Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Adaptive Policies for Markov Decision Processes

Apostolos N. Burnetas, Michael N. Katehakis,

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 22, No 1, February 1997
Printed in U.S.A.

OPTIMAL ADAPTIVE POLICIES FOR MARKOV DECISION PROCESSES

APOSTOLOS N. BURNETAS AND MICHAEL N. KATEHAKIS

In this paper we consider the problem of adaptive control for Markov Decision Processes. We give the explicit form for a class of adaptive policies that possess optimal increase rate properties for the total expected finite horizon reward, under sufficient assumptions of finite state-action spaces and irreducibility of the transition law. A main feature of the proposed policies is that the choice of actions, at each state and time period, is based on indices that are inflations of the right-hand side of the estimated average reward optimality equations.

1. Introduction. Consider a finite state and action Markovian Decision Process (MDP) with incomplete information. Under an irreducibility assumption for the unknown transition law, it is shown in Theorem 1 that there exists a class C_R of adaptive policies with optimal increase rate properties of the expected finite horizon reward, or equivalently optimal convergence properties of the expected average reward.

The ideas involved in this paper are a natural generalization of the work on the multi-armed bandit (MAB) problem in Lai and Robbins (1985a), Katehakis and Robbins (1995) and Burnetas and Katehakis (1996). The MAB problem, in the form studied therein, can be viewed as a one state MDP, with actions representing the population sampled in a period, and expected rewards that depend on unknown parameters. In Lai and Robbins (1985a), policies with the same optimality properties are obtained when for each population the rewards are generated by a density that depends on a single unknown parameter, as is the case of a *single parameter exponential* family. Simpler index policies were shown to be optimal in Katehakis and Robbins (1995), in the case of normal densities with known variance, and in Burnetas and Katehakis (1996), when rewards are generated by densities that depend on a vector of parameters, such as an arbitrary discrete distribution with known support.

The novel approach of this paper lies in the direct treatment of the Markov dynamics and the resulting form of an optimal policy that utilizes the state-action generated information efficiently. In addition, the parameter space is not restricted to be finite, and the transition law is assumed to be unknown except for its support. In previous related work the MDP model has been viewed as a bandit problem, with one bandit for each deterministic policy that is used over specified time intervals, such as the time between successive visits to a recurrent state; the single unknown parameter for each bandit being the average reward of the deterministic policy it represents. This was the approach taken in Fox and Rolph (1973), who first obtained "consistent" policies (i.e., policies for which the expected average reward converges to the optimal under the true unknown transition law) for MDPs using Robbins (1952), and in Agrawal, Teneketzis and Anantharam (1989a), Shimkin and Shwartz (1996) where

Received December 13, 1994; revised November 22, 1995 and June 27, 1996.
AMS 1991 subject classification. Primary: 90C40; secondary: 62L12.
OR/MS Index 1978 subject classification. Primary: Dynamic programming/optimal control/Markov decision processes; secondary: Statistics/Estimation.
Key words. Markov decision processes, dynamic programming, adaptive control, sequential estimation.

0364-765X/97/2201/0222/$05.00

adaptive policies with optimal increase rate were obtained for MDP and repeated games, respectively, using Lai and Robbins (1985a).

In comparison to the approach taken herein, that in Agrawal et al. (1989a) allows coupling between the uncertainties for different actions and states. However, it requires a combinatorial number of bandits and it places restrictions on the use of the state-action dependent information collected, which are necessary to maintain the independence of the bandits required by the MAB model of Lai and Robbins (1985a). These restrictions are termed "black box approach" in Lai and Yakowitz (1995), where a (non–Markov) Decision Process is also viewed as a MAB. A similar approach is followed by Graves and Lai (1995) for compact parameter spaces, general state spaces, and action sets while still assuming finiteness of the set of deterministic policies.

A direct treatment of the coupling between states and uncertainties along the lines of this paper is one of the interesting open problems in this area. Computation results and conjectures for the MDP problem with side constraints are contained in Burnetas and Katehakis (1995).

Other related work on the "multi-armed bandit" version of this problem includes the following: Chernoff (1967), Gittins (1979), Whittle (1980), Glazebrook (1991). For the problem of MDPs with incomplete information, consistent policies have been shown to exist under various conditions; c.f., Mandl (1974), Federgruen and Schweitzer (1981), Kolonko (1982), Kumar and Varaiya (1986a), Hernández-Lerma (1989), Fernández-Gaucherand, Arapostathis and Marcus (1993) and references therein. For adaptive policies with suboptimal increase rates we refer to Agrawal (1990) and references therein. For additional references on incomplete information MDP problems, c.f., Anantharam, Varaiya and Walrand (1987), Agrawal, Teneketzis and Anantharam (1989b), Anantharam et al. (1987), Borkar and Varaiya (1979), Kumar and Varaiya (1986b), Rieder (1975), Van Hee (1980), Schäl (1987), Milito and Cruz, Jr. (1992), White and Eldeib (1994) and Rieder and Weishaupt (1995).

The main result is stated and proved under the minimal and easy to verify assumptions of Markov dynamics, finite state and action spaces and observable irreducibility of the unknown transition law. These conditions can be relaxed when one makes more general modeling assumptions as in Mallows and Robbins (1964). For example, policies in C_R will have N-horizon regret of the order $O(\log N)$ when the state space is countable and even if the dynamics are not necessarily Markov provided that assumptions for: (i) the existence of limits of the estimates employed and (ii) the claims of Propositions 3, 4, and 5, of §6, hold; see remark in Katehakis and Robbins (1995) for the case of the MAB problem. This type of conditions are discussed in Lai and Yakowitz (1995).

To make the key ideas clear, we chose to restrict the discussion in §§2 to 6 to the case where the rewards are known, and the transition law is unknown; the modifications required for models with unknown reward structure are discussed in §7. In §2 we present the model and background on average reward MDPs. In §3 we construct the class of index policies and state the main theorem of the paper. Although the subsequent proofs are sometimes involved, it is very easy to compute and implement a policy in C_R. In §4 we obtain an expression for the N-horizon regret, and establish a large deviations type result that is used in the sequel. In §5 we show that for all policies with uniformly "fast" convergence properties the regret has an asymptotic lower bound. In §6 we prove that any policy in the index class attains this lower bound. In §7 we consider the following generalizations to the case of unknown rewards: In the first, the one step reward is a function of the current state and action and the next state visited, i.e., the expected one step reward is a function of the transition probabilities. In the second, the one step reward is independent of the next state, but also random, with a distribution that depends on unknown parameters. We also present the solution to the multi-armed bandit problem with discrete distributions. Proofs of intermediate lemmata are given in Appendix A; explanatory examples involving two and three state MDPs are given in Appendix B.

2. The incomplete information model. Consider a discrete time, finite state and action space MDP described by the quadruplet (S, A, R, P), where $S = \{1, 2, \dots, s\}$ is the state space, $A = \bigcup_{x \in S} A(x)$ is the action space, with $A(x)$ being the set of admissible actions in state x, $R = [r(x, a)]_{x \in S, a \in A(x)}$, is the reward structure and $P = [p_{xy}(a)]_{x, y \in S, a \in A(x)}$ is the transition law. The transition probability vectors $p_x(a)$ are unknown and belong to known sets $\Theta(x, a)$.

The statistical framework used in the sequel is as follows.

(a) For any fixed state-action pair (x, a) such that $a \in A(x)$, let the discrete random variable $Y_j(x, a) \in S$ denote the state visited immediately after the jth occurrence of (x, a). From the Markov property, $Y_j(x, a)$, $j = 1, 2, \dots$ are i.i.d. with distribution $p_x(a)$. Let $(\mathscr{Y}^{(n)}_{x,a}, \mathscr{B}^{(n)}_{x,a})$ denote the sample space of a realization $(Y_1(x, a), \dots, Y_n(x, a))$, $1 \leq n \leq \infty$. For any probability vector $p_x(a) \in \Theta(x, a)$, let $\mathbf{P}_{p_x(a)}$ be the probability measure on $\mathscr{B}^{(1)}_{x,a}$ generated by $p_x(a)$ and $\mathbf{P}^{(n)}_{p_x(a)}$ the measure on $\mathscr{B}^{(n)}_{x,a}$ generated by n independent replications of $Y_1(x, a)$. In the sequel $\mathbf{P}^{(n)}_{p_x(a)}$ will often be abbreviated by $\mathbf{P}_{p_x(a)}$. Expectations under $p_x(a)$ will be denoted by $\mathbf{E}_{p_x(a)}$.

(b) Let the random variables X_t, A_t, $t = 0, 1, \dots$ denote respectively the state of the process and the action taken in period t. A *history* ω_k is any feasible sequence of states and actions during the first k time periods, $\omega_k = x_0, a_0, \dots, x_{k-1}, a_{k-1}, x_k$, such that $a_t \in A(x_t)$, $t = 0, \dots, k - 1$. Let $(\Omega^{(k)}, \mathscr{F}^{(k)})$, $1 \leq k \leq \infty$ denote the sample space of histories ω_k, where $\Omega^{(k)}$ is the set of all histories ω_k and $\mathscr{F}^{(k)}$ the σ-field generated by Ω_k. *Events*, defined on $\mathscr{F}^{(k)}$ are denoted by capital letters. The complement of event B is denoted by $\bar{B}$.

A *policy* π is defined as a sequence $\{\pi_k\}$ of probability measures on $A = \bigcup_{x \in S} A(x)$ given ω_k, such that $\pi_k(A(X_k) | \omega_k) = 1$, for all periods $k \geq 0$ and histories ω_k. It represents a randomized law of selecting actions based on the entire observed history and the parameters of the problem. A policy π is *adaptive* if $\pi_k(\cdot | \omega_k)$ does not depend on knowledge of P. A policy π is *deterministic* if there exists a function $f : S \to A$, $k = 0, 1, \dots$, with $f(x) \in A(x)$, being the action taken in state $x = X_k$ for all k and ω_k. In this case π will be denoted by f. There is a one-to-one correspondence between the vectors in the product of the actions sets: $\mathscr{A} = \prod_{x \in S} A(x)$ and the deterministic policies, thus for simplicity $\mathscr{A}$ will also denote the class of the latter. The set of all policies will be denoted by C.

Probability and expectation under transition law P, policy $\pi \in C$ and starting state x_0 will be denoted by $\mathbf{P}^{\pi, P}_{x_0}$, $\mathbf{E}^{\pi . P}_{x_0}$. For notational expedience we may use the symbols $\pi(k)$ and $\pi(\omega_k)$ to denote the actions A_k taken under a policy π.

(c) Given a history ω_k, let $T_k(x, a)$ denote the number of occurrences of the pair (x, a) up to period k. Assume that there are estimators $\hat{p}^{T_k(x . a)}_x(a) = \hat{p}_x(Y_1(x, a), \dots, Y_{T_k(x, a)}(x, a))$, $k \geq 0$, of the transition probability vectors $p_x(a)$, with $\hat{p}^0_x$ arbitrary unless otherwise specified.

Any strongly consistent estimation procedure that satisfies the claim of Lemma 4 and its support satisfies the requirements of assumption (A) below, is sufficient for our analysis. Such an estimation procedure will be given in §2.3.

REMARK 1. Note the distinction between the policy dependent $(\Omega^{(k)}, \mathscr{F}^{(k)}, \mathbf{P}^{\pi, P}_{x_0})$ and policy independent $(\mathscr{Y}^{(n)}_{x,a}, \mathscr{B}^{(n)}_{x,a}, \mathbf{P}^{(n)}_{p_x(a)})$ probability spaces. However, since $\hat{p}^t_x(a)$ is a function of $Y_1(x, a), \dots, Y_t(x, a)$ only, it is easy to see by conditioning, that the following relations hold, for any sequence $F(k, t; x, a) \subseteq \Theta(x, a)$, $k, t \geq 1$,

$$\mathbf{P}_{x_0}^{\pi, P}\big(\hat{p}_x^{T_k(x, a)}(a) \in F(k, T_k(x, a); x, a), T_k(x, a) = t\big)$$

$$\leq \mathbf{P}_{p_x(a)}\big(\hat{p}_x^t(a) \in F(k, t; x, a)\big),$$

$$\mathbf{P}_{x_0}^{\pi, P}\big(\hat{p}_x^{T_k(x, a)}(a) \in F(k, T_k(x, a); x, a)\big)$$

$$\leq \mathbf{P}_{p_x(a)}\big(\hat{p}_x^t(a) \in F(k, t; x, a) \text{ for some } t \leq k\big).$$

2.1. Unobservable quantities related to the process. In this section we state the assumption regarding the parameter space of the unknown transition law and define several unobservable constants and sets, such as expected rewards, optimality equations, sets of critical actions etc., that are used in the subsequent analysis.

(i) *Transition law and parameter space.* Let $P = [p_{xy}(a)]_{x. y \in S, a \in A(x)}$ denote the unknown transition law. The parameter space for a probability vector $p_x(a)$ is:

$$\Theta(x, a) = \Big\{q \in \mathbb{R}^s : \sum_{y \in S} q(y) = 1, q(y) > 0, \forall y \in S^+(x, a) \text{ and } q(y)$$

$$= 0, \forall y \notin S^+(x, a)\Big\},$$

where $S^+(x, a) = S^+(x, a; P) = \{y \in S : p_{xy}(a) > 0\}$.

All assumptions made in this paper are summarized below as

ASSUMPTION (A).

(1) The sets S, $\{A(x)\}_{x \in S}$ are finite and $P \in \tilde{\mathscr{P}} = \prod_{(x. a)} \Theta(x, a)$.

(2) For all $x \in S$, $a \in \mathscr{A}(x)$, the sets $S^+(x, a)$ are independent of P, known and such that the transition matrices $P(f) = [p_{xy}(f(x))]_{x, y \in S}$, are irreducible, for all policies $f \in \mathscr{A}$.

Reference to assumption (A) is made only in the main theorem, for emphasis. It will be omitted in the statements of the propositions and lemmata, for simplicity.

If the parameter sets $\Theta(x, a)$ are restricted, by introducing additional information, then one can find adaptive policies with bounded regret; such cases are discussed in Lai and Robbins (1985b) for the MAB problem. The issue of how much of a restriction is allowable without violating the claim of Theorem 1 is discussed, in the context of the MAB problem, in Burnetas and Katehakis (1996); see necessary conditions and Examples 1 to 3 therein.

(ii) *Expected rewards and regret.* Let $V_N^\pi(x_0, P) = \mathbf{E}_{x_0}^{\pi, P} \sum_{t=0}^{N-1} r(X_t, A_t)$, $V_N(x_0, P) = \sup_{\pi \in C} V_N^\pi(x_0, P)$ denote respectively the expected total reward during the first N transitions under policy π, and the optimal N-horizon reward, as functions of P, i.e., when "the true value P is known to the experimenter." The loss or *regret* due to incomplete information, incurred in the N-horizon when a policy π is used is $R_N^\pi(x_0, P) = V_N(x_0, P) - V_N^\pi(x_0, P)$.

Let $g^\pi(x_0, P) = \liminf_{N \to \infty} V_N^\pi(x_0, P)/N$ denote the long run expected average reward under policy π and $g^*(x_0, P) = \lim_{N \to \infty} V_N(x_0, P)/N$. The existence of the last limit follows from standard dynamic programming results (Ross 1983).

(iii) *Optimality equations—restricted problems.* For a collection $\{D(x) \subseteq A(x)\}_{x \in S}$, let $\mathscr{D} = \prod_{x \in S} D(x) \subseteq \mathscr{A}$, and define the *restricted problem* $(P, \mathscr{D})$ as an MDP with state space S, action sets $\{D(x), x \in S\}$ and with transition law and reward structure given by the restrictions on $\mathscr{D}$ of P and R respectively.

Assumption (A) and standard MDP theory (c.f. Derman 1970, Ross 1983) imply that there exist a constant $g(P, \mathscr{D}) = \sup_{\pi \in C(\mathscr{D})} g^{\pi}(x_0, P)$ (independent of x_0) and a vector $h(P, \mathscr{D}) = [h(x; P, \mathscr{D})]_{x \in S}$ (defined up to an additive constant) satisfying the *average reward optimality equations*

$$(2.1) \qquad g(P, \mathscr{D}) + h(x; P, \mathscr{D}) = \mathscr{L}^*(x; P, \mathscr{D}), \qquad x \in S,$$

where

$$\mathscr{L}^*(x; P, \mathscr{D}) = \max_{a \in D(x)} \{\mathscr{L}(x, a; p_x(a), h(P; \mathscr{D}))\},$$

$$\mathscr{L}(x, a; q, h) = r(x, a) + qh, \; qh = \sum_{y \in S} q(y) h(y)$$

and $C(D)$ denotes the set of all policies for restricted problem $(P, \mathscr{D})$.

The set of deterministic policies of $(P, \mathscr{D})$, will be denoted by $\mathscr{D}$. When $\mathscr{D} = \{f\}$, we will use the abbreviations $g_f(P) = g(P, \{f\})$, $h_f(P) = h(P, \{f\})$.

Let $O(x; P, \mathscr{D}) = \{a \in D(y) : \mathscr{L}(x, a; p_x(a), h(P; \mathscr{D})) = \mathscr{L}^*(x; P, \mathscr{D})\}$ be the set of maximizing actions in (2.1), and $\mathscr{O}(P, \mathscr{D}) = \prod_{x \in S} O(x; P, \mathscr{D})$ be the set of optimal deterministic policies for problem $(P, \mathscr{D})$, i.e., $g_f(P) = g(P; \mathscr{D})$, $\forall f \in \mathscr{O}(P, \mathscr{D})$.

Using this terminology, the solution of the average reward optimality equations for the initial unrestricted problem is given by $g^*(P) = g(P, \mathscr{A}) = \sup_{\pi \in C} g^{\pi}(P)$, $h^*(P) = h(P, \mathscr{A})$, $\mathscr{L}^*(x; P) = \mathscr{L}^*(x; P, \mathscr{A})$ and $\mathscr{O}(P) = \mathscr{O}(P; \mathscr{A})$.

Let $\phi^*(x, a; P) = \mathscr{L}^*(x; P) - \mathscr{L}(x, a; p_x(a), h^*(P))$ denote the test quantity of the average reward optimality equations for the pair (x, a). Note that $\phi^*(x, a; P) > 0$, $\forall x$, $\forall a \notin O(x; P)$ and $O(x, P) = \{a \in A(x) : \phi^*(x, a; P) = 0\}$.

(iv) *Critical state—action pairs.* (a) For (x, a) such that $a \notin O(x; P)$ and for $q \in \Theta(x, a)$, let $Q = [Q_{x'y'}(a')] \in \tilde{\mathscr{P}}$ be a modification of the transition law P defined as follows

$$Q_{x'y'}(a') = \begin{cases} q(y'), & \text{if } x' = x,\ a' = a, \\ p_{x'y'}(a'), & \text{otherwise}. \end{cases}$$

When we need to stress the dependence of Q on x, a, P, q, we will write it as $Q(x, a; P, q)$.

(b) For (x, a) such that $a \notin O(x; P)$, let $\Delta\Theta(x, a; P) = \{q : O(x; Q(x, a; P, q)) = \{a\}\}$, denote the set of parameter values that make action a at state x uniquely optimal and define the set of *critical* state-action pairs $\mathbf{B}(P)$, for any $P \in \tilde{\mathscr{P}}$, as $\mathbf{B}(P) = \{(x, a) : a \notin O(x; P)$ and $\Delta\Theta(x, a; P) \neq \varnothing\}$.

(c) Let $\mathbf{I}(p, q) = \sum_{y \in S^+(x, a)} p(y) \log[p(y)/q(y)]$, denote the Kullback-Leibler information between vectors $p, q \in \Theta(x, a)$.

(d) Let $\mathbf{K}(x, a; P) = \inf\{\mathbf{I}(p_x(a), q) : q \in \Delta\Theta(x, a; P)\}$. For any (x, a), $\mathbf{K}(x, a; P)$ is zero if $a \in O(x; P)$; it is positive if $(x, a) \in \mathbf{B}(P)$, and it is infinite otherwise.

(e) Let $\mathbf{M}(P) = \sum_{(x, a) \in \mathbf{B}(P)} \phi^*(x, a; P)/\mathbf{K}(x, a; P)$.

In view of the foregoing definitions, a state-action pair (x, a) is critical if a is not an optimal action in state x under transition law P, and there exists a vector $q \in \Theta(x, a)$ with the following property. If the transition probability row $p_x(a)$ is replaced with q, while everything else remains unchanged, action a becomes the unique optimal action for state x under this modified transition law $Q(x, a; P)$.

For a critical pair (x, a), $\mathbf{K}(x, a; P)$ denotes the minimum distance, in the sense of Kullback-Leibler information, of the transition probability row $p_x(a)$ from the set $\Delta\Theta(x, a; P)$ of vectors q that would make a the optimal action for x. It is a measure of the importance of critical pair (x, a), because, as it will be shown in Theorem 2, a necessary condition for an adaptive policy to have a uniformly good behavior for all values of the unknown transition law, is that any critical pair (x, a) occurs a number of times asymptotically equal to $\log N/\mathbf{K}(x, a; P)$.

In this spirit, $\mathbf{M}(P)$ represents an aggregate measure of (estimation) importance of all critical pairs for the MDP problem as a function of the transition law P. Its relevance as such a measure is established by Theorem 1. Note that $\mathbf{M}(P) = 0$ only in the degenerate cases that all policies are optimal, i.e., $\phi^*(x, a; P) = 0$, for all state action pairs (x, a), and/or when there are no critical pairs ($\mathbf{B}(P) = \emptyset$), i.e., none of nonoptimal actions can be made optimal by changing only its own transition probability vector.

In Lemma 1 below we develop alternative characterizations for $\Delta\Theta(x, a; P)$, and $\mathbf{B}(P)$; they indicate the relationship of the critical state-action pairs with the average reward optimality equations. From this it follows that for any $(x, a) \in \mathbf{B}(P)$ the computation of $\mathbf{K}(x, a; P)$ (as a function of P) involves the minimization of a convex function subject to two linear constraints. The proof of the lemma is given in Appendix A, and graphical illustrations are given in Appendix B.

LEMMA 1. *For all $P \in \tilde{\mathscr{P}}$ and (x, a) such that $a \notin O(x; P)$, the following are true.*
(i) $\Delta\Theta(x, a; P) = \{q \in \Theta(x, a)\colon \mathscr{L}(x, a; q, h^*(P)) > \mathscr{L}^*(x; P)\}$.
(ii) $\Delta\Theta(x, a; P) \neq \emptyset$, *if and only if* $r(x, a) + \max_{y \in S^+(x, a)} h^*(y; P) > \mathscr{L}^*(x; P)$.
(iii) $\mathbf{B}(P) = \{(x, a)\colon a \notin O(x; P), r(x, a) + \max_{y \in S^+(x, a)} h^*(y; P) > \mathscr{L}^*(x; P)\}$.

2.2. Optimality criteria. Since $V_N(x_0, P)$ does not depend on π, maximization of $V_N^\pi(x_0, P)$ with respect to π is equivalent to minimization of $R_N^\pi(x_0, P) = V_N(x_0, P) - V_N^\pi(x_0, P)$. In general it is not possible to find an adaptive policy which minimizes $R_N^\pi(x_0, P)$ uniformly in P for any fixed N. However, the following definitions of optimality can be used.

A policy π will be called *uniformly fast convergent* (UF) if $R_N^\pi(x_0, P) = o(N^\alpha)$, as $N \to \infty$, $\forall \alpha > 0$, $\forall P \in \tilde{\mathscr{P}}$, $\forall x_0 \in S$.

A UF policy π_0 will be called *uniformly maximum convergence rate* (UM) if $\limsup_{N \to \infty} R_N^{\pi_0}(x_0, P))/R_N^\pi(x_0, P) \leq 1$, $\forall P \in \tilde{\mathscr{P}}$ such that $\mathbf{M}(P) > 0$, for all UF π, $\forall x_0 \in S$.

Let $C_F \supset C_M$ denote the classes of UF and UM policies respectively.

Because $\limsup_{N \to \infty} R_N^\pi(x_0, P)/N = \limsup_{N \to \infty}(V_N(x_0, P) - V_N^\pi(x_0, P))/N = g^*(P) - g(\pi, P) \geq 0$, classes C_F and C_M can be expressed in terms of the rate of convergence of $V_N^*(x_0, P)/N$ to $g^*(P)$ as follows.

If $\pi \in C_F$, then $|V_N^\pi(x_0, P)/N - g^*(P)| = o(N^{\alpha - 1})$, therefore $V_N^\pi(x_0, P)/N$ converges to $g^*(P)$ faster than $N^{\alpha-1}$, $\forall P \in \tilde{\mathscr{P}}$, $\forall \alpha > 0$. Note that a UF policy is also *consistent* in the sense of Robbins (1952) and Fox and Rolph (1973), i.e., $\lim_{N \to \infty} V_N^\pi(x_0, P)/N = g^*(P)$. This follows from the definition of UF, with $\alpha = 1$.

These definitions refer to properties with respect to the whole parameter space $\tilde{\mathscr{P}}$. A UM policy is "best"(i.e., it has maximum rate of convergence) among "good"(i.e., UF) policies for all values of $P \in \tilde{\mathscr{P}}$ for which the unobservable $\mathbf{M}(P)$ is not equal to zero. When $\mathbf{M}(P) = 0$, a UM policy possesses the UF property (i.e., it is still "good"). In Theorem 1 it is shown that UM policies exist by establishing that the class of index policies C_R constructed in the next section are UM. In view of Theorem 1, the rate of convergence for UM policies, is equal to $\mathbf{M}(P) \log N/N$ when $\mathbf{M}(P) \neq 0$, while if $\mathbf{M}(P) = 0$, $R_N^{\pi_0}(x_0, P) = o(\log N)$ as $N \to \infty$ for any UM policy π_0.

2.3. Estimators. Given a history ω_k, define the following statistics.

(i) Let $T_k(x)$, $T_k(x, a)$, $T_k(x, y, a)$ denote the number of visits to state x, the number of occurrences of the state-action pair (x, a) and the number of transitions from x to y under action a, during the first k transitions, i.e., $T_k(x) = \sum_{t=0}^{k-1} Z_t(x)$, $T_k(x, a) = \sum_{t=0}^{k-1} Z_t(x, a)$ and $T_k(x, y, a) = \sum_{t=0}^{k-1} Z_t(x, y, a)$, where $Z_t(x) = \mathbf{1}(X_t = x)$, $Z_t(x, a) = \mathbf{1}(X_t = x,\ A_t = a)$ and $Z_t(x, y, a) = \mathbf{1}(X_t = x,\ A_t = a,\ X_{t+1} = y)$.

(ii) Let $n_t(y; x, a) = \sum_{j=1}^{t} \mathbf{1}(Y_j(x, a) = y)$, $t \geq 1$. Note that $T_k(x, y, a) = n_{T_k(x,a)}(y; x, a)$.

(iii) Let $f_t(y; x, a) = n_t(y; x, a)/t$, for $t \geq 1$ and $f_0(y; x, a) = 1/|S^+(x, a)|$, where $|S|$ denotes the cardinality of any set S.

(iv) For $t \geq 0$, let $\hat{p}_x^t(a) = [\hat{p}_{xy}^t(a)]_{y \in S}$, where $\hat{p}_{xy}^t(a) = 0$ if $y \notin S^+(x, a)$, otherwise $\hat{p}_{xy}^t(a) = (1 - w_t) f_0(y; x, a) + w_t f_t(y; x, a)$, where $w_t = t/(|S^+(x, a)| + t)$.

(v) Let $\hat{P}^k = [\hat{p}_{xy}^{T_k(x,a)}(a)]$ denote the estimate of the transition law P, where we suppress the dependence of $\hat{p}_{xy}^{T_k(x,a)}(a)$ on $T_k(x, y, a)$ for notational simplicity. Note that under this estimation scheme, $\hat{P}^k \in \tilde{\mathscr{P}}$, for all ω_k, $k \geq 0$.

3. Index policies and the main theorem. At any period $k \geq 0$, given a history ω_k, with "current state" $X_k = x$, we have an estimate $\hat{P}^k$ for P. Using the estimate we obtain the solution $\hat{g}^k = g(\hat{P}^k, \mathscr{D}_k)$, $\hat{h}^k = h(\hat{P}^k, \mathscr{D}_k)$ of the restricted problem $(\hat{P}^k, \mathscr{D}_k)$. This problem is the observable MDP with transition law $\hat{P}^k$ and with action space $\mathscr{D}_k$ equal to the product of the "relatively frequently sampled" action sets (where $\log^2 b = (\log b)^2$):

$$D_k(x) = D_k(x; \omega_k) = \{a \in A(x); T_k(x, a) \geq \log^2 T_k(x)\}, \qquad x \in S.$$

The *index*, $\mathbf{U}$, of any action $a \in A(x)$ is defined as $\mathbf{U}(x, a; \omega_0) = \mathscr{L}(x, a; \hat{p}_x^0(a), \hat{h}^0)$, for $k = 0$, and

$$\text{(3.1)}\quad \mathbf{U}(x, a; \omega_k) = \sup_{q \in \Theta(x,a)} \{\mathscr{L}(x, a; q, \hat{h}^k) \colon \mathbf{I}(\hat{p}_x^{T_k(x,a)}(a), q) \leq \log k / T_k(x, a)\}, \quad \text{if } k \geq 1.$$

Let $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k) = \{a \in O(x; \hat{P}^k, \mathscr{D}_k)\colon\ T_k(x, a) < \log^2 T_k(x) + 1\}$, and let $\Gamma_2(x; \hat{P}^k, \mathscr{D}_k) = \{a \in A(x)\colon \mathbf{U}(x, a; \omega_k) = \max_{a' \in A(x)} \mathbf{U}(x, a'; \omega_k)\}$.

Define a class C_R of index policies such that at any point of time k when the state is $x = X_k$ the action taken is determined by the following rule.

(I) Solve the restricted problem $(\hat{P}^k, \mathscr{D}_k)$ and compute the set $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k)$.

(II) Take any action from $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k)$, if this set is equal to $O(x; \hat{P}^k, \mathscr{D}_k)$, otherwise compute the indices and take any action from the set $\Gamma_2(x; \hat{P}^k, \mathscr{D}_k)$.

To avoid repetitions we always assume that when $j = 0$ in a ratio of the form $\log k/j$, then the latter is equal to ∞. In the sequel the discussion is simplified by the use of the generic function

$$\mathbf{u}_{x,a}(p, h, \gamma) = \sup_{q \in \Theta(x,a)} \{\mathscr{L}(x, a; q, h) \colon \mathbf{I}(p, q) \leq \gamma\}$$

defined for all x, a and $p, q \in \Theta(x, a)$, $h \in \mathbb{R}^s$, $0 \leq \gamma \leq \infty$, and such that $\mathbf{U}(x, a; \omega_k) = \mathbf{u}_{x,a}(\hat{p}_x^{T_k(x,a)}(a), \hat{h}^k, \log k/T_k(x, a))$.

We next discuss the main ideas involved in the definition of the index policies. In every period, the estimated transition law can in principle be used to solve a set of average reward optimality equations and estimate an optimal policy. However, it is easy to see that this policy results in a positive probability of converging to a nonoptimal solution, hence its regret is unbounded; c.f. Robbins (1952), Fox and Rolph (1973). The remedy to this situation is to allow taking seemingly inferior actions from time to time. This should be done often enough to alleviate the possibility of estimation errors, but not so often as to affect the rate of convergence to the long run average reward. The main result of the paper is that the optimal tradeoff with respect to the convergence rate is achieved if all actions from critical state-action pairs (x, a) are taken a number of times asymptotically equal to $\log N/\mathbf{K}(x, a; P)$.

Policies in C_R attain this asymptotic behavior using the following mechanism. In period k, the solution of the average reward "estimated optimality equations" of the restricted problem provide estimates $\hat{g}^k, \hat{h}^k$ of the optimal expected average and differential rewards.

Instead of selecting an action that maximizes the right-hand side $\mathscr{L}(x, a; \hat{p}_x^{T_k(x,a)}(a), \hat{h}^k)$ of the "estimated optimality equations" for the current state x, the action with the largest value of the index is taken, unless $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k) = O(x; \hat{P}^k, \mathscr{D}_k)$, when a *forced selection* of an (any) action from $O(x; \hat{P}^k, \mathscr{D}_k)$ is made.

The index $\mathbf{U}(x, a; \omega_k)$ for action a is the inflation of the right-hand side of the "estimated" optimality equations, calculated as the maximum value of $\mathscr{L}(x, a; q, \hat{h}^k)$ in a confidence region specified as the subset of $\Theta(x, a)$ consisting of all probability vectors q that are close (in the Kullback-Leibler information sense) to the estimate $\hat{p}_x^{T_k(x,a)}(a)$, i.e., $\mathbf{I}(\hat{p}_x^{T_k(x,a)}(a), q) \leq \log k/T_k(x, a)$.

To see the role played by the forced selections, consider a time period k, when the state is x and current estimates implying that $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k) = O(x; \hat{P}^k, \mathscr{D}_k)$. Then, we have the situation that none of the optimal actions at state x of the restricted problem $(\hat{P}^k, \mathscr{D}_k)$ will be contained in the restricted set the next time state x will be visited, *unless* one of these actions is taken. In §6 it is shown that as a consequence of this forced selection scheme, the optimal solutions of the restricted problems have the following asymptotic monotonicity property: $\mathbf{P}_{x_0}^{\pi_0, P}[g(P, \mathscr{D}_{k+1}) \geq g(P, \mathscr{D}_k)] = 1 - o(1/k)$, as $k \to \infty$, $\forall \pi_0 \in C_R$.

It should be noted however that forced selections are introduced to handle the worst case contingency for a policy in C_R considered in Proposition 5. Computational experience in Burnetas and Katehakis (1995) indicates that this is a rare eventuality.

The restricted action sets serve the following purpose. In any period when solving the optimality equations, we consider only those actions for which the estimated transition probabilities are relatively accurate. Thus, asymptotic results can be established regarding the convergence of $\hat{g}^k, \hat{h}^k$ to $g^*(P)$ and $h^*(P)$, respectively. In addition, in §6 it is shown that $\mathbf{P}_{x_0}^{\pi_0, P}[\mathscr{D}_k \subseteq \mathscr{O}(P)] = 1 - o(1/k)$, $\forall \pi_0 \in C_R$. This implies that, for large k, the sets $D_k(x)$ will contain only optimal actions, with high probability. Therefore, there is a reduction in the computational effort in solving the average reward optimality equations for the restricted problem $(\hat{P}^k, \mathscr{D}_k)$, as k increases.

We discuss several properties of the index below.

By construction, $\mathbf{U}(x, a; \omega_k)$ is increasing in k and decreasing in $T_k(x, a)$. This implies that relatively undersampled actions are given a higher chance to be taken in every period.

For any $\gamma > 0$ and any vector $h \neq 0$, the supremum in $\mathbf{u}_{x,a}(p, h, \gamma)$ is attained, i.e., $\exists q^* \in \Theta(x, a)$, satisfying $\mathbf{I}(p, q) = \gamma$, such that $\mathbf{u}_{x,a}(p, h, \gamma) = \mathscr{L}(x, a; q^*, h)$. This is so because the set $\{q \in \Theta(x, a): \mathbf{I}(p, q) \leq \gamma\}$ is closed and convex, therefore, the supremum of the linear function $\mathscr{L}(x, a; q, h)$ is attained at a point on the boundary of this set.

In addition, $\mathbf{u}_{x,a}(p, h, \infty) = \sup_{q \in \Theta(x,a)} \mathscr{L}(x, a; q, h) = r(x, a) + \max_{y \in S^+(x,a)} h(y)$ (note that the supremum need not be attained in this case). Thus, if for some ω_k, $T_k(x, a) = 0$, then, $\mathbf{U}(x, a; \omega_k) = r(x, a) + \max_{y \in S^+(x,a)} \hat{h}^k(y)$.

In Appendix B we give two examples, for $|S^+(x, a)| = 2$ and 3 where the computation of $\mathbf{U}(x, a; \omega_k)$ (and the unobservable $\mathbf{K}(x, a; P)$) is presented graphically.

Finally, we discuss the two choices we made in the specification of C_R; both made for simplification of the proofs.

First, in the definition of $\mathbf{U}(x, a; \omega_k)$, $\log k/T_k(x, a)$ can be replaced by a function of the form $(\log k + f(\log k))/T_k(x, a)$, where $f(t)$ is any function of t with $f(t) = o(t)$ as $t \to \infty$, without affecting the optimality results. It can be shown that $\log T_k(x)/T_k(x, a)$ can also be used. This change introduces complications in the proofs and is omitted. The index $\mathbf{U}(x, a; \omega_k)$ is uniquely defined up to this equivalence.

Second, in the definition of the "relatively frequently sampled" action sets: $D_k(x)$ we chose to specify them by the relation $T_k(x, a) \geq \log^2 T_k(x)$; a careful reading of the proof of Proposition 5 (in §6) shows that any function $f(T_k(x))$ satisfying $\log t = o(f(t))$, and $f(t) = o(t)$ can be used instead of $\log^2 T_k(x)$ without affecting the results.

We next state the main theorem of this paper.

THEOREM 1. *Under assumption* (A) *the following claims are true.*

(i) $\liminf_{N \to \infty} R_N^{\pi}(x_0, P)/\log N \geq \mathbf{M}(P)$, $\forall \pi \in C_F$, $\forall P \in \tilde{\mathscr{P}}$.

(ii) $\limsup_{N \to \infty} R_N^{\pi_0}(x_0, P)/\log N \leq \mathbf{M}(P)$, $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$.

(iii) $R_N^{\pi_0}(x_0, P) = \mathbf{M}(P) \log N + o(\log N)$, *as* $N \to \infty$, $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$.

(iv) $C_R \subseteq C_M$.

PROOF. Claims (i) and (ii) are proved in Theorems 2 and 3, in §§5 and 6, respectively.

From (ii) $\forall \pi_0 \in C_R$, $R_N^{\pi_0}(x_0, P) = O(\log N) = o(N^{\alpha})$, $\forall \alpha > 0$, therefore, $C_R \subseteq C_F$, and using also claim (i), $\lim_{N \to \infty} R_N^{\pi_0}(x_0, P)/\log N = \mathbf{M}(P)$, thus (iii) follows.

To show the last claim we only need to divide $R_N^{\pi_0}(x_0, P)$ and $R_N^{\pi}(x_0, P)$ by $\mathbf{M}(P) \log N$, when $M(P) > 0$. □

Claim (i) of the theorem states that the regret of any UF policy is (asymptotically) larger than $\mathbf{M}(P) \log N$, because critical actions must be taken at least $\log N/\mathbf{K}(x, a; P)$ times. Part (ii) states that the regret of an index policy is (asymptotically) smaller than $\mathbf{M}(P) \log N$, which means that policies in C_R achieve the lower bound of the regret in the class of UF policies, thus they are UM.

In view of Theorem 1, it is instructive to compare the asymptotic expressions below. Equation (3.2) is an expression for the maximum expected finite horizon reward under complete information about P (c.f. Hernández-Lerma 1989), while equation (3.3) is the corresponding expression for expected finite horizon reward of a UM policy π_0, under incomplete information about P following from Theorem 1 (using also equation (3.2)).

$$V_N(x_0, P) = Ng^*(P) + O(1) \tag{3.2}$$

$$V_N^{\pi_0}(x_0, P) = Ng^*(P) - \mathbf{M}(P) \log N + o(\log N). \tag{3.3}$$

Regarding the asymptotic behavior of $V_N^{\pi_0}(x_0, P)$ with respect to the unobservable constant $\mathbf{M}(P)$, it follows from Theorem 1 that, for $P \in \tilde{\mathscr{P}}$ such that $\mathbf{M}(P) = 0$, $V_N^{\pi_0}(x_0, P) = Ng^*(P) + o(\log N)$, therefore $V_N^{\pi_0}(x_0, P)/N$ converges to $g^*(P)$ faster than $\log N/N$. However this does not necessarily represent the fastest rate of convergence. A modification of C_R, which guarantees a maximum convergence rate for this case is an open problem.

4. The regret function and state action frequencies.

In this section we consider an arbitrary policy $\pi \in C$ and establish two results. First, we develop an expression for the N-horizon regret $R_N^{\pi}(x_0, P)$ in terms of the expected state-action frequencies $T_N(x, a)$. Second, we develop asymptotic bounds for the distributions of the state frequencies under any $\pi \in C$.

In Proposition 1 below it is shown that the contribution to the regret of any nonoptimal action under any policy π is asymptotically proportional to the expected number of times this action is taken under π, with the proportionality factor equal to the test quantity $\phi^*(x, a; P)$.

PROPOSITION 1. *For any policy* $\pi \in C$,

$$R_N^{\pi}(x_0, P) = \sum_{x \in S} \sum_{a \notin O(x; P)} \mathbf{E}_{x_0}^{\pi, P}[T_N(x, a)] \phi^*(x, a; P) + O(1), \quad \textit{as } N \to \infty.$$

PROOF. It is well known (c.f. Veinott 1974, Hernández-Lerma 1989) that for all $P \in \tilde{\mathscr{P}}$ $V_N(x_0, P) = Ng^*(P) + h^*(x_0, P) + O(1)$ (as $N \to \infty$). Substituting this expression for $V_N(x_0, P)$ into the regret definition we obtain $R_N^{\pi}(x_0, P) = D_N^{\pi}(x_0, P) + O(1)$, where $D_N^{\pi}(x_0, P) = Ng^*(P) + h^*(x_0; P) - V_N^{\pi}(x_0, P)$.

Express $V_N^{\pi}(x_0, P)$ recursively as $V_N^{\pi}(x_0, P) = \mathbf{E}_{x_0}^{\pi, P}[\mathbf{E}_{A_0}[\mathbf{E}_{X_1}[r(x_0, A_0) + V_{N-1}^{\pi}(X_1, P)]]]$. This follows by conditioning on action A_0 and the state X_1 after the first transition.

Substituting the above expression into the definition of $D_N^{\pi}(x_0, P)$ yields

$$\begin{aligned} D_N^{\pi}(x_0, P) &= g^*(P) + h^*(x_0; P) - \mathbf{E}_{x_0}^{\pi, P}\big[\mathbf{E}_{A_0}[r(x_0, A_0)]\big] \\ &\quad + \mathbf{E}_{x_0}^{\pi, P}\big[\mathbf{E}_{A_0}\big[\mathbf{E}_{X_1}[(N-1) g^*(P) - V_{N-1}^{\pi}(X_1, P)]\big]\big]. \end{aligned}$$

After adding and subtracting the same quantity, $\mathbf{E}_{x_0}^{\pi, P}[\mathbf{E}_{A_0}[\mathbf{E}_{X_1}[h^*(X_1; P)]]]$ from the right-hand side, we obtain $D_N^{\pi}(x_0, P) = \mathbf{E}_{x_0}^{\pi, P}[\phi^*(x_0, A_0; P)] + \mathbf{E}_{x_0}^{\pi, P}[D_{N-1}^{\pi}(X_1, P)]$, and iterating down to $N = 1$,

$$D_N^{\pi}(x_0, P) = \sum_{t=0}^{N-1} \mathbf{E}_{x_0}^{\pi, P}[\phi^*(X_t, A_t; P)] = \sum_{x \in S} \sum_{a \in A(x)} \mathbf{E}_{x_0}^{\pi, P}[T_N(x, a)] \phi^*(x, a; P),$$

where the last equality follows from $\phi^*(X_t, A_t; P) = \sum_{x \in S, a \in A(x)} Z_t(x, a) \phi^*(x, a; P)$ and the definition of $T_N(x, a)$.

Since $\phi^*(x, a, P) = 0$, $\forall a \in O(x; P)$, the proposition follows. □

In the following proposition it is shown that, as a consequence of the irreducibility assumption, all states will be visited often, regardless of the policy used.

PROPOSITION 2. (i) *There exist* $A > 0$ *and* $\gamma > 0$ *such that* $\mathbf{P}_{x_0}^{\pi, P}[T_k(x) \le \rho k] \le Ae^{(\rho - \gamma)k}$, *for all* $x \in S$, $k \ge s = |S|$, $\rho > 0$, $\pi \in C$.

(ii) *For all* $\rho < \gamma$, $x \in S$, $\pi \in C$, $\mathbf{P}_{x_0}^{\pi, P}[T_k(x) \le \rho k] = o(1/k)$ *as* $k \to \infty$.

PROOF. We only prove (i), because then (ii) is immediate. The proof for the case that a fixed stationary policy is used is given in Ellis (1985). The proof below is for the case of general policies.

Fix a state $x \in S$. For any policy π and any $\rho > 0$, it follows from the Markov inequality that $\mathbf{P}_{x_0}^{x, P}[T_k(x) \le \rho k] = \mathbf{P}_{x_0}^{\pi, P}[e^{-T_k(x)} > e^{-\rho k}] \le e^{\rho k} \mathbf{E}_{x_0}^{\pi, P}[e^{-T_k(x)}]$.

Consider the expression $\mathbf{E}_{x_0}^{\pi, P}[e^{-T_k(x)}]$. Let $k \ge s =$ the cardinality of the state space. By conditioning on the history ω_{k-s} of states and actions until time $k - s$, we obtain

$$\mathbf{E}_{x_0}^{\pi,P}[e^{-T_k(x)}] = \mathbf{E}_{x_0}^{\pi,P}\left[\mathbf{E}_{\omega_{k-s}}^{\pi,P}[e^{-T_k(x)}]\right] = \mathbf{E}_{x_0}^{\pi,P}\left[e^{-T_{k-s}(x)}\mathbf{E}_{\omega_{k-s}}^{\pi,P}[e^{-(T_k(x)-T_{k-s}(x))}]\right].$$

In order to prove the proposition, it is sufficient to show that there exists an $\epsilon_1 < 1$ such that, for all policies π and histories $\omega_{k-s} = x_0, a_0, \dots, x_{k-s}$,

$$\mathbf{E}_{\omega_{k-s}}^{\pi.P}[e^{-(T_k(x)-T_{k-s}(x))}] \leq \epsilon_1. \tag{4.1}$$

Indeed, assume (4.1) is true. Then, $\mathbf{E}_{x_0}^{\pi,P}[e^{-T_k(x)}] \leq \epsilon_1 \mathbf{E}_{x_0}^{\pi,P}[e^{-T_{k-s}(x)}]$, and repeating the same argument, we obtain $\mathbf{E}_{x_0}^{\pi,P}[e^{-T_k(x)}] \leq \mathbf{E}_{x_0}^{\pi,P}[e^{-T_{k_0}(x)}]\epsilon_1^{\lfloor k/s \rfloor}$, where $k_0 = \mathrm{mod}(k, s)$. Therefore, the claim follows with $\gamma = -\log \epsilon_1/s$.

We now show (4.1). For any $k \geq s$, the possible values for $T_k(x) - T_{k-s}(x)$ are $j = 0, \dots, s$. Let $Q^{\pi,P}(j; \omega_{k-s}) = \mathbf{P}_{\omega_{k-s}}^{\pi,P}[T_k(x) - T_{k-s}(x) = j]$ and $W^{\pi,P}(\omega_{k-s}) = \mathbf{E}_{\omega_{k-s}}^{\pi,P}[T_k(x) - T_{k-s}(x)]$.

We first show that for any $P \in \tilde{\mathscr{P}}$ there exists $v(P) > 0$ such that $W^{\pi.P} \geq v(P)$, $\forall \pi$, $\forall \omega_{k-s}$.

The quantity $W^{\pi,P}(\omega_{k-s})$ denotes the expected number of visits to state x during the interval from $k - s$ to k, given the initial history and a policy π which takes actions based on the history alone and not on the (unknown) transition probabilities. Therefore, $W^{\pi,P}(\omega_{k-s}) \geq v_{k-s,k}(\omega_{k-s}; P)$, where $v_{k-s.s}(\omega_{k-s}; P)$ is the minimum expected number of visits to state x between $k - s$ to k, over all policies based on the history ω_{k-s} and full knowledge of P.

Consider an MDP with known transition law P, and one step cost equal to 1 if the current state is x and zero otherwise. Then $v_{k-s.k}(\omega_{k-s}; P)$ as defined above is equal to the minimum total cost of this process from time $k - s$ to k, given a history ω_{k-s}.

Since, $v_{k-s,k}(\omega_{k-s}; P)$ is the solution of a finite horizon dynamic programming problem with stationary transition probabilities and cost structure with initial state x_{k-s}, it follows that $v_{k-s,k}(\omega_{k-s}; P) = v_{0,s}(x_{k-s}; P)$. In addition, there exists a finite sequence $f = \{f_0, \dots, f_{s-1}\}$ of deterministic policies under which $v_{0,s}(x_{k-s}; P)$ is attained, i.e., $v_{0,s}(x_{k-s}; P) = v_{0,s}^f(x_{k-s}; P)$.

From Theorem 1 in Veinott (1974), by making state x absorbing, we have that $v_{0,s}^f(x_{k-s}; P) > 0$ for every sequence $f = \{f_0, \dots, f_{s-1}\}$. Because there is a finite number of such sequences, it follows that $v_{0,s}(x_{k-s}; P) > 0$, for all x_{k-s}.

Therefore, $W^{\pi,P}(\omega_{k-s}) \geq v(P) = \min_{x \in S} v_{0,s}(x; P)$, and since $W^{\pi,P}(\omega_{k-s}) = \sum_{j=0}^{s} jQ^{\pi,P}(j; \omega_{k-s}) \leq s(1 - Q^{\pi,P}(0; \omega_{k-s}))$, it follows that $Q^{\pi,P}(0; \omega_{k-s}) \leq 1 - v(P)/s$.

Thus, $\mathbf{E}^{\pi,P}[e^{-(T_k(x)-T_{k-s}(x))}|\omega_{k-s}] = \sum_{j=0}^{s} e^{-j}Q^{\pi,P}(j; \omega_{k-s}) \leq Q^{\pi,P}(0; \omega_{k-s}) + e^{-1}(1 - Q^{\pi,P}(0; \omega_{k-s})) \leq e^{-1} + (1 - e^{-1})(1 - v(P)/s) < 1$, and (4.1) follows with $\epsilon_1 = e^{-1} + (1 - e^{-1})(1 - v(P)/s)$. □

5. The asymptotic lower bound for the regret. In this section claim (i) of Theorem 1 is proved. This is accomplished by showing first, in Theorem 2(i), that under any UF policy π, the expected number of times that action a from a critical state-action pair $(x, a) \in \mathbf{B}(P)$ is selected, is asymptotically larger than $\log N/\mathbf{K}(x, a; P)$.

The main idea of this proof is the following.

Consider a pair (x, a) and two transition laws $P, Q \in \tilde{\mathscr{P}}$, such that $(x, a) \in \mathbf{B}(P)$, and $Q = Q(x, a; P, q)$. Because the UF policy π must satisfy the fast convergence property under both P and Q, it follows that, under Q, action a which is uniquely optimal for state x, must be selected "many" times, in the sense that for all $a' \neq a$, $\mathbf{E}_{x_0}^{\pi,Q}T_N(x, a') = o(N^\alpha)$, $\forall a > 0$.

However, probability and expectation under P and Q are related through the likelihood ratio $L(P, Q; \omega_N)$ defined for all histories of states and actions. Using a change of measure from P to Q and the properties of the likelihood ratio described in Lemma 2, it is shown in Theorem 2 that the lower bound for $T_N(x, a)$, under Q, translates exactly into $\log N/\mathbf{K}(x, a; P)$, under P.

For $P, Q \in \mathscr{P}$, let

$$L(P,Q;\omega_N) = \mathbf{P}_{x_0}^{\pi,P}[X_0, A_0, X_1, \dots, X_N]/\mathbf{P}_{x_0}^{\pi,Q}[X_0, A_0, X_1, \dots, X_N]$$

denote the likelihood ratio of P and Q given history ω_N.

For $k \ge 0$ and any $p, q \in \Theta(x, a)$ let $\Lambda_k(p, q) = \prod_{j=1}^{k} p_{Y_j(x,a)}/q_{Y_j(x,a)}$ denote the part of L corresponding to the history of transitions out of state x under action a (an empty product is equal to 1).

Lemma 2 states the precise relationship between L and Λ and a convergence property of Λ based on the law of large numbers. The proof is included in Appendix A. We use the notation $\lfloor b \rfloor$ for the integer part of a constant b.

LEMMA 2. (i) *If $Q = Q(x, a; P, q)$, then $L(P, Q; \omega_N) = \Lambda_{T_N(x,a)}(p_x(a), q)$.*

(ii) *Let b_N be an increasing sequence of positive constants such that $b_N \to \infty$, as $N \to \infty$. Then,*

$$\mathbf{P}_{p_x(a)}\Big(\max_{k \le \lfloor b_N \rfloor} \log \Lambda_k / b_N > (1+\delta)\mathbf{I}(p,q)\Big) = o(1) \quad \text{as } N \to \infty, \forall \delta > 0. \quad \square$$

THEOREM 2. (i) $\liminf_{N\to\infty} \mathbf{E}_{x_0}^{\pi,P} T_N(x, a)/\log N \ge 1/\mathbf{K}(x, a; P)$, $\forall (x, a) \in \mathbf{B}(P)$, $\forall \pi \in C_F$.

(ii) $\liminf_{N\to\infty} R_N^{\pi}(x_0, P)/\log N \ge \mathbf{M}(P)$, $\forall \pi \in C_F$, $\forall P \in \tilde{\mathscr{P}}$.

PROOF. We use arguments from the proof of Theorem 2 in Lai and Robbins (1985a), together with Proposition 2 and Lemma 2.

Since $\mathbf{K}(x, a; P) > 0$, $\forall (x, a) \in \mathbf{B}(P)$, the Markov inequality implies that $\forall \epsilon > 0$, $\mathbf{E}_{x_0}^{\pi,P} T_N(x, a)/\log N \ge (1 - \epsilon)/\mathbf{K}(x, a; P)\mathbf{P}_{x_0}^{\pi,P}[T_N(x, a)/\log N \ge (1 - \epsilon)/\mathbf{K}(x, a; P)]$, $\forall N > 1$.

Thus, to show the theorem, it suffices to prove that $\lim_{N\to\infty} \mathbf{P}_{x_0}^{\pi,P}[T_N(x, a)/\log N \ge (1 - \epsilon)/\mathbf{K}(x, a; P)] = 1$, or, equivalently,

$$\lim_{N\to\infty} \mathbf{P}_{x_0}^{\pi,P}\left[T_N(x,a) < \frac{(1-\epsilon)\log N}{\mathbf{K}(x,a;P)}\right] = 0, \quad \forall \epsilon > 0.$$

Let $\rho \in (0, \gamma)$, where γ is a constant that satisfies Proposition 2. Then,

$$\mathbf{P}_{x_0}^{\pi,P}\left[T_N(x,a) < \frac{(1-\epsilon)\log N}{\mathbf{K}(x,a;P)}\right]$$

$$\le \mathbf{P}_{x_0}^{\pi,P}[T_N(x) < \rho N] + \mathbf{P}_{x_0}^{\pi,P}\left[T_N(x) \ge \rho N, T_N(x,a) < \frac{(1-\epsilon)\log N}{\mathbf{K}(x,a;P)}\right].$$

From Proposition 2, $\mathbf{P}_{x_0}^{\pi,P}[T_N(x) < \rho N] = o(1)$. Therefore, to prove the theorem it suffices to show that $\mathbf{P}_{x_0}^{\pi,P}[T_N(x) \ge \rho N$, $T_N(x, a) < (1 - \epsilon)\log N/\mathbf{K}(x, a; P)] = o(1)$.

Let $\pi \in C_F$, $(x, a) \in \mathbf{B}(P)$ and $\delta = \epsilon/(2 - \epsilon) > 0$. By definition of $\mathbf{K}(x, a; P)$, there exists $q \in \Delta\Theta(x, a; P)$ such that $\mathbf{K}(x, a; P) \le \mathbf{I}(p_x(a), q) < (1 + \delta)\mathbf{K}(x, a; P)$. After simple algebra, $(1 - \epsilon)/\mathbf{K}(x, a; P) < (1 - \delta)/\mathbf{I}(p_x(a), q)$, thus it suffices to show that $\mathbf{P}_{x_0}^{\pi,P}\mathbf{A}_N^{\delta} = o(1)$, $\forall \delta > 0$, where $\mathbf{A}_N^{\delta} = \{\omega_N: T_N(x) \ge \rho N, T_N(x, a)/\log N < (1 - \delta)/\mathbf{I}(p_x(a), q)\}$.

Let $Q = Q(x, a; P, q)$, for q as defined above. Since $\pi \in C_F$, and $O(x; Q) = \{a\}$, it follows that

$$\mathbf{E}_{x_0}^{\pi, Q}[T_N(x) - T_N(x, a)] = \mathbf{E}_{x_0}^{\pi, Q}\left[\sum_{a' \neq a} T_N(x, a')\right] = o(N^{\alpha}), \qquad \forall a > 0.$$

Therefore, from the Markov inequality,

$$\mathbf{P}_{x_0}^{\pi, Q}\mathbf{A}_N^{\delta} \leq \mathbf{P}_{x_0}^{\pi, Q}[T_N(x) - T_N(x, a) \geq \rho N - \beta \log N]$$

$$\leq \frac{\mathbf{E}_{x_0}^{\pi, Q}[T_N(x) - T_N(x, a)]}{\rho N - \beta \log N} = \frac{o(N^{\delta/2})}{\rho N - \beta \log N} = o(N^{\delta/2 - 1}).$$

Let $\mathbf{C}_N^{\delta} = \{\omega_N\colon \log \Lambda_{T_N(x, a)}(p_x(a), q) \leq (1 - \delta/2)\log N\}$. Then, $\mathbf{P}_{x_0}^{\pi, P}\mathbf{A}_N^{\delta} = \mathbf{P}_{x_0}^{\pi, P}(\mathbf{A}_N^{\delta}\mathbf{C}_N^{\delta}) + \mathbf{P}_{x_0}^{\pi, P}(\mathbf{A}_N^{\delta}\overline{\mathbf{C}}_N^{\delta})$. Thus,

$$\mathbf{P}_{x_0}^{\pi, P}(\mathbf{A}_N^{\delta}\mathbf{C}_N^{\delta}) \leq e^{(1 - \delta/2)\log N}\mathbf{P}_{x_0}^{\pi, Q}(\mathbf{A}_N^{\delta}\mathbf{C}_N^{\delta})$$

$$\leq N^{1 - \delta/2}\mathbf{P}_{x_0}^{\pi, Q}\mathbf{A}_N^{\delta} = N^{1 - \delta/2}o(N^{\delta/2 - 1}) = o(1),$$

where the 1st inequality above is due to a change of measure transformation between P and Q using Lemma 2(i) and the property that on $\mathbf{C}_N^{\delta}$, $\Lambda_{T_N(x, a)}(p_x(a), q) \leq e^{(1 - \delta/2)\log N}$.

To show that $\mathbf{P}_{x_0}^{\pi, P}(\mathbf{A}_N^{\delta}\overline{\mathbf{C}}_N^{\delta}) = o(1)$, note that, from Remark 1:

$$\mathbf{P}_{x_0}^{\pi, P}(\mathbf{A}_N^{\delta}\overline{\mathbf{C}}_N^{\delta}) \leq \mathbf{P}_{p_x(a)}\left(\max_{k \leq \lfloor b_N \rfloor}\{\log \Lambda_k\} > (1 - \delta/2)\log N\right)$$

$$= \mathbf{P}_{p_x(a)}\left(\max_{k \leq \lfloor b_N \rfloor}\{\log \Lambda_k\}/b_N > \mathbf{I}(p_x(a), q)(1 + \delta/(2(1 - \delta)))\right),$$

where $b_N \equiv (1 - \delta)\log N/\mathbf{I}(p_x(a), q)$. Thus, the property follows from Lemma 2(ii). Therefore, $\mathbf{P}_{x_0}^{\pi, P}\mathbf{A}_N^{\delta} = o(1)$, as $N \to \infty$, and the proof of part (i) is complete. Claim (ii) follows from (i) and Proposition 1. □

6. Optimality of index policies. In this section claim (ii) of Theorem 1 is proved. The result is stated as Theorem 3 at the end of the section.

We will use the notation $\|v\|$ to denote the L_{∞}-norm of vector v. For any pair (x, a), let Z_k be the abbreviation for $Z_k(x, a) = \mathbf{1}\,(X_k = x,\ \pi_0(\omega_k) = a)$ and for any $\epsilon > 0$ define the random variables:

$$T_N^{(1)}(x, a; \epsilon) = \sum_{k=1}^{N-1} \mathbf{1}\left(Z_k = 1, \|\hat{h}^k - h^*\| \leq \epsilon, \mathbf{U}(x, a; \omega_k) \geq \mathscr{L}^*(x; P) - 2\epsilon\right),$$

$$T_N^{(2)}(x, a; \epsilon) = \sum_{k=1}^{N-1} \mathbf{1}\Big(Z_k - 1, \|\hat{h}^k - h^*\| \leq \epsilon, \mathscr{O}(\hat{P}^k, \mathscr{D}_k) \subseteq \mathscr{O}(P),$$

$$\mathbf{U}(x, a; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon\Big),$$

$$T_N^{(3)} = \sum_{k=1}^{N-1} \mathbf{1}\left(\|\hat{h}^k - h^*\| > \epsilon \text{ or } O(\hat{P}^k, \mathscr{D}_k) \nsubseteq \mathscr{O}(P)\right).$$

Note that $T_N^{(i)}$ represents the number of times during transitions 1 to $N - 1$ that

> An action a is taken in state x, while at the same time the estimate $\hat{h}^k$ is within ϵ from the optimal solution under complete information h^*, and the index of action a is within 2ϵ from the maximum right-hand side $\mathscr{L}^*(x, P)$ of the (unobservable) optimality equations for state x (for $i = 1$),
>
> Action a is taken in state x with its index being less than $\mathscr{L}^*(x; P) - 2\epsilon$, while the estimate $\hat{h}^k$ is within ϵ from the true (under P) optimal vector h^*, and all the optimal policies for the restricted problem $(\hat{P}^k, \mathscr{D}_k)$ are optimal under P (for $i = 2$),
>
> Either the estimate $\hat{h}^k$ is not within ϵ from the true optimal vector $h^* = h^*(P)$, or there are optimal policies for the restricted problem $(\hat{P}^k, \mathscr{D}_k)$ that are not optimal under P (for $i = 3$).

Asymptotic upper bounds for $\mathbf{E}_{x_0}^{\pi_0, P} T_N^{(i)}$, $i = 1, 2, 3$, are developed in Propositions 3, 4, 5, respectively. Then, the result of Theorem 3 follows, since (including the transition at time 0) $T_N(x, a) \leq 1 + \Sigma_{i=1}^3 T_N^{(i)}$ holds sample pathwise.

The main idea in the proof of Proposition 3 below, is that if a is a nonoptimal action (under P) in state x, then events specifying $T_N^{(1)}(x, a; \epsilon)$ can occur only if either $T_k(x, a)$ is sufficiently small in every intermediate period k (since $\mathbf{U}(x, a; \omega_k)$ is decreasing in $T_k(x, a)$), or if the estimate $\hat{p}_x^{T_k(x, a)}(a)$ of the probability row $p_x(a)$ is significantly different from the true value. The frequency of the first contingency is bounded using the simple counting argument in Lemma 3, and of the second by using a large deviations property of the transition probability estimates established in Lemma 4.

LEMMA 3. *Let Z_t be any sequence of* 0-1 *constants* (*or random variables*) *and let* $t_k = \Sigma_{t=0}^{k-1} \mathbf{1}\, \{Z_t = 1\}$. *Then,* $\forall c > 0$, $\Sigma_{k=1}^{N-1} \mathbf{1}\, \{Z_k = 1,\ t_k \leq c\} \leq c + 1$ (*pointwise if we have random variables*).

PROOF. Note that $\Sigma_{k=1}^{N-1} \mathbf{1}\, \{Z_k = 1,\ t_k = i\} \leq 1$, $\forall i = 0, \ldots, \lfloor c \rfloor$. Therefore, $\Sigma_{k=1}^{N-1} \mathbf{1}$ $\{Z_k = 1,\ t_k \leq c\} = \Sigma_{k=1}^{N-1} \Sigma_{i=0}^{\lfloor c \rfloor} \mathbf{1}\, \{Z_k = 1,\ t_k = i\} = \Sigma_{i=0}^{\lfloor c \rfloor} \Sigma_{k=1}^{N-1} \mathbf{1}\, \{Z_k = 1,\ t_k = i\} \leq \lfloor c \rfloor + 1 \leq c + 1$. □

In the proof of Lemma 4 and Proposition 3 we use the following generalizations of $\Delta\Theta(x, a; P)$ and $K(x, a; P)$ defined in §2.1.

For any $\epsilon \geq 0$ let $\Delta\Theta(x, a; P, \epsilon) = \{q \in \Theta(x, a)\colon \mathscr{L}(x, a; q, h^*(P)) > \mathscr{L}^*(x; P) - \epsilon\}$ and $\mathbf{J}(p_x(a); P, \epsilon) = \inf\{\mathbf{I}(p_x(a), q)\colon q \in \Delta\Theta(x, a; P, \epsilon)\}$.

LEMMA 4. (i) *For any pair* (x, a) *and* $\beta > 0$, *there exist constants* $C = C(x, a)$, $c = c(x, a)$, $t_0 > 0$, *such that* $\mathbf{P}_{p_x(a)}[|\hat{p}_{xy}^t(a) - p_{xy}(a)| > \beta] \leq Ce^{-ct}$, *for* $t \geq t_0$.

(ii) *For any* $\delta, \epsilon > 0$, $\mathbf{P}_{p_x(a)}[\mathbf{J}(\hat{p}_x^t(a); P, \epsilon) < \mathbf{J}(p_x(a); P, \epsilon) - \delta] = o(1/t)$, *as* $t \to \infty$.

PROOF. (i) Since $1 - w_t \to 0$, it follows from the definition of $\hat{p}_{xy}^t(a)$ that, for any $\beta > 0$, there exists $t_0 = t_0(\beta) < \infty$ such that $\mathbf{P}_{p_x(a)}[|\hat{p}_{xy}^t(a) - p_{xy}(a)| > \beta] \leq \mathbf{P}_{p_x(a)}[|f_t(y; x, a) - p_{xy}(a)| > \beta/2]$, $\forall t \geq t_0$. Because $f_t(y; x, a)$ is the average of i.i.d. Bernoulli random variables with mean $p_{xy}(a)$, it follows from Cramér's theorem on large deviations (c.f., Dembo and Zeitouni 1993, p. 27) that $\mathbf{P}_{p_x(a)}[|f_t(y; x, a) - p_{xy}(a)| > \beta/2] \leq Ce^{-ct}$ for some C, $c > 0$, therefore, part (i) follows.

(ii) From the continuity of $\mathbf{I}(p, q)$ in p and q and, hence, of $\mathbf{J}(p; P, \epsilon)$ in p, it follows that the event $\{\mathbf{J}(\hat{p}_x^t(a); P, \epsilon) < \mathbf{J}(p_x(a); P, \epsilon) - \delta\}$ implies the event: $\{\|\hat{p}_x^t(a) - p_x(a)\| > \eta\}$, for some $\eta = \eta(\delta) > 0$. Thus, as a consequence of part (i), $\mathbf{P}_{p_x(a)}[\mathbf{J}(\hat{p}_x^t(a); P, \epsilon) < \mathbf{J}(p_x(a); P, \epsilon) - \delta] = o(1/t)$, as $t \to \infty$. □

PROPOSITION 3. *$\forall \pi_0 \in C_R$, $\forall P \in \hat{\mathscr{P}}$ and $\forall x \in S$, $a \notin O(x,P)$, the following is true.*

$$\lim_{\epsilon \to 0} \limsup_{N \to \infty} \mathbf{E}_{x_0}^{\pi_0, P} T_N^{(1)}(x,a;\epsilon)/\log N \le 1/\mathbf{K}(x,a;P), \quad \textit{if } (x,a) \in \mathbf{B}(P),$$

$$\lim_{\epsilon \to 0} \limsup_{N \to \infty} \mathbf{E}_{x_0}^{\pi_0, P} T_N^{(1)}(x,a;\epsilon)/\log N = 0, \quad \textit{if } (x,a) \notin \mathbf{B}(P).$$

PROOF. Fix $x \in S$, $a \in A(x)$, $a \notin O(x;P)$. We will use the abbreviations: $t_k = T_k(x,a)$, $\hat{p}^{t_k} = \hat{p}_x^{T_k(x,a)}(a)$, $\Delta\Theta(\epsilon) = \Delta\Theta(x,a;P,\epsilon)$, $\Delta\Theta = \Delta\Theta(x,a;P,0)$, $\mathscr{L}(q) = \mathscr{L}(x,a;q,h^*(P))$, $\hat{\mathscr{L}}(q) = \mathscr{L}(x,a;q,\hat{h}^k)$, $\mathscr{L}^* = \mathscr{L}^*(x;P)$.

On the event $\{\|\hat{h}^k - h^*\| \le \epsilon\}$ it is true that

$$|\hat{\mathscr{L}}(q) - \mathscr{L}(q)| \le \epsilon, \quad \forall q. \tag{6.1}$$

We consider two cases.

Case 1. $\exists \epsilon_0 > 0$ such that $\Delta\Theta(\epsilon_0) = \varnothing$, i.e., $\mathscr{L}(q) \le \mathscr{L}^* - \epsilon_0$ $\forall q$. Then, for any $\epsilon < \epsilon_0/3$, it follows from (6.1) that $\hat{\mathscr{L}}(q) \le \mathscr{L}(q) + \epsilon \le \mathscr{L}^* - \epsilon_0 + \epsilon < \mathscr{L}^* - 2\epsilon$, therefore, $\mathbf{U}(x,a;\omega_k) < \mathscr{L}^* - 2\epsilon$, hence $T_N^{(1)}(x,a;\epsilon) = 0$.

Case 2. $\Delta\Theta(\epsilon) \ne \varnothing$, $\forall \epsilon > 0$. Note that the definitions of the functions $\mathbf{u}_{x,a}$ and $\mathbf{J}$ imply that they possess the following "duality" property. The inequality $\mathbf{u}_{x,a}(p, h^*(P), \gamma) > \mathscr{L}^*(P) - \epsilon$ is equivalent to $\mathbf{J}(p;P,\epsilon) \le \gamma$, $\forall \epsilon, \gamma > 0$.

Now, (6.1) implies that $\mathscr{L}(q) > \hat{\mathscr{L}}(q) - \epsilon$, $\forall q$, thus, $\mathbf{u}_{x,a}(\hat{p}^{t_k}; h^*(P), \log k/t_k) > \mathbf{u}_{x,a}(\hat{p}^{t_k}; \hat{h}^k, \log k/t_k) - \epsilon = \mathbf{U}(x,a;\omega_k) - \epsilon$. Therefore, on the event $\{\mathbf{U}(x,a;\omega_k) \ge \mathscr{L}^* - 2\epsilon\}$ it is true that $\mathbf{u}_{x,a}(\hat{p}^{t_k}; h^*(P), \log k/t_k) \ge \mathscr{L}^* - 3\epsilon$, thus, from the duality between $\mathbf{u}$ and $\mathbf{J}$, $\mathbf{J}(\hat{p}^{t_k}; P, 3\epsilon) \le \log k/t_k$.

Combining the above, $T_N^{(1)}(x,a;\epsilon) \le \sum_{k=1}^{N-1} \mathbf{1}\,(Z_k = 1, \mathbf{J}(\hat{p}^{t_k}; P, 3\epsilon) \le \log k/t_k)$.

Take any $\epsilon < [\mathscr{L}^* - \mathscr{L}(p_x(a))]/3$ and let $\mathbf{J}_\epsilon = \mathbf{J}(p_x(a); P, 3\epsilon)$, $\hat{\mathbf{J}}_\epsilon = \mathbf{J}(\hat{p}^{t_k}; P, 3\epsilon)$. Then, $\mathbf{J}_\epsilon > 0$ and $\forall \delta \in (0, \mathbf{J}_\epsilon)$ we have (sample path-wise) that

$$\begin{aligned}
T_N^{(1)}(x,a;\epsilon) &\le \sum_{k=1}^{N-1} \mathbf{1}\big(Z_k = 1, \hat{\mathbf{J}}_\epsilon \le \log k/t_k\big) \\
&= \sum_{k=1}^{N-1} \Big[\mathbf{1}\big(Z_k = 1, \mathbf{J}_\epsilon \le \log k/t_k, \hat{\mathbf{J}}_\epsilon \ge \mathbf{J}_\epsilon - \delta\big) \\
&\qquad + \mathbf{1}\big(Z_k = 1, \hat{\mathbf{J}}_\epsilon \le \log k/t_k, \hat{\mathbf{J}}_\epsilon < \mathbf{J}_\epsilon - \delta\big)\Big] \\
&\le \sum_{k=1}^{N-1} \Big[\mathbf{1}(Z_k = 1, t_k \le \log k/(\mathbf{J}_\epsilon - \delta)) + \mathbf{1}\big(Z_k = 1, \hat{\mathbf{J}}_\epsilon < \mathbf{J}_\epsilon - \delta\big)\Big] \\
&\le \log N/(\mathbf{J}_\epsilon - \delta) + 1 + \sum_{k=1}^{N-1} \mathbf{1}\big(Z_k = 1, \hat{\mathbf{J}}_\epsilon < \mathbf{J}_\epsilon - \delta\big),
\end{aligned}$$

where, for the last inequality we have used Lemma 3.

Taking expectations of the first and last terms and because δ is arbitrarily small, $\mathbf{E}_{x_0}^{\pi_0, P} T_N^{(1)}(x, a; \epsilon) \leq \log N / \mathbf{J}_\epsilon + 1 + \mathbf{E}_{x_0}^{\pi_0, P}[\sum_{k=1}^{N-1} \mathbf{1}(Z_k = 1, \hat{\mathbf{J}}_\epsilon < \mathbf{J}_\epsilon - \delta)]$. In addition,

$$\begin{aligned}
&\mathbf{E}_{x_0}^{\pi_0, P}\left[\sum_{k=1}^{N-1} \mathbf{1}\big(Z_k = 1, \hat{\mathbf{J}}_\epsilon < \mathbf{J}_\epsilon - \delta\big)\right] \\
&\quad = \mathbf{E}_{x_0}^{\pi_0, P}\left[\sum_{j=0}^{T_N(x,a)} \mathbf{1}\big(\mathbf{J}(\hat{p}^j; P, 3\epsilon) < \mathbf{J}(p_x(a); P, 3\epsilon) - \delta\big)\right] \\
&\quad \leq \mathbf{E}_{x_0}^{\pi_0, P}\left[\sum_{j=0}^{N} \mathbf{1}\big(\mathbf{J}(\hat{p}^j; P, 3\epsilon) < \mathbf{J}(p_x(a); P, 3\epsilon) - \delta\big)\right] \\
&\quad \leq \mathbf{E}_{p_x(a)}\left[\sum_{j=0}^{N} \mathbf{1}\big(\mathbf{J}(\hat{p}^j; P, 3\epsilon) < \mathbf{J}(p_x(a); P, 3\epsilon) - \delta\big)\right] = o(\log N),
\end{aligned}$$

where the second inequality follows from Remark 1 and the last equality from Lemma 4. Therefore, $\mathbf{E}_{x_0}^{\pi_0, P} T_N^{(1)}(x, a; \epsilon) \leq \log N / \mathbf{J}(p_x(a); P, 3\epsilon) + o(\log N)$.

To complete the proof, note that $\lim_{\epsilon \to 0} \mathbf{J}(p_x(a); P, 3\epsilon) = \mathbf{K}(x, a; P)$, if $\Delta\Theta \neq \varnothing$ and $\lim_{\epsilon \to 0} \mathbf{J}(p_x(a); P, 3\epsilon) = \infty$, if $\Delta\Theta = \varnothing$.

Indeed, the first claim is a direct consequence of the definitions of $(x, a) \in \mathbf{B}(P)$, $\Delta\Theta$, $\mathbf{J}$ and $\mathbf{K}$.

For a proof of the second claim, if $\Delta\Theta(\epsilon) \neq \varnothing$ and $\Delta\Theta = \varnothing$, then $\mathscr{L}^* = \sup_{q \in \Theta(x,a)}\{\mathscr{L}(q)\} < \infty$. For any $\gamma > 0$, let $u(\gamma) = u_{x,a}(p_x(a), h^*(P), \gamma)$, note that $\mathscr{L}^* > u(\gamma)$ and let $\epsilon_0 = (\mathscr{L}^* - u(\gamma))/2 > 0$. Then, $\mathscr{L}(q) \geq \mathscr{L}^* - \epsilon_0 = (\mathscr{L}^* + u(\gamma))/2 > u(\gamma)$, $\forall q \in \Delta\Theta(\epsilon_0)$. Hence, $\mathbf{I}(p_x(a), q) \geq \gamma$, $\forall q \in \Delta\Theta(\epsilon_0)$, otherwise we have the contradiction that $\exists q_0 \in \Delta\Theta(\epsilon_0)$ with $\mathscr{L}(q_0) \leq u(\gamma)$. In summary (and using the definition of $\mathbf{J}_\epsilon$), we have shown that $\forall \gamma > 0$, $\exists \epsilon_0 = \epsilon_0(\gamma)$ such that $\mathbf{J}_\epsilon > \gamma$; the proof of the claim is complete. □

In Proposition 4, below, the structure of an index policy is used to show that if a is a nonoptimal action (under P) in state x, events specifying $T_N^{(2)}(x, a; \epsilon)$ can occur only if the index of each optimal action a^* is also less than $\mathscr{L}^*(x; P) - 2\epsilon$. In Lemma 6 it is established that the probability of such an event occurring at any period k is $o(1/k)$. In proving Lemma 6, the definition of the index $\mathbf{U}(x, a^*, \omega_k)$ implies that a necessary condition for $\mathbf{U}(x, a^*; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon$ is that the transition probability estimate $\hat{p}_x(a^*)$ is sufficiently far from the true value $p_x(a^*)$. The proof that the probability of this latter event is $o(1/k)$ does not utilize the structure of the policy or the index. Instead it requires a change of measure transformation developed in Lemma 5 the proof of which is included in Appendix A.

Let

$$\begin{aligned}
\lambda(p; q_1, q_2) &= \mathbf{I}(p, q_2) - \mathbf{I}(p, q_1) \\
&= \sum_{y \in S^+(x,a)} p_y \log\big[q_{1_y}/q_{2_y}\big], \quad \text{for } x \in S,\ a \in A(x),\ p, q_1, q_2 \in \Theta(x, a).
\end{aligned}$$

LEMMA 5. (i) *Fix $x \in S$, $a \in A(x)$, $\epsilon > 0$ sufficiently small. Then $\forall h \in \mathbb{R}^s$, with $h(y) \neq h(y')$ for some $y, y' \in S^+(x, a)$, there exists a vector $q^0(\epsilon) \in \Theta(x, a)$ such that*

$\mathscr{L}(x, a; q^0(\epsilon), h) = \mathscr{L}(x, a; p_x(u), h) - \epsilon$ *and*

$$\{q \colon \mathscr{L}(x, a; q, h) < \mathscr{L}(x, a; p_x(a), h) - \epsilon\}$$
$$= \{q \colon \lambda(q; p_x(a), q^0(\epsilon)) < -\mathbf{I}(q^0(\epsilon), p_x(a))\}.$$

(ii) $\forall x \in S,\ a \in A(x),\ p, q \in \Theta(x, a),\ c, d > 0$ *and* $b_1, b_2, \gamma \in \mathbb{R}$,

$$\sum_{t=\lfloor(d \log k+\gamma)\rfloor}^{k} \mathbf{P}_p[\lambda(f_t; p, q) < -c + b_1/t, \mathbf{I}(f_t, q) > (\log k + b_2)/t]$$
$$= o(1/k), \quad as\, k \to \infty.$$

(iii) $\forall x \in S,\ a \in A(x),\ p, q \in \Theta(x, a),\ c, d > 0$ *and* $\beta, \gamma \in \mathbb{R}$,

$$\sum_{t=\lfloor(d \log k+\gamma)\rfloor}^{k} \mathbf{P}_p[\lambda(\hat{p}_x^t(a); p, q) < -c, \mathbf{I}(\hat{p}_x^t(a), q) > (\log k + \beta)/t]$$
$$= o(1/k), \quad as\, k \to \infty. \quad \square$$

For $p \in \Theta(x, a)$ and $\beta \in \mathbb{R}$, let $F_{kt}(p; \beta) = \{q \in \Theta(x, a)\colon\ \mathbf{I}(p, q) \leq (\log k + \beta)/t\}$. Note that $\mathbf{U}(x, a; \omega_k) = \max\{\mathscr{L}(x, a; q, \hat{h}^k)\colon\ q \in F_{k, T_k(x, a)}(\hat{p}_x^{T_k(x, a)}(a); 0)\}$. For $\epsilon > 0$, $\beta \in \mathbb{R}$, $h \in \mathbb{R}^s$, and $1 \leq t \leq k$, the events:

$$(6.2) \quad B_{kt}(x, a; \epsilon, \beta, h)$$
$$= \{\omega_k \colon \mathscr{L}(x, a; q, h) < \mathscr{L}(x, a; p_x(a), h) - \epsilon, \forall q \in F_{kt}(\hat{p}_x^t(a); \beta)\},$$

satisfy the claim of Lemma 6 below.

LEMMA 6. *For all* $x \in S,\ a \in A(x),\ \epsilon > 0,\ \beta \in \mathbb{R},\ h \in \mathbb{R}^s$,

$$\sum_{t=0}^{k} \mathbf{P}_{p_x(a)}[B_{kt}(x, a; \epsilon, \beta, h)] = o(1/k), \quad as\, k \to \infty.$$

PROOF. If $h(y) = h_1$, $\forall y \in S^+(x, a^*)$, then $\mathscr{L}(x, a; q, h) = r(x, a) + h_1 = \mathscr{L}(x, a; p_x(a), h)$, $\forall q$, therefore, $\mathbf{P}_{p_x(a)}[B_{kt}(x, a; \epsilon, \beta, h)] = 0$, $\forall t$ and the claim is true.

We next consider the case that $h(y) \neq h(y')$ for some $y, y' \in S^+(x, a)$. From Lemma 5(i), it follows that for $\epsilon > 0$ sufficiently small, there exists a probability vector $q^0 = q^0(\epsilon) \in \Theta(x, a)$ such that $\mathscr{L}(x, a; q^0, h) = \mathscr{L}(x, a; p_x(a), h) - \epsilon$ and

$$\{q \colon \mathscr{L}(x, a; q, h) < \mathscr{L}(x, a; p_x(a), h) - \epsilon\} = \{q \colon \lambda(p_x(a), q^0) < -c\},$$

where $c = \mathbf{I}(q^0, p_x(a))$. Hence, $B_{kt}(x, a; \epsilon, \beta, h) \subseteq B'_{kt}$, where $B'_{kt} = \{\lambda(q; p_x(a), q^0) < -c, \forall q \in F_{kt}(\hat{p}_x^t(a); \beta)\}$ and it is sufficient to prove that $\sum_{t=0}^{k} \mathbf{P}_{p_x(a)}[B'_{kt}] = o(1/k)$.

On the event B'_{kt} the following are true. First, since $\hat{p}_x^t(a) \in F_{kt}(\hat{p}_x^t(a); \beta)$, it follows that $\lambda(\hat{p}_x^t(a); p_x(a), q^0) < -c$. Second, since $\lambda(q^0; p_x(a), q^0) = -c$, it follows that $q^0 \notin F_{kt}(\hat{p}_x^t(a); \beta)$, i.e., $\mathbf{I}(\hat{p}_x^t(a), q^0) > (\log k + \beta)/t$.

Therefore, $B'_{kt} \subseteq B''_{kt}$, where $B''_{kt} = \{\lambda(\hat{p}_x^t(a); p_x(a), q^0) < -c,\ \mathbf{I}(\hat{p}_x^t(a), q^0) > (\log k + \beta)/t\}$. Let $\bar{I}(q^0) = \max_{q \in \Theta(x.\, a)} \mathbf{I}(q, q^0) = \max_{y \in S^+(x, a)} |\log q^0(y)| < \infty$. For $t < (\log k + \beta)/\bar{I}(q^0)$ it is true that $\mathbf{I}(\hat{p}_x^t(a), q^0) \leq \bar{I}(q^0) < (\log k + \beta)/t$, thus,

$B''_{kt} = \varnothing$. Therefore, in order to prove the lemma it suffices to show that $\sum_{t=\lfloor(\log k+\beta)/\bar{I}\rfloor}^{k} \mathbf{P}_{p_x(a)}[B''_{kt}] = o(1/k)$. The last relation follows from Lemma 5(iii) for sufficiently small ϵ. This completes the proof. □

We can now prove the following.

PROPOSITION 4. $\mathbf{E}_{x_0}^{\pi_0 \cdot P} T_N^{(2)}(x, a; \epsilon) = o(\log N)$, *as* $N \to \infty$, $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$, $x \in S$, $a \notin O(x; P)$ *and* $\epsilon > 0$ *sufficiently small.*

PROOF. Recall that $T_N^{(2)}(x, a; \epsilon) = \sum_{k=1}^{N-1} \mathbf{1}\,(A_k(x, a; \epsilon))$, where

$$A_k(x, a; \epsilon) = \Big\{Z_k = 1, \|\hat{h}^k - h^*\| \le \epsilon, \mathscr{O}\big(\hat{P}^k, \mathscr{D}_k\big) \subseteq \mathscr{O}(P),$$

$$\mathbf{U}(x, a; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon\Big\}.$$

To prove the proposition it is sufficient to show that $\mathbf{P}_{x_0}^{\pi_0, P} A_k(x, a; \epsilon) = o(1/k)$, as $k \to \infty$, $\forall \epsilon > 0$. Fix an $x \in S$ and an $a \notin O(x; P)$ as in the statement of the proposition. On the event $A_k(x, a; \epsilon)$ it is true that: $\Gamma_1(x; \hat{P}^k, \mathscr{D}_k) \subseteq \mathscr{O}(\hat{P}^k, \mathscr{D}_k) \subseteq O(x; P)$ and since $a \notin O(x; P)$, it follows that, on the event $A_k(x, a; \epsilon)$, $a \in \Gamma_2(x; \hat{P}^k, \mathscr{D}_k)$. Hence, on $A_k(x, a; \epsilon)$ the following holds: $\mathbf{U}(x, a^*; \omega_k) \le \mathbf{U}(x, a; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon$, $\forall a^* \in O(x; P)$. Thus, $A_k(x, a; \epsilon) \subseteq \bigcap_{a^* \in O(x; P)} A'_k(x, a^*; \epsilon)$, where

$$A'_k(x, a^*; \epsilon) = \Big\{\omega_k : \|\hat{h}^k - h^*\| \le \epsilon, \mathbf{U}(x, a^*; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon\Big\},$$

and it suffices to show that $\mathbf{P}_{x_0}^{\pi_0, P} A'_k(x, a^*; \epsilon) = o(1/k)$, as $k \to \infty$, for any $a^* \in O(x; P)$. Fix $a^* \in O(x; P)$ and $\epsilon > 0$, and note that $A'_k(x, a^*; \epsilon) = \bigcup_{t=0}^{k}(A'_k(x, a^*; \epsilon) \cap \{T_k(x, a^*) = t\})$.

On the event $A'_k(x, a^*; \epsilon) \cap \{T_k(x, a^*) = t\}$ the following are true. First, since $\|\hat{h}^k - h^*\| \le \epsilon$, it is true that $\mathscr{L}(x, a^*; q, h^*) \le \mathscr{L}(x, a^*; q, \hat{h}^k) + \epsilon$ for all q. Second, since $\mathbf{U}(x, a^*; \omega_k) < \mathscr{L}^*(x; P) - 2\epsilon$ and $T_k(x, a^*) = t$, it is true that $\mathscr{L}(x, a^*; q, \hat{h}^k) < \mathscr{L}^*(x; P) - 2\epsilon$ for all $q \in F_{kt}(\hat{p}_x^t(a^*); 0)$. Combining the above, it follows that $A'_k(x, a^*; \epsilon) \cap \{T_k(x, a^*) = t\} \subseteq B_{kt}(x, a^*; \epsilon, 0, h^*)$, where the events $B_{kt}(x, a^*; \epsilon, 0, h^*)$ were defined in (6.2). Thus,

$$\mathbf{P}_{x_0}^{\pi_0, P} A'_k(x, a^*; \epsilon) = \sum_{t=0}^{k} \mathbf{P}_{x_0}^{\pi_0, P}\big(A'_k(x, a^*; \epsilon) \cap \{T_k(x, a^*) = t\}\big)$$

$$\le \sum_{t=0}^{k} \mathbf{P}_{p_x(a^*)}{:}B_{kt}(x, a^*; \epsilon, 0, h^*) = o(1/k), \quad \text{as } k \to \infty,$$

where the first inequality is due to Remark 1, and the last equality is established by Lemma 6 with $a = a^*$, $h = h^*$, $\beta = 0$. □

In Proposition 5 it is shown that $\mathbf{E}_{x_0}^{\pi_0, P} T_N^{(3)}(\epsilon) = o(\log N)$, by showing that the event that in period k the solution $\hat{h}^k$ of the restricted problem $(\hat{P}^k, \mathscr{D}_k)$ differs from the optimal solution $h^*(P)$ by more than ϵ, has probability of order $o(1/k)$.

To prove this claim, we divide the first k time periods (for large k) into $m + 1$ subintervals, where $m = |\mathscr{A}|$ is the total number of deterministic policies, in such a way that the length of each subinterval increases linearly with k. The main step of the proof is to show that, with high probability, after the first subinterval, the transition probability estimates for all frequently taken actions will be sufficiently accurate, so that the following events will occur. If at the beginning of any subsequent subinterval the solution of the (observable) restricted problem is a nonoptimal policy for the (unobservable) unrestricted problem $(P, \mathscr{A})$, then, during this subinterval, the indices of the under-sampled actions will be large enough to induce taking at least one improving action in every state; this property is shown in Lemma 7 and employed in Lemma 9. Thus, at the end of the subinterval the new estimate of the optimal policy will be strictly better than in the beginning. This monotonicity property ensures that, even in the worst case, at the end of the last subinterval (i.e., at period k), a true optimal policy will have been identified.

Let $m = |\mathscr{A}|$ denote the total number of deterministic policies. Fix $\beta < 1/(m+1)$.

For $k > (1/(m+1) - \beta)^{-1}$, divide the time interval $\{0, \dots, k-1\}$ into $m+1$ subintervals $I_\nu^k = \{t\colon b_\nu^k \le t < b_{\nu+1}^k\}$, $\nu = 0, \dots, m$, where $b_0^k = 0$ and $b_\nu^k = k - (m + 1 - \nu)\lfloor k/(m+1) \rfloor$, $\nu = 1, \dots, m+1$. For $\nu = 0$, $b_1^k - b_0^k = k - m\lfloor k/(m+1)\rfloor \ge k/(m+1) > \beta k$, while for $\nu = 1, \dots, m$, $b_{\nu+1}^k - b_\nu^k = \lfloor k/(m+1) \rfloor > k/(m+1) - 1 > \beta k$, since $k > (1/(m+1) - \beta)^{-1}$. Therefore, the length of all subintervals I_ν^k is greater than βk, and $b_\nu^k > \nu \beta k$. In addition, $b_{\nu+1}^k \le k$, therefore, $b_{\nu+1}^k / b_\nu^k < 1/(\nu\beta) \le 1/\beta$ and $\log b_{\nu+1}^k - \log b_\nu^k < -\log \beta$, $\forall \nu = 1, \dots, m$.

For $\nu = 1, \dots, m$, let $\Delta_\nu(y, a) = T_{b_{\nu+1}^k}(y, a) - T_{b_\nu^k}(y, a)$, and $\Delta_\nu(y) = \sum_{a \in A(y)} \Delta_\nu(y, a)$, denote, respectively, the number of occurrences of the pair (y, a) and of state y, during subinterval I_ν^k.

Fix $\rho \in (0,1)$ such that Proposition 2 holds. For $k > (1/(m+1) - \beta)^{-1}$, and $\delta, \zeta > 0$ let

$$B_k(\zeta) = \Big\{\omega_k\colon \Delta_\nu(y) > \rho\beta k \text{ and } \big\| \hat{p}_y^{T_t(y,a)}(a) - p_y(a) \big\| \le \zeta,$$

$$\forall y \in S, \forall \nu = 1, \dots, m, \forall t \ge b_1^k, \forall a \in D_t(y)\Big\}.$$

$B_k(\zeta)$ denotes the event that during any interval I_ν^k all states are visited at least $\rho\beta k$ times, and also after period b_1^k, for all actions $a \in D_t(y)$, the estimates of the transition probabilities are within ζ of the true values.

In Lemma 7 below several intermediate properties of the estimation scheme and the index function are established in terms of the events $C_\nu^k(\delta)$, defined for $\delta > 0$ and $\nu = 1, \dots, m$. Since the proof of these properties is not dependent on the structure of the index policies, but instead uses mainly continuity arguments, it is included in Appendix A. Let

$$C_\nu^k(\delta) = \Big\{\omega_k\colon h(P, \mathscr{D}_t) = h\big(P, \mathscr{D}_{b_\nu^k}\big), \forall t \in I_\nu^k \text{ and } \mathbf{U}(y, a; \omega_{t_0})$$

$$\le \mathscr{L}\big(y, a; p_y(a), h\big(P; \mathscr{D}_{b_\nu^k}\big)\big) - \delta, \quad \text{for some } (y, a) \text{ and some } t_0 \in I_\nu^k\Big\}.$$

LEMMA 7. *$\forall \pi_0 \in C_R$ $\forall \delta$, $\epsilon > 0$, $\exists \zeta_1 = \zeta_1(\epsilon)$, $\zeta_2 = \zeta_2(\delta)$, $\zeta_3 = \zeta_3(\delta) > 0$, $k_0 = k_0(m, \delta)$ such that:*

(i) $B_k(\zeta) \subseteq \{\omega_k\colon |g_f(\hat{P}^t) - g_f(P)| < \epsilon, \|h_f(\hat{P}^t) - h_f(P)\| < \epsilon, \forall t \ge b_1^k, \forall f \in \mathscr{D}_t\}$, $\forall \zeta < \zeta_1$.

(ii) $B_k(\zeta) \subseteq \{\omega_k\colon \mathbf{U}(y, a; \omega_t) \le \mathscr{L}(y, a; p_y(a), h(P, \mathscr{D}_t)) + \delta, \forall t \ge b_1^k, y \in S, a \in D_t(y)\}$ $\forall \zeta < \zeta_2$, $k > k_0$.

(iii) $\mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta) C_\nu^k(\delta)] = o(1/k)$, *as* $k \to \infty$, $\forall \nu = 1, \dots, m$, $\forall \zeta < \zeta_3$.

LEMMA 8. *Let $\pi_0 \in C_R$. On the event $B_k(\zeta)$, $\forall \epsilon > 0$, $\exists \xi_1 = \xi_1(\epsilon) > 0$, $k_0 = k_0(m, \epsilon)$ such that, for $k > k_0$ the following are true.*

(i) *For $t \ge b_1^k$, any $y \in S$, and $a \in A(y)$, if $Z_t(y, a) = 1$ with $a = \pi_0(t) \in$*

$\Gamma_1(y;\hat{P}^t,\mathscr{D}_t)$, *then* $\sum_{j=1}^{\tau}\mathbf{1}(Z_{t+j}(y,a)=1,\ \ a\in\Gamma_1(y;\hat{P}^{t+j},\mathscr{D}_{t+j}))\le 1$, *where* $\tau=\rho\beta k/(2\log(\rho\beta k))$, $\forall\zeta>0$.

(ii) $\mathscr{O}(\hat{P}^t,\mathscr{D}_t)\subseteq\mathscr{O}(P,\mathscr{D}_t)$, *and* $\|h(\hat{P}^t,\mathscr{D}_t)-h(P,\mathscr{D}_t)\|<\epsilon$, $\forall t=b_1^k,\dots,k$, $\forall\zeta<\xi_1$.

(iii) $g(P,\mathscr{D}_{t+1})\ge g(P,\mathscr{D}_t)$, $\forall t=b_1^k,\dots,k$, $\forall\zeta<\xi_1$.

PROOF. (i) Fix $y\in S$, $a=\pi_0(t)\in\Gamma_1(y;\hat{P}^t,\mathscr{D}_t)$, $t\ge b_1^k$, let $n=T_t(y)$, $\sigma=n/(2\log n)$. Recall that $\Gamma_1(y;\hat{P}^t,\mathscr{D}_t)\subseteq D_t(y)$. Thus, from the definitions of $D_t(y)$ and $\Gamma_1(\cdot,\cdot,\cdot)$ we have $\log^2 n\le T_t(y,a)<\log^2 n+1$.

For n sufficiently large it is true that

$$\sum_{j=1}^{\sigma}\mathbf{1}\left(Z_{t+j}(y,a)=1,a\in\Gamma_1\left(y;\hat{P}^{t+j},\mathscr{D}_{t+j}\right)\right)$$

$$=\sum_{j=1}^{\sigma}\mathbf{1}\left(Z_{t+j}(y,a)=1,\log^2 T_{t+j}(y)\le T_{t+j}(y,a)<\log^2 T_{t+j}(y)+1\right)$$

$$\le\sum_{j=1}^{\sigma}\mathbf{1}\left(Z_{t+j}(y,a)=1,\log^2 n+1\le T_{t+j}(y,a)<\log^2(n+\sigma)+1\right)$$

$$\le\sum_{j=1}^{\sigma}\mathbf{1}\left(Z_{t+j}(y,a)=1,\log^2 n+1\le T_{t+j}(y,a)<\log^2 n+2\right)$$

$$=\sum_{j=1}^{\sigma}\mathbf{1}\left(Z_{t+j}(y,a)=1,T_{t+j}(y,a)=\lfloor\log^2 n\rfloor+2\right)\le 1.$$

The first inequality is due to: $T_{t+j}(y,a)\ge T_{t+1}(y,a)=T_t(y,a)+1\ge\log^2 n+1$ (since $Z_t(y,a)=1$) and $T_{t+j}(y)\le T_{t+\sigma}(y)\le T_t(y)+\sigma=n+\sigma$ (since, $j\le\sigma$).

The second inequality follows from the property that, for n sufficiently large, $\log^2(n+\sigma)<\log^2 n+1$. To see this, let $\upsilon(x)=2x\log(1+1/(2x))+\log^2(1+1/(2x))$ and note that $\upsilon(x)$ is increasing for $x\ge e$ and $\lim_{x\to\infty}\upsilon(x)=1$, thus, $\upsilon(\log n)<1$ and $\log^2 n+1>\log^2(n+n/(2\log n))$.

The last inequality follows from a counting argument similar to that in the proof of Lemma 3.

For all $\zeta>0$, on the event $B_k(\zeta)$ it is true that $n=T_t(y)\ge T_{b_1^k}(y)\ge\rho\beta k$, thus, $\sigma\ge\rho\beta k/(2\log(\rho\beta k))$. This completes the proof of (i).

In the proof of (ii) and (iii), below, we assume that $g_f(P)\ne g_{f'}(P)$ for at least one pair of policies $f,f'\in\mathscr{A}$, because otherwise all claims are trivially true.

(ii) Let $\delta_1=\min\{|g_f(P)-g_{f'}(P)|: f,f'\in\mathscr{A},\ g_f(P)\ne g_{f'}(P)\}>0$. Then, $\forall f,f'\in\mathscr{A}$, either $|g_f(P)-g_{f'}(P)|\ge\delta_1$, or $g_f(P)=g_{f'}(P)$. From Lemma 7(i), it follows that $\exists\zeta_1(\delta_1/2)>0$, $\zeta_1(\epsilon)>0$ such that for $\zeta<\min\{\zeta_1(\delta_1/2),\zeta_1(\epsilon)\}$ we have,

$$B_k(\zeta)\subseteq B_k'(\zeta)\equiv\Big\{\omega_k:\left|g_f(\hat{P}^t)-g_f(P)\right|<\delta_1/2,\left\|h_f(\hat{P}^t)-h_f(P)\right\|$$
$$<\epsilon,\forall f\in\mathscr{D}_t,t=b_1^k,\dots,k\Big\}.$$

Let $f\in\mathscr{O}(\hat{P}^t,\mathscr{D}_t)$. On the event $B_k'(\zeta)$ it is true that, $\forall f'\in\mathscr{D}_t\subseteq\mathscr{A}$,

$$g_f(P)>g_f(\hat{P}^t)-\delta_1/2\ge g_{f'}(\hat{P}^t)-\delta_1/2>g_{f'}(P)-\delta_1,$$

thus, $g_f(P) \geq g_{f'}(P)$, by the choice of δ_1. Therefore, $B'_k(\zeta) \subseteq \{\omega_k\colon \mathscr{O}(\hat{P}^t, \mathscr{D}_t) \subseteq \mathscr{O}(P, \mathscr{D}_t), \|h(\hat{P}^t, \mathscr{D}_t) - h(P, \mathscr{D}_t)\| < \epsilon \ \forall t = b_1^k, \dots, k\}$ and part (ii) follows with $\xi_1 = \xi_1(\epsilon) = \min\{\zeta_1(\delta_1/2), \zeta_1(\epsilon)\}$.

(iii) It suffices to show that for all $t \geq b_1^k$,

$$\mathscr{D}_{t+1} \cap \mathscr{O}\left(\hat{P}^t, \mathscr{D}_t\right) \neq \emptyset. \tag{6.3}$$

Indeed, assume (6.3) is true. From part (ii), for any $\epsilon > 0$ and $\zeta < \xi_1(\epsilon)$, on the event $B_k(\zeta)$ it is true that $\mathscr{O}(\hat{P}^t, \mathscr{D}_t) \subseteq \mathscr{O}(P, \mathscr{D}_t)$, and this, in conjunction with (6.3), implies that $\mathscr{D}_{t+1} \cap \mathscr{O}(P, \mathscr{D}_t) \neq \emptyset$, $\forall t \geq b_1^k$, thus $g(P, \mathscr{D}_{t+1}) \geq g(P, \mathscr{D}_t)$.

We next prove (6.3). Let $y = X_t(\omega_k)$. We will show that $D_{t+1}(y) \cap O(y; \hat{P}^t, \mathscr{D}_t) \neq \emptyset$.

Indeed, if $\Gamma_1(y; \hat{P}^t, \mathscr{D}_t) = O(y; \hat{P}^t, \mathscr{D}_t)$, then $\pi_0(t, \omega_k) = a \in \Gamma_1(y; \hat{P}^t, \mathscr{D}_t) \cap \mathscr{D}_t(y)$, i.e., $T_t(y, a) \geq \log^2 T_t(y)$ and $T_{t+1}(y, a) = T_t(y, a) + 1$, $T_{t+1}(y) = T_t(y) + 1$. Hence, using the inequality $\log^2(n) + 1 > \log^2(n + 1)$, we get $T_{t+1}(y, a) = T_t(y, a) + 1 \geq \log^2 T_t(y) + 1 > \log^2(T_t(y) + 1) = \log^2 T_{t+1}(y)$, i.e., $a \in D_{t+1}(y)$.

If $\Gamma_1(y; \hat{P}^t, \mathscr{D}_t) \neq O(y; \hat{P}^t, \mathscr{D}_t)$, then $\exists a \in O(y; \hat{P}^t, \mathscr{D}_t)$ and $a \notin \Gamma_1(y; \hat{P}^t, \mathscr{D}_t)$, i.e., $T_t(y, a) \geq \log^2(T_t(y)) + 1$. Then for this a, regardless of the value of $\pi_0(t, \omega_k)$, we have $T_{t+1}(y, a) \geq T_t(y, a) \geq \log^2(T_t(y)) + 1 > \log^2(T_t(y) + 1) = \log^2 T_{t+1}(y)$, i.e., $a \in D_{t+1}(y)$.

To complete the proof of (6.3), note that $D_{t+1}(z) = D_t(z)$, $\forall z \in S$, $z \neq y$. □

In Lemma 9 below we prove the following monotonicity property (expressed via events G_k): $\forall \nu = 1, \dots, m$, if $\mathscr{D}_{b_\nu^k}$ includes an optimal policy in the beginning of interval I_ν^k, then this will also be true in the beginning of the next interval, otherwise at least one policy strictly better than all policies in $\mathscr{D}_{b_\nu^k}$ will be included in $\mathscr{D}_{b_{\nu+1}^k}$. Formally, let

$$G_k = \left\{\omega_k\colon \text{either } g\left(P, \mathscr{D}_{b_{\nu+1}^k}\right) > g\left(P, \mathscr{D}_{b_\nu^k}\right), \forall \nu \geq 1 \text{ or}\right.$$

$$\left. g\left(P, \mathscr{D}_{b_{\nu+1}^k}\right) = g\left(P, \mathscr{D}_{b_\nu^k}\right) = g^*(P) \text{ for some } \nu\right\}.$$

Then we have:

LEMMA 9. *$\exists \xi_2 > 0$ such that $\mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta)\overline{G}_k] = o(1/k)$, as $k \to \infty$, $\forall \zeta < \xi_2$.*

PROOF. We assume that $g_f(P) \neq g_{f'}(P)$ for at least one pair $f, f' \in \mathscr{A}$, because otherwise the claim is trivially true. For $t > 0$ and $y \in S$ let

$$A^+(y; P, \mathscr{D}_t) = \left\{a \in A(y)\colon g(P, \mathscr{D}_t) + h(y; P, \mathscr{D}_t)\right.$$

$$\left. < \mathscr{L}\left(y, a; p_y(a), h(y, P; \mathscr{D}_t)\right)\right\},$$

denote the set of improving actions for the unrestricted (unobservable) problem $(P, \mathscr{A})$ given the solution, $g(P, \mathscr{D}_t) h(y; P, \mathscr{D}_t)_{y \in S}$, of the restricted (unobservable) problem $(P, \mathscr{D}_t)$.

According to this definition, the set of improving actions for the unrestricted problem $(P, \mathscr{A})$ given the solution for a specific policy f, is $A^+(y; P, \{f\}) = \{a \in A(y)\colon g_f(P) + h_f(y; P) < \mathscr{L}(y, a; p_y(a), h_f(y; P))\}$.

Let $\delta_2 = \min\{\mathscr{L}(y, a; p_y(a), h_f(y; P)) - g_f(P) - h_f(y; P)\colon f \in \mathscr{A}, y \in S, a \in A^+(y; P, \{f\}) \neq \emptyset\}$. Note that $\delta_2 > 0$.

For $\delta > 0$ define events: $E_k(\delta) = \bigcap_{\nu=1}^{m} \overline{C_\nu^k}(\delta)$.

On the event $E_k(\delta)$ it is true $\forall \nu = 1, \dots, m$, that either $\forall y \in S$, $\forall a \in A^+(y; P, \mathscr{D}_{b_\nu^k})$, $h(P, \mathscr{D}_t) = h(P, \mathscr{D}_{b_\nu^k})$ and $\mathbf{U}(y, a; \omega_t) > \mathscr{L}(y, a; p_y(a), h(P, \mathscr{D}_t)) - \delta$, $\forall t \in I_\nu^k$, or $h(P, \mathscr{D}_{t_0}) \neq h(P, \mathscr{D}_{b_\nu^k})$, for some $t_0 \in I_\nu^k$. We will refer to the first type of conditions as conditions $(\mathbf{E}_1)$.

For $\epsilon > 0$ choose: $\xi_1 = \xi_1(\epsilon)$ such that Lemma 8(ii) and (iii) holds; $\zeta_2(\delta)$, $\zeta_3(\delta)$ such that Lemma 7 holds; let $\xi_2 = \min(\xi_1, \zeta_2(\delta_2/2), \zeta_3(\delta_2/2))$.

To prove the lemma it is sufficient to show that, for $\zeta < \xi_2$ and for k sufficiently large, $B_k(\zeta)E_k(\delta_2/2) \subseteq G_k$, or equivalently,

$$B_k(\zeta)E_k(\delta_2/2)\overline{G}_k = \varnothing. \tag{6.4}$$

Indeed, suppose (6.4) is true. Then $\mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta)\overline{G}_k] \leq \mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta)\overline{E_k(\delta_2/2)}] \leq \sum_{\nu=1}^{m} \mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta)C_\nu^k(\delta_2/2)]$. It is shown in Lemma 7(iii) that, for $\zeta < \zeta_3(\delta_2/2)$, $\mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta)C_\nu^k(\delta_2/2)] = o(1/k)$. Thus, the proof will be complete, because $m = |\mathscr{A}|$ is fixed.

We next prove (6.4). Let $\zeta < \xi_2$. Note that for such ζ, Lemma 8 holds. On the event $B_k(\zeta)$ it is true from Lemma 8(iii), that $g(P; \mathscr{D}_t)$ is nondecreasing in t for $t \geq b_1^k$, thus, $B_k(\zeta)\overline{G}_k \subseteq B_k(\zeta) \bigcup_{\nu=1}^{m} F_\nu^k$, where

$$F_\nu^k = \left\{\omega_k : g\left(P; \mathscr{D}_{b_{\nu+1}^k}\right) = g\left(P; \mathscr{D}_{b_\nu^k}\right) < g^*(P)\right\}.$$

Therefore, it suffices to show that $B_k(\zeta)E_k(\delta_2/2)F_\nu^k = \varnothing$, $\forall \nu = 1, \dots, m$.

To prove the last relation, it suffices to show that for any $\nu = 1, \dots, m$,

$$B_k(\zeta)E_k(\delta_2/2)F_\nu^k \subseteq \left\{\omega_k : \exists y : \Delta_\nu(y) \geq \rho\beta k, \text{ and } \Delta_\nu(y) = O(\log^2 k)\right\}, \tag{6.5}$$

because, for large k, $\{\omega_k : \exists y : \Delta_\nu(y) \geq \rho\beta k, \Delta_\nu(y) = O(\log^2 k)\} = \varnothing$, since $\log^2 k = o(k)$.

We next prove (6.5). By definition of $A^+(y; P, \mathscr{D}_{b_\nu^k})$, on the event F_ν^k, $\exists y \in S$: $A^+(y; P, \mathscr{D}_{b_\nu^k}) \neq \varnothing$. Consider such a y and let

$$\Delta_\nu^1(y, a) = \sum_{t \in I_\nu^k} \mathbf{1}\left(Z_t(y, a) = 1, a \in \Gamma_1\left(y; \hat{P}^t, \mathscr{D}_t\right)\right),$$

$$\Delta_\nu^2(y, a) = \sum_{t \in I_\nu^k} \mathbf{1}\left(Z_t(y, a) = -1, a \in \Gamma_2\left(y; \hat{P}^t, \mathscr{D}_t\right)\right).$$

Then,

$$\Delta_\nu(y) = \Sigma_{a \in A(y)} \Delta_\nu^1(y, a) + \Sigma_{a \in A^-(y; P, \mathscr{D}_{bk_\nu})} \Delta_\nu^2(y, a) + \Sigma_{a \notin A^-(y; P, \mathscr{D}_{bk_\nu})} \Delta_\nu^2(y, a).$$

For term $\Delta_\nu^1(y, a)$, it follows from Lemma 8(i) that $\forall t \in I_\nu^k$, if $Z_{t+y}(y, a) = 1$, with $a \in \Gamma_1(y; \hat{P}^{t+j}, \mathscr{D}_{t+j})$ for two different values of $j = j_1, j_2 \leq \tau = \rho\beta k/(2\log(\rho\beta k))$, then this event cannot occur for any other value of $j \leq \tau$. Therefore, $\Sigma_{j=1}^{\tau} \mathbf{1}$ $(Z_{t+j}(y, a) = 1,\ a \in \Gamma_1(y; \hat{P}^{t+j}, \mathscr{D}_{t+j}) \leq 2$. Hence, $\Delta_\nu^1(y, a) \leq 2(b_{\nu+1}^k - b_\nu^k)/\tau \leq (4\log(\rho\beta k))/(m\rho\beta) = O(\log k)$, for all $a \in A(y)$. For term $\Delta_\nu^2(y, a)$ we need the following three observations.

(1) $A^+(y; P, \mathscr{D}_t) = A^+(y; P, \mathscr{D}_{b_\nu^k})$, $\forall t \in I_\nu^k \cup \{b_{\nu+1}^k\}$, on the event $B_k(\zeta)F_\nu^k$. Indeed, consider periods $t, t+1 \in I_\nu^k$, with $y = X_t$. Note that $\forall z \neq y$, $T_{t+1}(z) = T_t(z)$ and $T_{t+1}(z, a) = T_t(z, a)$, $D_{t+1}(z) = D_t(z)$. In addition, from the definition of events F_ν^k and Lemma 8(iii) it follows that $g(P; \mathscr{D}_{t+1}) = g(P; \mathscr{D}_t)$. Therefore, $g(P, \mathscr{D}_t)$, $h(P; \mathscr{D}_t)$ satisfy the optimality equations of $(P; \mathscr{D}_{t+1})$ for all $z \neq y$. This must also be true for state y, because if the equations are not satisfied for some $a \in D_{t+1}(y)$, it is possible to construct a policy $f \in \mathscr{D}_{t+1}$ such that $g(P; \mathscr{D}_{t+1}) > g(P; \mathscr{D}_t)$. Thus, $h(P; \mathscr{D}_{t+1}) = h(P; \mathscr{D}_t)$, and $A^+(y; P, \mathscr{D}_t) = A^+(y; P, \mathscr{D}_{t+1})$, due to the uniqueness of solution of the optimality equations.

(2) From the definition of the improving sets $A^+(y; P, \mathscr{D}_t)$ it follows that, if $g(P, \mathscr{D}_t) < g^*(P)$ for some t, then there exists at least one policy $f' \notin \mathscr{D}_t$ such that $g_{f'}(P) > g(P, \mathscr{D}_t)$, therefore there exists at least one state y such that $A^+(y; P, \mathscr{D}_t) \neq \varnothing$. Conversely, if $g(P, \mathscr{D}_t) = g^*(P)$, then $A^+(y; P, \mathscr{D}_t) = \varnothing$ for all y. In addition, since $g(P, \mathscr{D}_t)$ is the optimal solution of $(P, \mathscr{D}_t)$, if $a \in A^+(y; P, \mathscr{D}_t)$ for some $y \in S$, then $a \notin \mathscr{D}_t(y)$.

Now, using observation (1) we have that on the event $B_k(\zeta)F_\nu^k$, if $a \in A^+(y; P, \mathscr{D}_{b_\nu^k})$, then $a \in A^+(y; P, D_{b_{\nu+1}^k}(y))$, thus $a \notin D_{b_{\nu+1}^k}(y)$.

(3) On the event $B_k(\zeta)E_k(\delta_2/2)F_\nu^k$, for k large and ζ small, $D_t(y) \cap \Gamma_2(y; \hat{P}^t, \mathscr{D}_t) = \varnothing$, $\forall t \geq b_1^k$. To prove this, we will show that for $t \in I_\nu^k$, large k and any $y \in S$, there exist $a' \in A^+(y; P, \mathscr{D}_t) = A^+(y; P, \mathscr{D}_{b_\nu^k})$ such that $\mathbf{U}(y, a'; \omega_t) > \mathbf{U}(y, a; \omega_t)$, $\forall a \in D_t(y)$. Indeed, note first that on the event $B_k(\zeta)F_\nu^k$ the following relations hold for all $t \in I_\nu^k$, $y \in S$, $a \in D_t(y)$ and large k,

$$
\begin{aligned}
\mathbf{U}(y, a; \omega_t) &\leq \mathscr{L}\big(y, a; p_y(a), h(y, P; \mathscr{D}_t)\big) + \delta_2/2 \\
&\leq g(P; \mathscr{D}_t) + h(y, P; \mathscr{D}_t) + \delta_2/2 \\
&= g\big(P; \mathscr{D}_{b_\nu^k}\big) + h\big(y, P; \mathscr{D}_{b_\nu^k}\big) + \delta_2/2,
\end{aligned}
$$

where the first inequality is due to Lemma 7(ii), with $\zeta < \xi_2 \leq \zeta_2(\delta_2/2)$, the second is due to the optimality equations of the restricted problem, and the third is due to observation (1). In addition, on $B_k(\zeta)E_k(\delta_2/2)F_\nu^k$, the conditions $(\mathbf{E}_1)$ specifying the $E_k(\delta_2/2)$ and the choice of δ_2, imply $\forall a' \in A^+(y; P, \mathscr{D}_{b_\nu^k})$,

$$
\begin{aligned}
\mathbf{U}(y, a'; \omega_t) &> \mathscr{L}\big(y, a'; p_y(a'), h(y, P; \mathscr{D}_{b_\nu^k})\big) - \delta_2/2 \\
&\geq g\big(P; \mathscr{D}_{b_\nu^k}\big) + h\big(y, P; \mathscr{D}_{b_\nu^k}\big) + \delta_2/2, \qquad \forall t \in I_\nu^k.
\end{aligned}
$$

Thus, the observation follows using the definition of policy $\pi_0 \in C_R$.

When $a \in A^+(y; P, \mathscr{D}_{b_\nu^k})$ it follows from observation (2) that $a \notin D_{b_{\nu+1}^k}(y)$, i.e., $T_{b_{\nu+1}^k}(y, a) < \log^2 b_{\nu+1}^k$. Thus, $\Delta_\nu^2(y, a) \leq \Delta_\nu(y, a) \leq T_{b_{\nu+1}^k}(y, a) < \log^2 b_{\nu+1}^k \leq \log^2 k$. For any $a \notin A^+(y; P, \mathscr{D}_{b_\nu^k})$ we have

$$
\begin{aligned}
\Delta_\nu^2(y, a) &\leq \sum_{t \in I_\nu^k} \mathbf{1}\big(Z_t(y, a) = 1, a \notin D_t(y)\big) \\
&\leq \sum_{t \in I_\nu^k} \mathbf{1}\big(Z_t(y, a) = 1, T_t(y, a) < \log^2 t\big) \\
&\leq \sum_{t \in I_\nu^k} \mathbf{1}\big(Z_t(y, a) = 1, T_t(y, a) < \log^2 b_{\nu+1}^k\big) \leq \log^2 b_{\nu+1}^k < \log^2 k.
\end{aligned}
$$

The first inequality follows from observation (3), the second from the definition of $D_t(y)$ and the fourth from Lemma 3. Hence, $\Delta_\nu(y) = \sum_{a \in A(y)} \Delta_\nu^1(y, a) + \sum_{a \in A^+(y; P, \mathscr{D}_{bk_\nu})} \Delta_\nu^2(y, a) + \sum_{a \notin A^+(y; P, \mathscr{D}_{bk_\nu})} \Delta_\nu^2(y, a) = O(\log^2 k)$ and relation (6.5) follows if we recall that on the event $B_k(\zeta)$, $\Delta_\nu(y) \geq \rho\beta k$ for all y. □

We can now prove the following result:

PROPOSITION 5. $\mathbf{E}_{x_0}^{\pi_0, P}[T_N^{(3)}(\epsilon)] = o(\log N)$, *as* $N \to \infty$, $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$ *and* $\epsilon > 0$.

PROOF. Recall that: $T_N^{(3)}(\rho, \epsilon) = \sum_{k=1}^{N-1} \mathbf{1}(\overline{A_k})$, where $\overline{A_k}$ is the complement of the event: $A_k = \{\omega_k \colon \mathscr{O}(\hat{P}^k, \mathscr{D}_k) \subseteq \mathscr{O}(P), \|h(\hat{P}^k, \mathscr{D}_k) - h^*(P)\| \leq \epsilon\}$. We will prove the proposition by establishing the following claims: $\forall \epsilon > 0$, $\exists \zeta_1 = \zeta(\epsilon) > 0$ such that

$$B_k(\zeta_1) G_k \subseteq A_k, \tag{6.6}$$

and

$$\mathbf{P}_{x_0}^{\pi_0, P}\left[\overline{B_k(\zeta_1) G_k}\right] = o(1/k), \quad \text{as } k \to \infty. \tag{6.7}$$

We first prove (6.6). On the event G_k it is true $\forall \nu = 1, \ldots, m$, that, if $g(P, \mathscr{D}_{b_\nu^k}) < g^*(P)$, then $\exists f \in \mathscr{D}_{b_{\nu+1}^k}$: $g_f(P) > g(P, \mathscr{D}_{b_\nu^k})$. Since the total number of intervals I_ν^k is $m = |\mathscr{A}|$, at the end of the last interval it is true that $\mathscr{D}_k \cap \mathscr{O}(P) \neq \emptyset$, thus,

$$G_k \subseteq \{\omega_k \colon \mathscr{O}(P, \mathscr{D}_k) \subseteq \mathscr{O}(P), g(P, \mathscr{D}_k) = g^*(P), h(P, \mathscr{D}_k) = h^*(P)\}.$$

From Lemma 8(ii) it follows that $\forall \epsilon > 0$, $\exists \xi_1(\epsilon) > 0$ such that, $\forall \zeta < \xi_1(\epsilon)$, for sufficiently large k, $B_k(\zeta) \subseteq \{\omega_k \colon \mathscr{O}(\hat{P}^k, \mathscr{D}_k) \subseteq \mathscr{O}(P, \mathscr{D}_k), \|h(\hat{P}^k, \mathscr{D}_k) - h(P, \mathscr{D}_k)\| < \epsilon\}$. Thus, (6.6) follows.

To prove (6.7), note that, $\mathbf{P}_{x_0}^{\pi_0, P}[\overline{B_k(\zeta) G_k}] \leq \mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta) \overline{G}_k] + \mathbf{P}_{x_0}^{\pi_0, P}[\overline{B}_k(\zeta)]$. From Lemma 9, it follows that $\mathbf{P}_{x_0}^{\pi_0, P}[B_k(\zeta) \overline{G}_k] = o(1/k)$, $\forall \zeta < \xi_2$. In addition, $\mathbf{P}_{x_0}^{\pi_0, P}[\overline{B}_k(\zeta)] \leq \sum_{y \in S} \sum_{\nu=0}^{m} \mathbf{P}_{x_0}^{\pi_0, P}[\Delta_\nu(y) \leq \rho\beta k] + \sum_{y \in S} \mathbf{P}_{x_0}^{\pi_0, P}[\Delta_\nu(y) > \rho\beta k,\ \forall \nu = 0, \ldots, m,\ \|\hat{p}_y^{T_t(y, a)}(a) - p_y(a)\| > \zeta$, for some $t \geq b_1^k$, $a \in D_t(y)]$. From Proposition 2(ii), by conditioning on the state $X_{b_\nu^k}$, it follows that $\mathbf{P}_{x_0}^{\pi_0, P}[\Delta_\nu(y) \leq \rho\beta k] = o(1/k)$, as $k \to \infty$.

Furthermore, on the event $\{\Delta_\nu(y) > \rho\beta k, \forall \nu = 0, \ldots, m\}$ it is true that, for $t \geq b_1^k$, $T_t(y) \geq \rho\beta k \geq \rho\beta t$. Thus, conditioning on the possible values of $T_t(y) = \rho\beta t, \ldots, t$ and $T_t(y, a) = \log^2 j, \ldots, j$ for $a \in D_t(y)$, using Remarks 1, and Lemma 4(i) (with C_a, c_a to note the dependence of the constants on the action a) it follows that, for k sufficiently large,

$$\begin{aligned}
&\mathbf{P}_{x_0}^{\pi_0, P}\Big[\Delta_\nu(y) > \rho\beta k, \forall \nu = 0, \ldots, m, \left\|\hat{p}_y^{T_t(y, a)}(a) - p_y(a)\right\| > \zeta, \\
&\qquad\qquad\qquad\qquad \text{for some } t \geq b_1^k,\ a \in D_t(y)\Big] \\
&\quad \leq \sum_{a \in D_t(y)} \sum_{t = b_1^k}^{k} \sum_{j = \rho\beta t}^{t} \sum_{i = \log^2 j}^{j} \mathbf{P}_{p_x(a)}\left[\left|\hat{p}_{xy}^{l}(a) - p_{xy}(a)\right| > \zeta\right] \\
&\quad \leq \sum_{a \in D_t(y)} \sum_{t = b_1^k}^{k} \sum_{j = \rho\beta t}^{t} \sum_{l = \log^2 j}^{j} C_a e^{-c_a l} \leq \sum_{a \in D_t(y)} C_a / (1 - e^{-c_a}) k^2 e^{-c_a \log^2(\rho\beta k)} \\
&\quad = \sum_{a \in D_t(y)} C_a e^{-c_a \log^2(\rho\beta)} / (1 - e^{-c_a}) k^{2 - 2c_a \log(\rho\beta) - c_a \log k} = o(1/k).
\end{aligned}$$

Therefore, $\mathbf{P}_{x_0}^{\pi_0, P}[\bar{B}_k(\zeta)] = o(1/k)$, as $k \to \infty$. Letting $\zeta_1 < \min(\xi_1, \xi_2)$, the proposition follows. □

Theorem 3 is an immediate consequence of Proposition 3, 4, and 5.

THEOREM 3. (i) $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$ *and* $\forall x \in S$, $a \notin O(x, P)$,

$$\limsup_{N\to\infty} \mathbf{E}_{x_0}^{\pi_0, P} T_N(x, a)/\log N \le 1/\mathbf{K}(x, a; P), \quad \textit{if } (x, a) \in \mathbf{B}(P),$$

$$\limsup_{N\to\infty} \mathbf{E}_{x_0}^{\pi_0, P} T_N(x, a)/\log N = 0, \quad \textit{if } (x, a) \notin \mathbf{B}(P).$$

(ii) $\limsup_{N\to\infty} R_N^{\pi_0}(x_0, P)/\log N \le \mathbf{M}(P)$, $\forall \pi_0 \in C_R$, $\forall P \in \tilde{\mathscr{P}}$.

PROOF. Consider a pair (x, a) such that $a \notin O(x; P)$, and let $\epsilon > 0$. Recall that, $\forall \omega_N$, $T_N(x,a) = \sum_{k=1}^{N-1} Z_k(x,a) \le 1 + T_N^{(1)}(x, a; \epsilon) + T_N^{(2)}(x, a; \epsilon) + T_N^{(3)}(\epsilon)$, $\forall \epsilon > 0$. Part (i) follows from Proposition 3, 4 and 5 when we let $\epsilon \to 0$. Part (ii) is a consequence of (i) and Proposition 1. □

7. Extensions. In this section we consider variations and generalizations of the problem described in §2.

MODEL 1 (GENERAL MODEL). There exist parameter vectors $\theta(x, a)$ in parameter spaces $\Theta(x, a)$ with the following property. If $X_t = x$, $\pi_0(t) = a$, then the one step reward is a random variable Z_t with distribution $F(z; x, a, \theta(x, a))$ and the next state is a random variable with support $S^+(x, a)$ and distribution $\mathbf{P}[X_{t+1} = y | X_t = x, \pi_0(t) = a] = p_{xy}(a; \theta(x, a))$. Parameters $\theta(x, a)$ are generally unknown, while $S^+(x, a)$ and the functional form of the distribution $F(z; x, a, \theta(x, a))$ are known. Let $\underline{\theta} = [\theta(x, a)]_{x \in S, a \in A(x)}$.

MODEL 1.1. The model studied in §§2 to 6 is a special case of the above with $\theta(x, a) = p_x(a)$, $\underline{\theta} = P$, $F(z; x, a, \theta(x, a)) = F(z; x, a)$ and $\mathbf{E}[Z_t | X_t = x, \pi_0(t) = a] = \int z\, dF(z; x, a) = r(x, a)$, known.

MODEL 1.2. Consider model 1 with the one step reward Z_t realized after the transition in period t takes place. If $X_t = x$, $\pi_0(t) = a$, $X_{t+1} = y$, then $Z_t = r_{xy}(a)$, where $r_{xy}(a)$ are known numbers. This is also a special case of the general model, with $\theta(x, a) = p_x(a)$, $\underline{\theta} = P$, $\mathbf{P}[Z_t = r_{xy}(a) | X_t = x, \pi_0(t) = a, X_{t+1} = y] = p_{xy}(a)$, $\forall y \in S^+(x, a)$ and $r(x, a) = \mathbf{E}[Z_t | X_t = x, \pi_0(t) = a] = p_x(a) r_x(a)$, where $r_x(a) = (r_{xy}(a), y \in S)$.

In this case, all arguments used in the proofs of Model 1.1 go through, the only difference being that the right-hand side of the average reward optimality equations is now equal to $\mathscr{L}(x, a; q, h) = q(r_x(a) + h)$.

With this modification, policy π_0 defined in §3 is *UM*, the constants $\mathbf{K}(x, a; P)$ and $\mathbf{M}(P)$ being the same as in §2.1. For all (x, a), $a \notin O(x; P)$, $\Delta\Theta(x, a; P) \ne \varnothing$ if and only if $\max_{y \in S^+(x, a)} (r_{xy}(a) + h^*(y; P)) > \mathscr{L}^*(x; P)$. Also,

$$\mathbf{B}(P) = \left\{(x, a) \colon a \notin O(x; P), \max_{y \in S^+(x, a)} \left(r_{xy}(a) + h^*(y; P)\right) > \mathscr{L}^*(x; P)\right\}.$$

MODEL 1.3. Consider Model 1 where Z_t is independent of X_{t+1} and follows a discrete distribution with known finite support $S^r(x, a)$ and unknown parameter vectors $\theta^r(x, a)$. This model is also a special case of the general model with $\theta(x, a) = [\theta^r(x, a), \theta^p(x, a)]$, where $\theta^p(x, a) = p_x(a)$. In this case, all arguments used in the proofs of Model 1.1 go through, with the following modifications.

The right-hand side of the average reward optimality equations is now expressed as $\mathscr{L}(x, a; \theta, h) = \mathbf{E}_{\theta^r}[Z] + \theta^p h$.

The history includes the past observations of the rewards, i.e., $\omega_k = (X_0, \pi_0(0), Z_0, \dots, X_{k-1}, \pi_0(k-1), Z_{k-1}, X_k)$. The estimates of the transition probabilities $\hat{\theta}^p(x, a)$ are the same as in §2.3. In addition, estimates $\hat{\theta}^r(x, a)$ of the reward parameters can be defined, which satisfy a large deviations property analogous to Lemma 4.

The likelihood ratio takes the form $\Lambda_k(\theta_1, \theta_2) = \prod_{j=1}^{k} \theta_1^r(Z_j)\theta_1^p(Y_j)/(\theta_2^r(Z_j)\theta_2^p(Y_j))$. From the strong law of large numbers, $\log \Lambda_k(\theta_1, \theta_2)/k \to \mathbf{I}^r(\theta_1, \theta_2) = \mathbf{I}(\theta_1^r, \theta_2^r) + \mathbf{I}(\theta_1^p, \theta_2^p)$, under θ_1, where $\mathbf{I}(\cdot, \cdot)$ is the Kullback-Leibler information.

The constants $\mathbf{K}(x, a; \underline{\theta})$, $\mathbf{M}(\underline{\theta})$ and therefore, the indices $\mathbf{U}(x, a; \omega_k)$ have the same form as in §2.1, with $\mathbf{I}$ replaced by $\mathbf{I}^r$. Also, $\mathbf{B}(\underline{\theta}) = \{(x, a)\colon\ a \notin O(x; \underline{\theta}), \max S^r(x, a) + \max_{y \in S^+(x, a)} h^*(y; \underline{\theta}) + > \mathscr{L}(x; \underline{\theta})\}$.

Model 1.4 (Multi-Armed Bandit). Consider the problem of sequential sampling from n statistical populations, Π_a, with the objective to maximize the total expected reward over N periods. UM policies were first developed in Lai and Robbins (1985a) in the context of a version of this problem; for related work see Katehakis and Robbins (1995), Li and Zhang (1992), Yakowitz and Lowe (1991) and for the discounted version of the problem, see Gittins (1979), Katehakis and Derman (1985), Katehakis and Veinott Jr. (1987), Glazebrook (1991), Katehakis and Rothblum (1996), Weiss (1994).

Consider the case in which the rewards associated with population Π_a are i.i.d. random variables following a discrete distribution with known support $S^r(a) = \{r_1(a), \dots, r_s(a)\}$ and unknown probabilities $\theta(a) = [\theta_j(a)]_{j \in S^r(a)}$. This model violates condition (1.7) in Lai and Robbins (1985a). However, it is a special case of Model 1.3, with state space $S = \{1\}$, action set $A(1) = A = \{1, \dots, n\}$ and $\theta_{11}^p(a) = p_{11}(a) = 1$. In addition, $g^*(\underline{\theta}) = \max_a \mathbf{E}_{\theta^r(a)}[Z]$, $h^*(\underline{\theta}) = 0$ and $\mathbf{I}^r(\theta_1, \theta_2) = \mathbf{I}(\theta_1^r, \theta_2^r)$.

Further details and direct proofs are given in Burnetas and Katehakis (1996).

Acknowledgements. We are indebted to H. Robbins for his many suggestions and encouragement. We are also grateful to A. F. Veinott Jr., K. Glazebrook and two anonymous referees for their comments and recommendations regarding this paper. The first draft (June 1994) was presented at the Symposium on Stochastic Processes and Modeling in honor of Cyrus Derman, Center of Applied Probability, Columbia University.

Appendix A

Proof of Lemma 1. (i) It suffices to show that $\forall x \in S$, $a \notin O(x; P)$, and $q \in \Theta(x, a)$, $O(x; Q(x, a; P, q)) = \{a\}$ if and only if $\mathscr{L}(x, a; q, h^*(P)) > \mathscr{L}^*(x; P)$.

Let $Q = Q(x, a; P, q)$, where $q \in \Theta(x, a)$. A key property that will be used in the proof is the following. For fixed (x, a), any policy $f \in \mathscr{A}$ such that $f(x) \neq a$, gives rise to the same transition matrix under both P and Q, thus, $g_f(Q) = g_f(P)$ and $h_f(Q) = h_f(P)$.

First, suppose $\mathscr{L}(x, a; q, h^*(P)) > \mathscr{L}^*(x; P)$. Consider any $f \in \mathscr{O}(P)$. Since $a \notin O(x; P)$, $f(x) \neq a$, thus, $g_f(Q) = g^*(P)$, $h_f(Q) = h^*(P)$, and $\mathscr{L}(x, a; q, h_f(Q)) = \mathscr{L}(x, a; q, h^*(P)) > \mathscr{L}^*(x; P) = \mathscr{L}(x, f(x); p_x(f(x)), h_f(Q))$. Therefore, the average reward optimality equations are not satisfied for the pair (x, a) under Q, thus $g_f(Q) < g^*(Q)$. For any $f' \in \mathscr{A}$ such that $f'(x) \neq a$, $g_{f'}(Q) = g_{f'}(P) g_f(P) < g^*(Q)$, therefore, $O(x; Q(x, a; P, q)) = \{a\}$.

We next prove that if $\mathscr{L}(x, a; q, h^*(P)) \leq \mathscr{L}^*(x; P)$, then $\mathscr{O}(P) \subseteq \mathscr{O}(Q)$, therefore a is not the unique optimal action under Q, thus, $q \notin \Delta\Theta(x, a; P)$.

For all $f \in \mathscr{O}(P)$, $\mathscr{L}(x, f(x); p_x(f(x)), h_f(Q)) = \mathscr{L}^*(x; P)$ thus, $\mathscr{L}(x, a; q, h_f(Q)) = \mathscr{L}(x, a; q, h^*(P)) \leq \mathscr{L}^*(x; P)$. In addition, $\mathscr{L}(x', a'; Q_{x'}(a'), h_f(Q)) = \mathscr{L}(x', a'; p_{x'}(a'), h^*(P)) \leq \mathscr{L}^*(x'; P)$, if not both $x' = x$, $a' = a$. Therefore, f satisfies the optimality equations under Q, i.e., $f \in \mathscr{O}(Q)$. This completes the proof of (i).

(ii) From (i), it suffices to show that $\exists q \in \Theta(x,a)$: $\mathscr{L}(x,a;q,h^*(P)) > \mathscr{L}^*(x;P)$, if and only if $d > 0$, where $d = r(x,a) + \max_{y \in S^+(x,a)} h^*(y;P) - \mathscr{L}^*(x;P)$. If $h^*(y;P) = h$, $\forall y \in S^+(x,a)$, then $d = 0$ and, in addition, $\mathscr{L}(x,a;q,h^*(P)) = r(x,a) + h = \mathscr{L}(x,a;p_x(a),h^*(P)) < \mathscr{L}^*(x;P)$, $\forall q \in \Theta(x,a)$, therefore, $\Delta\Theta(x,a;P) = \varnothing$ and the claim is true.

We next consider the case where $h^*(y;P) \neq h^*(y';P)$ for at least two $y, y' \in S^+(x,a)$. Let $\overline{h^*} = \max_{y \in S^+(x,a)} h^*(y;P)$, $h^* = \min_{y \in S^+(x,a)} h^*(y;P)$.

If $d > 0$, define $q \in \overline{\Theta(}x,a)$ as follows.

$$q(y) = \begin{cases} \delta, & \text{if } y \in S^+(x,a),\ y \neq \bar{y}, \\ 1 - z\delta, & \text{if } y = \bar{y}, \\ 0, & \text{if } y \notin S^+(x,a), \end{cases}$$

for some $0 < \delta < \min\{1/z, d/zv\}$, where $v = \overline{h^*} - h^*$, $z = |S^+(x,a)| - 1$, and $\bar{y} \in \arg\max_{y \in S^+(x,a)} h^*(y;P)$. Then $\mathscr{L}(x,a;q,h^*(P) = r(x,a) + (\overline{1 - z\delta)h^*} + \Sigma_{y \neq \bar{y}} \delta h^*(y;P) = \mathscr{L}^*(x;P)$ $+ d - \delta\Sigma_{y \neq \bar{y}}(\overline{h^*} - h^*(y;P)) \geq \mathscr{L}^*(x;P) + d - \delta zv > \mathscr{L}^*(x;P)$, by the choice of δ. Therefore, $q \in \Delta\Theta(x,a;P)$.

Next, consider $d \leq 0$. Then, for all $q \in \Theta(x,a)$, it is true that $\mathscr{L}(x,a;q,h^*(P)) \leq r(x,a) + \overline{h^*} \leq \mathscr{L}^*(x;P)$, thus, $\Delta\Theta(x,a;P) = \varnothing$.

Part (iii) follows from (ii) and the definition of the set $\mathbf{B}(P)$. □

PROOF OF LEMMA 2. (i)

$$\begin{aligned} L(P,Q;\omega_N) &= \mathbf{P}_{x_0}^{\pi,P}[X_0, A_0, X_1, \ldots, X_N] / \mathbf{P}_{x_0}^{\pi,Q}[X_0, A_0, X_1, \ldots, X_N] \\ &= \prod_{k=0}^{N-1} \mathbf{P}_{x_0}^{\pi,P}[A_k | X_0, A_0, \ldots, X_{k-1}] P[X_{k+1} | X_k, A_k] / \\ &\quad \left(\mathbf{P}_{x_0}^{\pi,Q}[A_k | X_0, A_0, \ldots, X_{k-1}] Q[X_{k+1} | X_k, A_k] \right) \\ &= \prod_{k=0}^{N-1} P[X_{k+1} | X_k, A_k] / Q[X_{k+1} | X_k, A_k] = \prod_{j=1}^{T_N(x\ a)} p_{xY_j(x,a)}(a) / q_{Y_j(x,a)}. \end{aligned}$$

The cancellation of terms is due to the observation that policy π selects actions using only the history; therefore, given the history of states and actions up to time k, action A_k is taken with the same probability under both P and Q. Also, by the construction of $Q(x,a;P,q)$ we have that $Q[X_{k+1}|X_k, A_k] = P[X_{k+1}|X_k, A_k]$, if not both $X_k = x$ and $A_k = a$, thus, all terms except those corresponding to $(X_k, A_k) = (x,a)$ are canceled as well.

(ii) Since $Y_j(x,a)$, $j = 1, 2, \ldots$ are independent and identically distributed, it follows from the strong law of large numbers that $\log \Lambda_k(p_x(a), q)/k = 1/k\Sigma_{j=1}^k \log[p_{xY_j(x,a)}(a)/q_{Y_j(x,a)}] \to \mathbf{I}(p_x(a), q)$ a.s. $(\mathbf{P}_{p_x(a)})$, as $k \to \infty$, therefore, $\max_{k \leq \lfloor b_N \rfloor} \log \Lambda_k / b_N \to \mathbf{I}(p_x(a), q)$, a.s. $(\mathbf{P}_{p_x(a)})$ for any increasing sequence b_N with $b_N \to \infty$, and the assertion follows. □

PROOF OF LEMMA 5. (i) For $\mu \geq 0$ define $\tilde{q}(\mu) \in \Theta(x,a)$ as follows:

$$\tilde{q}_y(\mu) = \begin{cases} p_{xy}(a) e^{-\mu h(y)} / b(\mu), & y \in S^+(x,a), \\ 0, & \text{otherwise}, \end{cases}$$

where $b(\mu) = \Sigma_{y \in S^+(x,a)} p_{xy}(a) e^{-\mu h(y)}$. For $\epsilon, \mu > 0$ let $H_{\mathscr{L}}(\epsilon)$, $H_\lambda(\mu)$ be the hyperplanes

$$H_{\mathscr{L}}(\epsilon) = \{q \in \Theta(x,a) \colon \mathscr{L}(x,a;q,h) = \mathscr{L}(x,a;p_x(a),h) - \epsilon\},$$

$$H_\lambda(\mu) = \{q \in \Theta(x,a) \colon \lambda(q; p_x(a), \tilde{q}(\mu)) = -c(\mu)\},$$

where $c(\mu) = -\lambda(\tilde{q}(\mu); p_x(a), \tilde{q}(\mu)) = \mathbf{I}(\tilde{q}(\mu), p_x(a))$. Let $H_{\mathscr{L}}^-(\epsilon)$, $H_{\mathscr{L}}^+(\epsilon)$, $H_\lambda^-(\mu)$, $H_\lambda^+(\mu)$ denote the corresponding half spaces, i.e., $H_{\mathscr{L}}^-(\epsilon) = \{q \colon \mathscr{L}(x,a;q,h) < \mathscr{L}(x,a;p_x(a),h) - \epsilon\}$, etc. We will show that for all $\epsilon > 0$ sufficiently small there exists $\mu = \mu(\epsilon) > 0$ such that $H_{\mathscr{L}}^-(\epsilon) = H_\lambda^-(\mu(\epsilon))$. Then the lemma will follow with $q^0(\epsilon) = \tilde{q}(\mu(\epsilon))$.

Fix $\epsilon > 0$. For all $\mu > 0$, $H_{\mathscr{L}}(\epsilon)$ and $H_\lambda(\mu)$ are parallel hyperplanes, since, by construction of $\tilde{q}(\mu)$, $\forall q \in \Theta(x,a)$,

$$\begin{aligned} \lambda(q; p_x(a), \tilde{q}(\mu)) &= \sum_{y \in S^+(x,a)} q(y) \log \frac{p_{xy}(a)}{\tilde{q}_y(\mu)} \\ &= \mu q h + \log b(\mu) \\ &= \mu \mathscr{L}(x,a;q,h) - \mu r(x,a) + \log b(\mu). \end{aligned}$$

In addition, $\tilde{q}(\mu) \in H_\lambda(\mu)$, because $\lambda(\tilde{q}(\mu); p_x(a), \tilde{q}(\mu)) = -\mathbf{I}(\tilde{q}(\mu), p_x(a)) = -c(\mu)$.

Hence, for $H_{\mathscr{L}}(\epsilon) = H_{\lambda}(\mu(\epsilon))$, it suffices to choose $\mu(\epsilon)$ such that $\tilde{q}(\mu(\epsilon)) \in H_{\mathscr{L}}(\epsilon)$, i.e., $\mathscr{L}(x, a; \tilde{q}(\mu(\epsilon)), h) = \mathscr{L}(x, a; p_x(a), h) - \epsilon$.

To show that such a $\mu(\epsilon)$ exists, note that for $\mu = 0$ it is true that $b(0) = 1$ and $\tilde{q}(0) = p_x(a)$, thus, $\mathscr{L}(x,a; \tilde{q}(0), h) = \mathscr{L}(x, a; p_x(a), h)$. Also

$$\lim_{\mu \to \infty} \tilde{q}(\mu) = \tilde{q}(\infty) \equiv \begin{cases} p_{xy}(a)/\Sigma_{y\ \ h(y)=\underline{h}}\, p_{xy}(a), & \text{if } h(y) = \underline{h}, \\ 0, & \text{otherwise}, \end{cases}$$

where $\underline{h} = \min_y h(y)$. Thus, $\lim_{\mu \to \infty} \mathscr{L}(x, a; \tilde{q}(\mu), h) = \mathscr{L}_{\infty} \equiv r(x, a) + \underline{h} < \mathscr{L}(x, a; p_x(a), h)$. Therefore, for any $\epsilon < \mathscr{L}(x, a; p_x(a), h) - \mathscr{L}_{\infty}$, there exists $\mu(\epsilon) > 0$ such that $\mathscr{L}(x, a; \tilde{q}(\mu(\epsilon)), h) = \mathscr{L}(x, a; p_x(a), h) - \epsilon$, because $\mathscr{L}(x, a; \tilde{q}(\mu), h)$ is continuous in μ.

To prove that $H_{\mathscr{L}}^{-}(\epsilon) = H_{\lambda}^{-}(\mu(\epsilon))$, it suffices to show that $p_x(a) \in H_{\mathscr{L}}^{+}(\epsilon)$ and $p_x(a) \in H_{\mathscr{L}}^{+}(\mu(\epsilon))$. The first is immediate, while for the second we note that

$$\lambda(p_x(a); p_x(a), \tilde{q}(\mu(\epsilon)))$$

$$= \mathbf{I}(p_x(a), \tilde{q}(\mu(\epsilon))) > 0 > -\mathbf{I}(\tilde{q}(\mu(\epsilon)), p_x(a)) = -c(\mu(\epsilon)).$$

(ii) The proof is an extension of the one for Lemma 2 in Lai and Robbins (1985a). Fix a pair (x, a) and abbreviate $Y_j(x, a)$ with Y_j. For $q_1, q_2 \in \Theta(x, a)$, let $\Lambda_t(q_1, q_2) = \Pi_{j=1}^{t} q_{1, Y_j(x,a)}/q_{2, Y_j(x,a)}$; see also Lemma 2. Observe that

$$\log \Lambda_t(q_1, q_2) = \sum_{j=1}^{t} \log\left[q_{1, Y_j(x,a)}/q_{2, Y_j(x,a)}\right]$$

$$= \sum_{j=1}^{t} \sum_{y \in S} 1\left(Y_j(x, a) = y\right) \log\left[q_{1y}/q_{2y}\right] = t\lambda(f_t; q_1, q_2).$$

In addition, from the information inequality $-\mathbf{I}(p, q_1) \leq 0$ it follows that $\sup_{q_1} \lambda(p; q_1, q_2) = \lambda(p; p, q_2) = \mathbf{I}(p, q_2)$, therefore, $\exp\{t\mathbf{I}(f_t, q_2)\} = \exp\{t\lambda(f_t; f_t, q_2)\} = \sup_{q_1} \Lambda_t(q_1, q_2)$.

Hence, for all k, t,

$$\mathbf{P}_p\left[\lambda(f_t; p, q) < -c + b_1/t, \mathbf{I}(f_t, q) > (\log k + b_2)/t\right]$$

$$= \mathbf{P}_p\left[\Lambda_t(p, q) \leq e^{b_1} e^{-ct}, \sup_{p_1} \Lambda_t(p_1, q) > k e^{b_2}\right].$$

After a change of measure transformation between p and q we obtain

$$\mathbf{P}_p\left[\Lambda_t(p, q) \leq e^{b_1} e^{-ct}, \sup_{p_1} \Lambda_t(p_1, q) > e^{b_2} k\right] \tag{A.1}$$

$$\leq e^{b_1} e^{-ct} \mathbf{P}_q\left[\sup_{p_1} \Lambda_t(p_1, q) > e^{b_2} k\right].$$

Since $\Theta(x, a)R$ is a subset of a compact set, for any $\delta > 0$ there exists a finite collection of vectors $q^{(i)} \in \Theta(x, a)$, and neighborhoods $\mathscr{N}_i(\delta)$, $i = 1, \ldots, m$, such that $\cup_i \mathscr{N}_i(\delta) \supseteq \Theta(x, a)$, and $\mathscr{N}_i(\delta) = \{p_1 \in \Theta(x, a): \|p_1 - q^{(i)}\| < \delta\}$.

For all $i = 1, \ldots, m$ and $y \in S^{+}(x, a)$ and $\forall q \in \Theta(x, \alpha)$, it is true that $\sup_{p_1 \in \mathscr{N}_i(\delta)} p_{1_y}/q_y \leq (q_y^{(i)} + \delta)/q_y$, thus,

$$\mathbf{E}_q\left[\sup_{p_1 \in \mathscr{N}_i(\delta)} \frac{p_{1, Y_1}}{q_{Y_1}}\right] \leq \mathbf{E}_q\left[\frac{q_{Y_1}^{(i)} + \delta}{q_{Y_1}}\right] = 1 + |S^{+}(x, a)|\delta.$$

Therefore, for any $\epsilon > 0$, selecting $\delta < \epsilon/|S^- + (x,a)|$, we obtain $\mathbf{E}_q[\sup_{p_1 \in \mathscr{N}_t(\delta)} p_{1,Y_1}/q_{Y_1}] \leq 1 + \epsilon$, $\forall i = 1, \ldots, m$, thus,

$$\text{(A.2)}\quad \mathbf{P}_q\left[\sup_{p_1 \in \mathscr{N}_t(\delta)} \Lambda_t(p_1, q) > e^{b_2}k\right] \leq e^{-b_2}k^{-1}\mathbf{E}_q\left[\sup_{p_1 \in \mathscr{N}_t(\delta)} \Lambda_t(p_1, q)\right]$$

$$\leq e^{-b_2}k^{-1}\left(\mathbf{E}_q\left[\sup_{p_1 \in \mathscr{N}_t(\delta)} \frac{p_{1,Y_1}}{q_{Y_1}}\right]\right)^t \leq e^{-b_3}k^{-1}(1+\epsilon)^t,$$

where the first inequality follows from the Markov inequality, the second from the fact that $\sup_{p_1 \in \mathscr{N}_t(\delta)} \Lambda_t(q, r) \leq \prod_{j=1}^t \sup_{p_1 \in \mathscr{N}_t(\delta)} p_{1,Y_j}/q_{Y_j}$, and the third from the fact that random variables Y_j, $j \geq 1$, are independent and identically distributed.

Combining (A.1) and (A.2),

$$\mathbf{P}_p\left[\Lambda_t(p,q) \leq e^{b_1}e^{-ct}, \sup_{p_1} \Lambda_t(q_1, q) > e^{b_2}k\right]$$

$$\leq e^{b_1}e^{-ct}\sum_{i=1}^{m} \mathbf{P}_q\left[\sup_{p_1 \in \mathscr{N}_i(\delta)} \Lambda_t(p_1, q) > e^{b_2}k\right]$$

$$\leq me^{b_1 - b_2}k^{-1}e^{-ct}(1+\epsilon)^t.$$

Selecting ϵ so that $e^{-c}(1+\epsilon) < 1$, we obtain

$$\sum_{t = \lfloor(d \log k + \gamma)\rfloor} \mathbf{P}_p\left[\Lambda_t(p,q) \leq e^{b_1}e^{-ct}, \sup_{p_1 \in \Theta(x,a)} \Lambda_t(p_1, q) > b_2 k\right]$$

$$\leq m_1 k^{-1} \sum_{t = \lfloor(d \log k + \gamma)\rfloor} (e^{-c}(1+\epsilon))^t \leq m_1 k^{-1-d(c - \log(1+\epsilon))},$$

where $m_1 = me^{b_1 - b_2 - c\gamma}(1+\epsilon)^\gamma/(1 - e^{-c}(1+\epsilon))$. Since $d > 0$ and $c - \log(1+\epsilon) > 0$ (by selection of ϵ), it follows that $-1 - d(c - \log(1+\epsilon)) < -1$ and the proof is complete.

(iii) Recall the definitions $\lambda(p; q_1, q_2) = \mathbf{I}(p, q_2) - \mathbf{I}(p, q_1) = \sum_{y \in S^+(x,a)} p_y \log[q_{1_y}/q_{2_y}]$, and $\hat{p}_x^t(a) = (1 - w_t)/z + w_t \cdot f_t$, where $z = |S^+(x,a)|$ and $w_t = t/(t+z)$; note that $t(1 - w_t)/z = w_t$. From these it follows that $t\lambda(\hat{p}_x^t(a); q_1, q_2) = w_t[b(q_1, q_2) + t\lambda(f_t; q_1, q_2)]$, where $b(q_1, q_2) = \sum_y \log[q_{1_y}/q_{2_y}]$.

Hence, on the event $\{\lambda(\hat{p}_x^t(a); p, q) < -c\}$ it is true that $w_t(b(p,q) + t\lambda(f_t; p, q)) < -ct$, therefore, since $0 < w_t < 1$, $b(p,q) + t\lambda(f_t; p, q) < -ct/w_t < -ct$.

In addition, since $b(q_1, q_2) \leq b_0(q_2) := -\sum_y \log q_{2_y}$ and $\lambda(f_t; q_1, q_2) \leq \mathbf{I}(f_t. q_2)$, $\forall q_1$, it follows that $t\mathbf{I}(\hat{p}_x^t(a), q_2) \leq w_t[b_0(q_2) + t\mathbf{I}(f_t, q_2)]$.

Thus, on the event $\{\mathbf{I}(\hat{p}_x^t(a), q) > (\log k + \beta)/t\}$ it is true that $w_t(b_0(q) + t\mathbf{I}(f_t, q)) > \log k + \beta$, therefore, since $0 < w_t < 1$, $b_0(q) + t\mathbf{I}(f_t, q) > (\log k + \beta)/w_t > \log k + \beta$.

Combining the above,

$$\sum_{t = \lfloor(d \log k + \gamma)\rfloor}^{k} \mathbf{P}_p\left[\lambda\left(\hat{p}_x^t(a); p, q\right) < -c, \mathbf{I}\left(\hat{p}_x^t(a), q\right) > (\log k + \beta)/t\right]$$

$$\leq \sum_{t = \lfloor(d \log k + \gamma)\rfloor}^{k} \mathbf{P}_p\left[\lambda(f; p, q) < -c - b(p,q)/t, \mathbf{I}(f, q)\right.$$

$$\left. > (\log k + \beta - b_0(q))/t\right],$$

and the result follows from (ii), with $b_1 = -b(p_x(a), q)$ and $b_2 = \beta - b_0(q)$. □

Proof of Lemma 7. (i) For $t \geq b_1^k$ and $f \in \mathscr{D}_t$, write the policy evaluation equations for f under P and $\hat{P}^t$ as $\Pi_f(P)\underline{h}_f(P) = r(f)$, and $\Pi_f(\hat{P}^t)\underline{h}_f(\hat{P}^t) = r(f)$, respectively, where we have set $h_f(1; P) = h_f(1; \hat{P}^t) = 0$, and $\Pi_f(P), \Pi_f(\hat{P}^t)$ denote the corresponding invertible matrices, resulting from this normalization. These systems have unique solutions: $\underline{h}_f(P) = [g_f(P), h_f(2; P), \ldots, h_f(s; P)]$, $\underline{h}_f(\hat{P}^t) = [g_f(\hat{P}^t), h_f(2; \hat{P}^t), \ldots, h_f(s; \hat{P}^t)]$.

Let $\Delta = \Pi_f(\hat{P}^t) - \Pi_f(P)$ and $\delta^h = \underline{h}_f(\hat{P}^t) - \underline{h}_f(P)$. On the event $B_k(\zeta)$ it is true that $\|\Delta\| \le s\zeta$, where s is the dimension of the matrices (number of states). Thus, from Theorem 5.8 of Noble and Daniel (1977), with $A = \Pi_f(P)$ and $R = \Delta$, it follows that, for $\zeta < 1/(s\|(\Pi_f(P))^{-1}\|)$, $\|\delta^h\| \le \|(\Pi_f(P))^{-1}\| \, \|\underline{h}_f(P)\| s\zeta/(1 - \Pi_f(P))^{-1}\|s\zeta)$. Therefore, for any $\epsilon > 0$, selecting $\zeta \le \zeta_1(\epsilon) := \min_{f \in \mathscr{A}} \zeta_1(\epsilon, f)$, where $\zeta_1(\epsilon, f) = \epsilon/[s\|(\Pi_f(P))^{-1}\|(\|\underline{h}_f(P)\| + \epsilon)]$, it is true that $\|\delta^h\| < \epsilon$, $\forall f \in \mathscr{D}_t$.

(ii) For the proofs of (ii) and (iii), recall that under $\pi_0 \in C_R$, $\mathbf{U}(y, a; \omega_t) = \mathbf{u}_{y,a}(\hat{p}_y^{T_t(y,a)}(a), h(\hat{P}^t, \mathscr{D}_t), \log t/T_t(y, a))$.

Fix $\delta > 0$, $y \in S$, $a \in D_t(y)$. From (i) it follows that, for $\zeta \le \zeta_1(\delta/3)$, on the event $B_k(\zeta)$ it is true that $\|h(\hat{P}^t, \mathscr{D}_t) - h(P, \mathscr{D}_t)\| < \delta/3$. Thus, $|\mathscr{L}(y, a; q, h(\hat{P}^t, \mathscr{D}_t)) - \mathscr{L}(y, a; q, h(P, \mathscr{D}_t))| = |qh(\hat{P}^t, \mathscr{D}_t) - qh(P, \mathscr{D}_t)| < \delta/3$, $\forall q$, and

$$\text{(A.3)} \qquad \mathbf{U}(y, a; \omega_t) \le \mathbf{u}_{y,a}\big(\hat{p}_y^{T_t(y,a)}(a), h(P, \mathscr{D}_t); \log t/T_t(y, a)\big) + \delta/3.$$

Since $\mathbf{u}_{y,a}(p, h, \gamma)$ is continuous in p, there exists $\zeta_1'(\delta/3) > 0$, such that, for $\zeta \le \zeta_1'(\delta/3)$, on the event $B_k(\zeta)$ it is true that

$$\text{(A.4)} \qquad \mathbf{u}_{y,a}\big(\hat{p}_y^{T_t(y,a)}(a), h(P, \mathscr{D}_t), \log t/T_t(y, a)\big)$$
$$\le \mathbf{u}_{y,a}\big(p_y(a), h(P, \mathscr{D}_t), \log t/T_t(y, a)\big) + \delta/3.$$

In addition, since $\mathbf{u}_{y,a}(p, h, \gamma)$ is continuous in γ and $\mathbf{u}_{y,a}(p, h, 0) = r(y, a) + ph = \mathscr{L}(y, a; p, h)$, it follows that there exists $\gamma_1 > 0$ such that $\mathbf{u}_{y,a}(p_y(a), h_f(P), \gamma_1) \le \mathscr{L}(y, a; p_y(a), h_f(P)) + \delta/3$ for all $f \in \mathscr{A}$, therefore,

$$\text{(A.5)} \qquad \mathbf{u}_{y,a}\big(p_y(a), h(P, \mathscr{D}_t), \gamma_1\big) \le \mathscr{L}\big(y, a; p_y(a), h(P, \mathscr{D}_t)\big) + \delta/3.$$

Select $\zeta < \zeta_2(\delta) := \min(\zeta_1(\delta/3), \zeta_1'(\delta/3))$ and take k sufficiently large, so that $\log k/\log^2(\rho\beta k) < \gamma_1$. Then, for all $t \ge b_1^k$ and $a \in D_t(y)$ it is true that $\log t/T_t(y, a) \le \gamma_1$, and the claim follows from the combination of (A.3), (A.4) and (A.5).

(iii) Fix $\nu \in \{1, \dots, m\}$, $\delta > 0$, $y \in S$, $a \in A(y)$. Let $C_k = C_\nu^k(\delta)$. From (i) it follows that, for $\zeta \le \zeta_1(\delta/2)$, on the event $B_k(\zeta)$ it is true that $\|h(\hat{P}^t, \mathscr{D}_t) - h(P, \mathscr{D}_t)\| < \delta/2$. Thus, $|\mathscr{L}(y, a; q, h(\hat{P}^t, \mathscr{D}_t)) - \mathscr{L}(y, a; q, h(P, \mathscr{D}_t))| < \delta/2$, $\forall q \in \Theta(x, a)$. hence, $\mathbf{U}(y, a; \omega_t) = \mathbf{u}_{y,a}(\hat{p}_y^{T_t(y,a)}(a), h(\hat{P}^t, \mathscr{D}_t); \log t/T_t(y, a)) \ge \mathbf{u}_{y,a}(\hat{p}_\nu^{T_t(y,a)}(a), h(P, \mathscr{D}_t); \log t/T_t(y, a)) - \delta/2$, $\forall t \in I_\nu^k$.

In addition,

$$\mathbf{u}_{y,a}\big(\hat{p}_y^{T_t(y,a)}(a), h(P, \mathscr{D}_t); \log t/T_t(y, a)\big)$$
$$\ge \mathbf{u}_{y,a}\big(\hat{p}_y^{T_t(y,a)}(a); h\big(P, \mathscr{D}_{b_\nu^k}\big), \log b_\nu^k/T_t(y, a)\big)$$
$$\ge \mathbf{u}_{y.a}\big(\hat{p}_y^{T_t(y,a)}(a); h\big(P, \mathscr{D}_{b_\nu^k}\big), \big(\log b_{\nu-1}^k + \log \beta\big), /T_t(y, a)\big),$$

since $t > b_\nu^k$, on the event C_k it is true that $h(P, \mathscr{D}_t) = h(P, \mathscr{D}_{b_\nu^k})$, and by construction of subintervals I_ν^k, we have that $\log b_\nu^k \ge \log b_{\nu+1}^k + \log \beta$. Therefore, to complete the proof of (iii), it suffices to show that $\mathbf{P}_{x_0}^{\pi_0, P}[C_k'] = o(1/k)$, where

$$C_k' = \big\{\omega_k : \exists t \in I_\nu^k, \mathbf{u}_{y,a}\big(\hat{p}_y^{T_t(y,a)}(a); h\big(P, \mathscr{D}_{b_\nu^k}\big), \big(\log b_{\nu+1}^k + \log \beta\big)/T_t(y, a)\big)$$
$$< \mathscr{L}\big(y, a; p_y(a), h\big(P; \mathscr{D}_{b_\nu^k}\big)\big) - \delta/2\big\}.$$

Considering the possible values for $T_t(y, a) = 0. \dots, b_{\nu+1}^k$, and using the definition of $\mathbf{u}_{y,a}(\cdot, \cdot, \cdot)$, it follows that $C_k' \subseteq \bigcup_{j=0}^{b_{\nu+1}^k} C_{k,j}'$, where

$$C_{k,j}' = \big\{\mathscr{L}\big(y, a; q, h\big(P, \mathscr{D}_{b_\nu^k}\big)\big) < \mathscr{L}\big(y, a; p_y(a), h\big(P; \mathscr{D}_{b_\nu^k}\big)\big)$$
$$- \delta/2, \forall q \in F_{b_{\nu+1}^k, j}\big(\hat{p}_y^j(a); \log \beta\big)\big\}.$$

(Recall that $F_{b^k_{\nu+1},j}(\hat{p}^j_y(a);\log\beta) = \{q\colon \mathbf{I}(\hat{p}^j_y(a), q) \le \log(b^k_{\nu+1} + \log\beta)/j\}$.) Thus, considering all possible policies $f \in \mathscr{D}_{b^k_\nu} \subseteq \mathscr{A}$ corresponding to the solution of the restricted problem $(P; \mathscr{D}_{b^k_\nu})$ we obtain

$$\begin{aligned}\mathbf{P}^{\pi_0,P}_{x_0}[C'_k] &\le \sum_{j=0}^{b^k_{\nu+1}} \mathbf{P}^{\pi_0,P}_{x_0}[C'_{k,j}] \\ &\le \sum_{f\in\mathscr{A}} \sum_{j=0}^{b^k_{\nu+1}} \mathbf{P}_{p_y(a)}\Big\{\mathscr{L}\big(y,a;q,h_f(P)\big) < \mathscr{L}\big(y,a;p_y(a),h_f(P)\big) \\ &\qquad\qquad -\delta/2, \forall q \in F_{b^k_{\nu+1},j}\big(\hat{p}^j_y(a);\log\beta\big)\Big\} \\ &= o\big(1/b^k_{\nu+1}\big) = o(1/k), \quad (\text{as } k \to \infty).\end{aligned}$$

The second inequality follows using Remark 1. The first equality follows from Lemma 6 with t replaced by j, k replaced by $b^k_{\nu+1}$ and β replaced by $\log\beta$ and the fact that $m = |\mathscr{A}|$ is finite. The last equality follows from the observation that $b^k_{\nu+1} \sim (\nu+1)/(m+1)k$, as $k \to \infty$, $\forall \nu = 1, \dots, m$. Thus, (iii) follows with $\zeta_3(\delta) := \zeta_1(\delta/2)$. □

Appendix B. This appendix includes graphical illustrations for the computation of the constant $\mathbf{K}(x, a; P)$ and the index $\mathbf{U}(x, a; \omega_k)$. We fix a pair (x, a) with $x \in S = \{1, 2, \dots, s\}$, $a \in A(x)$ and consider the cases in which $|S^+(x,a)| = 2, 3$. Then, for any $P \in \tilde{\mathscr{P}}$, the parameter space $\Theta(x, a)$, for $p_x(a)$, is respectively equal to $\{(q_1, q_2)\colon q_1, q_2 > 0,\ q_1 + q_2 < 1\}$, and $\{(q_1, q_2, q_3)\colon q_1, q_2 > 0,\ q_1 + q_2 + q_3 = 1\}$. However, for graphical simplicity, we will consider the projections $\Theta'(x, a)$ of the parameter spaces that are respectively equal to $\{q_1\colon 0 < q_1 < 1\}$ and $\{q_1, q_2 > 0\colon q_1 + q_2 < 1\}$.

EXAMPLE 1. Take $S^+(x, a) = \{1, 2\}$ with $r(x, a) = 3$. In this example possible transitions from state x under action a are into states 1 and 2 only and $p_x(a) = (p_1, 1 - p_1)$ with $p_1 \in \Theta'(x, a) = \{q_1\colon 0 < q_1 < 1\}$.

(i) For the computation of $\mathbf{K}(x, a; P)$ (see Figure 1) consider a $P \in \tilde{\mathscr{P}}$ with $h^*_1 = h^*(1; P) = 0$, $h^*_2 = h^*(2; P) = -4$, $\mathscr{L}^*(x; P) = \mathscr{L}^* = 2$, $p_1 = 1/3$. We have:

(a) $\Delta\Theta(x, a) = \{q = (q_1, 1 - q_1)\colon\ r(x, a) + h^*_1 q_1 + h^*_2(1 - q_1) > \mathscr{L}^*\}$; its projection is $\Delta\Theta'(x, a) = \{q_1\colon 3/4 < q_1 < 1\} \ne \varnothing$,

(b) $(x, a) \in \mathbf{B}(P)$ (since $\Delta\Theta(x, a) \ne \varnothing$),

(c) $\mathbf{I}(p_1, q_1) = 1/3\log(1/3q_1) + 2/3\log(2/(3/(1 - q_1)))$; it attains its minimum value (equal to 0) at $q_1 = p_1$,

(d) $\mathbf{K}(x, a; P) = \inf_{(q_1,q_2)\in\Delta\Theta(x,a:P)}\{\mathbf{I}((p_1, 1 - p_1), (q_1, q_2))\} = \inf_{q_1\in\Delta\Theta'(x,a\ P)}\{\mathbf{I}(p_1, q_1)\} = \mathbf{I}(p_1, q^0_1) = 0.384$.

(ii) In the computation of $\mathbf{U}(x, a; \omega_k)$ (see Figure 2) we postulate an observed history ω_k with $k = 150$ and $T_k = T_k(x, a) = 40$, $\hat{p}_x(a) = (\hat{p}_1, 1 - \hat{p}_1)$, $\hat{p}_1 = 1/2$, $\hat{h}_1 = \hat{h}^k(1) = 0$, $\hat{h}_2 = \hat{h}^k(2) = -2$. Then we have:

(a) with $\beta = \log k/T_k = 0.1253$, $F(k, T_k) = \{q \in \Theta(x, a)\colon \mathbf{I}(\hat{p}_x(a), q) \le \beta\}$, and its one-dimensional projection is $F'(k, T_k) = \{q_1 \in \Theta'(x, a)\colon 1/2\log(1/2q_1) + 1/2\log(1/(2(1 - q_1))) \le \beta\} = [0.265, 0.735] \ne \varnothing$,

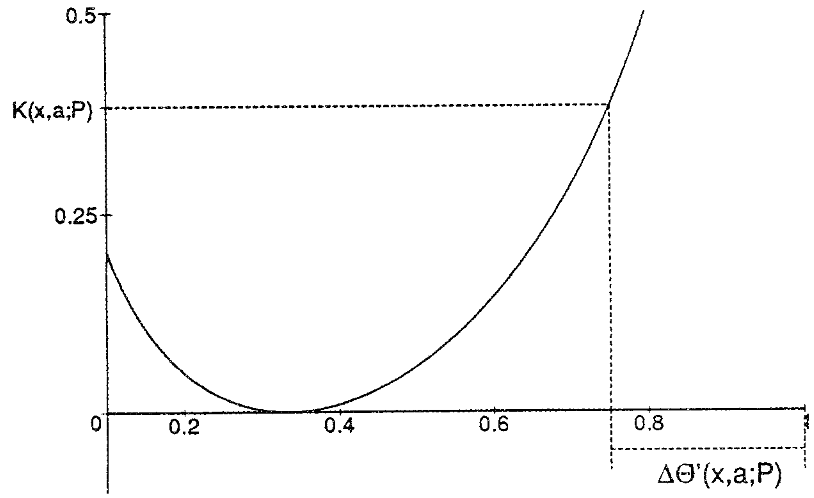

FIGURE 1. Computation of $\mathbf{K}(x, a; P)$ in Example 1.

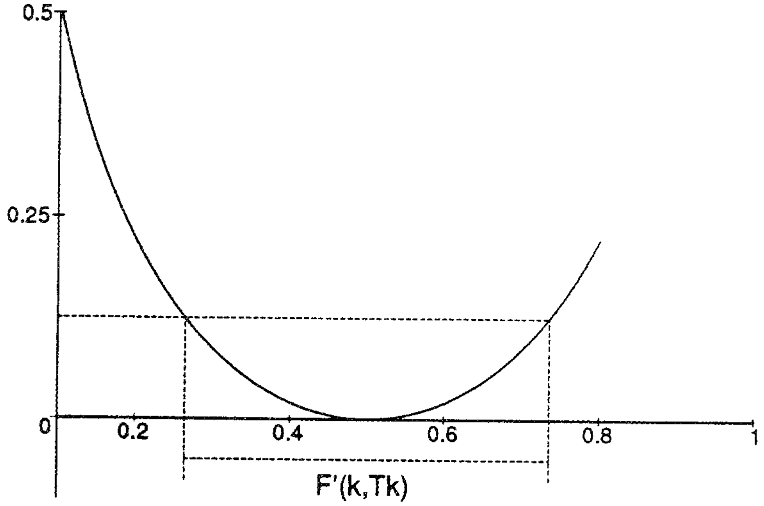

FIGURE 2. Computation of $\mathbf{U}(x, a; \omega_k)$ in Example 1.

(b) $\mathscr{L}(x, a; q, \hat{h}^k) = r(x, a) + q\hat{h}^k = 3 - 2(1 - q_1) = 1 + 2q,$

(c) $\mathbf{U}(x, a; \omega_k) = \max\{1 + 2q_1 \colon q_1 \in F'(k, T_k)\} = 1 + 2 \cdot 0.735 = 2.471.$

EXAMPLE 2. Take $S^+(x, a) = \{1, 2, 3\}$, with $r(x, a) = 2$. In this example $p_x(a) = (p_1, p_2, 1 - p_1 - p_2)$ with $(p_1, p_2) \in \Theta'(x, a) = \{(q_1, q_2) > (0, 0) \colon q_1 + q_2 < 1\}$.

(i) We compute $\mathbf{K}(x, a; P)$ (see Figure 3) for a $P \in \tilde{\mathscr{P}}$ with $h_1^* = h^*(1; P) = 0$, $h_2^* = h^*(2; P) = -4$, $h_3^* = h^*(3; P) = 6$, $\mathscr{L}^*(x; P) = \mathscr{L}^* = 6$ and $p_1 = 1/3$, $p_2 = 1/4$. We have:

(a) $\Delta\Theta(x, a) = \{q = (q_1, q_2, 1 - q_1 - q_2) \colon\ r(x, a) + h_1^* q_1 + h_2^* q_2 + h_y^*(1 - q_1 - q_2) > \mathscr{L}^*$, and its two-dimensional projection: $\Delta\Theta'(x, a) = \{0 < q_1,\ q_2 < 1 \colon 3q_1 + 5q_2 < 1\}$ is represented by the shaded region in Figure 3.

(b) $(x, a) \in \mathbf{B}(P)$ (since $\Delta\Theta(x, a) \neq \emptyset$),

(c) $\mathbf{I}(p, q) = 1/3 \log(1/3q_1) + 1/4 \log(1/4q_2) + 5/12 \log(5/(12(1 - q_1 - q_2)))$ attains its minimum value (equal to 0) at $q = p$, and it has convex contours one of which is shown in Figure 3,

(d) $\mathbf{K}(x, a; P) = \inf\{\mathbf{I}(p, q) \colon\ (q_1, q_2) \in \Delta\Theta'(x, a; P)\} = 0.223$ corresponds to the contour of $\mathbf{I}(p, q)$ tangent to the hyperplane $3q_1 + 5q_2 = 1$ at $q_1^0 = 0.18$, $q_2^0 = 0.092$.

(ii) In the computation of $\mathbf{U}(x, a; \omega_k)$ (see Figure 4) we postulate an observed history ω_k with $k = 150$ and $T_k = T_k(x, a) = 50$, $\hat{p}_x(a) = (\hat{p}_1, \hat{p}_2, 1 - \hat{p}_1 - \hat{p}_2)$, $\hat{p}_1 = 1/4$, $\hat{p}_2 = 1/3$, $\hat{h}_1 = \hat{h}^k(1) = 0$, $\hat{h}_2 = \hat{h}^k(2) = -2$, $\hat{h}_3 = \hat{h}^k(3) = 8$.

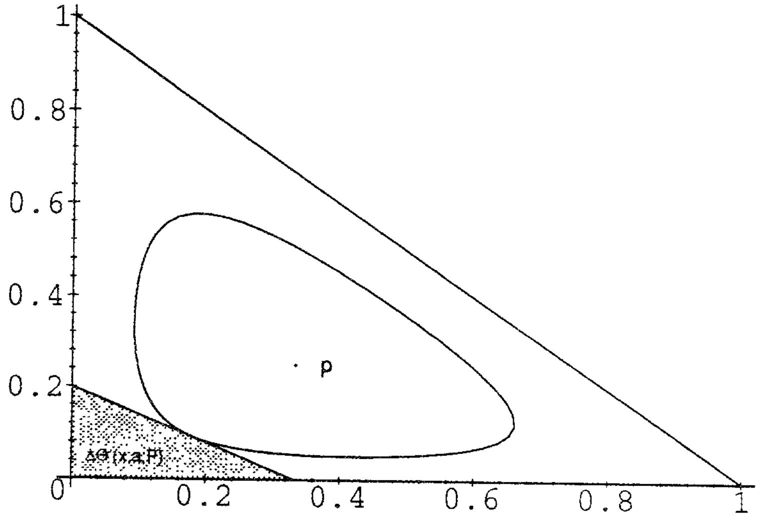

FIGURE 3. Computation of $\mathbf{K}(x, a; P)$ in Example 2.

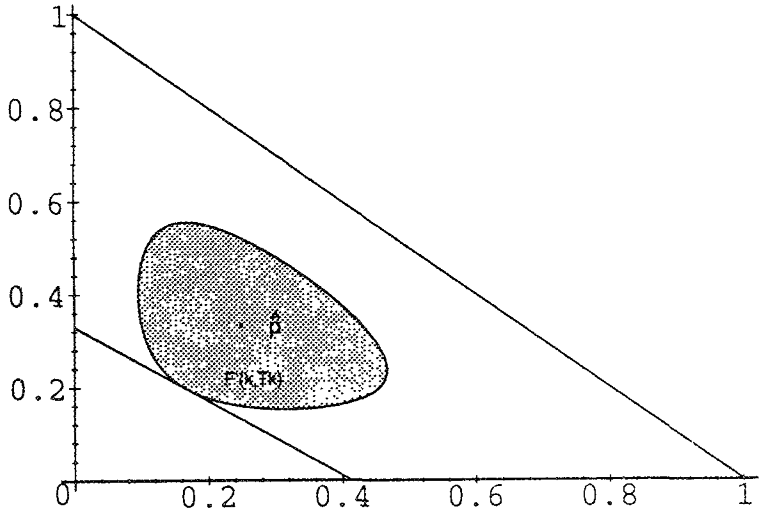

FIGURE 4. Computation of $\mathbf{U}(x, a; \omega_k)$ in Example 2.

Then we have:

(a) with $\beta = \log k/T_k = 0.1002$, $F(k, T_k) = \{q \in \Theta(x, a)\colon \mathbf{I}(\hat{p}_x(a), q) \le \beta\}$, and $F'(k, T_k) = \{(q_1, q_2) \in \Theta'(x, a)\colon 1/4 \log(1/4q_1) + 1/3 \log(1/3q_2) + 5/12 \log(5/(12(1 - q_1 - q_2))) = \beta\}$ (shaded convex region in Figure 4).

(b) $\mathscr{L}(x, a; q, \hat{h}^k) = r(x, a) + q\hat{h}^k = 10 - 8q_1 - 10q_2$,

(c) $\mathbf{U}(x, a; \omega_k) = \max\{10 - 8q_1 - 10q_2\colon (q_1, q_2) \in F'(k, T_k)\} = \mathscr{L}(x, a; q^*, \hat{h}^k)$, where q^* is the tangent point of the boundary of $F'(k, T_k)$ and a hyperplane of the form $8q_1 + 10q_2 = \text{const}$. In this example, $q_1^* = 0.168$, $q_2^* = 0.196$ and $\mathbf{U}(x, a; \omega_k) = 6.696$.

Note that there are two parallel hyperplanes tangent to the convex region $F'(k, T_k)$, one of which maximizes and the other minimizes $\mathscr{L}(x, a; q, \hat{h}^k)$. The index $\mathbf{U}(x, a; \omega_k)$ corresponds to the former hyperplane. Computationally, $\mathbf{U}(x, a; \omega_k)$ is the solution of a nonlinear programming problem with linear objective function and one convex and one linear constraint.

REMARK. When the MAB problem is viewed as a one-state MDP (see Model 1.4 of §7), the calculation of $\mathbf{U}$ presented in Examples 1 and 2 corresponds to the cases where the rewards are random variables that can take two and three values, respectively. In the first case, the values of $\mathbf{U}(1, a; \omega_k) = \mathbf{U}(1, a; k. T_k(a))$ are equal to the index $g_{k\ T_k(a)}$ obtained in Lai and Robbins (1985a); c.f., function g_{nt} on page 9, therein.

References

Agrawal, R. (1990). Adaptive control of Markov chains under the weak accessibility condition, in *29th Conf. on Decision and Control*, IEEE, 1426–1431.

____, D. Teneketzis, V. Anantharam (1989a). Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space. *IEEE Trans. Automat. Control* **AC-34** 1249–1259.

____, ____, ____ (1989b). Asymptotically efficient adaptive allocation schemes for controlled i.i.d. process: Finite parameter space. *IEEE Trans. Automat. Control* **AC-34** 258–267.

Anantharam, V., P. Varaiya, J. Walrand (1987). Asymptotically efficient allocation rules for multi-armed bandit problem with multiple plays. Part II: Markovian rewards. *IEEE Trans. Automat. Control* **AC-32** 975–982.

Borkar, V., P. Varaiya (1979). Adaptive control of Markov chains, I: Finite parameter set. *IEEE Trans. Automat. Control* **AC-24** 953–958.

Burnetas, A. N., M. N. Katehakis (1995). Efficient estimation and control for Markov processes, in 34th Conf. on Decision and Control, IEEE, 1402–1407.

____, ____ (1996). Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.* **17** 122–142.

Chernoff, H. (1967). Sequential models for clinical trials, *in* 'Proc. Fifth Berkeley Symp. Math. Statist.', University of California Press, 805–812.

Dembo, A., O. Zeitouni (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett.

Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

Ellis, R. S. (1985). *Entropy, Large Deviations and Statistical Mechanics*. Springer Verlag,

Federgruen, A., P. Schweitzer (1981). Nonstationary Markov decision problems with converging parameters. *J. Optim. Theory Appl.* **34** 207–241.

Fernández-Gaucherand, E., A. Arapostathis, S. I. Marcus (1993). Analysis of an adaptive control scheme for a partially observed controlled Markov chain. *IEEE, Trans. Automat. Control* **38**, 987–993.

Fox, B. L., J. E. Rolph (1973). Adaptive policies for Markov renewal programs. *Ann. Statist.* **1** 334–341.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Stat. Soc. Ser. B* **41** 335–340.

Glazebrook, K. D. (1991). Competing Markov decision processes. *Ann. Oper. Res.* **29** 537–564.

Graves, T., T. L. Lai (1995). Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM J. Control Optim.* (to appear).

Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes*, Springer-Verlag.

Katehakis, M. N., C. Derman (1985). Computing optimal sequential allocation rules in clinical trials, *in* J. van Ryzin, ed., 'Adaptive Statistical Procedures and Related Topics', Vol. 8, I.M.S. Lecture Notes–Monograph Series, pp. 29–39.

——, H. Robbins (1995). Sequential choice from several populations. *Proc. Nat. Acad. Sci. USA* 8584–8585.

——, U. Rothblum (1996). Finite state multi-armed bandit problems: Sensitive-discount, average-reward and average-overtaking optimality. *Ann. Appl. Probab.* **6** 1024–1034.

——, A. F. Veinott Jr. (1987). The multi-armed bandit problem: Decomposition and computation. *Math. Oper. Res.* **12** 262–268.

Kolonko, M. (1982). Strongly consistent estimation in a controlled Markov renewal model. *J. Appl. Probab.* **19** 532–545.

Kumar, P. R., P. Varaiya (1986a). *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice Hall.

——, —— (1986b). *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall.

Lai, T. L., H. Robbins (1985a). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6** 4–22.

——, —— (1985b). Asymptotically optimal allocation of treatments in sequential experiments. T. J. Santner and A. C. Tamhane, eds. *Design of experiments: Ranking and selection: Essays in honor of Robert E. Bechhofer.* Vol. 56, M. Dekker, pp. 127–142.

——, S. Yakowitz (1995). Nonparametric bandit methods. *IEEE Trans. Automat. Control* **40** 1199–1209.

Li, Z., C. Zhang (1992). Asymptotically efficient allocation rules for two Bernoulli populations. *J. Roy. Statist. Soc. B.* **54**, 609–616.

Mallows, C. L., H. Robbins (1964). Some problems of optimal sampling strategy. *J. Math. Anal. Appl.* **8** 90–103.

Mandl, P. (1974). Estimation and control in Markov chains. *Adv. Appl. Probab.* **6** 40–60.

Milito, R. A., J. B. Cruz Jr. (1992). A weak contrast function approach to adaptive semi-Markov decision models, S. Tzafestas and C. Watanabe, eds., *Stochastic Large Scale Engineering Systems*, Marcel Dekker, 253–278.

Noble, B., J. W. Daniel (1977). *Applied Linear Algebra*. 2nd ed., Prentice Hall.

Rieder, U. (1975). Bayesian dynamic programming. *Adv. Appl. Probab.* **7** 330–348.

——, J. Weishaupt (1995). Customer scheduling with incomplete information. *P.E.I.S.* **7** 269–284.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–536.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming*, Academic Press.

Schäl, M. (1987). Estimation and control in discounted stochastic dynamic programming. *Stochastics* **20** 51–71.

Shimkin, N., A. Shwartz (1996). Asymptotically efficient adaptive strategies in repeated games part II: Asymptotic optimality. *Math. Oper. Res.* **21** 487–512.

Van Hee, K. M. (1980). Markov decision processes with unknown transition law: The average return case, D. R. Hartley, L. C. Thomas, eds., *Return Developments in Markov Decision Processes*, Academic Press, 227–244.

Veinott, A. F. J. (1974). Markov decision chains, R. E. G. B. Dantzig, ed., *Studies in Optimization*, Mathematical Association of America, 124–159.

Weiss, G. (1994). *On almost optimal preemptive scheduling of stochastic jobs*, Technical Report, Georgia Institute of Technology, Atlanta, Georgia.

White, C. C., H. K. Eldeib (1994). Markov decision processes with imprecise transition probabilities. *Oper. Res.* **42** 739–749.

Whittle, P. (1980). Multi-armed bandits and the gittins index, *J. R. Statist. Soc. B.* **42** 143–149.

Yakowitz, S., W. Lowe (1991). Nonparametric bandit methods. *Ann. Oper. Res.* **28** 297–312.

A. N. Burnetas: Department of Operations Research, Case Western Reserve University, Cleveland, Ohio 44106; e-mail: atb4@po.cwru.edu

M. N. Katehakis: Faculty of Management and RUTCOR, Rutgers University, Newark, New Jersey 07102; e-mail: mnk@andromeda.rutgers.edu