-T}=i\\ &\qquad X_{k+T}=j, 1J\Big)+\Pr\Big(X_k \neq \hat{X}^{\text{known ends}}_{k\,|\,k+\Delta}\\ &\qquad t_c \in T_b \,|\, X_{k-T}=i, X_{k+T}=j, 1J\Big)\\ &\quad\geq \Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T}=i, X_{k+T}=j, 1J)\\ &\qquad+\Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T}=i, X_{k+T}=j, 1J) \quad (21)\end{aligned}$$

where the hypothesis testing problem discussed in Section IV-A has been invoked. The inequality arises because estimating the jump incorrectly is a sufficient but unnecessary condition for

[3]Note that this definition does not exclude the possibility of having jumps outside of $T_f \cup T_b$.

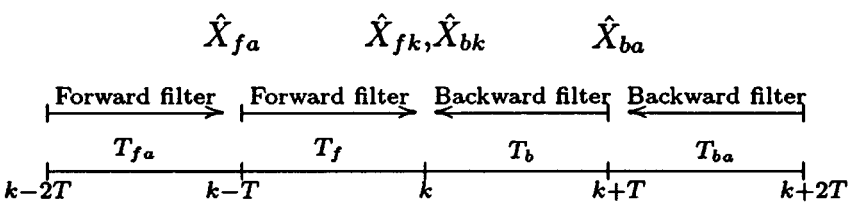

Fig. 2. Suboptimal smoothing scheme.

$X_k \neq \hat{X}_{k\,|\,k+\Delta}^{\text{known ends}}$. From (20) and (21), the lower bound is then

$$\begin{aligned}\Pr(X_k \neq \hat{X}_{k\,|\,k+\Delta}) &\geq \sum_i \sum_{j\neq i} \Pr(X_{k-T}=i, X_{k+T}=j, 1J) \\ &\quad \times \{\Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T}=i, X_{k+T}=j, 1J) \\ &\quad + \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T}=i, X_{k+T}=j, 1J)\}. \end{aligned} \tag{22}$$

C. Upper Bound on Smoothing Error

In this section, we will introduce a sub-optimal smoothing scheme (Fig. 2) in order to derive an upper bound to the smoothing error probability. Initially, we will restrict discussions to two-state Markov chains only. Consider the following log-likelihood ratios

$$Z_k = \sum_{j=k-T}^{k} \log \frac{c_{y_j 1}}{c_{y_j 2}} \tag{23a}$$

$$Z_{k-T} = \sum_{j=k-2T}^{k-T-1} \log \frac{c_{y_j 1}}{c_{y_j 2}}. \tag{23b}$$

Equations (23a) and (23b) are the analogs of a forward filter and predictor operating on $T_f = [k-T, k]$ and $T_{fa} = [k-2T, k-T-1]$, respectively, and are used to derive sub-optimal estimates of X_k and X_{k-T}. That is, if $Z_k \geq 0$, then let the estimate of X_k be 1, else set it to 2, and similarly for the estimate of X_{k-T}. Note that as $\epsilon \to 0$, the probability of a jump over a single time step is $O(\epsilon)$; hence, justifying estimating X_{k-T} using measurements up to only time $k-T-1$. There are analogous expressions for estimation of X_k and X_{k+T} using measurements over the intervals to the right of k and to the right of $k+T$. For the moment, we will leave $\hat{X}_{bk}$ and $\hat{X}_{ba}$ out of consideration.

Denote the suboptimal estimate obtained from the forward "filter" operating on T_f as $\hat{X}_{fk}$ [i.e., using the log-likelihood (23a)], and likewise, denote the estimate of X_k corresponding to a backward "predictor" operating on $T_b = [k+1, k+T]$ as $\hat{X}_{bk}$. A suboptimal smoothed estimate of X_k—call it $\hat{X}_{k\,|\,k+T}^{\text{sub}}$—will be obtained by combining $\hat{X}_{fk}$ and $\hat{X}_{bk}$ in the manner to be described. If $\hat{X}_{fk} = \hat{X}_{bk}$, then the probability of error in estimating X_k is $O(\epsilon^{1+\delta})$, $\delta > 0$ and arbitrary when we set $\hat{X}_{k\,|\,k+T}^{\text{sub}} = \hat{X}_{fk}$. However, if $\hat{X}_{fk} \neq \hat{X}_{bk}$, since there is no prior reason for supposing that $\hat{X}_{fk}$ is more reliable than $\hat{X}_{bk}$ (or vice versa), additional calculations are required to determine

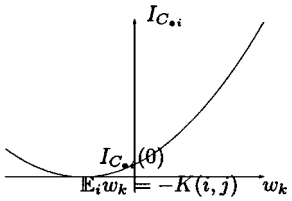

Fig. 3. Large deviations rate function.

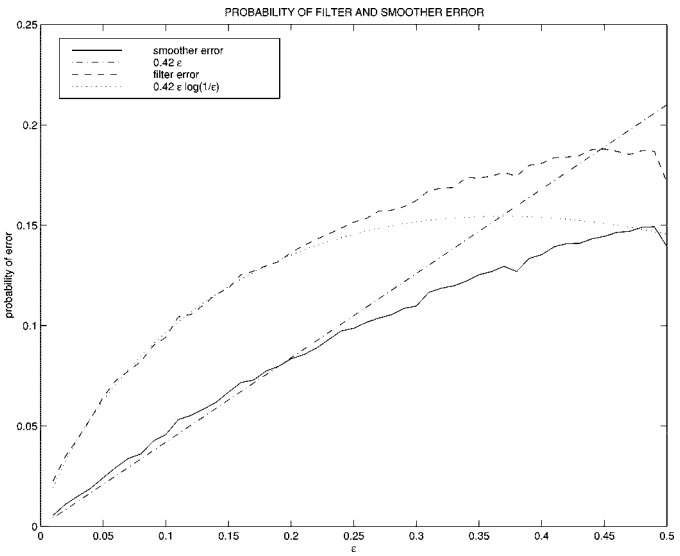

Fig. 4. Probability of filter and smoother error for small ϵ for matrices (38).

$\hat{X}_{k\,|\,k+T}^{\text{sub}}$. As a preliminary observation, we note that the probability of $\hat{X}_{fk}$ [which is obtained using (23a)] being in error if there is no change of state in T_f is $O(\epsilon^{1+\delta})$, $\delta > 0$ (see Appendix A) and similarly for $\hat{X}_{bk}$; from this, we can conclude that errors are largely the result of jump(s) in $T_f \cup T_b$—in fact, the main contribution to smoothing error results from having a single jump in $T_f \cup T_b$.

To determine the reliability of $\hat{X}_{fk}$ over $\hat{X}_{bk}$, let us augment the intervals T_f and T_b by T_{fa} and $T_{ba} = [k+T+1, k+2T]$ (Fig. 2). A forward "predictor" [see (23b)] operates on T_{fa} and a backward "predictor" on T_{ba} to estimate X_{k-T} and X_{k+T}. When $\hat{X}_{fk} \neq \hat{X}_{bk}$, the estimates $\hat{X}_{fa}$ and $\hat{X}_{ba}$ are used as the terminating states in the supplementary hypothesis testing problem over $T_f \cup T_b$ to determine whether the (single) jump occurs before or after time k. If $\hat{X}_{fk}$ is in error (due to a jump in T_f), the state from which the jump has occurred is not known exactly but can be estimated as $\hat{X}_{fa}$. Likewise, should $\hat{X}_{bk}$ be in error (due to a jump in T_b), the state to which the jump occurs can be estimated as $\hat{X}_{ba}$. Of course, $\hat{X}_{fa}$ and $\hat{X}_{ba}$ will, on occasion, be in error. In addition, there may be more than

one jump in $T_f \cup T_b$. Nevertheless, as a basis for determining a suboptimal estimate (which in theory can be anything), there is clearly a heuristic basis for the following rule.

1) If $\hat{X}_{fk} = \hat{X}_{bk}$, then choose $\hat{X}^{\text{sub}}_{k\,|\,k+T} = \hat{X}_{fk}$.
2) If $\hat{X}_{fk} \neq \hat{X}_{bk}$, decide whether a jump occurred in T_f or T_b using the supplementary hypothesis test discussed in Section IV-A. In applying the supplementary test, the end states are taken to be $\hat{X}_{fa}$ and $\hat{X}_{ba}$. If the supplementary test suggests that a jump has occurred in T_f, then $\hat{X}_{bk}$ is likely to be the correct estimate of X_k; therefore, set $\hat{X}^{\text{sub}}_{k\,|\,k+T} = \hat{X}_{bk}$; alternatively, if the test indicates the jump is in T_b, then set $\hat{X}^{\text{sub}}_{k\,|\,k+T} = \hat{X}_{fk}$.

Now, let $0J$, $1J$, and MJ denote the events that the state has no jump, one jump, and multiple jumps, respectively, in the interval $[k-2T, k+2T]$. The overall smoothing error in term of these events is then

$$\begin{aligned}\Pr(\text{Error}) &= \Pr(\text{Error}, 0J) + \Pr(\text{Error}, 1J) \\ &\quad + \Pr(\text{Error}, MJ). \end{aligned} \tag{24}$$

We will now consider the contribution of each possibility.

1) Case 1: Multiple Jumps: From (10), the probability of multiple jumps in $[k-2T, k+2T]$ is

$$\begin{aligned}\Pr(MJ) &= O((4\epsilon T)^2 + (4\epsilon T)^3 + (4\epsilon T)^4 + \cdots) \\ &= O((\epsilon T)^2) = O\left(\epsilon^2 \log^2 \frac{1}{\epsilon}\right).\end{aligned}$$

Since $\Pr(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, MJ) \leq \Pr(MJ)$, we then have

$$\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, MJ\Big) = O\left(\epsilon^2 \log^2 \frac{1}{\epsilon}\right). \tag{25}$$

2) Case 2: Zero Jump: In this situation, we need only consider the interval $T_f \cup T_b$. Suppose that $X_i = 1$ for $i \in T_f \cup T_b$. $\hat{X}_{fk} = 2$ will be declared if and only if $Z_k < 0$ and the probability of such an occurrence is bounded by (see Appendix A)

$$\begin{aligned}&\Pr(X_k \neq \hat{X}_{fk} \,|\, X_k = 1, 0J) \\ &\quad = \Pr(Z_k < 0 \,|\, X_k = 1, 0J) \leq \epsilon^{1+\delta}.\end{aligned}$$

$\delta > 0$ and arbitrary. Evidently, the probability of error is

$$\begin{aligned}&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 0J\Big) \\ &\quad = \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} = \hat{X}_{bk}, 0J\Big) \\ &\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk}, 0J\Big) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{fk}, X_k \neq \hat{X}_{bk}, 0J) \\ &\qquad + \Pr(X_k \neq \hat{X}_{fk}, 0J) + \Pr(X_k \neq \hat{X}_{bk}, 0J) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{fk}, X_k \neq \hat{X}_{bk} \,|\, 0J) \\ &\qquad + \Pr(X_k \neq \hat{X}_{fk} \,|\, 0J) + \Pr(X_k \neq \hat{X}_{bk} \,|\, 0J) \\ &\quad = O(\epsilon^{1+\delta}).\end{aligned} \tag{26}$$

3) Case 3: One Jump: Based on the four time intervals as indicated in Fig. 2, the error in this situation can be rewritten as the sum of four components:

- $\Pr(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_{fa})$;
- $\Pr(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_f)$;
- $\Pr(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_b)$;
- $\Pr(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_{ba})$

where the notation means only one jump in the given interval and none outside of it (but still within $[k-2T, k+2T]$). In the following derivations, we will limit our discussions to the forward "filters" corresponding to (23a) and (23b) only since the same arguments apply to the backward estimates by symmetry. We will also assume without loss of generality, $X_{k-T} = 2$, and $X_{k+T} = 1$. For each term, there are essentially three ways that an error can arise.

1) $\hat{X}_{fk}$ and $\hat{X}_{bk}$ simultaneously in error;
2) error in $\hat{X}_{fk}$ and not in $\hat{X}_{bk}$ but $\hat{X}^{\text{sub}}_{k\,|\,k+T}$ set to $\hat{X}_{fk}$;
3) error in $\hat{X}_{bk}$ and not in $\hat{X}_{fk}$, but $\hat{X}^{\text{sub}}_{k\,|\,k+T}$ set to $\hat{X}_{bk}$.

a) Jump in T_{fa}: For this situation, the error bounds are derived using the same reasoning as in the zero jump case. That is, with

$$\begin{aligned}&\Pr(X_k \neq \hat{X}_{fk} \,|\, X_k = 1, 0J \text{ in } T_f) \leq \epsilon^{1+\delta} \\ &\Pr(X_k \neq \hat{X}_{bk} \,|\, X_k = 1, 0J \text{ in } T_b) \leq \epsilon^{1+\delta}\end{aligned}$$

the upper bound to the error when a jump takes place in T_{fa} is therefore

$$\begin{aligned}&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_{fa}\Big) \\ &\quad \leq \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} = \hat{X}_{bk}, 1J \text{ in } T_{fa}\Big) \\ &\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T} = \hat{X}_{fk} \\ &\qquad\quad \hat{X}_{fk} \neq \hat{X}_{bk}, 1J \text{ in } T_{fa}\Big) \\ &\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T} = \hat{X}_{bk} \\ &\qquad\quad \hat{X}_{fk} \neq \hat{X}_{bk}, 1J \text{ in } T_{fa}\Big) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{fk}, X_k \neq \hat{X}_{bk}, 0J \text{ in } T_f \cup T_b) \\ &\qquad + \Pr(X_k \neq \hat{X}_{fk}, 0J \text{ in } T_f) \\ &\qquad + \Pr(X_k \neq \hat{X}_{bk}, 0J \text{ in } T_b) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{fk}, X_k \neq \hat{X}_{bk} \,|\, 0J \text{ in } T_f \cup T_b) \\ &\qquad + \Pr(X_k \neq \hat{X}_{fk} \,|\, 0J \text{ in } T_f) \\ &\qquad + \Pr(X_k \neq \hat{X}_{bk} \,|\, 0J \text{ in } T_b) = O(\epsilon^{1+\delta}).\end{aligned} \tag{27}$$

b) Jump in T_{ba}: The arguments are the same as those above. It can be seen that the error bound is

$$\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_{ba}\Big) = O(\epsilon^{1+\delta}). \tag{28}$$

c) Jump in T_f: Denote t_c as the true time of jump and $\hat{t}_c$ as its MAP estimate. When $\hat{X}_{fk} = \hat{X}_{bk}$, we have

$$\begin{aligned}&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} = \hat{X}_{bk}, 1J \text{ in } T_f\Big) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{bk}, 0J \text{ in } T_b) \\ &\quad \leq \Pr(X_k \neq \hat{X}_{bk} \,|\, 0J \text{ in } T_b) = O(\epsilon^{1+\delta}).\end{aligned} \tag{29}$$

Suppose now that $\hat{X}_{fk} \neq \hat{X}_{bk}$. We note also that $\Pr(X_{k-T} \neq \hat{X}_{fa})$ and $\Pr(X_{k+T} \neq \hat{X}_{ba})$ are $O(\epsilon^{1+\delta})$ since the state X is constant over T_{fa} and T_{ba} by our assumption. The next task is to determine whether the jump occurs in $T_f \cup T_b$ before or after k. That is, if the jump occurs in T_f, then $\hat{X}_{bk}$ is the more reliable estimate to use for $\hat{X}^{\text{sub}}_{k\,|\,k+T}$; conversely, if the jump occurs in T_b, then $\hat{X}_{fk}$ is the appropriate estimate to use for $\hat{X}^{\text{sub}}_{k\,|\,k+T}$. The error is then

$$\begin{aligned}
&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk}, 1J \text{ in } T_f\Big) \\
&\quad = \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk}, X_{k-T} = \hat{X}_{fa}, \\
&\qquad X_{k+T} = \hat{X}_{ba}, 1J \text{ in } T_f\Big) \\
&\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk} \\
&\qquad X_{k-T} \neq \hat{X}_{fa} \text{ or } X_{k+T} \neq \hat{X}_{ba}, 1J \text{ in } T_f\Big) \\
&\quad \leq \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk}, X_{k-T} = \hat{X}_{fa} \\
&\qquad X_{k+T} = \hat{X}_{ba}, 1J \text{ in } T_f\Big) \\
&\qquad + \Pr(1J \text{ in } T_f, X_{k-T} \neq \hat{X}_{fa} \text{ or } X_{k+T} \neq \hat{X}_{ba}) \\
&\quad \leq \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk}, \\
&\qquad X_{k-T} = \hat{X}_{fa}, X_{k+T} = \hat{X}_{ba}, 1J \text{ in } T_f\Big) \\
&\qquad + \Pr(X_{k-T} \neq \hat{X}_{fa} \,|\, 0J \text{ in } T_{fa}) \\
&\qquad + \Pr(X_{k+T} \neq \hat{X}_{ba} \,|\, 0J \text{ in } T_{ba}) \\
&\quad = \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk} \\
&\qquad X_{k-T} = \hat{X}_{fa}, X_{k+T} = \hat{X}_{ba}, 1J \text{ in } T_f\Big) + O(\epsilon^{1+\delta}).
\end{aligned} \tag{30}$$

The first term in (30) is, in fact, $O(T^{-1}) = O([\log(1/\epsilon)]^{-1})$ by appealing to Section IV-A and the definition of T. In other words, from (29) and (30)

$$\begin{aligned}
&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_f\Big) \\
&\quad \leq \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, \hat{X}_{fk} \neq \hat{X}_{bk} \\
&\qquad X_{k-T} = \hat{X}_{fa}, X_{k+T} = \hat{X}_{ba}, 1J \text{ in } T_f\Big) + O(\epsilon^{1+\delta}) \\
&\quad = \Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T} = \hat{X}_{fa} \neq X_{k+T}\hat{X}_{ba} \\
&\qquad = 1J \text{ in } T_f \cup T_b) + O(\epsilon^{1+\delta}).
\end{aligned} \tag{31}$$

d) Jump in T_b: The same arguments as the immediately prior case apply here. The bound is therefore

$$\begin{aligned}
&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_b\Big) \\
&\quad \leq \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T} = \hat{X}_{fa} \neq X_{k+T} = \hat{X}_{ba} \\
&\qquad 1J \text{ in } T_f \cup T_b) + O(\epsilon^{1+\delta}).
\end{aligned} \tag{32}$$

Using (27) and (28), when there is a single jump in $[k - 2T, k + 2T]$, the error rate can be written as

$$\begin{aligned}
&\Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J\Big) \\
&\quad = \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_f\Big) \\
&\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_b\Big) + O\left(\epsilon^2 \log \frac{1}{\epsilon}\right)
\end{aligned} \tag{33}$$

since the probability of a single jump in an interval of length T is $O(\epsilon T) = O(\epsilon \log(1/\epsilon))$. Consequently, from (25) and (26) and (31)–(33), the overall upper bound to the error for the two-state HMM is

$$\begin{aligned}
&\Pr(X_k \neq \hat{X}_{k\,|\,k+T}) \\
&\quad \leq \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}\Big) \\
&\quad = \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_f\Big) \\
&\qquad + \Pr\Big(X_k \neq \hat{X}^{\text{sub}}_{k\,|\,k+T}, 1J \text{ in } T_b\Big) + O\left(\epsilon^2 \log \frac{1}{\epsilon}\right) \\
&\quad \leq \{\Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T} = 1 \\
&\qquad X_{k+T} = 2, 1J \text{ in } T_f \cup T_b) \\
&\qquad + \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T} = 1 \\
&\qquad X_{k+T} = 2, 1J \text{ in } T_f \cup T_b)\} \\
&\qquad \times \Pr(X_{k-T} = 1, X_{k+T} = 2, 1J \text{ in } T_f \cup T_b) \\
&\qquad + \{\Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T} = 2 \\
&\qquad\quad X_{k+T} = 1, 1J \text{ in } T_f \cup T_b) \\
&\qquad\quad + \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T} = 2 \\
&\qquad\quad X_{k+T} = 1, 1J \text{ in } T_f \cup T_b)\} \\
&\qquad \times \Pr(X_{k-T} = 2, X_{k+T} = 1, 1J \text{ in } T_f \cup T_b) \\
&\qquad + O(\epsilon^{1+\delta}).
\end{aligned} \tag{34}$$

To generalize the upper bound in (34) to multistate HMMs, it is necessary to modify the definition of the data interval T since this is the part of the argument that explicitly relies on the states concerned. Let us consider T derived for specific states i, j. From (49), it is clear that for fixed ϵ, δ

$$\Pr(\text{False Alarm}) \leq \exp\left[-T \inf_{w_k > 0} I_{C_{\bullet i}}(w_k)\right] = \epsilon^{1+\delta}$$

where $w_k = \log(c_{y_k j}/c_{y_k i})$, and $I_{C_{\bullet i}}(\cdot)$ is the rate function evaluated with respect to the distribution $C_{\bullet i}$. It follows also that for fixed δ, if T is increased, the false alarm rate will decrease at a faster rate as $\epsilon \to 0$ than if a smaller data interval is used. Consequently, to ensure at least the same probability of false alarm for multistate HMMs as the two-state case, we will redefine T to be the largest of such intervals. First, let

$$\Lambda^*_{ij} = \sup_{\theta} \left[-\log \mathbb{E}_i \left(\frac{c_{y_k j}}{c_{y_k i}}\right)^{\theta}\right].$$

Then, the appropriate data length T is

$$T = \max_{ij} \frac{1+\delta}{\Lambda^*_{ij}} \log \frac{1}{\epsilon}$$

with $\delta > 0$ and arbitrary.

V. Overall Smoothing Error

By combining the results from Sections IV-B and IV-C, we now state the first main result of this paper.

Lemma V.1: Consider a discrete-time and discrete-state hidden Markov model with state variables X_k and observations Y_k, where the subscript k denotes time. The transition probability and observation matrices are parametrized as in (3) and (4), respectively. In addition, assume that the KL divergence (5) exists, and $K(i,j) > 0$ for all $i \neq j \in \{1,2,\ldots,N\}$.

Denote by $\Pr(X_k \neq \hat{X}_{k\,|\,k+T})$ the probability of smoothing error in optimally estimating the state of the HMM at time k using a smoothing lag of T, where

$$T = \max_{ij} \frac{1+\delta}{\Lambda^*_{ij}} \log \frac{1}{\epsilon}, \quad \text{and}$$

$$\Lambda^*_{ij} = \sup_{\theta} \left[-\log \mathbb{E}_i \left(\frac{c_{y_k j}}{c_{y_k i}} \right)^{\theta} \right].$$

In addition, denote by t_c the time at which the state makes a transition in the event that there is only a single transition over the interval $[k-T, k+T]$. As $\epsilon \to 0$, we have

$$\begin{aligned}
&\Pr(X_k \neq \hat{X}_{k\,|\,k+T}) \\
&\quad = \sum_i \sum_{j \neq i} \Pr(X_{k-T} = i, \\
&\qquad X_{k+T} = j, 1J \text{ in } T_f \cup T_b) \\
&\qquad \times \{\Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T} = i \\
&\qquad\quad X_{k+T} = j, 1J \text{ in } T_f \cup T_b) \\
&\qquad\quad + \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T} = i \\
&\qquad\quad X_{k+T} = j, 1J \text{ in } T_f \cup T_b)\} + O(\epsilon^{1+\delta}) \qquad (35)
\end{aligned}$$

for some $\delta > 0$ and where $T_f = [k-T, k]$ and $T_b = [k+1, k+T]$ and $\hat{t}_c$ denotes the MAP estimate of t_c.

Proof: The result of the lemma is self-evident from the lower and upper bounds derived in Sections IV-B and IV-C, respectively. ■

The second and related result is as follows.

Theorem V.1: Adopt the same assumptions as Lemma V.1. The asymptotic smoothing error as $\epsilon \to 0$ is

$$\Pr(X_k \neq \hat{X}_{k\,|\,k+T}) = \left(\sum_i \pi_i \sum_{j \neq i} \gamma(i,j) \lambda_{ji} \right) \epsilon (1 + o(1)) \qquad (36)$$

where $\Pi = (\pi_i)$ is the stationary distribution of the Markov chain, and the $\gamma(i,j)$ depend on just the conditional distributions $C_{\bullet i}$ and $C_{\bullet j}$.

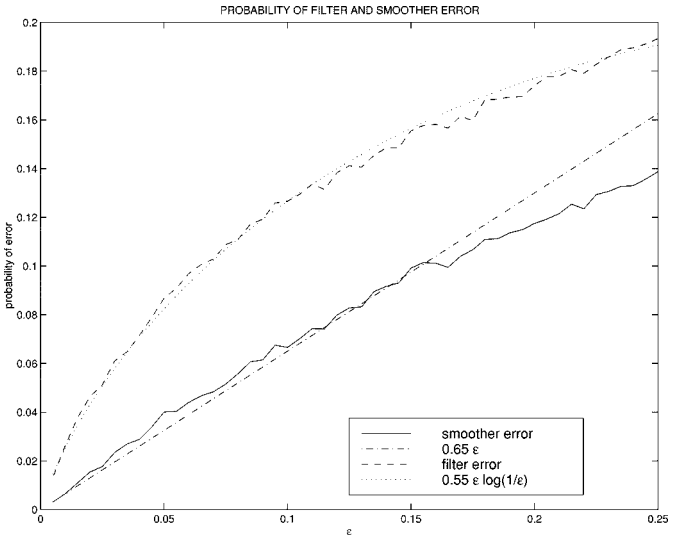

Fig. 5. Probability of filter and smoother error for small ϵ for matrices (39).

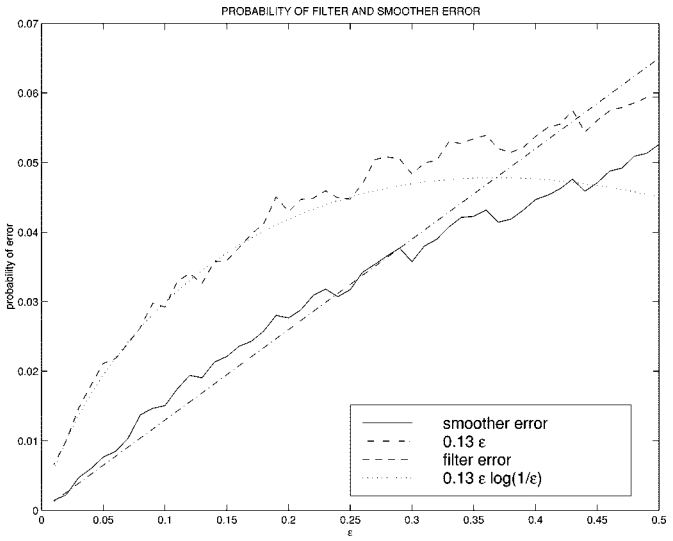

Fig. 6. Probability of filter and smoother error for small ϵ for matrices (40).

Proof: We will show first that (35) of Lemma V.1 actually implies (as $\epsilon \to 0$)

$$\begin{aligned}
&\Pr(X_k \neq \hat{X}_{k\,|\,k+T}) \\
&\quad = (1 + o(1)) \sum_i \sum_{j \neq i} \Pr(X_{k-T} = i \\
&\qquad X_{k+T} = j, 1J \text{ in } T_f \cup T_b) \\
&\qquad \times \{\Pr(\hat{t}_c \in T_b, t_c \in T_f \,|\, X_{k-T} = i \\
&\qquad\quad X_{k+T} = j, 1J \text{ in } T_f \cup T_b) \\
&\qquad\quad + \Pr(\hat{t}_c \in T_f, t_c \in T_b \,|\, X_{k-T} = i \\
&\qquad\quad X_{k+T} = j, 1J \text{ in } T_f \cup T_b)\}.
\end{aligned}$$

To prove this is so, we will establish that the double summation on the right side of (35) is $O(\epsilon)$. This involves a dissection of the products in each summand. Consider first terms such as

$\Pr(X_{k-T}=i, X_{k+T}=j, 1J \text{ in } T_f \cup T_b)$, which can be evaluated as

$$\begin{aligned}&\Pr(X_{k-T}=i, X_{k+T}=j, 1J \text{ in } T_f \cup T_b)\\ &\quad=\sum_{p=k-T}^{k+T-1}\Pr(X_{k-T}\ldots=X_p=i\\ &\qquad X_{p+1}\ldots=X_{k+T}=j)\\ &\quad=\sum_{\ell=1}^{2T}\{(1-\epsilon\lambda_{jj})^{2T-\ell}(1-\epsilon\lambda_{ii})^{\ell-1}\}\epsilon\lambda_{ji}\\ &\qquad\times\Pr(X_{k-T}=i).\end{aligned}$$

We will now show that this term is $\epsilon T(1+o(1))$ as $\epsilon\to 0$ (or, equivalently, $T\to\infty$). It is then immediate that the expression before the $O(\epsilon^{1+\delta})$ term on the right side of (35) is $O(\epsilon)$ since the other terms in the products making up each summand are $O(1/T)$ by Theorem IV.1.

Let us assume without loss of generality that $\lambda_{ii}\geq\lambda_{jj}$. It can be seen that

$$\begin{aligned}&2T(1-\epsilon\lambda_{ii})^{2T-1}\\ &\quad\leq\sum_{\ell=1}^{2T}\{(1-\epsilon\lambda_{jj})^{2T-\ell}(1-\epsilon\lambda_{ii})^{\ell-1}\}\\ &\quad\leq 2T(1-\epsilon\lambda_{jj})^{2T-1}.\end{aligned}$$

Recall that $T=[(1+\delta)/\Lambda^*]\log(1/\epsilon)$. In order to determine the behavior of $(1-\epsilon\lambda_{jj})^{2T-1}$ as $\epsilon\to 0$, we will examine the function

$$f(x)=\left(1-\frac{\lambda_{jj}}{x}\right)^{\frac{2(1+\delta)}{\Lambda^*}\log x-1} \tag{37}$$

by letting $\epsilon=1/x$. Since $\lim_{x\to\infty}(\log x)/x=0$ and $\lim_{x\to\infty}x\ln(1-\lambda_{jj}/x)=-\lambda_{jj}$, we have

$$\begin{aligned}&\lim_{x\to\infty}\ln f(x)\\ &\quad=\lim_{x\to\infty}\left(\frac{2(1+\delta)}{\Lambda^*}\log x-1\right)\ln\left(1-\frac{\lambda_{jj}}{x}\right)\\ &\quad=\lim_{x\to\infty}\left(\frac{2(1+\delta)\log x}{\Lambda^* x}-\frac{1}{x}\right)\left[x\ln\left(1-\frac{\lambda_{jj}}{x}\right)\right]=0.\end{aligned}$$

Consequently, $\lim_{\epsilon\to 0}(1-\epsilon\lambda_{jj})^{2T-1}=1$, and the same arguments apply to the term $(1-\epsilon\lambda_{ii})^{2T-1}$. As $\epsilon\to 0$, it can be seen that $\sum_{\ell=1}^{2T}\{(1-\epsilon\lambda_{jj})^{2T-\ell}(1-\epsilon\lambda_{ii})^{\ell-1}\}=2T(1+o(1))$, and

$$\begin{aligned}&\Pr(X_{k-T}=i, X_{k+T}=j, 1J \text{ in } [k-T, k+T])\\ &\quad=2\epsilon\lambda_{ji}\Pr(X_{k-T}=i)T(1+o(1)).\end{aligned}$$

The final step of the proof is accomplished by appealing to stationarity and substituting $\Pi=(\pi_i)$ and recalling that the probability of error in localizing the time of jump to the appropriate half interval is $\gamma(i,j)/2T$ when starting from state i and terminating at state j. ■

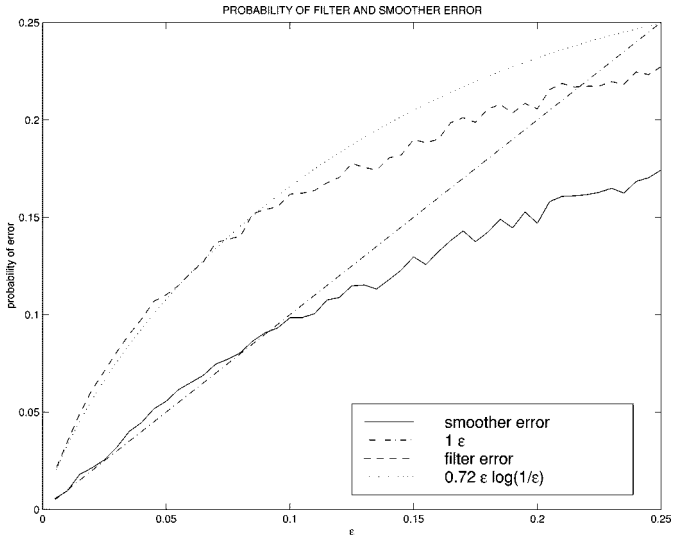

Fig. 7. Probability of filter and smoother error for small ϵ for matrices (41).

VI. Simulations

A series of simulations have been carried out, principally to demonstrate the order of magnitudes in the filtering and smoothing error probabilities for HMMs. It is seen that in general there is close agreement between theoretical predictions and experimental results.

In the simulations, we have used a two-state Markov chain with two possible discrete observations. We used the following A, C combination as the basic system and alternatively varied the A and C matrices independently. For the first set of results (Figs. 4 and 5), the system matrices are

$$A=\begin{bmatrix}1-0.2\epsilon & 0.4\epsilon\\ 0.2\epsilon & 1-0.4\epsilon\end{bmatrix}\quad C=\begin{bmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{bmatrix}. \tag{38}$$

This was followed by varying the A matrix (see Figs. 6 and 7) as

$$A=\begin{bmatrix}1-0.45\epsilon & 0.4\epsilon\\ 0.45\epsilon & 1-0.4\epsilon\end{bmatrix}\quad C=\begin{bmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{bmatrix} \tag{39}$$

$$A=\begin{bmatrix}1-0.32\epsilon & 0.04\epsilon\\ 0.32\epsilon & 1-0.04\epsilon\end{bmatrix}\quad C=\begin{bmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{bmatrix}. \tag{40}$$

Last, the C matrix was varied while keeping A constant

$$A=\begin{bmatrix}1-0.2\epsilon & 0.4\epsilon\\ 0.2\epsilon & 1-0.4\epsilon\end{bmatrix}\quad C=\begin{bmatrix}0.8 & 0.4\\ 0.2 & 0.6\end{bmatrix}. \tag{41}$$

For each combination of A, C, i.e., at a fixed ϵ, ten sets of state and observation sequences were generated, with 10 000 data points in each set. The smoothing lag chosen is 150, which was observed to be sufficient for the range of ϵ used in the present simulations to obtain the least smoothing error. The results presented are the averages of the smoothed estimates for each set.

VII. Conclusion

In this paper, we have extended the approach taken by Khasminskii *et al.* [7] to derive the asymptotic smoothing error as $\epsilon\to 0$ via a series of hypothesis testing problems. It is seen

that as $\epsilon \to 0$, the smoothing error is $O(\epsilon)$, compared with $O(\epsilon \log(1/\epsilon))$ for the filtering error. This also means that as $\epsilon \to 0$, the relative improvement of smoothing over filtering is $O(\log(1/\epsilon))$, which can be significant for small ϵ. These theoretical predictions were observed in our simulations.

APPENDIX A
LARGE DEVIATION RESULTS

In this section, we summarize certain key results from large deviations theory (see also [10] and [11]).

For a scalar random variable x with distribution $p(x)$, let us define the following:

Definition A.1 (Moment Generating Function):

$$M_p(\theta) = \mathbb{E}_p\{\exp[\theta x]\} \tag{42}$$

where $\mathbb{E}_p$ denotes expectation with respect to $p(x)$.

Definition A.2 (Large Deviation Rate Function):

$$I_p(x) = \sup_\theta [\theta x - \log M_p(\theta)]. \tag{43}$$

Some key properties:

1) $I_p(x)$ is convex.
2) $\min_x I_p(x) = I_p(m) = 0$, where $m = \mathbb{E}_p(x)$.

Theorem A.1 (Cramér's Theorem): Let S_n denote the sample average of n i.i.d. random variables x_i, $i = 1, 2, \ldots, n$, each x_i distributed according to $p(x)$, that is

$$S_n = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

If the moment generating function $M_p(\theta) < \infty$ is defined for all θ, then for every closed subset $F \subset \mathbb{R}$

$$\lim_{n\to\infty} \sup \frac{1}{n} \log \Pr(S_n \in F) \le - \inf_{x\in F} I_p(x) \tag{44}$$

and every open subset $G \subset \mathbb{R}$

$$\lim_{n\to\infty} \inf \frac{1}{n} \log \Pr(S_n \in G) \ge - \inf_{x\in G} I_p(x) \tag{45}$$

with $M_p(\theta)$ and $I_p(x)$, as defined in (42) and (43) for $p(x)$.

Remark A.1: For a sequence of i.i.d. random variables, (44) actually holds for every n and not just for n large enough, i.e., for $n > 0$, $\Pr(S_n \in [a, b]) \le \exp[-nI_p(a)]$ when $a > \mathbb{E}_p[x]$, and $\Pr(S_n \in [a, b]) \le \exp[-nI_p(b)]$ for the case $b < \mathbb{E}_p[x]$. This upper bound is known as the Chernoff bound in communications theory; see also ([10, ch. 7]).

For the following results, let us consider a binary hypothesis testing problem:

H_0: $\{x_1, x_2, \ldots, x_n\}$ i.i.d. sequence, with density $q(x)$;
H_1: $\{x_1, x_2, \ldots, x_n\}$ i.i.d. sequence, with density $p(x)$.

Denote $\alpha_n = \Pr(\text{Declare } H_1 \,|\, H_0)$ as the probability of a false alarm and $\beta_n = \Pr(\text{Declare } H_0 \,|\, H_1)$ as probability of a miss. In a Neyman–Pearson test, the observed normalized log-likelihood ratio is compared with a threshold γ_n

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i)}{q(x_i)} = \begin{cases} >\gamma_n & \text{decide } H_1 \\ \le \gamma_n & \text{decide } H_0. \end{cases}$$

Note that if $q(\cdot)$ is the true distribution, then

$$\lim_{n\to\infty} S_n = \mathbb{E}_q \log \frac{p(\cdot)}{q(\cdot)} = -K(q, p)$$
$$\lim_{n\to\infty} S_n = \mathbb{E}_p \log \frac{p(\cdot)}{q(\cdot)} = K(p, q)$$

if $p(\cdot)$ is the true distribution, where the expectation is taken over the appropriate densities.

The exponential rates of α_n and β_n for the Neyman–Pearson test with a constant threshold γ is given by (see [11, ch. 3]) the follwing theorem.

Theorem A.2: The Neyman–Pearson test with a constant threshold $\gamma \in (-K(q, p), K(p, q))$ independent of n satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = -I_q(\gamma) < 0 \tag{46a}$$
$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -I_p(\gamma) = \gamma - I_q(\gamma) < 0 \tag{46b}$$

where $I_q(\gamma)$ and $I_p(\gamma)$ are the large deviation rate functions with respect to the distributions $q(\cdot)$ and $p(\cdot)$, respectively, evaluated at the threshold γ.

Let us now consider the Bayes probability of error, which is defined as [see also (9)] $P_n \stackrel{\text{def}}{=} \alpha_n \Pr(H_0) + \beta_n \Pr(H_1)$. For the best achievable Bayes error, we have ([11, , ch. 3]) the following theorem.

Theorem A.3: Consider a hypothesis testing problem consisting of hypotheses H_0 and H_1 with the Bayes probability of error defined above. If $0 < \Pr(H_0) < 1$, then

$$\inf_S \lim_{n\to\infty} \frac{1}{n} \log P_n = -I_q(0) \tag{47}$$

where $I_q(0) = (\sup_\theta [\theta x - \log M_q(\theta)])_{x=0} = -\log M_q(0)$. In other words, the best asymptotic Bayes error is achieved by a Neyman–Pearson test with threshold $\gamma = 0$.

A. Explanation for T

In this section, we will apply Cramér's theorem and show that the definition of T in [7] is related to maintaining the false alarm rate to be $O(\epsilon^{1+\delta})$, $\delta > 0$ for a simple hypothesis problem, which we review now.

Given a set of observations $\mathcal{Y} = \{y_1, y_2, \ldots, y_T\}$, let us consider the following hypothesis testing problem.

H_0: $\{y_1, y_2, \ldots, y_T\}$ is a sequence of i.i.d. data, each of law $C_{\bullet i} = \{\Pr(Y_k = \ell \,|\, X_k = i)\}$;
H_1: $\{y_1, y_2, \ldots, y_T\}$ is a sequence of i.i.d. data, each of law $C_{\bullet j} = \{\Pr(Y_k = \ell \,|\, X_k = j)\}$;

where $k \in \{1,2,\ldots,T\}$, and $\ell \in \{1,2,\ldots,M\}$. Denote the log-likelihood ratio as

$$Z_T = \sum_{k=1}^{T} \log \frac{c_{y_k j}}{c_{y_k i}}. \tag{48}$$

From Theorem A.3, it is seen that to achieve the best Bayes error the threshold is 0; hence, if $Z_T < 0$, then H_0 will be declared, and if $Z_T > 0$, the alternative hypothesis H_1 will be declared instead. We further assume that H_1, i.e., state j, is the hypothesis we wish to detect.

1) Fixing False Alarm Rate: The false alarm rate $\Pr(Z_T > 0 \,|\, H_0)$ can be set to a desired level, say $\Pr(Z_T > 0 \,|\, H_0) \leq \epsilon^{1+\delta}$, $\delta > 0$ and ϵ small. By subsequently applying the Chernoff bound (Remark A.1), a sufficient data length $T(\epsilon)$ can be determined such that this error probability is attained. That is

$$\Pr(Z_T > 0 \,|\, H_0) \leq \exp\left[-T \inf_{w_k > 0} I_{C_{\bullet i}}(w_k)\right] = \epsilon^{1+\delta} \tag{49}$$

with $I_{C_{\bullet i}}(\cdot)$ the rate function, evaluated with respect to the distribution $C_{\bullet i}$, and $w_k = \log(c_{y_k j}/c_{y_k i})$. Finally, using the convexity of the rate function (see Fig. 3) and the fact that $\mathbb{E}_i \log(c_{y_k j}/c_{y_k i}) = -K(i,j) < 0$, we have

$$\inf_{w_k > 0} I_{C_{\bullet i}}(w_k) = I_{C_{\bullet i}}(0) = \sup_{\theta} \left[-\log \mathbb{E}_i \left(\frac{c_{y_k j}}{c_{y_k i}}\right)^{\theta}\right] = \Lambda^*. \tag{50}$$

Hence

$$T = \frac{(1+\delta)}{\Lambda^*} \log \frac{1}{\epsilon}. \tag{51}$$

Appendix B
Detection of Location of Jump

Suppose that at time $k-T$, the Markov chain is in state i, and at a later unknown time t_c, the state changes to state j, $t_c \in [k-T+1, k+T]$. Using Definition IV.1, we have

$$S(t) = \sum_{\ell=k-T}^{t-1} \log \frac{c_{y_\ell i}}{c_{y_\ell i}} + \sum_{\ell=t}^{k+T} \log \frac{c_{y_\ell j}}{c_{y_\ell i}} = \sum_{\ell=t}^{k+T} \log \frac{c_{y_\ell j}}{c_{y_\ell i}}. \tag{52}$$

Now, a necessary but not sufficient condition for estimating $\hat{t}_c > t_c$ is $S(\hat{t}_c = t_c + J) > S(t_c)$ for all $t_c + J \leq k+T$. For $J = 1$, this amounts to the inequality

$$\Pr(\hat{t}_c = t_c + 1) \leq \Pr(S(t_c+1) > S(t_c)) = \Pr\left(\log \frac{c_{y_{t_c} j}}{c_{y_{t_c} i}} < 0\right). \tag{53}$$

We note that since the true distribution is $C_{\bullet j}$ over $[t_c, t_c+J-1]$, $\mathbb{E}_j \log(c_{y_k j}/c_{y_k i}) = K(j,i) > 0$. Now, using the Chernoff bound (see Appendix A) for the general case $k-T < t_c + J \leq k+T$, we have

$$\begin{aligned}\Pr(\hat{t}_c = t_c + J) &\leq \Pr(S(t_c+J) > S(t_c)) \\ &= \Pr\left(\frac{1}{J} \sum_{\ell=t_c}^{t_c+J-1} \log \frac{c_{y_\ell j}}{c_{y_\ell i}} < 0\right) \\ &\leq \exp\left[-J I_{C_{\bullet j}}(0)\right]\end{aligned} \tag{54}$$

where $w_\ell = \log(c_{y_\ell j}/c_{y_\ell i})$, $I_{C_{\bullet j}}(0) = \inf_{w_\ell < 0} \sup_\theta [\theta w_\ell - \log M_{C_{\bullet j}}(\theta)]$, using convexity of the rate function.

The analogous condition for $k - T \leq t_c - J < k + T$ is

$$\begin{aligned}\Pr(\hat{t}_c = t_c - J) &\leq \Pr(S(t_c-J) > S(t_c)) \\ &= \Pr\left(\frac{1}{J} \sum_{\ell=t_c-J}^{t_c-1} \log \frac{c_{y_\ell j}}{c_{y_\ell i}} > 0\right) \\ &\leq \exp\left[-J I_{C_{\bullet i}}(0)\right]\end{aligned} \tag{55}$$

where $w_\ell = \log(c_{y_\ell j}/c_{y_\ell i})$, $I_{C_{\bullet i}}(0) = \inf_{w_\ell > 0} \sup_\theta [\theta w_\ell - \log M_{C_{\bullet i}}(\theta)]$. In this situation, the observations are of law $C_{\bullet i}$ on $[t_c - J, t_c - 1]$, and hence, $\mathbb{E}_i \log(c_{y_\ell j}/c_{y_\ell i}) = -K(i,j) < 0$.

Appendix C
Inequality for Calculation of Lower Bound

In this section, we will prove the relationship

$$\Pr(\hat{t}_c < t_c \,|\, t_c = k+1) > \alpha$$

where $\alpha > 0$ is some constant independent of T as $T \to \infty$. This inequality is crucial in determining a lower bound to the detection error in Section IV-A.

Using (54) and (55), we have for arbitrarily large but fixed T and D, $D < T$

$$\Pr(|\hat{t}_c - t_c| \geq D \,|\, t_c = k+1) \leq \sum_{j=D}^{T} 2\exp[-jI^*] \tag{56}$$

where $I^* = \min(I_{C_{\bullet j}}(0), I_{C_{\bullet i}}(0))$. Since (56) is bounded as $T \to \infty$, there obviously exists an D (independent of T) such that

$$\Pr(|\hat{t}_c - t_c| < D \,|\, t_c = k+1) \geq 0.9. \tag{57}$$

We will now focus on the conditional probability

$$\Pr(\hat{t}_c < t_c \,|\, t_c = k+1, |\hat{t}_c - t_c| < D)$$

and define $S_M(t) = \sum_{\ell=t}^{k+D-1} \log(c_{y_\ell j}/c_{y_\ell i}) = \sum_{\ell=t}^{k+D-1} \xi_\ell$. As $|\hat{t}_c - t_c| < D$ by assumption, we then have

$$\hat{t}_c = \arg\max_t S_M(t) = \arg\max_t S(t).$$

Since ξ_ℓ is an i.i.d. random variable and $\Pr(\xi_\ell > 0) > 0$ and $\Pr(\xi_\ell \leq 0) > 0$ for $\ell = k-D+1, k-D+2, \ldots, k+D-1$ because $K(i,j) = -\mathbb{E}_i \xi_\ell = -\mathbb{E}_i \log(c_{y_\ell j}/c_{y_\ell i}) > 0$,

$K(j,i) = \mathbb{E}_j \xi_\ell > 0$ and $\sum_m c_{mn} = 1$, there exists, with nonzero probability, an event Γ such that

$$\Gamma = \{\xi_{k-D+1} \le 0, \ldots, \xi_{t_c-2} \le 0 \quad \xi_{t_c-1} > 0, \ldots, \xi_{k+D-1} > 0\}.$$

It is clear that under this event, $\hat{t}_c = \arg\max S_M(t) = t_c - 1$. In other words

$$\begin{aligned} &\Pr(\hat{t}_c < t_c \,|\, t_c = k+1, |\hat{t}_c - t_c| < D) \\ &= \sum_{i=k-D+2}^{k} \Pr(\hat{t}_c = i \,|\, t_c = k+1, |\hat{t}_c - t_c| < D) \\ &\ge \Pr(\hat{t}_c = t_c - 1 \,|\, t_c = k+1, |\hat{t}_c - t_c| < D) \\ &\ge \Pr(\Gamma)(> 0). \end{aligned} \tag{58}$$

Consequently, from (57) and (58), we have

$$\Pr(\hat{t}_c < t_c \,|\, t_c = k+1) \ge 0.9 \Pr(\Gamma) = \alpha$$

where α is a constant independent of T.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[2] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, 2nd ed. New York: Springer-Verlag, 1994.

[3] L. Shue, B. D. O. Anderson, and S. Dey, "Exponential stability of filters and smoothers for hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 2180–2194, Aug. 1998.

[4] D. Clements and B. D. O. Anderson, "A nonlinear fixed-lag smoother for finite-state Markov processes," *IEEE Trans. Inform. Theory*, vol. IT–21, pp. 446–452, 1975.

[5] B. D. O. Anderson and S. Chirarattananon, "Smoothing as an improvement on filtering: A universal bound," *Electron. Lett.*, vol. 7, no. 18, pp. 524–525, 1971.

[6] J. B. Moore and K. L. Teo, "Smoothing as an improvement on filtering in high noise," *Syst. Contr. Lett.*, vol. 8, no. 1, pp. 51–54, 1986.

[7] R. Khasminskii and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stoch. Process. Their Appl.*, vol. 63, no. 1, pp. 1–10, 1996.

[8] G. Golubev and R. Khasminskii, "Asymptotic optimal filtering for a hidden Markov model," *Math. Methods Statist.*, vol. 7, no. 2, pp. 192–209, 1998.

[9] B. D. O. Anderson and T. Kailath, "Forward and backward models for finite-state Markov processes," *Adv. Appl. Prob.*, vol. 11, no. 1, pp. 118–133, 1979.

[10] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 1990.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1988.

Louis Shue was born in Vietnam in 1971. He received the B.Sc. degree in 1994 and the B.E. degree in 1996 (both with first-class honors) from Monash University, Australia, and the Ph.D. degree in systems engineering from the Australian National University, Canberra.

He is currently a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. His current research interests include speech and image processing, small target tracking in infrared imagery, and hidden Markov models.

Brian D. O. Anderson (S'62–M'66–SM'74–F'75) was born in Sydney, Australia, and received his undergraduate education at the University of Sydney, with majors in pure mathematics and electrical engineering. He received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

Following completion of his education, he worked in industry in Silicon Valley and served as a faculty member in the Department of Electrical Engineering at Stanford. He was Professor of electrical engineering at the University of Newcastle, Callaghan, Australia, from 1967 to 1981 and is now Professor of Systems Engineering at the Australian National University, Canberra, and Director of the Research School of Information Sciences and Engineering. His interests are in control and signal processing.

Dr. Anderson is a Fellow of the Royal Society, the Australian Academy of Science, and the Australian Academy of Technological Sciences and Engineering. He is an an Honorary Fellow of the Institution of Engineers, Australia. He holds doctorates (honoris causa) from the Universite Catholique de Louvain, Belgium, Swiss Federal Institute of Technology, Zurich, the University of Sydney, and the University of Melbourne. He served a term as President of the International Federation of Automatic Control from 1990 to 1993 and is currently President of the Australian Academy of Science.

Franky De Bruyne was born in Deinze, Belgium, in 1969. He received the B.E.E. and Ph.D. degrees form the Université Catholique de Louvain, Louvain la Neuve, Belgium, in 1992 and 1996, respectively.

He was a Research Fellow with the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, from 1996 to 1999. Currently, he is a Research Engineer with the Pulp and Paper Group, Siemens Belgium, Huizingen. His main interests are modeling and modeling for control design.