Proceedings of the 34th
Conference on Decision & Control
New Orleans, LA - December 1995

Geometric convergence of filters for hidden Markov models

Rene K. Boel
Vakgroep Elektrische Energietechniek, Universiteit Gent
Technologiepark-Zwijnaarde, B-9052 Ghent, Belgium
Tel:+32-9-2645658 Fax:+32-9-2645840 Email:Rene.Boel@rug.ac.be

John B. Moore and Subhrakanti Dey
Department of Systems Engineering
Research School of Information Sciences and Engineering
Australian National University, Canberra ACT 0200, Australia

Abstract

Hidden Markov models have proved suitable for many interesting applications which can be modelled using some unobservable finite state Markov process, influencing measured signals. This can be used to describe bursty telecommunications traffic, or the faults in a complicated systems, for modelling the activity in neurons, for modelling speech patterns, etc. In all these applications, one has to estimate the unobservable underlying state of the Markov process, using the observed signals. Optimal recursive filters are well known for this estimation problem. Recently risk sensitive filters for the same problem have also been obtained. An important question in studying the quality of such filters is the rate at which arbitrarily assigned initial conditions are forgotten. In this paper we show that the effect of initial conditions on these filters dies out geometrically fast under very reasonable observability assumptions. The proof is given in the simplest case of finite state space and of a finite, quantised, observations space. However the method can be extended to more general models by continuity arguments.

1 Introduction

Estimation of the state sequence of a finite-state Markov chain observed under memoryless noise, given the observations, is known as the standard hidden Markov model (HMM) estimation problem. The filters used in practical applications to achieve this task will have to work in real time, in a recursive way. In [6] the optimal filter - in the sense of minimum mean squared error, is derived and shown to be of a recursive form, consisting of a state transition update and a measurement update for the conditional distribution of the state of the hidden Markov model. The conditional distribution of the state at time $k+1$ depends on the conditional distribution of the state at time k and on the new observation at time $k+1$. To start this filter, one must make an initial guess at time 0. One important property of a good filter is that the effect of this initial guess dies out geometrically fast. This insures that a mistake in the initial guess will not make the filter give wrong results forever. Intuitively this also guarantees that the effect of round-off errors, outliers, or any other causes of failure of the filter at a small number of points in time, will not corrupt the estimation results forever. The proof of this geometric convergence property is the goal of this paper.

In this paper, we do not try to solve this problem in its general form. Rather we try to give a very simple, in fact, trivial, proof of this property for the optimal filter for the case where the state space is finite, and the measurements are quantised as elements of a finite set. First of all we note that the conditional distributions form a Markov process (*Kunita* [4]). Moreover this Markov process is irreducible in a certain sense, provided the hidden Markov process is irreducible and the observations are informative, i.e. no two states generate the same probability over the set of observations. Next we show that the update of the conditional distribution of the state at time k to the conditional distribution of the state at time $k+1$ is just a multiplication by the transition operator, a matrix which has all eigenvalues strictly inside the unit circle, except for a single eigenvalue 1. The eigenvalue 1 simply expresses the fact that the conditional distribution is a vector of terms summing to 1 before and after the update. Within the subset of normalised vectors, the state update is actually a strict contraction. Finally the conditional probability of the state at time $k+1$, given the observed signal up to time k is updated by using the new measurement Y_{k+1}. We show that on the average this update operator is a contraction operator.

A new class of robust filters known as risk-sensitive filters has been developed in [9] [7]. In [7], it has been developed for hidden Markov models with continuous-range observation space. In this paper, we derive the risk-sensitive filter for hidden Markov models with finite-discrete observation space. Stability results are also derived for these filters. Also, smoothing filters for HMM show the same exponential forgetting. In each of these cases one assumes that the correct parameters of the hidden Markov model are known, and used in the design of the filter. The meth-

0-7803-2685-7/95 $4.00 © 1995 IEEE

ods to prove the result can also be applied to filters when these parameters are not known, but are replaced by some a priori estimate or where they are estimated on-line. Then the conditional distribution is no longer a Markov process. However one can show that the hidden Markov state, the conditional distribution (and in the case of an adaptive estimator the parameter estimate) together form a Markov process which is geometrically ergodic under certain reasonable conditions.

In the next section of this paper we introduce the model, and the optimal filter, in detail. In section 3 we prove the geometric convergence. In section 4, we derive the risk-sensitive recursive estimates and the optimizing risk-sensitive filter for hidden Markov models with finite-discrete states and finite-discrete observations. We also present the ergodicity results for these recursive estimates.

2 Hidden Markov model

Consider a first-order homogeneous finite-state Markov process X_k with state space $\mathcal{S}_X$, and transition matrix A defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Also define $\mathcal{F}_k^0 \triangleq \sigma(X_0, \ldots, X_k)$ and the corresponding complete filtration $\{\mathcal{F}_k\}$. Without loss of generality we can take $N = \sharp\mathcal{S}_X$ and denote the N elements of $\mathcal{S}_X$ by $e_n = (0 \ldots 0\ 1\ 0 \ldots 0)'$, the unit vectors in $\mathbf{R}^N$. The elements of the transition matrix

$$a_{ij} = \mathcal{P}(X_{k+1} = e_i \mid X_k = e_j), \quad i, j \in \{1, 2, \ldots, N\}$$

represent the probability of reaching the state e_i at time $k + 1$ given that the state at time k is e_j. Of course, $a_{ij} > 0$ and $\sum_{i=1}^N a_{ij} = 1$, $\forall j \in \{1, \ldots, N\}$. This leads to the following simple representation of the process X_k (see [11], [12], [6]):

$$X_{k+1} = AX_k + V_{k+1} \tag{1}$$

where $E(V_{k+1} \mid \mathcal{F}_k) = 0$, i.e. V_k is a $(\mathcal{P}, \mathcal{F}_k)$-martingale increment sequence. The linear form of (1) is simply a result of the special structure of the space $\mathcal{S}_X$: on the set $\{0, 1\}$ all functions are linear functions. To keep the derivation simple in the next section we assume that the Markov process X_k with transition matrix A is aperiodic and irreducible (see e.g. Meyn and Tweedie [5])

At each time instant k a signal Y_k is observed. This observation takes values in the finite space $\mathcal{S}_Y = \{f_1, f_2, \ldots, f_M\}$ where without loss of generality the values f_m can be represented as a unit vector (all components 0 except for a single 1) in $\mathbf{R}^M$. Below we will abuse notation and sometimes write $Y_k = m$ when we mean $Y_k = f_m$. The value of the random variable Y_k depends on the state X_k and on a noise term W_k. This dependence can be expressed as follows

$$Y_k = CX_k + W_k \tag{2}$$

where W_k forms an independent increment sequence, and the elements of the matrix C are defined as the conditional probabilities

$$c_{mn} = \mathcal{P}(Y_k = f_m \mid X_k = e_n)$$

It is obvious that $c_{mn} > 0$, and $\sum_{m=1}^M c_{mn} = 1$.

Define $\{\mathcal{Y}_k\}$ to be the complete filtration generated by $\sigma(Y_0, \ldots, Y_k)$ and $\{\mathcal{G}_k\}$ to be the complete filtration generated by $\mathcal{G}_k^0 \triangleq \sigma(X_0, \ldots, X_k, Y_0, \ldots, Y_{k-1})$. It is easy to see that $E[Y_k \mid X_k] = CX_k$ and hence $E[W_k \mid \mathcal{G}_k] = 0$, so that W_k is a $(\mathcal{P}, \mathcal{G}_k)$- martingale increment. In order to calculate the conditional expectation $E(f(X_k) \mid \mathcal{Y}_k)$ for any function f, we need the N-vector $\hat{p}_k$ representing the conditional distribution of the state X_k given the observations $Y_0, Y_1, \ldots, Y_k$:

$$\hat{p}_k(i) = \mathcal{P}(X_k = e_i \mid \mathcal{Y}_k) \tag{3}$$

It is known (see e.g. *Elliott et al.* [6]) that these conditional distributions can be calculated recursively, i.e. $\hat{p}_{k+1} = F(A.\hat{p}_k, Y_{k+1})$ for a well defined function F.

This recursive transformation consists of two steps. The function F transforms the conditional probability $A.\hat{p}_k$ of X_{k+1} given the observations up to time k, into the conditional probability of the same state X_{k+1} given observations up to time $k + 1$. This measurement update is given explicitly by:

$$F(q, Y_{k+1}) = \frac{diag(C_{Y_{k+1},\cdot}) \,.\, q}{\sum_{n=1}^N C_{Y_{k+1},n} . q(n)} \tag{4}$$

Note that this measurement update consists of a linear transformation $C_{Y_{k+1},n}.q(n)$ (the numerator of (4)), followed by a normalisation, expressed by the denominator.

The second step of the recursive transformation calculates the conditional distribution of X_{k+1} from the conditional distribution of X_k, given the same set of observations $Y_0, Y_1, \ldots, Y_k$. This is simply a multiplication by the transition matrix A. Since A is a stochastic matrix the updated probability $\hat{p}_{k+1}$ is automatically normalised, if $\hat{p}_k$ is normalised to one. This step is therefore a linear transformation.

We assume now that the elements of the matrices A and C are known. Then *Kunita* [4] has shown that the stochastic process $\hat{p}_k$ is a Markov process, with as state space the subset of $\mathbf{R}^N$ of normalised vectors: $\mathcal{S}_p = \{q \in \mathbf{R}^N, \sum_{n=1}^N q(n) = 1\}$. The transition probabilities for $\hat{p}_k$ can be written down explicitly, since there are only $\mathbf{M}$ possible values in the continuous space $\mathbf{R}^N$

which $\hat{p}_{k+1}$ can take, given a value of $\hat{p}_k$. These M values correspond to the M values which Y_{k+1} can take.

$$\hat{p}_{k+1}=\frac{diag(c_{mn}).A.\hat{p}_k}{\sum_{n=1}^{N} c_{mn}.A.\hat{p}_k(n)} \qquad (5)$$

with probability $\mathcal{P}(Y_{k+1}=f_m \mid \hat{p}_k)=(C.A.\hat{p}_k)(m)=\sum_{n,l} c_{mn}.a_{nl}\hat{p}_k(l)$.

The initial condition for the recursive filter is given by the initial value $\hat{p}_0$ of this Markov process. The question whether the filter forgets initial conditions geometrically fast is thus equivalent to the question whether $\hat{p}_k$ is a geometrically ergodic Markov process.

Clearly the state space $\mathcal{S}_p$ cannot be completely ergodic. Let the conditional distribution after a measurement update be such that we are (almost) certain that the state is e_n. After the next state update this becomes $A.e_n=A_{,n}$, the n-th column of A. Whatever the initial distribution $\hat{p}_0$, after one step the state $\hat{p}_k$ will be inside the convex hull $co(A)=\{\sum_n \lambda_n A_{,n} \mid \sum_n \lambda_n=1\}$. All states $\hat{p}_k$ outside this convex hull are transient, and physically not meaningful. Hence it makes sense to limit the state space $\mathcal{S}_p$ to its subset $F(co(A))$, the set of points reachable form a vector inside $co(A)$ after one measurement update step. This reduced state space will be assumed from now on.

3 Geometric ergodicity of the estimator

Before checking the geometric ergodicity of the Markov process $\hat{p}_k$, we first have to see whether the process is irreducible. Since the state space is continuous, it cannot be irreducible in the sense of reaching any state from any other state, with positive probability. However consider any open subset O within the reduced state space $\mathcal{S}_p$, and any initial state $\hat{p}_0$. When all the columns of C are different, there exists a finite k and a sequence of observations $Y_0, Y_1, \ldots, Y_k$ which occurs with non-zero probability, such that $\hat{p}_k$ is in the open set O. This is called forward accessibility in [5], or strong irreducibility [2]. Since the state space $\mathcal{S}_p$ is compact, this forward accessibility essentially guarantees ergodicity of the Markov process $\hat{p}_k$. However as explained in the introduction we need that the estimate only depends the most recent observations $Y_0, Y_1, \ldots, Y_k$ in order to obtain a good filter. To prove this we need to show geometrically fast forgetting.

Remark 1 Even if A is not irreducible, the Markov process $\hat{p}_k$ may have distributions converging to an equilibrium distribution independent of the initial distribution. Consider as an example the case where $A=I$, i.e. the underlying state remains constant. We are then actually using the HMM filter as an identifier. It is known that the estimate converges w.p. 1 to the correct state as soon as the columns of C are all different.

Consider now the effect of the state transition step. The multiplication of the intermediate probability $\mathcal{P}(X_k \mid Y_0, Y_1, \ldots, Y_k, Y_{k+1})$ by the matrix A has as effect that the probability is coming closer towards the equilibrium distribution $\pi=A.\pi$ of the Markov process X_k. The eigenvalue 1 of the matrix A has as left eigenvector the vector with all 1's, as right eigenvector the equilibrium distribution. This insures that the sum of the elements of the distribution is always normalised (sums to 1). All the other eigenvalues are strictly less than 1 in absolute valuc by thc Frobenius theorem [3] for positive matrices. Hence within $\mathcal{S}_p$ the distance between two distributions $\mid A.p - A.\tilde{p} \mid$ after a state transition update is strictly less than the distance $\mid p-\tilde{p} \mid$ between the vectors before the update. This state transition update is a strict contraction operator.

Consider now the measurement update step $F(Y_k, q)$. We have to show that the distance between two (conditional) distributions gets reduced, on the average, by this transformation. Given the conditional distribution q the distribution of the observations Y_{k+1} is $\mathcal{P}(Y_{k+1}=f_m \mid q)=\sum_{n=1}^{N}\sum_{\ell=1}^{N} c_{mn}.a_{n\ell}q(\ell)=(C.A.q)(m)$. This is exactly the normalising factor in the denominator of the measurement update equation. However, it is still difficult to calculate $E \mid F(Y_{k+1}, p)-F(Y_{k+1}, \tilde{p}) \mid$ and compare it to $\mid p-\tilde{p} \mid$, because two different normalising factors are involved. What can be calculated is the change in normed distance when $\tilde{q}=q+\delta q$, i.e. the effect of a small perturbation. To obtain this difference calculate the average Jacobian with respect to q of the transformation $F(Y_{k+1}, q)$, and take the average over all possible values of Y_{k+1}, given q. This derivation leads to the following expression for the (i,j)-th element:

$$\sum_m [c_{mi}.\delta_{ij} - \frac{c_{mi}.c_{mj}.q(i)}{\sum_l c_{ml}.q(l)}]$$

In matrix form this gives an identity matrix minus a complicated matrix.

To get some insight in the conditions for this matrix to be contracting, take first the case of a hidden Markov model with only two unobservable states $\mathcal{S}_X=\{e_1, e_2\}$, and two signal values ($M=2$). Then the normalised conditional distribution can be written as $\hat{p}_k=(\hat{p}_k(1), 1-\hat{p}_k(1))'$. Hence the state space for the Markov process of conditional distributions can be reduced to a subset of the interval $[0, 1]$. Both the state transition update and the measurement update can be reduced to one-dimensional recursions, and the Jacobian to be calculated reduces to a simple derivative (rewriting F as a function of $\hat{p}_k(1)$ only). The derivative is explicitly calculated as

$$E(\Delta)=1-\frac{(c_{11}-c_{12})^2.\hat{p}_k(1).(1-\hat{p}_k(1))}{\sum_m (c_{m1}.\hat{p}_k(1)+(1-c_{m1}).(1-\hat{p}_k(1)))}$$

where $\Delta = \left| \frac{\partial F}{\partial p}(Y_{k+1}, \hat{p}_k(1)) \right|$. This is always less than 1 as soon as the observability condition $c_{11} \neq c_{12}$ is satisfied. This condition is evidently necessary since otherwise the observations would carry no information whatsoever about the unobserved state.

In the case with $N=2$ and $M \geq 2$ we find a similar expression for the magnitude of the contraction, involving a product of all the difference $c_{m1} - c_{k1}$. In fact whenever $N=2$ we can give a necessary and sufficient condition for exponential stability of the HMM filter. This condition states *Bitmead and Boel* [1] that

$$E[log\lambda_2(A) \cdot \left| \frac{\partial F}{\partial p}(Y_{k+1}, \hat{p}_k(1)) \right|] < 0 \qquad (6)$$

where $\lambda_2(A)$ is the second largest eigenvalue of A (strictly less than 1). Of course this condition is not easily verifiable since it is in general very difficult to evaluate the expectation. For $N > 2$ there is no such necessary and sufficient condition, because the Jacobian of the transformation F, even reduced to a subspace of dimension $N-1$, is a matrix. The derivation of (6) depends strongly on the commutativity of the different update steps.

However it is still possible to obtain geometric ergodicity of $\hat{p}_k$ by simply showing that is a contraction (not necessarily strict) (see e.g. *Bougerol* [2]). It suffices e.g. to calculate the eigenvalues of $E(\Delta.\Delta')$, and prove that there is at most a simple eigenvalue 1, while all the other eigenvalues are strictly less than 1. Simple observability conditions on C guaranteeing this property will be a topic of further research. The rate of forgetting initial conditions (or any other past data) is then at least as fast as $\lambda_2(A)$. This may however be a pessimistic estimate.

4 Risk-sensitive filtering for hidden Markov models

4.1 Problem Definition

Consider the signal model defined by (1) and (2). Our problem objective is to find an estimate $\hat{X}_k$ of X_k, where $\hat{X}_k \in \mathbb{R}^N$, such that the following criterion is satisfied,

$$\hat{X}_k = \underset{\zeta \in \mathbb{R}^N}{\operatorname{argmin}} J_k(\zeta), \quad J_k(\zeta) = E[\theta \exp(\theta \Psi_{0,k}(\zeta)) \mid \mathcal{Y}_k] \quad \forall k=0, 1, \ldots \qquad (7)$$

where $\theta (>0)$ is the risk-sensitive parameter and

$$\Psi_{0,k}(\zeta) = \hat{\Psi}_{0,k-1} + \frac{1}{2}(X_k - \zeta)' Q_k (X_k - \zeta), \quad Q_k \geq 0 \ \forall k \qquad (8)$$

where

$$\Psi_{m,n} \triangleq \frac{1}{2} \sum_{i=m}^{n} (X_i - \hat{X}_i)' Q_i (X_i - \hat{X}_i).$$

Remark 2 In [7], it has been considered that $\hat{X}_k \in \mathcal{S}_X$. To avoid a technical problem which will be explained in the next section, we assume here that $\hat{X}_k \in \mathbb{R}^N$.

4.2 Change of Measure and Reformulated Cost Index

Define $Y_k^i = \langle Y_k, f_i \rangle$, where $Y_k = (Y_k^1, \ldots, Y_k^M)$ such that for each $k \in \mathbb{N}$, exactly one component is equal to 1, the remainder being 0. Define a new measure $\bar{\mathcal{P}}$ where $\{Y_k\}, k \in \mathbb{N}$ is a sequence of *i.i.d* random variables and

$$\bar{\mathcal{P}}(Y_k^j = 1) = \frac{1}{M}$$

. Let $c_k = CX_k$ and $c_k^i = \langle c_k, f_i \rangle$. Also define

$$\bar{\lambda}_k = \prod_{i=1}^{M} (M c_k^i)^{Y_k^i}, \quad \bar{\Lambda}_k = \Pi_{l=0}^{k} \bar{\lambda}_l$$

If we set the Radon-Nikodym derivative $\frac{d\mathcal{P}}{d\bar{\mathcal{P}}} |_{\mathcal{G}_k} = \bar{\Lambda}_k$, then under $\mathcal{P}$,

$$E[Y_k \mid \mathcal{G}_k] = CX_k$$

Using a version of Bayes' Theorem, we have

$$E[\theta \exp(\theta \Psi_{0,k}(\zeta)) \mid \mathcal{Y}_k] = \frac{\bar{E}[\bar{\Lambda}_k \theta \exp(\theta \Psi_{0,k}(\zeta)) \mid \mathcal{Y}_k]}{\bar{E}[\bar{\Lambda}_k \mid \mathcal{Y}_k]} \qquad (9)$$

Hence, we work under $\bar{\mathcal{P}}$ where the modified problem objective is to determine $\hat{X}_k$ ($\in \mathbb{R}^N$) such that

$$\hat{X}_k = \underset{\zeta \in \mathbb{R}^N}{\operatorname{argmin}} \bar{E}[\bar{\Lambda}_k \theta \exp(\theta \Psi_{0,k}(\zeta)) \mid \mathcal{Y}_k] \qquad (10)$$

4.3 Recursive estimates

Definition 3 *Define the measure $\alpha_k(j)$ to be the unnormalised information state such that*

$$\alpha_k(j) = \bar{E}[\bar{\Lambda}_{k-1} \theta \exp(\theta \hat{\Psi}_{0,k-1}) \langle X_k, e_j \rangle \mid \mathcal{Y}_{k-1}] \qquad (11)$$

Remark 4 Note that $\alpha_k(j)$ can be interpreted as an information state of an augmented plant where the state includes the actual state of the system and part of the risk-sensitive cost. For details, see [8].

Lemma 5 *The information state $\alpha_k = (\alpha_k(1), \ldots, \alpha_k(N))'$ obeys the following recursion*

$$\alpha_{k+1} = A \mathcal{D}_k' \mathcal{B}_k' \alpha_k \qquad (12)$$

where

$$\mathcal{B}_k = diag\{Mc_1(Y_k), \ldots, Mc_N(Y_k)\}$$

$$\mathcal{D}_k = diag\left\{\exp\left(\frac{\theta}{2}(e_1-\hat{X}_k)'Q_k(e_1-\hat{X}_k)\right),\right.$$
$$\left.\ldots,\exp\left(\frac{\theta}{2}(e_N-\hat{X}_k)'Q_k(e_N-\hat{X}_k)\right)\right\}$$

and $c_i(Y_k)=c_{ji}$ *if* $Y_k=f_j$, $\forall i \in \{1,\ldots,N\}$, $j \in \{1,\ldots,M\}$.

Proof The proof can be carried out in the same way as it has been done for continuous-range observations in [7]. □

Remark 6 Note here that the information state filter is linear and finite-dimensional.

Note 7 Normalization:
Define the normalized recursive estimates by $\hat{\alpha}_{k+1}$. It can be easily shown that

$$\hat{\alpha}_{k+1}=\frac{A\mathcal{D}'_k\mathcal{B}'_k\hat{\alpha}_k}{\sum_{i=1}^{N} Mc_i(Y_k)\exp\left(\frac{\theta}{2}(e_i-\hat{X}_k)'Q_k(e_i-\hat{X}_k)\right)\hat{\alpha}_k(i)} \quad (13)$$

Theorem 8 *The optimizing estimate* $\hat{X}_k$ *is given by*

$$\hat{X}_k = \underset{\xi\in\mathbb{R}^N}{\operatorname{argmin}} \sum_{j=1}^{N}\prod_{i=1}^{M}(Mc_k^i)^{Y_k^i} \times \exp\left(\frac{\theta}{2}(e_j-\xi)'Q_k(e_j-\xi)\right)\alpha_k(j) \quad (14)$$

Proof Again, the proof is exactly similar to that one given in [7] and hence not given here. □

Remark 9 It should be obvious from the convex nature of the expression on the R.H.S of (14) that $\hat{X}_k$ exists and is unique.

4.4 Geometric ergodicity of the recursive risk-sensitive filter for a 2-state M-output symbol HMM

From the results derived in the previous section, we see that the normalized risk-sensitive estimates for a 2-state M-output symbol HMM are given by the following recursion

$$\hat{\alpha}_{k+1}=AF_k(Y_k,\hat{\alpha}_k) \quad (15)$$

where $\hat{\alpha}_k=(\hat{\alpha}_k(1)\ \ 1-\hat{\alpha}_k(1))'$ and $F_k(Y_k,\hat{\alpha}_k)$ is a nonlinear vector function given by

$$F_k(Y_k,\hat{\alpha}_k) = \frac{1}{\sum_{i=1}^{2} c_i(Y_k)\exp\left(\frac{\theta}{2}(e_i-\hat{X}_k)'Q_k(e_i-\hat{X}_k)\right)\hat{\alpha}_k(i)} \times diag\left\{c_1(Y_k)\exp\left(\frac{\theta}{2}(e_1-\hat{X}_k)'Q_k(e_1-\hat{X}_k)\right),\right.$$
$$\left. c_2(Y_k)\exp\left(\frac{\theta}{2}(e_2-\hat{X}_k)'Q_k(e_2-\hat{X}_k)\right)\right\}\hat{\alpha}_k \quad (16)$$

This recursion can be broken into 2 steps of transformation, a nonlinear mapping followed by a linear mapping. In section 3, the linear transformation has already been shown to be a strict contraction due to the fact that A is a transition probability matrix. Hence, we just deal with the nonlinear transformation and derive the condition under which it will be a contraction in an averaging sense.

Theorem 10 *The necessary and sufficient condition for the nonlinear mapping* $F_k : \mathbb{R}^M \times \mathbb{R}^2 \to \mathbb{R}^2$ *to be a contraction* $\forall k \in \mathbb{N}$ *in an averaging sense is given by*

$$\sum_{m=1}^{M} |g_m(\hat{\alpha}_k(1),\theta)| < M, \quad \forall k \quad (17)$$

where

$$g_m(\hat{\alpha}_k(1),\theta) = [c_{m1}\exp(\sigma_k^2(1))\,\hat{\alpha}_k(1) + c_{m2}\exp(\sigma_k^2(2))\,(1-\hat{\alpha}_k(1))]^{-2} \times c_{m1}c_{m2}\exp(\sigma_k^2(1)+\sigma_k^2(2)) [1+\theta\hat{\alpha}_k(1)(1-\hat{\alpha}_k(1)) \times \{(e_2-e_1)'Q_k\frac{\partial\hat{X}_k}{\partial\hat{\alpha}_k(1)}|_{Y_k,\hat{\alpha}_k(1)}\}]$$

and

$$\sigma_k^2(i)=\frac{\theta}{2}(e_i-\hat{X}_k)'Q_k(e_i-\hat{X}_k),\ i=1,2.$$

Proof The proof is omitted here but can be found in [13]. □

Remark 11 It is obvious from (14) is that $\hat{X}_k$ is a function of $\hat{\alpha}_k(1)$, although the functional relationship is not explicitly known. This prevents us from obtaining any further simplification of the condition (17) given above. We assume that for a given Y_k, $\hat{X}_k=L(\hat{\alpha}_k(1))$ where $L \in C^1(\mathbb{R})$. We also assume $\left|\frac{\partial\hat{X}_k}{\partial\hat{\alpha}_k(1)}\right|_{Y_k,\hat{\alpha}_k(1)} < \infty$. It is for this reason that we chose $\hat{X}_k \in \mathbb{R}^N$, rather than $\hat{X}_k \in \mathcal{E}$, as mentioned before.

Corollary 12 *A sufficient condition for the nonlinear mapping* $F_k : \mathbb{R}^M \times \mathbb{R}^2 \to \mathbb{R}^2$ *to be a contraction* $\forall k \in \mathbb{N}$ *in an averaging sense is*

$$|g_m(\hat{\alpha}_k(1),\theta)| < 1, \quad \forall m \in \{1,\ldots,M\} \quad (18)$$

where $\theta \neq 0$.

Proof The proof is immediate from Theorem 10. □

Remark 13 It should be noted here, that for $\theta=0$, when the risk-neutral filter for HMMs is obtained, (see [7]) this sufficient condition implies observability and hence is much

stricter than the observability condition obtained in Section 3. In fact, this condition might not be satisfied for all values of $\hat{\alpha}_k(1) \in (0, 1)$, for a given a set of parameters A, C.

Remark 14 It should be also noted that for $\theta \neq 0$, none of the conditions in Theorem 10 or Corollary 12 implies observability. In fact, even if $c_{m1}=c_{m2}$, $\forall m \in \{1, \ldots, M\}$, the conditions (17) and (18) can be satisfied provided

$$\left| (e_2 - e_1)' Q_k \frac{\partial \hat{X}_k}{\partial \hat{\alpha}_k(1)} |_{Y_k, \hat{\alpha}_k(1)} \right| < \gamma^2(\theta, \hat{\alpha}_k(1))$$

Note that observability is not a necessary condition for the geometric ergodicity of the recursive estimates in neither the risk-neutral nor the risk-sensitive case. But, it is correct to observe that observability is not a necessary condition for the nonlinear mapping $F_k : \mathbb{R}^M \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ to be a contraction, as opposed to the case of risk-neutral HMM filters in Section 3.

Remark 15 Risk-sensitive filters for hidden Markov models with N=2 and a continuous-range observation space $\mathbb{R}^p$ have been derived in [7]. The general condition for the corresponding nonlinear mapping to be a contraction can be similarly derived where the derivation involves an integration over the range of the observation process rather than a summation as in (17).

Also, it is not difficult to see that a necessary and sufficient condition similar to (6) can be obtained for the exponential stability of the recursive risk-sensitive estimates. However, the most fundamental observation that can be made from the above results is that for sufficiently large values of θ, $|g_m(\hat{\alpha}_k(1), \theta)|$ can be $\geq 1, \forall m$ and hence none of the conditions (17) and (18) would be satisfied. In other words, the risk-sensitive filter may become unstable, i.e., a small change in the initial conditions may result in an instability of the risk-sensitive filter. This restriction on θ has been also observed in [9] [10] for the case of risk-sensitive filters for linear Gauss-Markov models. It has been seen that sufficiently large values of θ may make a certain matrix negative definite, resulting in the non-existence of the solution of a certain Riccatti equation. Simulation studies for risk-sensitive filters for HMMs with continuous-range observations [7] have also shown that risk-sensitive filters lose their robustness against uncertain noise environments for sufficiently large values of θ. However, there is yet no general theory of choosing θ before starting the estimation process such that the stability of risk-sensitive filters would be guaranteed throughout.

References

[1] R. Bitmead and R. Boel: "On Stochastic Convergence of infinite Products of Random Matrices and its Role in Adaptive Estimation Theory", Proceedings IFAC Symposium on Identification and System Parameter Estimation, York, 1985

[2] P. Bougerol: "Théorèmes limite pour les systèmes linéaires á coefficients markoviens", Probability Theory and related fields, vol.78, 1988, pp.193-211, and "Comparaisons des exposants de Lyapunov des processus markoviens multiplicatifs", Annales de l'Institut Henri Poincaré, vol. 24, 1988, pp.439-489.

[3] A. Berman and R. Plemmons: Nonnegative matrices in the Mathematical Sciences, SIAM Classics in Applied Mathematics, 1994.

[4] H. Kunita: "Asymptotic Behaviour of Nonlinear Filtering Errors of Markov Processes", Journal of Multivariate Analysis, vol. 1, 1971, pp. 365-393.

[5] S. Meyn and R. Tweedie: Markov Chains and Stochastic Stability, Springer Verlag, 1993.

[6] R. J. Elliott, L. Aggoun and J. B. Moore, *Hidden Markov Models: Estimation and Control*, (Springer-Verlag, Application of Mathematics Series, New York, 1994).

[7] S. Dey and J. B. Moore, "Risk-sensitive filtering and smoothing for Hidden Markov Models," *Systems and Control Letters*, accepted for publication, 1994.

[8] J. B. Moore, R.J. Elliott and S. Dey, "Risk-sensitive Generalizations of Minimum Variance Estimation and Control," *Proc. of the IFAC Symposium on Nonlinear Control System Design (NOLCOS)*, California, June 1995, to appear.

[9] S. Dey and J. B. Moore, "Risk-sensitive filtering and smoothing via Reference Probability Methods", *Proc. of the American Control Conference*, Seattle, June 1995, to appear.

[10] J. L. Speyer, C. Fan and R. N. Banavar, "Optimal Stochastic Estimation with Exponential Cost Criteria," *Proceedings of the 31st Conference on Decision and Control*, Vol. 2, pp. 2293-2298, Dec. 1992.

[11] A. Segall, "Recursive Estimation from Discrete-Time Point Processes," *IEEE Trans. on Info: Theory*, Vol. IT-22, No. 4, pp. 422-431, July 1976.

[12] R. K. Boel, "Discrete-time Martingales in Filtering and Stochastic Control," Technical Report, The University of Trondheim, The Norwegian Institute of Technology, Division of Engineering Cybernetics, March 1976.

[13] S. Dey, *Topics in Robust Nonlinear Estimation and Control*, PhD Thesis, The Australian National University, forthcoming.