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

WP13 5:10

Efficient Estimation and Control for Markov Processes

Apostolos N. Burnetas
Department of Operations Research
Case Western Reserve University
Cleveland, OH 44106
atb4@po.cwru.edu

Michael N. Katehakis
Faculty of Management
and RUTCOR
Rutgers University
Newark, NJ 07102
mnk@andromeda.rutgers.edu

Abstract

We consider the problem of sequential control for a finite state and action Markovian Decision Process with incomplete information regarding the transition probabilities $P \in \tilde{\mathcal{P}}$. Under suitable irreducibility assumptions for $\tilde{\mathcal{P}}$, we construct adaptive policies that maximize the rate of convergence of realized rewards to that of the optimal (non adaptive) policy under complete information. These adaptive policies are specified via an easily computable index function, of states, controls and statistics, so that one takes a control with the largest index value in the current state in every period.

1. Introduction

Consider a discrete time, Markovian Decision Process with state space $S = \{1, 2, \ldots, s\}$, finite action space $A = \bigcup_{x \in S} A(x)$, where $A(x)$ is the set of admissible actions in state x, and transition law $P = [p_{xy}(a)]_{x,y \in S, a \in A(x)}$. The transition probability vectors $p_x(a)$ are unknown and belong to known sets $\Theta(x, a)$. Let $\tilde{\mathcal{P}} = \{P \ s.t. \ p_{xy}(a) \in \Theta(x, a)\}$. Under suitable irreducibility assumptions for $\tilde{\mathcal{P}}$, we construct a class C_R of adaptive policies that maximize the rate of convergence of realized rewards to that of the optimal (non adaptive) policy under complete information. These adaptive policies are specified via an easily computable index function, of states, controls and statistics, so that one takes a control with the largest index value in the current state in every period.

The ideas involved in this paper represent generalizations of the theory developed in [2], [3], [8] and [10]. Related work is that of [6], [12], [1] and [9]. Main differences of our work from the later are: i) we use the Markovian structure of the problem (i.e., an "open box" methodology) and ii) optimal policies are specified via easily computable indices so that one takes (and thus samples) actions, or controls, with the largest index value in every current state and period, instead of sampling policies (from the class of all deterministic policies) at the end of cycles.

For computational issues for MDPs see [11], [5], [13] and references therein.

The paper is organized as follows. In section 2 we formulate the problem of control of a Markov process under side constraints. In section 3 we present the index policies and in section 4 we give initial computation results for a queuing control problem first introduced in the context of optimal replacement in [4].

2. The Partial Information Model

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 j^{th} occurrence of (x, a). From the Markov property, $Y_j(x, a), j = 1, 2, \ldots$ are i.i.d. with distribution $p_x(a)$.

(b) Let the random variables X_t, A_t, $t = 0, 1, \ldots$ 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, $\omega_k = x_0, a_0, \ldots x_{k-1}, a_{k-1}, x_k$, such that $a_t \in A(x_t), t = 0, \ldots, k-1$. Let $(\Omega^{(k)}, \mathcal{F}^{(k)})$, $1 \leq k \leq \infty$ denote the sample space of histories ω_k, where $\Omega^{(k)}$ is the set of all histories ω_k and $\mathcal{F}^{(k)}$ the σ–field generated by Ω_k. *Events*, defined on $\mathcal{F}^{(k)}$ are denoted by capital letters. The complement of event B is denoted by $\overline{B}$.

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 generally 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 *stationary Markov* if $\pi_k(\cdot|\omega_k) = \pi_0(\cdot|x_k)$, for all k, ω_k. A policy π is *deterministic* if there exist functions $f_k : S \to A, k = 0, 1, \ldots$, with $f_k(x) \in A(x)$, such that for all k and ω_k, $\pi_k(\{f(X_k)\}|\omega_k) = 1$. In this case π is also denoted by $\{f_k\}$. Let C denote

0-7803-2685-7/95 $4.00 © 1995 IEEE

the set of all policies, and C_M, C_D the set of all stationary Markov and stationary deterministic policies, respectively.

Probability and expectation under transition law P, policy $\pi \in C$ and starting state x_0 will be denoted by $\mathbf{P}_{x_0}^{\pi,P}, \mathbf{E}_{x_0}^{\pi,P}$.

2.1. Unobservable quantities.

i) Transition Law and Parameter Space. Let $P = [p_{xy}(a)]_{x,y\in S, a\in A(x)}$ denote the unknown transition law and let $\mathcal{P}$ denote the set of all such P. We make the following assumption.

Assumption (A). For all $x \in S, a \in A(x)$, the sets $S^+(x,a) = \{ y \in S : p_{xy}(a) > 0 \}$ 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 $f \in C_D$.

Let $\tilde{\mathcal{P}}$ be the subset of $\mathcal{P}$ which consists of the transition laws P satisfying **(A)**. In the sequel we take $\tilde{\mathcal{P}} = \prod_{(x,a)} \Theta(x,a)$, where

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

denotes the parameter space for the probability vector $p_x(a)$.

ii) Performance Criteria subject to side constraints. The performance of a policy π is characterized by a reward structure $R = [r(x,a),\ x \in S, a \in A(x)]$, and m distinct cost structures $C_i = [c_i(x,a),\ x \in S, a \in A(x)]$, $i = 1,\ldots,m$, where $r(x,a)$ denotes the expected one step reward and $c_j(x,a)$ the expected one step cost corresponding to cost criterion i, for the state action pair (x,a).

Let $V_n^{\pi,P}(x_0) = \mathbf{E}_{x_0}^{\pi,P} \sum_{t=0}^{N-1} r(X_t, A_t)$ denote the expected total reward during the first n transitions under policy π, and $g_0^\pi(P) = \underline{\lim}_{n\to\infty} V_n^{\pi,P}(x_0)/n$ the expected average reward. Note that, because of Assumption (A), $g_0^\pi(P)$ is independent of the starting state x_0.

Similarly, let

$$g_i^\pi(P) = \overline{\lim_{n\to\infty}} \mathbf{E}_{x_0}^{\pi,P} \sum_{t=0}^{N-1} c_i(X_t, A_t)$$

denote the expected average cost for criterion i, $i = 1,\ldots,m$.

The optimality criterion is maximization of the expected average reward subject to constraints on the expected average costs. For $v \in \mathbb{R}^m$ let $C_F(P,v) = \{\pi \in C : g_i^\pi(P) \le v_i, i = 1,\ldots,m\}$ denote the set of feasible policies.

A policy $\pi^* \in C_F(P,v)$ is optimal if $g_0^{\pi^*}(P) = g^*(P,v)$, where $g^*(P,v) = \sup\{g_0^\pi(P) : \pi \in C_F(P,v)\}$.

Since P is unknown, $g^*(P,v)$ is an unobservable quantity. It represents the maximum expected average reward (subject to the constraints for the expected average costs), that can be attained "if the transition law P is known to the experimenter".

In the sequel vector v is considered known and fixed, and the dependence of all the problem quantities on v is not explicitly denoted.

iii) Unobservable Linear Programming solutions. Let $\mathcal{A} = \prod_{x\in S} A(x)$ denote the cartesian product of the action sets. There is a one to one correspondence between the vectors in $\mathcal{A}$ and the stationary and deterministic policies $f \in C_D$.

For a collection $\{D(x) \subseteq A(x),\ x \in S\}$, let $\mathcal{D} = \prod_{x\in S} D(x) \subseteq \mathcal{A}$. For any $\mathcal{D} \subseteq \mathcal{A}$ define the *restricted problem* $(P,\mathcal{D})$ as a markovian decision process with state space S, action sets $\{D(x),\ x \in S\}$ and transition law equal to the restriction of P on $\mathcal{D}$. There is a one to one correspondence between the vectors in $\mathcal{D}$ and the stationary and deterministic policies of the restricted problem $(P,\mathcal{D})$.

For a restricted problem $(P,\mathcal{D})$, let $LP(P,\mathcal{D})$ denote the following linear programming problem with variables $\sigma(x,a)$, $x \in S, a \in D(x)$, where $\delta_{xy} = 1$ if $x = y$ and 0 otherwise.

$$\begin{array}{lll}
\max & \sum_{x\in S}\sum_{a\in D(x)} r(x,a)\sigma(x,a) & \\
\text{st} & \sum_{x\in S}\sum_{a\in D(x)} \sigma(x,a) & = 1 \\
 & \sum_{x\in S}\sum_{a\in D(x)} (\delta_{xy} - p_{xy}(a))\sigma(x,a) & = 0, \quad y \in S \\
 & \sum_{x\in S}\sum_{a\in D(x)} c_i(x,a)\sigma(x,a) & \le v_i, \quad i = 1,\ldots,m \\
 & \sigma(x,a) \ge 0 &
\end{array}$$

Because of the first constraint, $LP(P,\mathcal{D})$ is always bounded. If it is also feasible, then let $(\sigma(x,a;P,\mathcal{D}),\ x \in S, a \in D(y))$ denote the optimal solution and $O(x;P,\mathcal{D}) = \{a \in D(x) : \sigma(x,a;P,\mathcal{D}) > 0\}$; let $(\gamma(P,\mathcal{D}),\ h(P,\mathcal{D}),\ \lambda(P,\mathcal{D}))$ be the optimal dual solution, where $h(P,\mathcal{D}) = (h(x;P,\mathcal{D}),\ x \in S)$, and $\lambda(P,\mathcal{D}) = (\lambda_i(P,\mathcal{D}), i = 1,\ldots,m)$. Also let $\sigma^*(P) = \sigma(P,\mathcal{A})$ and $(\gamma^*(P), h^*(P), \lambda^*(P)) = (\gamma(P,\mathcal{A}), h(P,\mathcal{A}),\ \lambda(P,\mathcal{A}))$ be the optimal primal and dual solutions, respectively, of the unrestricted problem $(P,\mathcal{A})$.

Using this terminology, the optimal average reward $g^*(P)$ of the initial constrained markovian decision process is equal to $\gamma^*(P) + \sum_{i=1}^m \lambda_i^*(P,\mathcal{A})v_i$. In addition, an optimal, in general randomized, stationary Markov policy π^* can be obtained from $\sigma^*(P)$ by setting $\pi^*(a|x) = \sigma^*(x,a;P)/\sum_{a'\in A(x)} \sigma^*(x,a';P)$ (c.f. [7]).

Let $\phi^*(x,a;P) = \phi(x,a;p_x(a), \gamma^*(P), h^*(P), \lambda^*(P))$ denote the marginal reward of pair (x,a) for

$LP(P,\mathcal{A})$, where, for $\gamma \in \mathbb{R}$, $q,h \in \mathbb{R}^s$, and $\lambda \in \mathbb{R}^m$, $\phi(x,a;q,g,h,\lambda) = \gamma + h(i) + \sum_{i=1}^m \lambda(i)c_i(x,a) - r(x,a) - \sum_{y\in S} q(y)h(y)$. Then, the optimality conditions for $LP(P,\mathcal{A})$, can be written as $\phi^*(x,a;P) \geq 0, \forall(x,a)$.

iv) Critical State–Action Pairs. **(a)** For (x,a) such that $a \notin O(x;P)$, let $\mathbf{\Delta\Theta}(x,a;P) = \{q \in \Theta(x,a) : \phi(x,a;q,\gamma^*(P),h^*(P),\lambda^*(P)) < 0\}$, denote the set of values q of the transition probability row for pair (x,a) under which a belongs to an optimal (possibly randomized) policy.

(b) Define the set of *critical* state - action pairs for any $P \in \tilde{\mathcal{P}}$, as $\mathbf{B}(P) = \{(x,a) : a \notin O(x;P)$ and $\mathbf{\Delta\Theta}(x,a;P) \neq \emptyset\}$.

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

(d) For $(x,a) \in B(P)$ let $\mathbf{K}(x,a;P) = \inf\{\mathbf{I}(p_x(a),q) : q \in \mathbf{\Delta\Theta}(x,a;P)\}$.

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

2.2. Optimality Criteria. In the present context only adaptive policies are available. Let $R_n^{\pi,P}(x_0) = ng^*(P) - V_n^{\pi,P}(x_0)$ represent the loss or regret, due to partial information, incurred in the n–horizon expected reward when a policy π is used. Maximization of $V_n^{\pi,P}(x_0)$ with respect to π is equivalent to minimization of $R_n^{\pi,P}(x_0)$. In general it is not possible to find an adaptive policy which minimizes $R_n^{\pi,P}(x_0)$ uniformly in P. Therefore, a different definition of optimality is required.

A policy π will be called *uniformly feasible* (UF) if $\pi \in C_F(P), \forall P \in \tilde{\mathcal{P}}$ such that $C_F(P) \neq \emptyset$.

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

A UC policy π will be called *uniformly fast convergent* (UFC) if the following stronger condition holds: $ng^*(P) - V_n^{\pi,P}(x_0) = o(n^\alpha)$, as $N \to \infty$, $\forall\alpha > 0$, $\forall P \in \tilde{\mathcal{P}}$, $\forall x_0 \in S$.

A UFC policy π_0 will be called *uniformly maximum convergence rate* ($UMCR$) if $\overline{\lim}_{N\to\infty}(ng^*(P) - V_n^{\pi_0,P}(x_0))/(ng^*(P) - V_n^{\pi,P}(x_0)) \leq 1$, $\forall P \in \tilde{\mathcal{P}}$ such that $\mathbf{M}(P) > 0$, for all UFC π , $\forall x_0 \in S$. Note that according to this definition a $UMCR$ policy has maximum rate of convergence only for those values of the parameter space for which $\mathbf{M}(P) > 0$; when $\mathbf{M}(P) = 0$ it is UFC.

Let $C_{UF} \supset C_{UC} \supset C_{UFC} \supset C_{UM}$ denote the classes of UF, UC, $UFC$and $UMCR$ policies, respectively.

Remark 1. Because $\overline{\lim}_{n\to\infty}(ng^*(P) - V_n^{\pi,P}(x_0))/n = g^*(P) - g_0^\pi(P) \geq 0$, $\forall\pi \in C_{UF}$, classes C_{UC}, C_{UFC}, C_{UM} can be expressed in terms of the rate of convergence of $V_n^{\pi,P}(x_0)/n$ to $g^*(P)$ as follows.

If $\pi \in C_{UC}$, then $\lim_{n\to\infty} V_n^{\pi,P}(x_0)/n = g^*(P)$ for all P. No claim regarding the rate of convergence can be made.

If $\pi \in C_{UFC}$, then it is also true that $|V_n^{\pi,P}(x_0)/n - g^*(P)| = o(n^{(\alpha-1)})$, therefore $V_n^{\pi,P}(x_0)/n$ converges to $g^*(P)$ faster than $n^{(\alpha-1)}$, $\forall P \in \tilde{\mathcal{P}}$, $\forall\alpha > 0$.

If $\pi \in C_{UM}$, then, for all $P \in \tilde{\mathcal{P}}$ with $\mathbf{M}(P) > 0$ the rate of convergence of $V_n^{\pi,P}(x_0)/n$ to $g^*(P)$ is the maximum among all policies in C_{UFC} and exactly equal to $\mathbf{M}(P)\log n/n$.

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) Let $\forall t \geq 0$, $\hat{p}_x^t(a) = [\hat{p}_{xy}^t(a)]_{y\in S}$, where $\hat{p}_{xy}^t(a) = 0$ if $y \notin S^+(x,a)$ and

$\hat{p}_{xy}^t(a) = (1-w_t)f_0(y;x,a) + w_t f_t(y;x,a)$, and $w_t = t/(|S^+(x,a)| + t)$, otherwise.

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 this estimation scheme ensures that assumption **(A)** is satisfied.

3. UMCR Index policies

In this section two classes C_R^0, C_R^1 of adaptive policies are defined, and it is shown that if $m = 0$, that is, the markovian decision process is unconstrained, then $C_R \subset C_{UM}$, whereas if $m > 0$ a conjecture is made for C_R^1.

Given a history ω_k, define the following.

i) The restriction $\mathcal{D}_k \subseteq \mathcal{A}$ as the product of the "relatively frequently sampled" action sets:

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

ii) Let $\hat{\sigma}^k = \sigma(\hat{P}^k, \mathcal{D}_k)$ be the primal and $\hat{\gamma}^k =$

$\gamma(\hat{P}^k, \mathcal{D}_k)$, $\hat{h}^k = h(\hat{P}^k, \mathcal{D}_k)$, $\hat{\lambda}^k = \lambda(\hat{P}^k, \mathcal{D}_k)$ the dual solution of the restricted problem $(\hat{P}^k, \mathcal{D}_k)$.

iii) For $x \in S, a \in A(x)$ define the **index** $\mathbf{U}(x, a; \omega_k)$ as

$$\mathbf{U}(x, a; \omega_k) = \inf_{q \in \Theta(x,a)} \{\phi(x, a; \hat{p}_x^{T_k(x,a)}(a), \hat{\gamma}^k, \hat{h}^k, \hat{\lambda}^k)) : \mathbf{I}(\hat{p}_x^{T_k(x,a)}(a), q) \leq \log T_k(x)/T_k(x, a)\},$$

for $k \geq 1$, where a ratio of the form $\log k/0$ is assumed equal to ∞.

iv) For $x \in S$ let

$$\Gamma_1(x; \hat{P}^k, \mathcal{D}_k) = \{ a \in O(x; \hat{P}^k, \mathcal{D}_k) : T_k(x, a) < \log^2 T_k(x) + 1 \},$$

$$\Gamma_2(x; \hat{P}^k, \mathcal{D}_k) = \{ a \in A(x) : \mathbf{U}(x, a; \omega_k) = \min_{a' \in A(x)} \mathbf{U}(x, a'; \omega_k) \}.$$

3.1. Unconstrained case. If $m = 0$, then the primal solution $\sigma^*(P, \mathcal{D})$ satisfies the following: $\forall\, x \in S, \exists a = f(x) \in D(x)$, such that $\sigma^*(x, a; P, \mathcal{D}) > 0$ and $\sigma^*(x, a'; P, \mathcal{D}) = 0$, for $a' \neq f(x)$, i.e., there exists a deterministic optimal policy $f \in \mathcal{D}$. In addition, the dual solution of any problem $(P, \mathcal{D})$ consists of γ^*, h^* only.

In the unconstrained case define the class C_R^0 of *index policies* which, at time k and state $X_k = x$, take any action from $\Gamma_1(x; \hat{P}^k, \mathcal{D}_k)$, if $\Gamma_1(x; \hat{P}^k, \mathcal{D}_k) = O(x; \hat{P}^k, \mathcal{D}_k)$, and any action from $\Gamma_2(x; \hat{P}^k, \mathcal{D}_k)$ otherwise.

Remark 2. **(a)** For all (x, a), $\mathbf{U}(x, a; \omega_k) \leq \phi(x, a; \hat{p}_x^{T_k(x,a)}(a), \hat{\gamma}^k, \hat{h}^k)$, i.e., $\mathbf{U}(x, a; \omega_k)$ represents a deflation of the marginal reward of the restricted problem. In addition, $\mathbf{U}(x, a; \omega_k)$ is decreasing in k and increasing in $T_k(x, a)$, thus giving higher chance to the "under sampled" actions to be selected.

(b) We refer to the case in which at time k : $\Gamma_1(x; \hat{P}^k, \mathcal{D}_k) = O(x; \hat{P}^k, \mathcal{D}_k)$, as a *forced selection* instance. The idea of the forced selections is that if in some period k it is detected that none of the optimal actions of the restricted problem $(\hat{P}^k, \mathcal{D}_k)$ will be contained in the restricted set the next time state x will be visited, one of these actions is taken in order to remain in the restricted set. As a consequence of the forced selection scheme, the optimal solutions of the restricted problems have the following asymptotic monotonicity property: $\mathbf{P}_{x_0}^{\pi_0, P}[\gamma(P, \mathcal{D}_{k+1}) \geq \gamma(P, \mathcal{D}_k)] = 1 - o(1/k)$, as $k \to \infty$, for all $\pi_0 \in C_R^0$.

(c) In [2] it is shown that $\mathbf{P}_{x_0}^{\pi_0, P}[D_k(x) \subseteq O(x; P), \forall x \in S] = 1 - o(1/k)$, for all $\pi_0 \in C_R^0$. 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 $(\hat{P}^k, \mathcal{D}_k)$, as k increases.

The next theorem is proved along the same lines as Theorem 1 in [2].

Theorem 1 *For $m = 0$,*

1. $\underline{\lim}_{n \to \infty} \mathbf{E}_{x_0}^{\pi, P} T_N(x, a)/\log n \geq 1/\mathbf{K}(x, a; P)$, $\forall (x, a) \in \mathbf{B}(P)$, $\forall \pi \in C_{UFC}$.
2. $\forall \pi_0 \in C_R^0$, $\forall P \in \tilde{\mathcal{P}}$ *and* $\forall\, x \in S, a \notin O(x, P)$,
   (a) *if* $(x, a) \in \mathbf{B}(P)$, *then*
   $$\overline{\lim_{n \to \infty}} \mathbf{E}_{x_0}^{\pi_0, P} T_N(x, a)/\log n \leq 1/\mathbf{K}(x, a; P),$$
   otherwise
   $$\overline{\lim}_{n \to \infty} \mathbf{E}_{x_0}^{\pi_0, P} T_N(x, a)/\log n = 0.$$
   (b) $R_n^{\pi_0, P}(x_0) = \mathbf{M}(P) \log n + o(\log n)$.
3. $C_R^0 \subseteq C_{UM}$.

3.2. Constrained case. In the constrained case $(m \geq 1)$, define a class C_R^1 of policies as follows. At time k and state $X_k = x$, if $\Gamma_1(x; \hat{P}^k, \mathcal{D}_k) = O(x; \hat{P}^k, \mathcal{D}_k)$, take any action from $\Gamma_1(x; \hat{P}^k, \mathcal{D}_k)$. Otherwise, if there exists $a \in \Gamma_2(x; \hat{P}^k, \mathcal{D}_k)$, $a \notin O(x; \hat{P}^k, \mathcal{D}_k)$, then take any such action. If no such a exists, then choose randomly among $a \in O(x; \hat{P}^k, \mathcal{D}_k)$, with randomization probabilities $q(a) = \hat{\sigma}^k(a)/\sum_{a' \in O(x; \hat{P}^k, \mathcal{D}_k)} \hat{\sigma}^k(a')$.

In current work we aim to prove the following theorem, which we state as a conjecture.

Theorem 2 *For $m \geq 1$, the claims of Theorm 1 hold for any adaptive policy $\pi \in C_R^1 \subseteq C_{UM}$.*

4. A Queueing/Reliability Application

Consider a discrete time queueing system with a single server and a buffer with capacity $s - 1$. State $i, i = 0, 1, \ldots, s$, corresponds to i customers in the system, including the server. Customers arrive at the system according to the following mechanism. If the system is in state i, $i = 0, \ldots, s - 1$, at the beginning of a period, then Q_{ij} denotes the probability that there will be j arrivals during this period, $j = 0, \ldots, s - 1$, and Q_{is} the probability that the number of arrivals will be at least s. If more arrivals than the empty spaces in the system occur, the additional arrivals do not join the system, but rather they are lost. The probabilities Q_{ij} are unknown.

There are two actions available in states $i = 1, \ldots, s - 1$. Action 1 corresponds to the server being idle,

and action 2 to server working. It is assumed that if the server works during a period, it is able to serve all the customers waiting in the system and all the customers arriving during the period, thus there will be no customers in the beginning of the next period. In states 0 and s, only actions 1 and 2, respectively, are available.

Reward $r(i,a)$ represents profit from either serving the waiting customers if action 2 is taken, or from the server working on different jobs if action 1 is taken. Cost $c(i,a)$ represents operational expenses.

The objective is to find a policy for switching the server on and off, so that the expected average profit is maximized subject to a constraint on the expected average cost.

This queueing system is a special case of the model described in section 2, with state space $S = \{0,\ldots,s\}$, action sets $A(0) = \{1\}$, $A(1) = \cdots = A(s-1)\{1,2\}$, $A(s) = \{2\}$, and transition law $p_{0j}(1) = Q_{0j}, j = 0,\ldots,s$, $p_{ij}(1) = Q_{ij}, i = 1,\ldots,s-1$ $j = 0,\ldots,s$, and $p_{i0}(2) = 1, i = 1,\ldots,s$.

The model described above has the following machine replacement interpretation as well. Consider a machine that can be in one of states $0,\ldots,s$, at the beginning of a period, where the state denotes the level of damage, with state 0 corresponding to a new machine and state s to a nonoperational machine. Actions 1 and 2 represent operation and replacement of the machine, respectively, and Q_{ij} is equal to the probability that a machine in state i in the beginning of a period, if not replaced, will be in state j in the beginning of the next period. A new machine is never replaced, and a nonoperational machine is always replaced. The quantities $r(i,a), c(i,a)$ represent expected one period profits and operational costs. The unconstrained version of the reliability model is included in [14], and also in [4], where uniformly consistent policies are developed for the case of unknown Q_{ij}.

In this section a 4-state ($s = 3$) instance of this problem is simulated for both the unconstrained and constrained cases, for a total number of 1000 transitions in each case. The (unknown) arrival probabilities matrix is taken equal to

$$Q = \begin{bmatrix} 0.2 & 0.3 & 0.2 & 0.3 \\ 0 & 0.6 & 0.1 & 0.3 \\ 0 & 0 & 0.4 & 0.6 \end{bmatrix},$$

the rewards $r(0,1) = 15$, $r(1,1) = 9$, $r(1,2) = 2$, $r(2,1) = 11$, $r(2,2) = 4$, $r(3,2) = -3$, the costs $c(0,1) = 10$, $c(1,1) = 12$, $c(1,2) = 15$, $c(2,1) = 14$, $c(2,2) = 18$, $c(3,2) = 20$, and the upper bound on the expected average cost $v = 13$.

Figures 1 and 2 below describe the evolution of the average reward, for the unconstrained and constrained cases respectively. The optimal solution under complete information is equal to $g^* = 8.83$ for the unconstrained and $g^* = 8.76$ for the constrained problem.

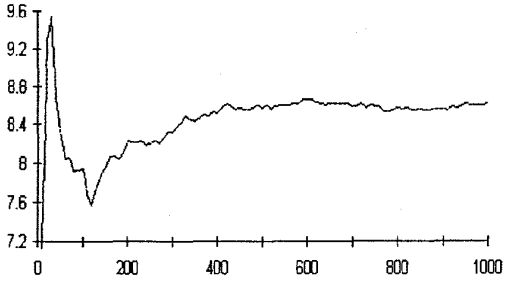

Figure 1. Unconstrained Case

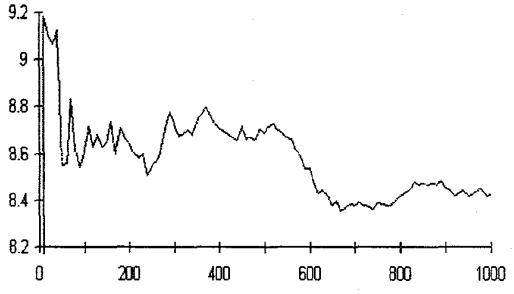

Figure 2. Constrained Case

References

[1] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymtotically efficient adaptive allocation schemes for controlled Markov chains : Finite parameter space. *IEEE Trans. A.C.*, 34:1249–1259, 1989.

[2] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for dynamic programming. Technical report, Rutgers University, 1994.

[3] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Adv. Appl. Math.*, to appear.

[4] C. Derman. On optimal replacement rules when changes of state are Markovian. In R. Bellman,

editor, *Mathematical Optimization Techniques.* Uniersity of California Press, 1963.

[5] E. A. Feinberg and A. Shwartz. Constrainted Markov decision models with weighted discounted criteria. *Math. Op. Res.*, 20(2):302–320, 1995.

[6] B. L. Fox and J. E. Rolph. Adaptive policies for Markov renewal programs. *Ann. Stat.*, 1:334–341, 1973.

[7] L. C. M. Kallenberg. *Linear programming and finite Markovian control problems.* Mathematisch Centrum, second edition, 1983.

[8] M. N. Katehakis and H. Robbins. Sequential choise from several populations. *Proc. Nat. Acad. Sci. USA*, to appear.

[9] T. L. Lai and S. Yakowitz. Nonparametric bandit methods. Technical Report, Stanford University, Statistics Department, Stanford, California, 1995.

[10] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6:4–22, 1985.

[11] A. M. Makowski and A. Shwartz. Implementation issues for Markov decision processes. *IMA Stoch. Diff. Syst. Stoch. Control Theor. Appl.*, 10:323–327, 1988.

[12] R. A. Milito and J. B. Cruz, Jr. A weak contrast function approach to adaptive semi–Markov decision models. In S. Tzafestas and C. Watanabe, editors, *Stochastic Large Scale Engineering Systems*, pages 253–278. Marcel Dekker, 1992.

[13] M. L. Puterman. *Markov Decision Processes.* J. Wiley, 1994.

[14] S. M. Ross. *Introduction to Stochastic Dynamic Programming.* Academic Press, 1983.