

On Near Optimality of the Set of Finite-State Controllers for Average Cost POMDP

Huizhen Yu

Department of Computer Science, P.O. Box 68, FIN-00014 University of Helsinki, Finland,
janey.yu@cs.helsinki.fi

Dimitri P. Bertsekas

Laboratory for Information and Decision Systems, Massachusetts Institute of Technology,
 Cambridge, Massachusetts 02139, USA, dimitrib@mit.edu

We consider the average cost problem for partially observable Markov decision processes (POMDP) with finite state, observation, and control spaces. We prove that there exists an ϵ -optimal finite-state controller (FSC) functionally independent of initial distributions for any $\epsilon > 0$, under the assumption that the optimal \liminf average cost function of the POMDP is constant. As part of our proof, we establish that if the optimal \liminf average cost function is constant, then the optimal \limsup average cost function is also constant, and the two are equal. We also discuss the connection between the existence of nearly optimal finite-history controllers and two other important issues for average cost POMDP: the existence of an average cost that is independent of the initial state distribution, and the existence of a bounded solution to the constant average cost optimality equation.

Key words: partially observable Markov decision processes; finite-state and control models; average cost criterion; optimality conditions

MSC2000 subject classification: Primary: 93E20, 93C41; secondary: 90C40, 93C55

OR/MS subject classification: dynamic programming/optimal control; partially observable Markov decision processes

History: Received May 22, 2006; revised May 21, 2007.

1. Introduction.

1.1. Average cost POMDP. We focus on the classical problem of a time-homogeneous partially observable Markov decision process (POMDP) in discrete time, with finite state, observation, and control spaces, and with the average cost criterion. Most of our results also hold for the case where the state space is finite, but the observation and control spaces may be infinite, while the cost per stage is bounded. This will be discussed later. We denote by S_t, Y_t, U_t the state, observation, and control at time $t = 0, 1, \dots$, which take values in finite sets \mathcal{S}, \mathcal{Y} , and \mathcal{U} , respectively. The state evolves from S_t to S_{t+1} , according to transition probabilities $p(S_{t+1} | S_t, U_t)$. At each time $t \geq 1$, an observation Y_t is generated according to transition probabilities $p(Y_t | S_t, U_{t-1})$ (we assume that at time 0, there is no observation, so Y_0 is a dummy variable). A randomized control U_t is generated with probabilistic dependence on the past history

$$H_t = (U_0, Y_1, \dots, U_{t-1}, Y_t).$$

The states are thus not observable to the controller. There is a per-stage cost $g(S_t, U_t)$ associated with applying control U_t at state S_t . We consider minimization with respect to the long-run average cost criterion, which we will now define more precisely.

For $t \geq 0$, the history space at time t is defined recursively by

$$\mathcal{H}_0 = \emptyset, \quad \mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{U} \times \mathcal{Y}.$$

A policy π is defined as a collection $(\mu_t)_{t \geq 0}$, where $\mu_t(U_t | H_t)$ is a probability distribution over the control space \mathcal{U} that depends on the history H_t . An initial state distribution ξ and a policy π induce a stochastic process $\{(S_t, Y_t, U_t)_{t \geq 0}\}$ with joint probability distribution $\mathbb{P}^{\xi, \pi}$, with respect to which expected costs of π are defined. The set of all such policies is denoted by Π . It is the common set of admissible policies of the POMDP for any given initial distribution. Denoting by E_ξ^π the expectation with respect to $\mathbb{P}^{\xi, \pi}$, we define the expected k -stage cost $J_k^\pi(\xi)$ of a policy $\pi \in \Pi$ for an initial distribution ξ by

$$J_k^\pi(\xi) = E_\xi^\pi \left\{ \sum_{t=0}^{k-1} g(S_t, U_t) \right\}.$$

Since the limit of $(1/k)J_k^\pi(\xi)$ as $k \rightarrow \infty$ does not necessarily exist, we define the *liminf* and *limsup average costs* of π by

$$J_-^\pi(\xi) = \liminf_{k \rightarrow \infty} \frac{1}{k} J_k^\pi(\xi), \quad J_+^\pi(\xi) = \limsup_{k \rightarrow \infty} \frac{1}{k} J_k^\pi(\xi), \quad (1)$$

which are the asymptotically best and worst, respectively, long-run average costs under π . The *optimal liminf* and *limsup average cost functions* are defined by

$$J_-^*(\xi) = \inf_{\pi \in \Pi} J_-^\pi(\xi), \quad J_+^*(\xi) = \inf_{\pi \in \Pi} J_+^\pi(\xi). \quad (2)$$

Following usual convention, we view $J_+^*(\xi)$ as the optimal average cost function and aim to find a policy that minimizes $J_+^\pi(\xi)$ (such a policy may depend on ξ). Note, however, that both $J_-^*(\xi)$ and $J_+^*(\xi)$ will be of interest in our analysis. We will say that policy π is ϵ -*liminf* or ϵ -*limsup optimal* for ξ if $J_-^\pi(\xi) \leq J_-^*(\xi) + \epsilon$ or $J_+^\pi(\xi) \leq J_+^*(\xi) + \epsilon$, respectively. We will simply say that π is ϵ -*liminf* or ϵ -*limsup optimal* if the respective relations hold for all ξ .

The analysis of the POMDP just defined is usually based on the well-known conversion to a time-homogeneous Markov decision process (MDP) in belief space, i.e., in the space $\mathcal{P}(\mathcal{S})$ of probability distributions over the state space. In this context, we view as state ξ_t , the distribution of S_t conditioned on the current history H_t , also referred to as *belief state at time t*. Assuming that there are n states numbered $1, \dots, n$, we may view a belief state ξ as an n -dimensional vector with components ξ^1, \dots, ξ^n . An important equation, referred to as the *constant average cost optimality equation*, is

$$\lambda + h(\xi) = \min_{u \in \mathcal{U}} [\xi' \bar{g}(u) + E[h(\phi(\xi, u, Y_1))]], \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \quad (3)$$

where λ is a scalar, h is a function of ξ , $\bar{g}(u)$ is the n -dimensional vector whose components are $g(1, u), \dots, g(n, u)$, the expectation $E[\cdot]$ is with respect to Y_1 conditioned on ξ and $U_0 = u$, and $\phi(\xi, u, Y_1)$ is the conditional distribution of state S_1 , given that the initial state distribution is ξ , initial control u is applied, and observation Y_1 is obtained (the expectation $E[\cdot]$ can be expressed in closed form from the state and observation transition probabilities).¹

It is well known² (see, e.g., Theorem 9.1.3 or Theorem 9.1.2(C), Puterman [15]) that when the constant average cost optimality Equation (3) admits a solution (λ, h) with h being a bounded function over $\mathcal{P}(\mathcal{S})$, then the optimal limsup and liminf costs are both equal to λ , and are independent of the initial state distribution ξ . Furthermore, if $\mu^*(\xi)$ attains the minimum in the right-hand side of this equation, the corresponding stationary policy μ^* (stationary with respect to the belief states) is average cost optimal. However, the available sufficient conditions for existence of a bounded solution (e.g., Ross [16], Platzman [14], Runggaldier and Stettner [17], Fernández-Gaucherand et al. [9], Hsu et al. [10]) tend to be specialized, and are often hard to verify. Indeed, examples indicate that for POMDP, the dependence of the optimal average cost on ξ is a far more complex issue (see §3), and is not yet fully understood. This is in contrast with finite-state MDP, for which the existence of solutions of the optimality equations, and the independence of the optimal average cost on the initial state are known to be related fundamentally to the chain structure of the MDP.

From a computational point of view, solving a POMDP by exact dynamic programming (DP) methods for continuous state-space MDP can be very difficult, even if there were no analytical difficulties as aforementioned. The reason is that the dimension of the belief space $\mathcal{P}(\mathcal{S})$ can be very high (so DP is hard to apply in belief space), while the size of the history space increases exponentially over time (so DP cannot be applied in history space for a long-length horizon). In addition, these methods rely on computing the conditional state distributions, which is generally intractable for POMDP with a large number of states.

Thus, approximations based on a cost or policy parametrization involving a finite set of parameters may be useful. Our purpose in this paper is to show the adequacy of one such policy parametrization, the finite-state controller (FSC) approach, which we describe in the next section. FSCs induce finite-state Markov chains in a POMDP, and thus simplify the average cost problem both analytically and computationally. Furthermore, they can be used for reinforcement learning in a partially observable environment where no knowledge of the model or the initial distribution is assumed known, and the only requirement is that the cost at each stage is a random variable observable to the controller.

¹ We remind the reader that more generally, the average cost optimality equations are a pair of nested equations, which reduce to Equation (3) when the optimal average cost is constant. This topic is, however, beyond the scope of this paper.

² To see this, note first that the statement is true for a countable space MDP and then note that for every initial state distribution, a finite space POMDP can be viewed as a countable space MDP on the set of belief states that are reachable from the initial belief state.

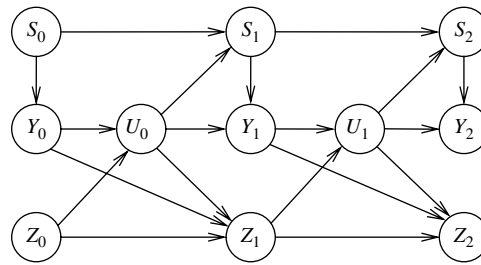


FIGURE 1. The graphical model of a POMDP controlled by an FSC, with directed edges showing the probabilistic dependence structure of the random variables. The random variables S_t and Y_t are the states and observations of the POMDP, and Z_t are the internal states of the FSC, which evolve in a Markov way.

1.2. FSCs. Roughly speaking, an FSC is a finite-state probabilistic automaton that accepts as inputs the observations Y_t and produces as outputs the controls U_t . The internal state of the FSC evolves probabilistically depending on the recent history, and the controller outputs a randomized control depending on its internal state. The internal state may involve estimates of certain quantities or predictions of certain events that are considered important for the control task—these can even be approximations of the conditional state distributions; see Aberdeen and Baxter [1], for instance. A simple special case is to use the internal state as a memory device that stores a finite-length portion of the recent history. The corresponding policy π can thus be represented by $\pi = (\mu_t)_{t \geq 0}$, where μ_t maps the recent history of length k , say, to a probability distribution on the control space. We refer to such a policy as a *finite-history controller*. An FSC is, however, both more efficient and more general than this finite-history controller, in that it can potentially *memorize* an infinitely long history by memorizing a finite number of key events (see the examples of §3).

Mathematically, a POMDP controlled by an FSC can be represented by a graphical model³ as shown in Figure 1. Denote the internal states of the FSC at time t by Z_t . The overall system is described by the transition model of the POMDP and the transition model of the FSC, which consists of the transition probabilities for the internal states, $P(Z_{t+1} | Z_t, Y_t, U_t)$, and the conditional probabilities for controls, $P(U_t | Z_t, Y_t)$. If each of the former transition probabilities assigns probability 1 to a unique internal state Z_{t+1} , and each of the latter conditional probabilities assigns probability 1 to a unique control U_t , we say that the FSC is *deterministic*. An important feature of a POMDP with an FSC, is that the joint process $\{(S_t, Y_t, Z_t, U_t)_{t \geq 0}\}$ of the states, observations, controls, and the internal states of the controller is a time-homogeneous finite-state Markov chain. As a result, the asymptotic behavior of the POMDP with an FSC can be understood and analyzed based on finite-state MDP theory (even though the states S_t are not observable). For instance, the average cost function (on the joint state space) together with the so-called differential cost, or bias, function satisfy a pair of average cost equations for an MDP, and the liminf and limsup average cost functions are equal.

Computationally, finding an FSC that is optimal over all FSC of given structure, is possible by using some parametric optimization technique. In particular, the transition probabilities of an FSC may contain an adjustable parameter vector θ ; for example, θ may consist of some or all of the transition probabilities of the FSC. For any fixed value of its parameter vector θ , an FSC $\pi(\theta)$ yields an average cost $J^{\pi(\theta)}(\xi)$, which may be optimized over θ if desired. This optimization can be attempted using one of the several methods proposed recently for optimization in policy space. For example, the gradient of $J^{\pi(\theta)}(\xi)$ with respect to θ can be estimated from sample trajectories, and gradient-based methods can be used for policy improvement (see, e.g., Jaakkola et al. [11], Baxter and Bartlett [2], Meuleau et al. [13], Aberdeen and Baxter [1], Yu [18]). This optimization over θ , however, is beyond our scope.

1.3. Scope of the paper. In this paper, we focus on the question whether for any $\epsilon > 0$, there exists an ϵ -optimal FSC π_ϵ , functionally independent of the initial distribution, i.e., one such that

$$J^{\pi_\epsilon}(\xi) \leq J_+^*(\xi) + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

We prove that this is true, assuming that the optimal liminf average cost is independent of ξ . The ϵ -optimal controller obtained in our proof is, in fact, a finite-history controller, and can be chosen to be deterministic. This

³ The graphical model is a way of specifying the conditional independence structure of random variables by a graph. Here, a directed acyclic graph $G = (V, E)$ with vertex set $V = \{X_1, \dots, X_n\}$ and edge set E indicates that the joint distribution of the random variables (X_1, \dots, X_n) is of the product form $\prod_{X \in V} p(X | X_{pa})$, where X_{pa} denotes the parent vertices of vertex X , i.e., vertices that are adjacent to the incoming edges of X . A reference to graphical models is Lauritzen [12].

proof also shows the existence of an ϵ -optimal finite-history controller, functionally independent of the initial distribution, in the case where the observation and control spaces are infinite (a finite-history controller is not an FSC if the observation and/or the control space are infinite).

As part of our analysis, we show that if the optimal liminf average cost is constant, then the optimal limsup average cost is also constant and is equal to the optimal liminf average cost, which is a new result. Furthermore, the optimal average costs remain unchanged if we restrict policies to be deterministic rather than randomized. On the other hand, when the optimal liminf average cost depends on the initial distribution ξ , an ϵ -optimal controller will likely also have to depend on ξ , so an FSC will typically be inappropriate in this case. For a simple example, suppose there are two states that are absorbing regardless of the control (so S_i stays constant), but the choice of u that minimizes the cost per stage $g(s, u)$ is different for different s . Then, if the observations provide no information about the system state, there can be no ϵ -optimal FSC for sufficiently small ϵ .

The existence of the various ϵ -optimal and ϵ -liminf optimal policies, independent of the initial distribution, may also be viewed as necessary conditions for the optimal average cost to be a constant and the constant average cost optimal equation to admit a bounded solution. They complement the existing results on the average cost POMDP problem.

We finally note that our line of analysis is unrelated to commonly used methods of proof for showing the existence of bounded solutions of the optimality equations and the existence of stationary optimal policies. Instead, our proofs are based on concavity and convexity properties of various cost functions and on methods of convex analysis.

The paper is organized as follows. In §2, we prove our main results. In §3, we illustrate some aspects of our analysis through examples. In §4, we discuss other implications of our results.

Throughout the paper, by a near-optimal policy, we mean a policy that is near optimal for all initial state distributions ξ , unless we explicitly mention that it is near optimal for a given ξ . Recall also that policies in Π , including the FSC we will consider, do not functionally depend on the initial state distribution.

2. Main results. In this section, we show that if J_-^* is a constant function, then $J_-^* = J_+^*$ and there exists an ϵ -optimal FSC that is functionally independent of the initial distribution. This result (Corollary 2.1) is obtained as a consequence of the existence of an ϵ -optimal finite-history controller (Theorem 2.1), which does not require the assumption of finiteness of the observation and the control spaces. The latter result is obtained after proving two intermediate results: (1) the existence of an ϵ -liminf optimal policy (see Proposition 2.1) and (2) the existence of an ϵ -liminf optimal finite-history policy (see Proposition 2.2), both functionally independent of the initial state distribution.

First, we prove some basic concavity and convexity properties of the average cost functions. We show, in particular, that for all policies $\pi \in \Pi$, $J_-^\pi(\cdot)$ is a concave function, and $J_+^\pi(\cdot)$ is a convex function, and the optimal liminf average cost $J_-^*(\cdot)$ is a concave function. The concavity of $J_-^*(\cdot)$ has not been observed in the POMDP literature, to our knowledge.

LEMMA 2.1. *The liminf average cost function $J_-^\pi(\cdot)$ is concave for all $\pi \in \Pi$, and the optimal liminf average cost function $J_-^*(\cdot)$ is also concave.*

PROOF. Clearly, $(1/k)J_k^\pi(\xi)$ is a linear function of ξ , so that

$$\frac{1}{k}J_k^\pi(\xi) = \sum_{i=1}^m \alpha_i \frac{1}{k}J_k^\pi(\xi_i)$$

for any convex combination $\xi = \sum_{i=1}^m \alpha_i \xi_i$. By the nonnegativity of α_i and the inequality

$$\liminf_{n \rightarrow \infty} (a_n + b_n) \geq \liminf_{n \rightarrow \infty} a_n + \liminf_{n \rightarrow \infty} b_n,$$

which holds for any scalar sequences $\{a_n\}$ and $\{b_n\}$, we have

$$J_-^\pi(\xi) = \liminf_{k \rightarrow \infty} \sum_{i=1}^m \alpha_i \frac{1}{k}J_k^\pi(\xi_i) \geq \sum_{i=1}^m \alpha_i \liminf_{k \rightarrow \infty} \frac{1}{k}J_k^\pi(\xi_i) = \sum_{i=1}^m \alpha_i J_-^\pi(\xi_i),$$

so $J_-^\pi(\cdot)$ is concave. Since $J_-^*(\xi) = \inf_{\pi \in \Pi} J_-^\pi(\xi)$ by definition, $J_-^*(\cdot)$ is concave, being the pointwise infimum of concave functions. \square

REMARK 2.1. We cannot assert that $J_+^*(\cdot)$ is necessarily concave or convex. If $J_+^*(\cdot) = J_-^*(\cdot)$, then $J_+^*(\cdot)$ is concave by the preceding lemma; however, whether the optimal liminf and limsup average costs are equal for a POMDP is still an open question (see Remark 2.5). On the other hand, although one can show similar to Lemma 2.1 that $J_+^\pi(\cdot)$ is convex, using the inequality

$$\limsup_{n \rightarrow \infty} (a_n + b_n) \leq \limsup_{n \rightarrow \infty} a_n + \limsup_{n \rightarrow \infty} b_n,$$

$J_+^*(\xi)$ is defined as $\inf_{\pi \in \Pi} J_+^\pi(\xi)$, so being the pointwise infimum of convex functions, it may be neither convex nor concave.

REMARK 2.2. Another interesting fact, which we will not use in our analysis, is that $J_-^*(\cdot)$ and $J_+^*(\cdot)$ are Lipschitz continuous with a Lipschitz constant $C = 2 \sup_{s,u} |g(s, u)|$. This is because of the fact that for all $\pi \in \Pi$, the functions $J_-^\pi(\cdot)$ and $J_+^\pi(\cdot)$ are Lipschitz continuous with the same Lipschitz constant, as can be easily shown.

We now show that when $J_-^*(\cdot)$ is a constant function, an ϵ -liminf optimal controller exists (which is independent of the initial state distribution). This is the first key step in proving the near optimality of FSC. The intuition behind the analysis is as follows. The function $J_-^\pi(\cdot)$ is bounded below by $J_-^*(\cdot)$, the pointwise infimum of $J_-^\pi(\cdot)$ over $\pi \in \Pi$. However, $J_-^\pi(\cdot)$ is concave and π can be chosen, so that $J_-^\pi(\hat{\xi}) \leq J_-^*(\hat{\xi}) + \delta$ for some interior point $\hat{\xi}$ of $\mathcal{P}(\mathcal{S})$, and for any $\delta > 0$. Thus, if $J_-^*(\cdot)$ is constant, the concave function $J_-^\pi(\cdot)$ cannot be simultaneously bounded below by $J_-^*(\cdot)$ and be *too different* from $J_-^*(\cdot)$, implying that π is near-liminf optimal for all initial distributions ξ . This observation is stated and proved formally in the following proposition. Abusing notation, we use J_-^* to denote the constant value of the function $J_-^*(\cdot)$, when the latter is constant.

PROPOSITION 2.1. *If $J_-^*(\cdot)$ is a constant function, then for any $\epsilon > 0$, there exists a history-dependent randomized policy $\pi \in \Pi$ that is ϵ -liminf optimal, i.e.,*

$$J_-^\pi(\xi) \leq J_-^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

The proof starts with a lemma, which uses some basic notions of convex analysis such as the relative interior and relative boundary of a convex set.⁴ We denote by \mathfrak{R}^n the n -dimensional Euclidean space with the standard norm $\|\cdot\|$.

LEMMA 2.2. *Let X be a convex and compact subset of \mathfrak{R}^n , with relative interior and relative boundary denoted by $\text{ri}(X)$ and D , respectively. Let $f: X \rightarrow \mathfrak{R}, f(x) \geq 0$ be a nonnegative concave function. Then, for any $\hat{x} \in \text{ri}(X)$ and any $x \in X$,*

$$f(x) \leq C_{\hat{x}} f(\hat{x}), \quad \text{where } C_{\hat{x}} = \left(\frac{\max_{z \in X} \|z - \hat{x}\|}{\min_{z \in D} \|z - \hat{x}\|} + 1 \right).$$

PROOF. For any $\hat{x} \in \text{ri}(X)$ and $x \in X$, let $r(x)$ be the intersection point of the relative boundary D and the ray that starts at x and passes through \hat{x} . By the concavity and nonnegativity of f , we have

$$\begin{aligned} f(\hat{x}) &\geq \frac{\|\hat{x} - r(x)\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(x) + \frac{\|x - \hat{x}\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(r(x)) \\ &\geq \frac{\|\hat{x} - r(x)\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(x). \end{aligned}$$

Hence

$$f(x) \leq \frac{\|x - \hat{x}\| + \|\hat{x} - r(x)\|}{\|\hat{x} - r(x)\|} f(\hat{x}) \leq \left(\frac{\max_{z \in X} \|z - \hat{x}\|}{\min_{z \in D} \|z - \hat{x}\|} + 1 \right) f(\hat{x}),$$

and the claim follows. \square

⁴ Let X be a subset of \mathfrak{R}^n . The *affine hull* $\text{aff}(X)$ of X is defined to be $S + \bar{x}$, where \bar{x} is an arbitrary point in X and S is the subspace spanned by $X - \bar{x}$. A point $x \in X$ is called a *relative interior point* if there exists an open neighborhood $\mathcal{N}(x)$ of x such that $\mathcal{N}(x) \cap \text{aff}(X) \subset X$. The set of relative interior points of X is called the *relative interior* of X , and is denoted by $\text{ri}(X)$. A convex set always has a nonempty relative interior. Let $\text{cl}(X)$ be the closure of X . The set $\text{cl}(X) \setminus \text{ri}(X)$ is called the *relative boundary* of X .

PROOF OF PROPOSITION 2.1. Since the state space \mathcal{S} is finite, $\mathcal{P}(\mathcal{S})$ is a convex and compact set in \mathfrak{R}^n . Let D be the relative boundary of $\mathcal{P}(\mathcal{S})$. Pick an arbitrary $\hat{\xi}$ in the relative interior of $\mathcal{P}(\mathcal{S})$, and let $C_{\hat{\xi}}$ be defined as in Lemma 2.2.

For any $\epsilon > 0$, let $\pi \in \Pi$ be a policy such that $J_{-}^{\pi}(\hat{\xi}) \leq J_{-}^{*} + \epsilon/C_{\hat{\xi}}$. By Lemma 2.1, $J_{-}^{\pi}(\cdot)$ is concave, and by definition of J_{-}^{*} , we have $J_{-}^{\pi}(\xi) \geq J_{-}^{*}$ for all $\xi \in \mathcal{P}(\mathcal{S})$. Applying Lemma 2.2 to the concave and nonnegative function $J_{-}^{\pi}(\xi) - J_{-}^{*}$, we have

$$J_{-}^{\pi}(\xi) - J_{-}^{*} \leq C_{\hat{\xi}}(J_{-}^{\pi}(\hat{\xi}) - J_{-}^{*}) \leq C_{\hat{\xi}} \frac{\epsilon}{C_{\hat{\xi}}} \leq \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

i.e., $J_{-}^{\pi}(\xi) \leq J_{-}^{*} + \epsilon$ and π is ϵ -liminf optimal. \square

We now use Proposition 2.1 and its line of proof to show that the set of finite-history policies contains a near-liminf optimal policy. First, we prove a lemma that will allow us to truncate at some finite stage a certain history-dependent randomized policy that is near-liminf optimal, and form another sufficiently good policy that only uses the corresponding truncated finite history. The idea is that for the policy π of Proposition 2.1 and some sufficiently large k , the linear k -stage cost $(1/k)J_k^{\pi}$ is almost bounded below by the almost constant function J_{-}^{π} and is close to J_{-}^{π} at some interior point $\hat{\xi}$ of $\mathcal{P}(\mathcal{S})$, so by appealing again to Lemma 2.2, it is uniformly close to J_{-}^{*} .

For all states $s \in \mathcal{S}$, we denote by $e_s \in \mathcal{P}(\mathcal{S})$ the distribution that assigns probability 1 to s , i.e., $e_s(\{s\}) = 1$.

LEMMA 2.3. *If $J_{-}^{*}(\cdot)$ is a constant function, then for any $\epsilon > 0$, there exist a policy $\pi \in \Pi$ and an integer k (depending on π) such that*

$$\left| \frac{1}{k} J_k^{\pi}(\xi) - J_{-}^{*} \right| \leq \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

PROOF. The proof bears similarity with the proof of Proposition 2.1. Pick an arbitrary $\hat{\xi}$ in the relative interior of $\mathcal{P}(\mathcal{S})$, and let $C_{\hat{\xi}}$ be defined as in Lemma 2.2. For any $\epsilon > 0$, let $\delta = \epsilon/3C_{\hat{\xi}} < \epsilon/3$. By Proposition 2.1, we can choose a policy $\pi \in \Pi$ such that

$$J_{-}^{\pi}(\xi) \leq J_{-}^{*} + \delta, \quad \forall \xi \in \mathcal{P}(\mathcal{S}). \quad (4)$$

Since \mathcal{S} is finite and by definition $J_{-}^{\pi}(\cdot)$ is the pointwise liminf of the functions $\{(1/k)J_k^{\pi} \mid k \geq 1\}$, there exists K_1 such that

$$\frac{1}{k} J_k^{\pi}(e_s) \geq J_{-}^{\pi}(e_s) - \delta \geq J_{-}^{*} - \delta, \quad \forall s \in \mathcal{S}, \quad k \geq K_1.$$

Since J_k^{π} is a linear function, it follows that

$$\frac{1}{k} J_k^{\pi}(\xi) \geq J_{-}^{*} - \delta, \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \quad k \geq K_1. \quad (5)$$

For the $\hat{\xi}$ that we picked earlier in the proof, by the definition of $J_{-}^{\pi}(\cdot)$ and Equation (4), there exists $K_2 > K_1$, and a $k \geq K_2$ such that

$$\frac{1}{k} J_k^{\pi}(\hat{\xi}) \leq J_{-}^{\pi}(\hat{\xi}) + \delta \leq J_{-}^{*} + 2\delta. \quad (6)$$

By our choice of k and Equation (5), $(1/k)J_k^{\pi}(\xi) - J_{-}^{*} + \delta$ is a nonnegative function that is concave (since it is linear). Therefore, applying Lemma 2.2, we have

$$\frac{1}{k} J_k^{\pi}(\xi) - J_{-}^{*} + \delta \leq C_{\hat{\xi}} \left(\frac{1}{k} J_k^{\pi}(\hat{\xi}) - J_{-}^{*} + \delta \right) \leq 3C_{\hat{\xi}}\delta \leq \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

where the second inequality is because of Equation (6) and the third inequality is because of our choice of δ . Combining the above relation with Equation (5), we thus have

$$J_{-}^{*} - \epsilon \leq \frac{1}{k} J_k^{\pi}(\xi) \leq J_{-}^{*} + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

and the proof is complete. \square

For any $\epsilon > 0$, consider the policy π and the integer k of Lemma 2.3. We will construct a corresponding finite-history policy that is ϵ -liminf optimal. Let $\bar{k} = k$. If $\pi = (\mu_t)_{t \geq 0}$, we form a policy $\bar{\pi} = (\mu'_t)_{t \geq 0}$ by repeating the control rules of π for every \bar{k} -stage interval as follows. For any t , define $m(t) = \text{mod}(t, \bar{k})$, and define

$$\mu'_t(U_t | h_t) = \mu_{m(t)}(U_t | \delta_{m(t)}(h_t)),$$

where $\delta_{m(t)}: \mathcal{H}_t \rightarrow \mathcal{H}_{m(t)}$ maps a length- t history h_t to a length- $m(t)$ history by extracting the last length- $m(t)$ segment of h_t . Since π is ϵ -liminf optimal uniformly for all initial distributions, the policy $\bar{\pi}$ is also ϵ -liminf optimal, and indeed ϵ -optimal, as the following proposition and theorem indicate.

PROPOSITION 2.2. *If $J^*(\cdot)$ is a constant function, then for any $\epsilon > 0$, there exist an integer \bar{k} (depending on ϵ) and an ϵ -liminf optimal policy $\bar{\pi} \in \Pi$ such that $\bar{\pi}$ at each stage depends functionally only on the history of the most recent \bar{k} stages.*

PROOF. Consider the POMDP controlled by $\bar{\pi}$ defined in the discussion preceding the statement of the proposition. Let $\{\xi_t\}$ be the random process of conditional distributions of states S_t given H_t ; the observed history up to time t and before control U_t has been applied. The average of the $n\bar{k}$ -stage cost of $\bar{\pi}$ for an integer n can be written as

$$\begin{aligned} \frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) &= \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{\bar{k}} E_{\xi_0}^{\bar{\pi}} \left\{ \sum_{t=i\bar{k}}^{(i+1)\bar{k}-1} g(S_t, U_t) \right\} \\ &= \frac{1}{n} \sum_{i=0}^{n-1} E_{\xi_0}^{\bar{\pi}} \left\{ \frac{1}{\bar{k}} J_{\bar{k}}^{\bar{\pi}}(\xi_{i\bar{k}}) \right\}. \end{aligned}$$

By Lemma 2.3, the \bar{k} -stage average cost of $\bar{\pi}$ is uniformly within ϵ of the optimal J^* ; hence

$$\frac{1}{\bar{k}} J_{\bar{k}}^{\bar{\pi}}(\xi_{i\bar{k}}) \leq J^* + \epsilon$$

and

$$\frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) \leq \frac{1}{n} \sum_{i=0}^{n-1} (J^* + \epsilon) = J^* + \epsilon. \quad (7)$$

Consequently,

$$J_-^{\bar{\pi}}(\xi_0) \leq \liminf_{n \rightarrow \infty} \frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) \leq J^* + \epsilon;$$

i.e., the liminf average cost of $\bar{\pi}$ is also uniformly close to the optimal J^* . \square

The preceding proof implies more than what the proposition claims. Notice that by Equation (7), we also have

$$\limsup_{n \rightarrow \infty} \frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) \leq J^* + \epsilon.$$

On the other hand, for all positive integers $m < \bar{k}$, the total $(n\bar{k} + m)$ -stage cost $J_{n\bar{k}+m}^{\bar{\pi}}(\xi_0)$ differs from the total $n\bar{k}$ -stage cost $J_{n\bar{k}}^{\bar{\pi}}(\xi_0)$ by a finite amount, so we have

$$J_-^{\bar{\pi}}(\xi_0) = \liminf_{n \rightarrow \infty} \frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) \leq \limsup_{n \rightarrow \infty} \frac{1}{n\bar{k}} J_{n\bar{k}}^{\bar{\pi}}(\xi_0) = J_+^{\bar{\pi}}(\xi_0).$$

It thus follows that

$$J_-^{\bar{\pi}}(\xi) \leq J_+^{\bar{\pi}}(\xi) \leq J^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \quad (8)$$

and hence $\bar{\pi}$ is also ϵ -limsup optimal and

$$J_-^* = J_+^*. \quad (9)$$

It is worth noting here that the preceding analysis makes use of the finiteness of the state space, but does not make use of the finiteness of the observation and control spaces. So the associated results, including Propositions 2.1 and 2.2 and relevant discussions, hold for possibly infinite observation and control spaces. We will comment on this point later.

Under our assumption that the state, observation, and control spaces are finite, the near-optimal finite-history controller $\bar{\pi}$ in Proposition 2.2 can be viewed as an FSC with the internal state memorizing the current stage number modulo \bar{k} and the most recent length- \bar{k} sample path with $k \leq \bar{k}$. Hence Proposition 2.2 and the discussions above yield the existence of an ϵ -optimal FSC.

Before stating these conclusions as a theorem, we extend our analysis further to assert the near-optimality of deterministic finite-history controllers and deterministic FSC. We have so far considered randomized policies.

If instead we restrict policies to be deterministic, history dependent, a corresponding deterministic analog of the preceding analysis can be shown, using essentially the same proof. In particular, let $\bar{J}_-^*(\cdot)$ and $\bar{J}_+^*(\cdot)$ denote the liminf and limsup cost functions that are optimal over all deterministic, history-dependent policies. Then, if $\bar{J}_-^*(\cdot)$ is a constant function, we have $\bar{J}_-^* = \bar{J}_+^*$ and there exists a deterministic FSC, functionally independent of the initial distribution, that is, ϵ -optimal among the deterministic policies. Thus, if $J_-^*(\cdot) = \bar{J}_-^*(\cdot)$, then the assumption of $J_-^*(\cdot)$ being a constant function implies the existence of a deterministic finite-history controller, functionally independent of the initial distribution, that is ϵ -optimal among all policies. Indeed, it is true that $J_-^*(\xi) = \bar{J}_-^*(\xi)$ for all ξ , whether $J_-^*(\cdot)$ is constant or not. This follows from Feinberg’s result (see Theorem 3.2B and §4.4 of Feinberg [7]), applied to the belief MDP derived from the POMDP, which states that for any given initial distribution and randomized policy, there exists a deterministic policy that has smaller or equal liminf average cost. Thus we have the theorem as follows.

THEOREM 2.1. *If $J_-^*(\cdot)$ is a constant function, then $J_-^* = J_+^* = \bar{J}_-^* = \bar{J}_+^*$, and for any $\epsilon > 0$ there exists a finite-history controller, functionally independent of the initial distribution, that is ϵ -optimal. Moreover, this finite-history controller can be chosen to be deterministic.*

COROLLARY 2.1. *If $J_-^*(\cdot)$ is a constant function, then for any $\epsilon > 0$ there exists an FSC, functionally independent of the initial distribution, that is ϵ -optimal. Moreover, this FSC can be chosen to be deterministic.*

REMARK 2.3. Since Feinberg’s paper [7] is couched on a more general framework, it may be useful to give a proof that does not rely on his results, for the part of Theorem 2.1 relating to deterministic finite-history controllers. So we provide here an alternative proof of $J_-^* = \bar{J}_-^*$ using the assumption that J_-^* is constant. By Lemma 2.3, we have for some $\pi \in \Pi$ and integer k ,

$$\frac{1}{k} J_k^\pi(\xi) \leq J_-^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Thus the optimal k -stage cost function satisfies

$$\frac{1}{k} J_k^*(\xi) \leq \frac{1}{k} J_k^\pi(\xi) \leq J_-^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

We now take the optimal k -stage policy, which is deterministic and depends on the initial distribution, and apply it for every k -stage time interval. The resulting infinite-stage policy, call it π_d , uses at the beginning of each k -stage time interval the current conditional state distribution, which is computed from the conditional state distribution at the beginning of the preceding k -stage time interval, and the information collected during the k -stage time interval. By the same argument, as in the proof of Proposition 2.2, we have

$$J_-^{\pi_d}(\xi) \leq J_-^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Since ϵ can be arbitrarily small and

$$J_-^* \leq \bar{J}_-^*(\xi) \leq J_-^{\pi_d}(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

we conclude that $J_-^* = \bar{J}_-^*$.

REMARK 2.4. Propositions 2.1 and 2.2 and Theorem 2.1 apply to the more general case of a finite-state space and general observation and control spaces. In that case, a policy $\pi \in \Pi$ is a collection of stochastic transition kernels $(\mu_t)_{t \geq 0}$, where $\mu_t(du_t | h_t)$ is a probability measure on the control space for each $h_t \in \mathcal{H}_t$, the history space, and $\mu_t(A | h_t)$ is a measurable function on \mathcal{H}_t for each Borel-measurable set A of the control space. Here, measurability on the history space is with respect to either the Borel sigma algebra or the universal sigma algebra. Furthermore, some additional technical conditions need to be imposed on the cost per stage and/or the state transition probabilities, such as semicontinuity or semianalyticity. We note that while these measurability notions are needed here for the mathematical validity of the optimality equations and the existence of measurable optimal or near-optimal policies (see Bertsekas and Shreve [3] or Dynkin and Yushkevich [5]), they do not affect basic properties such as linearity, concavity, and convexity that form the basis of our analysis. Thus the proofs of Propositions 2.1 and 2.2 and Theorem 2.1 apply to general observation and control spaces with no modification necessary other than starting out with a mathematical framework that deals adequately with the technical measurability issues.

REMARK 2.5. To our knowledge, it is an open question whether $J_-^*(\xi) = J_+^*(\xi)$ or $J_+^*(\xi) = \bar{J}_+^*(\xi)$ for all ξ without the assumption that $J_-^*(\cdot)$ is constant. Indeed, there is an example of an average cost MDP with a countable state space and a finite control space, where the optimal limsup average cost is strictly larger than the

optimal \liminf average cost. Moreover, in this example, there exist no policies simultaneously \liminf and \limsup optimal, and the optimal \limsup average cost over deterministic policies is strictly larger than the optimal \limsup average cost over randomized policies (see Dynkin and Yushkevich [5], Chapter 7, Example 3; Feinberg [6], §5). However, there are no known examples of a finite-state POMDP, or equivalently, a belief MDP derived from a finite-state POMDP, where these happen.

Related to the preceding remark, there are two interesting results in the MDP literature, pointed out to us by the reviewer. In addition to the \liminf and \limsup average cost criteria, Bierth [4] considers two more criteria:

$$\inf_{\pi \in \Pi} E_{\xi}^{\pi} \left\{ \liminf_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} g(S_t, U_t) \right\}, \quad \inf_{\pi \in \Pi} E_{\xi}^{\pi} \left\{ \limsup_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} g(S_t, U_t) \right\}, \quad (10)$$

which are a lowerbound of $J_{-}^{*}(\cdot)$ and upperbound of $J_{+}^{*}(\cdot)$, respectively. Bierth [4] shows that the optimal average costs under the four criteria are all equal for finite-state MDP with arbitrary measurable action sets, while the weaker equality of the optimal \liminf and \limsup average costs is first proved by Feinberg [6] for the same class of MDP. For POMDP with either of the two criteria in (10), (as well as for MDP), the optimal value over all policies is the same as the one over the deterministic policies by the argument of convexity criteria in Feinberg [7].⁵

3. Examples. We first demonstrate by an example of an ϵ -optimal policy as in Proposition 2.1 and an FSC as in Corollary 2.1. In this example, the optimal average cost function of the POMDP is constant, but the constant average cost optimality Equation (3) does not have a bounded solution.

EXAMPLE 3.1 (PLATZMAN [14]). The POMDP has two states $\{1, 2\}$, three observations $\{1, 2, 3\}$, and three actions $\{1, 2, 3\}$. Under any action, the state remains the same; i.e.,

$$p(S_{t+1} = i | S_t = i, U_t = -) = 1, \quad i = 1, 2,$$

where the symbol “−” represents any action. By applying actions, one can gain information or rewards, but these two goals are *mutually exclusive*. Under action 1 or 2, the observation 3 is generated, which bears no information of the state. Under action 3, the correct state is observed with probability $p > 1/2$; i.e.,

$$p(Y_t = i | S_t = i, U_{t-1} = 3) = p, \quad i = 1, 2.$$

The per-stage costs are

$$\begin{aligned} g(1, 1) &= g(2, 2) = -1, \\ g(1, 2) &= g(2, 1) = g(1, 3) = g(2, 3) = 0. \end{aligned}$$

Thus, if the state can be guessed correctly, the corresponding noninformative action is best and produces a cost of -1 .

Consider a policy that applies action 3 with a diminishing frequency, and applies action 1 or 2 otherwise, depending on whether state 1 or 2 is more likely according to the current belief. Then, the state will eventually be guessed with very high probability, and the corresponding average cost can be arbitrarily close to -1 . Thus the optimal average cost function is constant and equals -1 . However, the constant average cost Equation (3) does not have a bounded solution. To see this, note that if it did, then there would exist an optimal deterministic stationary policy, which would associate with each belief ξ a unique optimal action. Clearly, for any ξ with $0 < \xi^1 < 1$ and $0 < \xi^2 < 1$, actions 1 or 2 cannot be optimal. But if action 3 is optimal for all ξ in the relative interior of the distribution space (i.e., the set of ξ with $\xi^i > 0$, $i = 1, 2$), then since starting from such ξ , the distribution always remains in the relative interior, action 3 is applied all the time, thereby incurring an average cost of zero, which cannot be optimal. We thus arrive at a contradiction, showing that Equation (3) does not admit a bounded solution. The preceding discussion also shows that neither does the optimality equation admit an unbounded solution in the sense defined and analyzed by Fernández-Gaucherand et al. [9].⁶

⁵ Essentially, this is because in POMDP for any initial distribution, the probability measure on the space of $\{(S_t, Y_t, U_t)_{t \geq 0}\}$ induced by a history-dependent policy can be viewed as a convex combination of the probability measures induced by deterministic policies. A proof follows from the argument of Theorem 1 of Feinberg [8] and Theorem 3.2 and §4.2 of Feinberg [7].

⁶ In Fernández-Gaucherand et al. [9], the solution may be unbounded on the space $\mathcal{P}(\mathcal{S})$, but it is bounded when restricted to a countable set of belief states that are reachable from a given initial state distribution, with the bound also depending on the initial distribution.

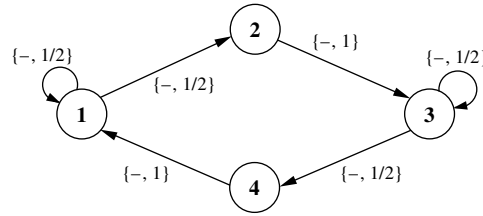


FIGURE 2. The symbolic representation of the MDP of Example 3.2. Possible state transitions are indicated by directed edges with {action, probability} values, and the symbol “-” stands for either action a or b .

For this example, the near-liminf optimal history-dependent policy as in Proposition 2.1 can be chosen to be optimal as follows. Choose a sequence of times t_k at which action 3 will be applied infinitely often and with diminishing frequency; i.e.,

$$\lim_{k \rightarrow \infty} t_k = \infty, \quad \lim_{t \rightarrow \infty} \frac{\max\{k \mid t_k \leq t\}}{t} = 0.$$

At time t , let $n_t(1)$ be the number of times that observation 1 is obtained up to time t , and $n_t(2)$ the number of times that observation 2 occurred up to time t . At time $t \neq t_k$, apply action 1 if $n_t(1) \geq n_t(2)$, and apply action 2 otherwise. By the law of large numbers, it is easy to show that such a policy has average cost -1 , and is therefore optimal.

The near-optimal FSC in Corollary 2.1 can be chosen as follows. During the first K stages, the controller applies action 3, and counts the number of times that observations 1 and 2 occurred. At time K , if more “1”s than “2”s have been observed, then it applies action 1 for the rest of the time; otherwise, it applies action 2 for the rest of the time. By the law of large numbers, clearly, for any $\epsilon > 0$, there exists a K sufficiently large such that the policy is ϵ -optimal. \square

Our next example relates to the assumption of a constant optimal average cost used in our analysis. This example shows that even in a very simple POMDP, an uncontrolled Markov chain, the optimal average cost can depend on the initial state distribution. By contrast, the optimal average cost would be constant in this example if the states of this Markov chain were observable.

EXAMPLE 3.2. The MDP has four states $\{1, 2, 3, 4\}$ and two actions $\{a, b\}$, and its transition structure is shown in Figure 2, where we use the symbol “-” to represent any action. Under any policy, the state process is a Markov chain with transition probabilities as follows: $p(1 \mid 1, -) = 1/2$, $p(2 \mid 1, -) = 1/2$; $p(3 \mid 2, -) = 1$; $p(3 \mid 3, -) = 1/2$, $p(4 \mid 3, -) = 1/2$; $p(1 \mid 4, -) = 1$. Clearly the Markov chain is recurrent and aperiodic.

The observation structure is such that the states $\{1, 3\}$ are indistinguishable and $\{2, 4\}$ are indistinguishable. In particular, let the observation space be $\{c, d\}$ and let $p(c \mid 1, -) = p(c \mid 3, -) = 1$; $p(d \mid 2, -) = p(d \mid 4, -) = 1$. Thus, if we know the initial state, the state process can be inferred from the observations as if it were completely observable.

We define the per-stage costs as $g(1, a) = g(3, b) = 1$ and all the other per-stage costs to be zero. It follows that if we start from an initial distribution ξ with $\xi^1 = 1$ or an initial distribution with $\xi^3 = 1$, then the optimal average cost is zero, while if we start from an initial distribution ξ with $\xi^1 = \xi^3 = 1/2$, say, then the optimal average cost is strictly greater than zero.

In this example, near-optimal control is not possible without taking into account the initial state distribution. The necessary condition for a constant optimal average cost function as given in Proposition 2.1, i.e., the existence of a near-optimal policy that has no functional dependence on the initial distribution, is violated.

The state evolution of this POMDP example is uncontrolled. See Yu [19] for another controlled POMDP example that has a nonconstant optimal average cost function and has an associated completely observable MDP that is recurrent and aperiodic (under every stationary and deterministic policy). \square

The next example demonstrates that an FSC is more general and efficient than a finite-history controller, and it can potentially memorize an infinitely long history by memorizing a finite number of key events. In particular, for some initial state distributions and small ϵ , there exists an ϵ -optimal FSC, but there is no ϵ -optimal finite-history controller.

EXAMPLE 3.3. We modify the preceding example by adding three more states, all transient, designated by 0, $(1, 2)$, $(3, 4)$, respectively. All the state transitions from these states are uncontrolled. From state 0, the next state is $(1, 2)$ or $(3, 4)$ with equal probability, and from state $(1, 2)$, the next state is 1 or 2 with equal probability, and similarly from state $(3, 4)$, the next state is 3 or 4 with equal probability. We define the observations of states $(1, 2)$ and $(3, 4)$ to be the state itself.

Consider starting from state 0. By letting the internal state memorizing which of the two states (1, 2) or (3, 4) occurs, an FSC can be chosen to be optimal for initial state 0, while no finite-history controllers can be ϵ -optimal in this case. To see this, note that by memorizing a fixed length of the recent history, no matter how long the length is, with probability 1 the finite-history controller will eventually not be able to distinguish state one from three, and therefore it will apply the same control rule in both states. \square

4. Concluding remarks. In this paper, we have explored the existence of near-optimal FSCs in a POMDP. Such controllers are computationally important because they depend on a finite-dimensional approximation of the available history, and can be optimized by choice of a finite set of parameters. They are also theoretically important, because, as we have shown in this paper, their existence is connected to the fundamental question of existence of an optimal average cost that is independent of the initial state distribution, and the related question of existence of a bounded solution to the constant average cost optimality equation.

Our results can be interpreted in various ways. They seem to reinforce the intuitive idea that if the optimal average cost of a POMDP is independent of the initial state distribution, then information becomes increasingly obsolete as time progresses. They also suggest in some sense that it might be relatively strong to assume for a POMDP that the constant average cost DP equation admits a bounded solution, and that we should study the more general case.

An interesting open question relates to whether it is possible to characterize information structures for POMDP, which guarantee that finite histories are adequate for near-optimal control. Another interesting research direction is to derive conditions guaranteeing the existence of near-optimal FSC for a single given initial distribution (rather than all distributions) when the optimal average cost of the POMDP is not constant.

Acknowledgments. The authors thank Professor Eugene Feinberg for helpful comments regarding his papers. (Feinberg [6, 7]), and other suggestions. The authors also thank Professor Sanjoy Mitter, Professor John Tsitsiklis, and Professor Vivek Borkar for stimulating and helpful discussions. This work was done while Huizhen Yu was a Ph.D. student at the Laboratory for Information and Decision Systems, M.I.T. This work was supported by National Science Foundation Grant ECS-0218328.

References

- [1] Aberdeen, D., J. Baxter. 2001. Internal-state policy-gradient algorithms for infinite-horizon POMDPs. Technical report, RSISE, Australian National University, Canberra, Australia.
- [2] Baxter, J., P. L. Bartlett. 2001. Infinite-horizon policy-gradient estimation. *J. Artificial Intelligence Res.* **15** 319–350.
- [3] Bertsekas, D. P., S. Shreve. 1978. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, New York.
- [4] Bierth, K. J. 1987. An expected average reward criterion. *Stochastic Process. Appl.* **26** 123–140.
- [5] Dynkin, E. B., A. A. Yushkevich. 1979. *Controlled Markov Processes*. Springer-Verlag, New York.
- [6] Feinberg, E. A. 1980. An ϵ -optimal control of a finite Markov chain with an average reward criterion. *Theory Probab. Appl.* **25** 70–81.
- [7] Feinberg, E. A. 1982. Controlled Markov processes with arbitrary numerical criteria. *Theory Probab. Appl.* **27** 486–503.
- [8] Feinberg, E. A. 1982. Nonrandomized Markov and semi-Markov strategies in dynamic programming. *Theory Probab. Appl.* **27** 116–126.
- [9] Fernández-Gaucherand, E., A. Arapostathis, S. I. Marcus. 1991. On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Ann. Oper. Res.* **29** 439–470.
- [10] Hsu, S.-P., D.-M. Chuang, A. Arapostathis. 2006. On the existence of stationary optimal policies for partially observed MDPs under the long-run average cost criterion. *Systems Control Lett.* **55** 165–173.
- [11] Jaakkola, T. S., S. P. Singh, M. I. Jordan. 1995. Reinforcement learning algorithm for partially observable Markov decision problems. *Proc. Neural Inform. Processing Systems Conf., Denver, CO*. MIT Press, Cambridge, MA.
- [12] Lauritzen, S. L. 1996. *Graphical Models*. Oxford University Press, Oxford, UK.
- [13] Meuleau, N., L. Peshkin, K.-E. Kim, L. P. Kaelbling. 1999. Learning finite-state controllers for partially observable environment. *Proc. 15th Conf. Uncertainty in Artificial Intelligence, Stockholm, Sweden*. Morgan Kaufmann, San Francisco.
- [14] Platzman, L. K. 1980. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM J. Control Optim.* **18**(4) 362–380.
- [15] Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., New York.
- [16] Ross, S. M. 1968. Arbitrary state Markovian decision processes. *Ann. Math. Statist.* **39**(6) 2118–2122.
- [17] Runggaldier, W. J., L. Stettner. 1994. *Approximations of Discrete Time Partially Observable Control Problems, Applied Mathematics Monographs*, Vol. 6. Giardini Editori e Stampatori, Pisa, Italy.
- [18] Yu, H. 2005. A function approximation approach to estimation of policy gradient for POMDP with structured policies. *Proc. 21st Conf. Uncertainty in Artificial Intelligence, Edinburgh, UK*. AUAI Press.
- [19] Yu, H. 2006. Approximate solution methods for partially observable Markov and semi-Markov decision processes. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.