Estimation of Hidden Markov Models for Partially Observed Risk Sensitive Control Problems

Bernard Frankpitt[1] and John S. Baras[2]
Dept. of Electrical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract

We look at the problem of estimation for partially observed, risk-sensitive control problems with finite state, input and output sets, and receding horizon. We describe architectures for risk sensitive controllers, and estimation, and we state conditions under which both the estimated model converges to the true model, and the control policy will converge to the optimal risk sensitive policy.

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 receding 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.* [6] treats a situation similar to that treated here with different methods. The methods that we use are based on existing work in stochastic approximation. In particular we use a recursive estimation scheme based on Krishnamurthy and Moore [7], and an approach from Arapostathis and Marcus [1] along with theorems from Benveniste *et al.* [4] 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 [6].

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 $\{e_1, \dots, e_N\}$ 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

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

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

$$B_{ij} = P(y_k = e_j \mid x_k = e_i).$$

The combined state and output process $\{x_k, y_k\}$ generates a filtration $\{\mathcal{G}_k\} \subset \mathcal{F}$, and a second filtration $\{\mathcal{O}_k\}$ is defined as the filtration generated by the sequence of pairs (u_{l-1}, y_l), $l \leq k$. $\mathcal{O}_k$ can be interpreted as a specification of the information available to an output feedback controller from the record of past plant inputs and outputs. 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

[1] Martin Marietta Chair in Systems Engineering.
e-mail: baras@isr.umd.edu

[2] e-mail: frankpit@isr.umd.edu

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 receding horizon control policy with 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_k is a measurable function on the σ-algebra $\mathcal{O}_k$. Each control policy $\mu = \mu_0, \mu_1, \ldots, \mu_{K-1}$ induces a probability distribution on $\mathcal{F}_K$ with density

$$P^\mu(u_{0,K-1}, x_{0,K}, y_{0,K}) = \langle x_K, By_K \rangle \langle x_0, \pi_0 \rangle \\ \times \prod_{l=0}^{K-1} \langle x_l, A_{u_l} x_{l+1} \rangle \langle x_l, By_l \rangle \langle u_l, \mu_l \rangle. \quad (1)$$

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

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

P^u is absolutely continuous with respect to $P^\dagger$ and has Radon-Nykodym derivative

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

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

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

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

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 $\{\mathcal{O}_k\}$, translating the cost to a functional on the information state, and then using dynamic programming to compute the optimal control. An appropriate choice of the information state at time k is the expected value of the accrued cost at time k, conditioned with respect to the σ-algebra $\mathcal{O}_k$, and expressed as a distribution over the state set X.

$$\sigma_k^\gamma(x) = \mathbf{E}^\dagger \Bigg[I_{\{x_k = x\}} \Lambda_k \\ \exp \left(\frac{1}{\gamma} \sum_{l=0}^{k-1} \phi(x_l, u_l) \right) \mid \mathcal{O}_k \Bigg]. \quad (3)$$

The information state process satisfies a linear recursion on ${\mathbb{R}^+}^N$

$$\sigma_k = \Sigma(u_{k-1}, y_k) \sigma_{k-1}, \quad (4)$$

with

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

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

$$\mathcal{J}^\gamma(\mu) = \mathbf{E}^\dagger \left[\langle \sigma_K^\gamma(\cdot), \exp(\phi_f(\cdot)/\gamma) \rangle \right]. \quad (5)$$

The value function associated with the finite-time, state-feedback control problem on the information state recursion (4) with cost function (5) is

$$S^\gamma(\sigma, l) = \\ \min_{\mu_l \cdots \mu_{K-1} \in \mathcal{M}} \mathbf{E}^\dagger \left[\langle \sigma_K^\gamma(\cdot), \phi_f(\cdot) \rangle \mid \sigma_l^\gamma = \sigma \right], \\ 0 \leq l < K. \quad (6)$$

The associated dynamic programming equation is

$$\begin{cases} S^\gamma(\sigma, l) = \min_{\mu_l \in \mathcal{M}} \mathbf{E}^\dagger \left[S^\gamma(\Sigma^\gamma(u_l, y_{l+1})\sigma, l+1) \right] \\ S^\gamma(\sigma, K) = \langle \sigma(\cdot), \phi_f(\cdot) \rangle. \end{cases} \quad (7)$$

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 (6) is the unique solution to the dynamic programming equation (7). Conversely, assume that S^γ is the solution of the dynamic programming equation (7) and suppose that μ^* is a policy such that for each $l = 0, \ldots, k-1$, $\mu_l^* = \bar{\mu}_l^*(\sigma_l^\gamma) \in \mathcal{M}$, where $\bar{\mu}_l^*(\sigma)$ achieves the minimum in (7). 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 [5].

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

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 (7) 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 k the information state recursion (4) is used with a record of the previous Δ observations and control values, and a predetermined initial value $\sigma_{k-\Delta}$ to compute the current value of the information state. The optimal probability distribution $\mu(\sigma_k)$ is selected, and a random procedure governed by this distribution is used to produce a control value u_k.

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 [7] 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_k 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 $\hat{A}_u$ and $\hat{B}$ 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 λ^0 minimizes the Kullback Leibler measure

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

Here $f(y_{0,k} \mid \lambda)$ is used to denote the distribution function induced by the parameter λ on the sequence of random variables $y_{0,k}$. 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

$$\begin{aligned} Q_k(\theta', \theta) &= \mathbf{E}[\log f(x_{0,k}, y_{0,k} \mid \theta) \mid y_{0,k}, \theta'] \\ \bar{Q}_k(\theta', \theta) &= \mathbf{E}[Q_k(\theta', \theta) \mid \theta^0] \end{aligned} \tag{8}$$

Krishnamurthy and Moore show that $\bar{Q}_k(\theta', \theta) > \bar{Q}_k(\theta', \theta')$ implies that $J(\theta) > J(\theta')$, and proceed to write down the stochastic gradient algorithm[1]

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

Where I_k is the Fisher information matrix for the combined state and output process

$$I_k(\theta_k) = -\partial^2 Q_{k+1} / \partial \theta^2 |_{\theta = \theta_k},$$

and $Q_{k+1}(\theta_k, \theta)$ is the empirical estimate for $Q(\theta_k, \theta)$ based on the first k 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 $\alpha_k = f(x_{k-\Delta} \mid y_{0,k-\Delta})$ which is calculated with the recursion

$$\alpha_k(j) = \frac{\sum_i \langle e_j, \hat{B} y_{l-\Delta} \rangle \hat{A}_{u_{l-\Delta-1};ij} \alpha_{k-1}(i)}{\sum_j \sum_i \langle e_j, \hat{B} y_{l-\Delta} \rangle \hat{A}_{u_{l-\Delta-1};ij} \alpha_{k-1}(i)}, \tag{9}$$

$\beta_k = f(y_{k-\Delta+1,k} \mid x_{k-\Delta})$ is computed with the backwards recursion

$$\beta_l(i) = \sum_j \beta_{l+1}(j) \hat{A}_{u_{l+1};ij} \langle e_i, \hat{B} y_{l-\Delta+1} \rangle.$$

$\zeta_k = f(x_{k-\Delta}, x_{k-\Delta-1} | y_{0,k})$ and $\gamma_k = f(x_{k-\Delta} \mid y_{0,k})$ are given in terms of α and β by

$$\begin{aligned} \zeta_{l|K,\Lambda_k}(i,j) &= \frac{\alpha_{l-\Delta-1}(i) A_{u_{l-\Delta-1};ij} \beta_{l-\Delta-1}(j)}{\sum_{i,j} \alpha_{l-\Delta-1}(i) A_{u_{l-\Delta-1};ij} \beta_{l-\Delta-1}(j)} \\ \gamma_{l|K,\Lambda_k}(i) &= \frac{\sum_j \beta_{l-\Delta}(j) A_{u_{l-\Delta;ij}} \alpha_{l-\Delta}(i)}{\sum_i \sum_j \beta_{l-\Delta}(j) A^{u_{l-\Delta-1};ij} \alpha_{l-\Delta}(i)}. \end{aligned}$$

and the empirical estimates of state frequency and pair frequency are given by the random variables $Z_k = 1/(k-\Delta) \sum \zeta_l \delta_{u_l}$ and $\Gamma_k = 1/(k-\Delta) \sum \gamma_l \delta_{y_l}$.

The result of the formal derivation is an algorithm that can be written in the standard form

$$\theta_{k+1} = \theta_k + \frac{1}{k} H(X_k, \theta_k) \tag{10}$$

[1] The θ_k are actually constrained to lie on Θ

where $X = \{x_k, u_{k-\Delta,k}, y_{k-\Delta,k}, \alpha_{k,k-1}, Z_k, \Gamma_k\}$ is a Markov chain, and the parts of H that correspond to the updates of A_u and B are given by

$$\frac{\frac{\hat{A}^2_{u;ij}}{Z_l(i,j)} \left(\sum_{r=1}^{N} \frac{\hat{A}^2_{u;ir}}{Z_l(i,r))} \left(\frac{\zeta_{k|K,\Lambda_k}(i,j)}{\hat{A}_{u;ij}} - \frac{\zeta_{k|K,\Lambda_k}(i,r)}{\hat{A}_{u;ir}} \right) \right)}{\sum_{r=1}^{N} \frac{\hat{A}^2_{u;ir}}{Z_l(i,r)}} \tag{11}$$

and

$$\frac{\frac{\hat{B}^2_{im}}{\Gamma_l(i,m)} \left(\sum_{r=1}^{N} \frac{\hat{B}^2_{ir}}{\Gamma_l(i,r)} \left(\frac{\gamma_{k|K,\Lambda_k}(i)\delta(y_k{=}f_m)}{\hat{B}_{im}} - \frac{\gamma_{k|K,\Lambda_k}(i)\delta(y_k{=}f_r)}{\hat{B}_{ir}} \right) \right)}{\sum_{r=1}^{N} \frac{\hat{B}^2_{ir}}{\Gamma_l(i,r)}} \tag{12}$$

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 (10) is governed by the following theorem.

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

$$\mu(y_{k-\Delta,k}, u_{k-\Delta-1,k-1})\{u\} > 0 \qquad \text{for all } u \in U \tag{13}$$

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 $X \in \mathcal{X}$

$$P_{n,X,a}\{\theta_k \text{ converges to } F\} \geq 1 - B \sum_{k=n+1}^{\infty} 1/k^{1+\gamma} \tag{14}$$

where θ_k is the sequence that is computed by the recursion (10)

The proof of the theorem is a non-trivial application of the results from part II, chapters 1 and 2 of Benveniste *et al.* [4]. Similar results 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 [8] who also use the theory from [4]. 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 ensures that conditional expectations associated with the evolution of the Markov chain are Lipschitz continuous with respect to the parameter estimates.

The central feature of the theory in [4] is the Poisson equation associated with the chain X_k

$$(I - \Pi_\theta)\nu_\theta = H(\cdot, \theta) - h_\theta$$

The function h_θ is the generator for the ODE that governs the asymptotic behavior, and regularity of the solutions ν_θ ensures that the sequence θ_k converges to a trajectory of the 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. The first major task in proving Theorem 4 is to establish that bounds of the form

$$|\Pi^n_\theta g(X_1) - \Pi^n_\theta g(X_2)| \leq K_1 L_g \rho^n$$
$$|\Pi^n_\theta g(X) - \Pi^n_{\theta'} g(X)| \leq K_2 L_g |\theta - \theta'|$$

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 second major task is establishing that the ODE converges asymptotically to the maximum likelihood estimate. To accomplish this a Lyapunov function type argument is used. An appropriate choice of Lyapunov function in this case is the function $U(\theta) = \bar{Q}(\theta^0, \theta)$. Arguments similar to those used by Baum and Petrie [3] to show prove regularity properties of $J(\theta)$ are used to establish the required local properties for $U(\theta)$.

5 Conclusions and Future Work

This paper presents a combined control-estimation algorithm for a hidden Markov model plant. It follows from the structural properties of the value function that the value function is a continuous function of the plant parameters A_u and B. Consequently convergence of the parameter estimates ensures convergence of the value function and convergence of the control policy to the optimal policy within the prescribed set.

We see the results that we present here as preliminary. Maximum likelihood techniques do not perform well when the number of parameters being estimated increases and the domains of attraction shrink. We are looking at approaches that bypass the model estimation stage and work directly with the estimation of the information state recursion for the separated controller.

References

[1] Aristotle Arapostathis and Steven I. Marcus. Analysis of an identification algorithm arising in

the adaptive estimation of markov chains. *Mathematics of Control, Signals and Systems*, 3:1–29, 1990.

[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*, 1993.

[3] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic fuctnions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1996.

[4] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, 1990. Translation of "Algorithmes adaptatifs et approximations stochastiques", Masson, Paris, 1987.

[5] 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 20742 U.S.A., 1996. obtainable from http://www.isr.umd.edu.

[6] Emmanuel Fernandéz-Gaucherand, Aristotle Arapostathis, and Steven I. Marcus. Analysis of an adaptive control scheme for a partially observed controlled Markov chain. *IEEE Transactions on Automatic Control*, 38(6):987–993, 1993.

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

[8] Francois Le Gland and Laurent Mevel. Geometric ergodicity in hidden Markov models. Technical Report No. 1028, IRISA/INRIA, Campus de Beaulieu, 35042 Rennes Cédex, France., 1996.