KYBERNETIKA — VOLUME *34* (1998), NUMBER 6, PAGES 739 – 746

ESTIMATION OF HIDDEN MARKOV MODELS FOR A PARTIALLY OBSERVED RISK SENSITIVE CONTROL PROBLEM[1]

Bernard Frankpitt and John S. Baras

This paper provides a summary of our recent work on the problem of combined estimation and control of systems described by finite state, hidden Markov models. We establish the stochastic framework for the problem, formulate a separated control policy with risk-sensitive cost functional, describe an estimation scheme for the parameters of the hidden Markov model that describes the plant, and finally indicate how the combined estimation and control problem can be re-formulated in a framework that permits an application of stochastic approximation techniques to the proof of asymptotic convergence of the estimator.

1. INTRODUCTION

Risk sensitive control of hidden Markov models has become a topic of interest in the control community largely in response to a paper by Baras and James [2] which shows that, in the small noise limit, risk sensitive control problems on hidden Markov models become robust control problems for non-deterministic finite state machines. This paper presents results that are part of a program to extend the work of Baras and James to cover situations where the plant is unknown. We consider the combined estimation and control problem for a class of controllers that implement randomized control strategies that approximate optimal risk-sensitive control on a finite horizon.

Problems of combined estimation and control have a long history, and the LQG case is standard material for stochastic control texts. Treatment of controlled hidden Markov models is more recent, the work of Fernández–Gaucherand et al [5] treats a situation similar to that treated here with different methods. The methods that we use are based on

[1]Research supported in part by the National Science Foundation Engineering Research Centers Program, NSFD CDR 8803012, and by the Lockheed Martin Chair in Systems Engineering.

existing work in stochastic approximation. In particular we use a recursive estimation scheme based on Krishnamurthy and Moore [6], and an approach from Arapostathis and Marcus [1] along with theorems from Benveniste et al [3] to prove convergence of the estimation scheme. The difference between this work and the preceding work is that by considering randomized strategies we can show convergence of the model estimate and the control without recourse to special reset conditions that are required in [5].

This paper is divided into five sections: the remainder of this section introduces the notation that we use, the second section describes the controller architecture, the third describes the estimator, the fourth states and discusses the convergence results, and the fifth presents some conclusions and directions for future work.

The Markov chains that are used in this paper are discrete-time finite-valued stochastic processes defined on an abstract probability space $(\Omega, \mathcal{F}, P)$. The finite state space is represented by the unit vectors $\{e_1, \ldots, e_N\}$ of $\mathbb{R}^N$ and the finite input space, U, is represented by the unit vectors in $\mathbb{R}^P$. If the input at time l has the value u_l, then the state transition matrix for the Markov chain has entries

$$A_{u_l;ij} = P(x_{l+1} = e_j \,|\, x_l = e_i, u_l).$$

The finite set of outputs Y is represented by the unit vectors in $\mathbb{R}^M$, and the transition matrix from state to output is given by

$$B_{ij} = P(y_l = e_j \,|\, x_l = e_i).$$

The combined state, input and output process $\{x_l, u_l, y_l\}$ generates a filtration $\{\mathcal{F}_l\} \subset \mathcal{F}$ in the usual way, and the process formed by combining input and output only generates a smaller filtration $\{\mathcal{Y}_l\} \subset \{\mathcal{F}_l\}$ on $\mathcal{F}$. In general, probability distributions on finite sets will be represented as vectors, expectations as inner products in Euclidean spaces of the appropriate dimensions, and probability kernels on finite spaces will be represented as matrices.

Let $\mathcal{M}$ denote the space of probability distributions on the finite set U, and $\mathcal{M}_\eta$, $0 \leq \eta \leq 1/P$ denote the compact subset of distributions that satisfy $\mu\{u\} \geq \eta$ for all $u \in U$. A control policy for a finite horizon of length K is a specification of a sequence of probability distributions on $\mu_0, \mu_1, \ldots \mu_{K-1} \in \mathcal{M}$. A control policy is an output feedback policy if each distribution μ_l is a measurable function on the σ-algebra $\mathcal{Y}_l$. Each control policy $\mu = \mu_0, \mu_1, \ldots, \mu_{K-1}$ induces a probability distribution on $\mathcal{F}_K$ with density

$$P^u(x_{0,K}, y_{0,K}) = \langle x_K, By_K \rangle \langle x_0, \pi_0 \rangle \prod_{l=0}^{K-1} \sum_{u \in U} \langle x_l, A_u x_{l+1} \rangle \langle x_l, By_l \rangle \langle u, \mu_l \rangle. \qquad (1)$$

where π_0 is the probability distribution for the random variable x_0. It is

convenient here to define an additional probability measure on Ω

$$P^{\dagger}(x_{0,K}, y_{0,K}) = \frac{1}{M}\langle x_0, \pi_0\rangle \prod_{l=0}^{K-1} \sum_{u\in U} \frac{1}{M}\langle x_l, A_u x_{l+1}\rangle \langle u, \mu_l\rangle.$$

P^u is absolutely continuous with respect to $P^{\dagger}$ and has Radon–Nikodym derivative

$$\left.\frac{dP^u}{dP^{\dagger}}\right|_{\mathcal{G}_K} = \Lambda_K = \prod_{l=0}^{K} M\langle x_l, By_l\rangle.$$

In addition, the output process y_l is i.i.d. with respect to $P^{\dagger}$ and has uniform marginal distributions $P^{\dagger}\{y_l = e_j\} = 1/M$.

2. CONTROLLER ARCHITECTURE

A risk sensitive control problem is defined on a hidden Markov model by specifying a cost functional with an exponential form. Given a running cost, $\phi(x, u)$, which is a function of both the state and the input, and a final cost $\phi_f(x)$, which is a function of the state only, the finite horizon, risk sensitive cost, associated with the control policy μ, with risk γ and horizon K is the functional

$$\mathcal{J}^{\gamma}(\mu) = \mathbf{E}\left[\exp\frac{1}{\gamma}\left(\phi_f(x_K) + \sum_{l=0}^{K-1}\phi(x_l, u_l)\right)\right]. \tag{2}$$

Expressed in terms of expectations with respect to the $P^{\dagger}$ measure, the cost is

$$\mathcal{J}^{\gamma}(\mu) = \mathbf{E}^{\dagger}\left[\Lambda_K \exp\frac{1}{\gamma}\left(\phi_f(x_K) + \sum_{l=0}^{K-1}\phi(x_l, u_l)\right)\right].$$

Optimal output feedback controls are computed by defining an information state that is a process adapted to the filtration $\{\mathcal{Y}_l\}$, translating the cost to a functional on the information state, and then using dynamic programming with respect to the information state dynamics to compute the optimal control. An appropriate choice of the information state at time l is the expected value of the accrued cost at time l, conditioned with respect to the σ-algebra $\mathcal{Y}_l$.

$$\sigma_l^{\gamma}(x) = \mathbf{E}^{\dagger}\left[I_{\{x_l = x\}}\Lambda_l \exp\left(\frac{1}{\gamma}\sum_{k=0}^{l-1}\phi(x_k, u_k)\right) \mid \mathcal{Y}_l\right].$$

The information state dynamics is described by a linear recursion on $\mathbb{R}^{+N}$

$$\sigma_l = \Sigma(u_{l-1}, y_l)\sigma_{l-1}, \tag{3}$$

with

$$\Sigma(u,y) = M\,\mathbf{diag}(\langle\cdot, By\rangle)\,A^{\top}(u)\,\mathbf{diag}(\exp(1/\gamma\,\phi(\cdot,u))).$$

The risk sensitive cost is expressed as a functional on the information state process by the formula

$$\mathcal{J}^{\gamma}(\mu) = \mathbf{E}^{\dagger}\left[\langle \sigma_K^{\gamma}(\cdot), \exp(\phi_f(\cdot)/\gamma)\rangle\right]. \tag{4}$$

The value function associated with the finite-time, state-feedback control problem on the information state recursion (3) with cost function (4) is

$$S^{\gamma}(\sigma,l) = \min_{\mu_l \ldots \mu_{K-1} \in \mathcal{M}} \mathbf{E}^{\dagger}\left[\langle \sigma_K^{\gamma}(\cdot), \phi_f(\cdot)\rangle \,|\, \sigma_l^{\gamma} = \sigma\right], \qquad 0 \le l < K. \tag{5}$$

The associated dynamic programming equation is

$$\begin{cases} S^{\gamma}(\sigma,l) & = & \min_{\mu_l \in \mathcal{M}} \mathbf{E}^{\dagger}\left[S^{\gamma}(\Sigma^{\gamma}(u_l, y_{l+1})\sigma, l+1)\right] \\ S^{\gamma}(\sigma,K) & = & \langle \sigma(\cdot), \phi_f(\cdot)\rangle. \end{cases} \tag{6}$$

An induction argument along the lines of that used by Baras and James [2] proves the following theorem.

Theorem 1. **The value function S^{γ} defined by (5) is the unique solution to the dynamic programming equation (6). Conversely, assume that S^{γ} is the solution of the dynamic programming equation (6) and suppose that μ^* is a policy such that for each $l = 0, \ldots, K-1$, $\mu_l^* = \overline{\mu}_l^*(\sigma_l^{\gamma}) \in \mathcal{M}$, where $\overline{\mu}_l^*(\sigma)$ achieves the minimum in (6). Then μ^* is an optimal output feedback controller for the risk-sensitive stochastic control problem with cost functional (2).**

The following structural properties are analogous to those proved by Fernández–Gaucherand and Marcus [4].

Theorem 2. **At every time l the value function $S^{\gamma}(\sigma,l)$ is convex and piecewise linear in the information state $\sigma \in I\!R^{+N}$. Furthermore, the information state is invariant under homothetic transformations of $I\!R^{+N}$.**

The randomized policies taking values in $\mathcal{M}_{\eta}$ approximate deterministic policies in the following way.

Theorem 3. **Let S_{η} denote the value function for the optimal control problem when the policy is restricted so that $\mu_l \in \mathcal{M}_{\eta}$ for all $0 \le l \le K-1$, then $S_0 = S$ is a deterministic policy,**

$$\frac{S_{\eta}(\sigma,l) - S_0(\sigma,l)}{1 + |\sigma|} \to 0$$

uniformly on $\mathbb{R}^{N^+} \times \{0, \ldots, K\}$, and the optimal policies converge $\mu_\eta^* \to \mu^*$.

The controller architecture that we propose is based on a moving window. Theorem 2 is used with the dynamic programming equation (6) to compute the value function for the finite horizon problem with horizon K. along with the values of the optimal output feedback distributions $\mu^*(\sigma)$. At each time l the information state recursion (3) is used with a record of the previous Δ observations and control values, and a predetermined initial value $\sigma_{l-\Delta}$ to compute the current value of the information state. The optimal probability distribution $\mu(\sigma_l)$ is selected, and a random procedure governed by this distribution is used to produce a control value u_l.

3. ESTIMATOR ARCHITECTURE

The estimator architecture is a maximum likelihood estimator. The recursive algorithm is derived by following the formal derivation that Krishnamurthy and Moore [6] give for a stochastic gradient scheme that approximates a maximum likelihood estimator for a hidden Markov model. The resulting algorithm is well described as a recursive version of the expectation maximization algorithm of Baum and Welch. Let θ_l denote an estimate for the parameters that determine the probabilistic structure of the hidden Markov chain. The components of θ, which are the entries of the transition matrices, are constrained to lie in a linear submanifold Θ by the requirement that the estimates $\widehat{A}_u$ and $\widehat{B}$ be stochastic matrices. Gradients and Hessians taken with respect to θ will be thought of as linear and bilinear forms on the tangent space to Θ.

A maximum likelihood estimator for a hidden Markov model with parameterization θ^* minimizes the Kullback–Leibler measure

$$J(\theta) = \mathbf{E}[\log f(y_{0,l} \,|\, \theta) \,|\, \theta^*].$$

Here $f(y_{0,l} \,|\, \theta)$ is used to denote the distribution function induced by the parameter θ on the sequence of random variables $y_{0,l}$. It turns out that $J(\theta)$ is not an easy quantity to calculate, however an equivalent condition can be stated in terms of the functions

$$\begin{aligned} Q_l(\theta', \theta) &= \mathbf{E}[\log f(x_{0,l}, y_{0,l} \,|\, \theta) \,|\, y_{0,l}, \theta'] \\ \overline{Q}_l(\theta', \theta) &= \mathbf{E}[Q_l(\theta', \theta) \,|\, \theta^*] \end{aligned} \tag{7}$$

Krishnamurthy and Moore show that $\overline{Q}_l(\theta', \theta) > \overline{Q}_l(\theta', \theta')$ implies that $J(\theta) > J(\theta')$, and proceed to write down the stochastic gradient algorithm[2]

$$\theta_{l+1} = \theta_l + I_{l+1}^{-1}(\theta_l) \left. \frac{\partial Q_{l+1}(\theta_l, \theta)}{\partial \theta} \right|_{\theta=\theta_l}$$

[2]The θ_l are actually constrained to lie on Θ.

Where I_l is the Fisher information matrix for the combined state and output process $I_l(\theta_l) = -\partial^2 Q_{l+1}/\partial\theta^2|_{\theta=\theta_l}$, and $Q_{l+1}(\theta_l, \theta)$ is the empirical estimate for $Q(\theta_l, \theta)$ based on the first l observations.

The central part of the estimator is a finite buffer containing the last Δ values of the input and output processes (the length is chosen to be the same as the length of the controller buffer in order to simplify the presentation). This buffer is used to update smoothed recursive estimates of the various densities from which the function Q and its derivatives are calculated. These densities are $\alpha_l = f(x_{l-\Delta} \,|\, y_{0,l-\Delta})$ which is calculated with the recursion

$$\alpha_l(j) = \frac{\sum_i \langle e_j, \widehat{B} y_{l-\Delta}\rangle \widehat{A}_{u_{l-\Delta-1};ij} \alpha_{l-1}(i)}{\sum_{i,j} \langle e_j, \widehat{B} y_{l-\Delta}\rangle \widehat{A}_{u_{l-\Delta-1};ij} \alpha_{l-1}(i)}, \tag{8}$$

$\beta_l = f(y_{l-\Delta+1,l} \,|\, x_{l-\Delta})$ is computed with the backwards recursion

$$\beta^k(i) = \sum_j \beta^{k+1}(j) \widehat{A}_{u_{k+1};ij} \langle e_i, \widehat{B} y_{k+1}\rangle.$$

The finite recursion is recalculated for each time l starting with $k = l$, and finishing with $k = l - \Delta$ and β_l takes the value $\beta_l = \beta^{l-\Delta}$. Estimates of the conditional densities $\zeta_l = f(x_{l-\Delta}, x_{l-\Delta-1}|y_{0,l})$ and $\gamma_l = f(x_{l-\Delta} \,|\, y_{0,l})$ are given in terms of α_{l-1}, α_l, β_{l-1} and β_l by

$$\begin{aligned} \zeta_l(i,j) &= \frac{\alpha_{l-1}(i) \widehat{A}_{u_{l-\Delta-1};ij} \beta_{l-1}(j)}{\sum_{i,j} \alpha_{l-1}(i) \widehat{A}_{u_{l-\Delta-1};ij} \beta_{l-1}(j)} \\ \gamma_l(i) &= \frac{\sum_j \beta_l(j) \widehat{A}_{u_{l-\Delta};ij} \alpha_l(i)}{\sum_{i,j} \beta_l(j) \widehat{A}_{u_{l-\Delta};ij} \alpha_l(i)}, \end{aligned}$$

and the empirical estimates of the joint state-next state pair frequency and state-output pair frequency are given by the recursive estimators

$$\begin{aligned} Z_l^u &= Z_{l-1}^u + (1-\rho)\delta_u(u_{l-\Delta-1}) \left(\zeta_l - Z_{l-1}^u\right) \\ \Gamma_l &= \Gamma_{l-1} + (1-\rho) \left(\gamma_l y_{l-\Delta}^\top - \Gamma_{l-1}\right) \end{aligned}$$

with $0 \ll \rho < 1$.

The result of the formal derivation is an algorithm that (after some work) can be written in the form of a standard stochastic approximation problem:

$$\theta_{l+1} = \theta_l + \frac{1}{l} H(X_l, \theta_l). \tag{9}$$

where $X = \{x_l, u_{l-\Delta,l}, y_{l-\Delta,l}, \alpha_{l,l-1}, Z_l, \Gamma_l\}$ is a Markov chain, and the parts of H that correspond to the updates of $\widehat{A}_u$ and $\widehat{B}$ are given by

$$\delta_u(u_{l-\Delta}) \frac{\frac{\widehat{A}_{u;ij}^2}{Z_l(i,j)} \left(\sum_{r=1}^N \frac{\widehat{A}_{u;ir}^2}{Z_l(i,r)} \left(\frac{\zeta_l(i,j)}{\widehat{A}_{u;ij}} - \frac{\zeta_l(i,r)}{\widehat{A}_{u;ir}} \right) \right)}{\sum_{r=1}^N \frac{\widehat{A}_{u;ir}^2}{Z_l(i,r)}}$$

and

$$\frac{\frac{\widehat{B}^2_{im}}{\Gamma_l(i,m)}\left(\sum_{r=1}^N \frac{\widehat{B}^2_{ir}}{\Gamma_l(i,r)}\left(\frac{\gamma_l(i)\delta_{e_m}(y_{l-\Delta})}{\widehat{B}_{im}} - \frac{\gamma_l(i)\delta_{e_r}(y_{l-\Delta})}{\widehat{B}_{ir}}\right)\right)}{\sum_{r=1}^N \frac{\widehat{B}^2_{ir}}{\Gamma_l(i,r)}}$$

respectively.

4. CONVERGENCE OF ESTIMATES

Let $P_{n:x,a}$ denote the distribution of (X_{n+k}, θ_{n+k}) when $X_n = x$, and $\theta_n = a$ then the convergence of the estimation algorithm (9) is governed by the following theorem.

Theorem 4. **If the matrices A and B are primitive, and the policies μ satisfy**

$$\mu(y_{k-\Delta,k}, u_{k-\Delta-1,k-1})\{u\} > 0 \qquad \textbf{for all } u \in U.$$

Then, there exists a neighborhood system $\mathcal{N}$ of θ^0 such that for any $F \in \mathcal{N}$, and for any compact set $Q \subset \Theta$ there exists a constants $B > 0$ and $\lambda \in [1/2, 1]$ such that for all $a \in Q$ and all $X \in \mathcal{X}$

$$P_{n,X,a}\{\theta_k \textbf{ converges to } F\} \geq 1 - B \sum_{k=n+1}^{\infty} 1/k^{1+\gamma}$$

where θ_k is the sequence that is computed by the recursion (9)

The p r o o f of the theorem is a non-trivial application of the results from part II, chapters 1 and 2 of Benveniste et al [3] in which the authors present an analysis of the ODE method for proving convergence of stochastic approximation algorithms. Results similar to Theorem 4 are proved for a related problem by Arapostathis and Marcus in [1] who use stochastic approximation results of Kushner, and then, in greater generality, by Le Gland and Mevel [7] who also use the theory from [3]. The major difference between the problems treated in the works cited and the problem treated here is the introduction of control to give a combined control-estimation problem. From the point of view of the stochastic approximation analysis the control policy affects the transition kernels of the underlying Markov chain, by introducing a dependency on the current estimates. The restriction made in the premise of the theorem on the space of randomized control policies allows the control policy to be incorporated into the Markov chain X_l in a way that ensures good ergodic properties for the transition kernel of X_l.

The ODE method relies on the use of a martingale convergence argument to prove convergence of the iterates of the stochastic approximation algorithm to the trajectories of an associated ODE. The central feature

of the treatment of Benveniste et al [3] is the use of regular solutions ν_θ to the Poisson equation

$$(I - \Pi_\theta)\nu_\theta = H(\cdot, \theta) - h_\theta \tag{10}$$

to provide the necessary martingale. The kernel Π_θ in (10) is the transition kernel for the chain X_l, and the function h_θ is the generator for the associated ODE. When applying the theory, ν_θ does not have to be calculated explicitly, its existence and regularity can be inferred from ergodic properties of the transition kernel Π_θ for chain X_k. Most of the effort in the proof is expended in establishing that bounds of the form

$$\begin{aligned} |\Pi_\theta^n g(X_1) - \Pi_\theta^n g(X_2)| &\leq K_1 L_g \rho^n \\ |\Pi_\theta^n g(X) - \Pi_{\theta'}^n g(X)| &\leq K_2 L_g |\theta - \theta'| \end{aligned}$$

hold for any Lipschitz function g and for all θ, θ', X_1 and X_2, where K_1 and K_2 are constants, and $0 \leq \rho < 1$. The condition on the admissible control strategies in the premise of Theorem 4 is key to establishing the second bound.

The other important task in the proof of Theorem 4 is establishing that the ODE converges asymptotically to the maximum likelihood estimate. To accomplish this a Lyapunov function argument is used. An appropriate choice of Lyapunov function in this case is the function $U(\theta) = \overline{Q}(\theta^0, \theta)$.

5. CONCLUSIONS AND FUTURE WORK

This paper presented an overview of the work that we are doing on the problem of combined estimation and control for systems that can be described by finite state hidden Markov models. We see the results that we present here as preliminary. Techniques which are based on the minimization of a relative entropy function, such as the estimation technique described here, do not perform well when the number of parameters being estimated increases and the domains of attraction shrink. The implication of this observation is that without additional a-priori assumptions our proposed control architecture is only practical for systems that can be modeled with a small state-space. Acknowledging this constraint, we see our work proceeding in three ways. We are looking at applications to systems that are likely to benefit from controllers which assume small state-spaces, we are considering how to incorporate a-priori structural assumptions about the plant into the frame-work that this paper presents, and finally we are looking for approaches that bypass the model estimation stage entirely and work directly with the estimation of the information state recursion for the separated controller.

(Received April 8, 1998.)

REFERENCES

[1] **A. Arapostathis and S. I. Marcus: Analysis of an identification algorithm arising in the adaptive estimation of Markov chains. Mathematics of Control, Signals and Systems** *3* **(1990),1–29.**

[2] **J. S. Baras and M. R. James: Robust and Risk–Sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models, to be published.**

[3] **A. Benveniste, M. Métivier and P. Priouret: Adaptive Algorithms and Stochastic Approximations. Springer–Verlag, Berlin 1990. Translation of "Algorithmes adaptatifs et approximations stochastiques", Masson, Paris 1987.**

[4] **E. Fernandéz–Gaucherand and S. I. Marcus: Risk–Sensitive Optimal Control of Hidden Markov Models: Structural Results. Technical Report TR 96-79, Institute for Systems Research, University of Maryland, College Park, Maryland 1996.**

[5] **E. Fernandéz–Gaucherand, A. Arapostathis and S. I. Marcus: Analysis of an adaptive control scheme for a partially observed controlled Markov chain. IEEE Trans. Automat. Control** *38* **(1993), 6, 987–993.**

[6] **V Krishnamurthy and J. B. Moore: On–line estimation of hidden Markov model parameters based on the. IEEE Trans. Signal Processing** *41* **(1993), 8, 2557–2573.**

[7] **F. Le Gland and L. Mevel: Geometric Ergodicity in Hidden Markov Models. Technical Report No. 1028, IRISA/INRIA, Campus de Beaulieu, Renees 1996.**

Mr. Bernard Frankpitt and Dr. John S. Baras, Institute for Systems Research and Department of Electrical Engineering, University of Maryland, College Park, MD 20742. U. S. A.

e-mails: frankpit@isr.umd.edu, baras@isr.umd.edu