Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Adaptive Optimization of Markov Reward Processes

Enrique Campos-Náñez and Stephen D. Patek

***Abstract*— We consider the problem of optimizing the average reward of Markov chains controlled by two sets of parameters 1) a set of tunable parameters and 2) a set of fixed but unknown parameters. We study the convergence characteristics of recursive estimation procedures based on the observation of regenerative cycles. We also provide sufficient conditions for the convergence to local optima of existing simulation-based optimization procedures under parameter certainty, in order to achieve simultaneous optimal selection of the tunable parameters and identification of the unknown parameters. To illustrate our approach, we discuss an algorithm which exploits the gradient of the likelihood of an observed regenerative cycle and its application to a regenerative simulation-based algorithm introduced in [1]. Our results are illustrated numerically in a problem of optimal pricing of services in a multi-class loss network.**

I. Introduction

Dynamic programming models have long been recognized as the right way to characterize many stochastic research allocation and control problems. Unfortunately, the dynamic programming algorithm itself suffers from the so called “curse of dimensionality”. Recent efforts to address dynamic programming scalability issues have focused on the development of simulation-based techniques, such as neuro-dynamic programming or approximate dynamic programming [2], reinforcement learning [3]–[6], and actor-critic methods [7]. Many of these techniques resort to parameterization in order to develop compact representations of one, or sometimes both, of two elements: the cost-to-go functions, and the policies themselves. These methods are simulation-based in the sense that they recur to simulation to 1) improve the cost-to-go representations in the first case (see [2], [6]), and 2) improve policies, typically through the estimation of the gradient of the performance function with respect to the control actions (see [8], [9], and relevant to the control of Markov chains the works in [1], [10]–[17]).

While some of the algorithms mentioned above are implicitly adaptive, in the sense that they do not require knowledge of the system model, (e.g. general stochastic approximations [18], [19]), they exhibit slow convergence due to biased gradient estimates, or large variances. This problem can be alleviated by incorporating knowledge of the system model, for example by exploiting regenerative structure to eliminate bias [20].

On the other hand, the problem of adaptive control of Markov chains has been also been studied, but only few and limited studies of model-based on-line algorithms are available. Most of the available works address the problem when a finite number of possible models is available, as in [21]–[25]. Other authors, overcome this restriction, but assume that the optimal policies for each model have been precomputed [26]. Others, embed the estimation process in a value iteration procedure [27], which is not efficient computationally when the state spaces are large. The work in [28], [29] estimates Q-factors, a technique that is adaptive since no system model is in principle needed, but requires the action set to be finite in order to build a Q-factor for each state-action pair, which quickly becomes impractical as the state space increases. In some specific cases, the Q-factors can be approximated via a neural network or other architectures [2], with limited success.

In this paper we study the problem of adaptive optimization of parameterized Markov reward processes. Specifically, we study on-line optimization with respect to the average reward criterion of Markov reward processes, where the transition probabilities, as well as the expected reward per stage, are functions of two sets of parameters: 1) a set of tunable parameters, and 2) a set of fixed but unknown parameters. We provide sufficient convergence criteria for estimation procedures that are based on the observation of regenerative cycles, and study the use of such estimation procedures to carry out simultaneous optimization under parameter certainty for existing simulation-based algorithms, such as the algorithm of [1]. Sufficient conditions are also provided for the simultaneous optimization-estimation algorithms. We illustrate the use of these criteria in an application to optimal pricing of multi-class loss networks for several scenarios of parameter uncertainty.

The paper is organized as follows. In Section II we present the problem of adaptive Markov control, and address the problem of estimation of unknown parameters. In Section III we discuss the use of the estimation algorithms in the algorithms of [1], and provide sufficient conditions for the convergence of the adaptive scheme to a local optimized/true parameter value. One of the key benefits of this adaptive algorithm is that it can be implemented in an on-line fashion given its small memory and computation requirements. In Section IV we introduce the problem of pricing services in a multi-class loss network, and verify the conditions for the convergence of the adaptive algorithm introduced in Section III. This section ends with a numerical example, where the adaptive algorithm is used to find optimal prices in a scenario

Enrique Campos-Náñez is with the Department of Engineering Management and Systems Engineering, The George Washington University, 1776 G Street Washington, DC, 20052, USA ecamposn@gwu.edu

Stephen D. Patek is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904, USA patek@virginia.edu

0-7803-9568-9/05/$20.00 ©2005 IEEE

where the service rates are unknown.

II. Adaptive Control of Markov Chains

Consider a discrete-time Markov chain $\{i_n\}_{n=0}^{\infty}$, with finite state space $S = \{1, 2, \ldots, N\}$. The evolution of the system depends on a continuous parameter vector $u \in \Re^K$. The system behavior depends also on a set of unknown, but fixed parameters $\theta^* \in \Theta \subset \Re^l$, where Θ is a compact set. In this way, the transition probabilities

$$p_{ij}(u, \theta^*) = P(i_{n+1} = j \mid i_n = i, u, \theta^*),$$

are functions of the control vector u, and the parameters θ^*. We let $P(u, \theta^*)$ denote the matrix with entries $p_{ij}(u, \theta^*)$. Also, the expected cost per transition observed at state i, denoted by $g_i(u, \theta^*)$ is a function of both u, and θ^*.

The optimization problem to be addressed is to find the *average reward problem* defined as

$$\lambda^*(i) = \max_{u \in \Re^K} \liminf_{T \to \infty} \frac{1}{T} E \left[\sum_{n=0}^{T} g_{i_n}(u, \theta^*) \middle| i_0 = i \right]. \quad (1)$$

The problem is twofold. On one hand it is necessary to estimate the unknown parameter θ^*; on the other hand, the optimization problem has to be solved. While is possible to estimate the parameters offline, and solve the optimization problem using this knowledge, this paper will focus on tackling the two problems at the same time, in order to obtain a greater degree of adaptation.

A. Structural Assumptions

To set the stage for a discussion of the adaptive algorithm, it is convenient at this point to state the following structural assumptions, which are similar to those commonly used in the adaptive Markov control literature (see for example [22]).

Assumption 1: The Markov chain defined by $P(u, \theta)$ is irreducible for all values of $u \in \Re^K$ and $\theta \in \Theta$, and aperiodic for all $u \in \Re^K$.

An immediate consequence of this assumption is that for all $u \in \Re^K$ and $\theta^* \in \Theta$, all states are positive recurrent and the steady-state distribution $\pi(u, \theta^*)$ exists, and hence the average reward associated with control vector u and unknown parameter vector θ^*, is given by

$$\lambda(u, \theta^*) = \sum_{i \in S} \pi(u, \theta^*) g_i(u, \theta^*).$$

Assumption 2: The functions $P(u, \theta)$, and $g_i(u, \theta)$ are twice differentiable with bounded first and second derivatives both in u, and θ.

An immediate consequence of Assumption 2 is that both g and p are Lipschitz continuous on θ.

Assumption 3 (Observability): For all $i, j \in S$, either $p_{ij}(u, \theta) = 0$ for all $u \in \Re^K$, and $\theta \in \Theta$ or $p_{ij}(u, \theta) \geq \epsilon > 0$, for all $u \in \Re^K$, and $\theta \in \Theta$.

This assumption is commonly used in the Adaptive Markov Control literature. It states that if a transition can be observed at a particular control vector u, and vector θ^*, and it has to be observable for any other values for u and θ^* with some probability uniformly bounded below by $\varepsilon > 0$ (see for example [22]).

Finally, with respect to an estimation process, we assume the following properties of an estimator function $\mathcal{E}$. Later in this section we illustrate two instances of such functions, and illustrate its use in a network control problem in Section IV-B.

Assumption 4: Given a regenerative cycle $\{i_0, i_1, \ldots, i_T = i_0\}$, the function

$$\mathcal{E}(u, \tilde{\theta}) = \mathcal{E}(u, \tilde{\theta}; i_0, i_1, \ldots, i_T),$$

and its expected value $e(u, \tilde{\theta}) = E\left[\mathcal{E}(u, \tilde{\theta})\right]$ are such that

1) Function $\mathcal{E}(u, \tilde{\theta})$ is Lipschitz continuous with respect to the estimate $\tilde{\theta}$ for all values $u \in \Re^K$.
2) The ODE associated with the expected direction of $\mathcal{E}(u, \tilde{\theta})$
$$\dot{\theta} = e(u, \theta),$$
is asymptotically stable around θ^* for all $u \in \Re^K$.

The following are some examples of such functions $\mathcal{E}(u, \tilde{\theta})$.

Example 1: **(Ideal Estimator)** Suppose that there exists a function $\bar{\theta}(u; i_0, \ldots, i_T)$ such that

1) It provides an unbiased estimate, i.e. $E[\bar{\theta}(u; i_0, \ldots, i_T)] = \theta^*$.
2) It has a bounded second moment $E[\|\bar{\theta}(u; i_0, \ldots, i_T)\|^2] < \infty$.

Under this conditions, we can define $\mathcal{E}(u, \tilde{\theta}; i_0, i_1, \ldots, i_T) = \bar{\theta}(u; i_0, \ldots, i_T) - \tilde{\theta}$, which satisfies Assumption 4, since $E[\mathcal{E}(u, \tilde{\theta})] = \theta^* - \tilde{\theta}$, and its second moment is bounded.

Example 2: **(Maximum Likelihood Estimator)** Given a regenerative cycle $\{i_0, i_1, \ldots, i_T = i_0\}$, a maximum likelihood estimator of θ^* is any solution to the optimization problem

$$\max_{\tilde{\theta} \in \Theta} \prod_{i=1}^{T} p_{i_{n-1} i_n}(u, \tilde{\theta}).$$

Consider the function

$$\mathcal{L}(u, \tilde{\theta}) = \mathcal{L}(u, \tilde{\theta}; i_0, i_1, \ldots, i_T) = \frac{1}{T} \sum_{k=1}^{T} \frac{\nabla_{\tilde{\theta}} p_{i_{n-1} i_n}(u, \tilde{\theta})}{p_{i_{n-1} i_n}(u, \tilde{\theta})},$$

which is the gradient with respect to the $\tilde{\theta}$ of the logarithm of the likelihood of the path under $\tilde{\theta}$, up to the factor $\frac{1}{T}$. As we will show in the next session, Assumption 2 implies that the function $\mathcal{E}(u, \tilde{\theta}) = \mathcal{L}(u, \tilde{\theta})$ is Lipschitz continuous. In some cases that will be discussed in Section IV-B, function $\mathcal{L}(u, \tilde{\theta})$ also satisfies the condition of asymptotic stability of Assumption 4.

B. Preliminaries

First, we note that functions $p_{ij}(u, \theta)$ and $g_i(u, \theta)$ are Lipschitz continuous, as a consequence of our Assumption 2 (see [30]). Second, we show that our estimator function $\mathcal{L}(u, \tilde{\theta})$ is also Lipschitz continuous.

Lemma 1: Under Assumptions 1-3, we have that $\mathcal{L}(u_m, \tilde{\theta}_m)$ is Lipschitz continuous.

Proof: Since each term of the sum $\mathcal{L}(u_m, \tilde{\theta}_m)$ is a function with continuous and bounded derivative, we can rewrite

$$\|\mathcal{L}(u_m, \theta_2) - \mathcal{L}(u_m, \theta_1)\| \leq K\|\theta_2 - \theta_1\|,$$

for all θ_1,$\theta_2 \in \Theta$. More specifically, we have that the difference $\|\mathcal{L}(u_m, \theta_2) - \mathcal{L}(u_m, \theta_1)\| =$

$$\begin{aligned}&\frac{1}{T_m}\| \sum_{n=t_m}^{t_{m+1}} L_{i_{n-1}i_n}(u_m, \theta_2) - L_{i_{n-1}i_n}(u_m, \theta_1)\| \\ &\leq \frac{T_m}{T_m} K\|\theta_2 - \theta_1\| = K\|\theta_2 - \theta_1\|,\end{aligned}$$

where the inequality follows from the fact that $L_{ij}(u, \theta)$ is differentiable with bounded derivative. ■

C. Estimation

Suppose we have access to observations of a system as described in the previous section. Select a state $i^* \in S$, such that i^* is recurrent for all $u \in \Re^K$, and all $\tilde{\theta}$. Let t_m mark the times of regeneration. We observe a regenerative cycles $\{i_{t_m}, i_{t_m+1}, \ldots, i_{t_{m+1}}\}$, realized under the control parameters u_m, and consider the stochastic approximation recursion

$$\tilde{\theta}_{m+1} = \Pi_\Theta \left[\tilde{\theta}_m + \gamma_m \mathcal{E}(u_m, \tilde{\theta}_m)\right], \tag{2}$$

where $\mathcal{E}(u_m, \tilde{\theta}_m) = \mathcal{E}(u_m, \tilde{\theta}_m; i_{t_m}, i_{t_m+1}, \ldots, i_{t_{m+1}})$ is a function satisfying Assumption 4, and γ_m is a set of stepsizes satisfying the so-called standard conditions:

Assumption 5: The stepsizes γ_m are such that

$$\sum_m \gamma_m = \infty, \quad \sum_m \gamma_m^2 < \infty.$$

Corollary 1: Under Assumptions 1-4, and 5, the recursion of Eqn. 2 is such that

$$\tilde{\theta}_m \to \theta^*,$$

as $m \to \infty$, and it does so at a rate of $O(\frac{1}{\sqrt{m}})$.

Proof: Under Assumptions 1-5, and 4, the update direction function holds the Assumptions on the result by [31] for Lipschitz functions. With the additional assumption of uniqueness of the stability of u^* under, we have the desired result. ■

Another consequence relevant to the use of this estimation procedure in an optimization context, is the summability of the errors, which is our following result.

Corollary 2: If Assumptions 1-5 hold, then

$$\sum_m \gamma_m \|\tilde{\theta}_m - \theta^*\| < \infty.$$

Proof: The fact that $E[\|\tilde{\theta}_m - \theta^*\|] < \frac{C}{\sqrt{m}}$ for some constant C, results in $\sum_m E[\gamma_m \|\tilde{\theta}_m - \theta^*\|] < \infty$, which together with the boundedness of the second moment, i.e.

$$\sum_m \gamma_m^2 \|\tilde{\theta}_m - \theta^*\| < \infty,$$

implies the desired result. ■

The following result will be useful in the following.

Lemma 2: Under Assumptions 1, we have that

$$E[\mathcal{L}(u, \tilde{\theta})] = \sum_{i \in S} \pi_i(u) \sum_{j \in S} p_{ij}(u, \theta^*) \frac{\nabla_{\tilde{\theta}} p_{ij}(u, \tilde{\theta})}{p_{ij}(u, \tilde{\theta})}, \tag{3}$$

where $\pi(u) = (\pi_1(u), \cdots, \pi_N(u))$ is the steady state distribution of the chain under control u.

Proof: We can make use of the uniform irreducibility of the Markov chain by defining K different Markov control process with the same states and transition probabilities, but the expected cost per stage for the k-th is denoted

$$\tilde{g}_i^k(u, \tilde{\theta}) = \sum_{j \in S} \frac{p_{ij}(u, \theta^*)}{p_{ij}(u, \tilde{\theta})} \frac{\partial p_{ij}(u, \tilde{\theta})}{\partial \tilde{\theta}_k}.$$

The average reward attained by such reward processes is given by

$$E[\mathcal{L}(u, \tilde{\theta})]_k = \sum_{i \in S} \pi_i(u) \tilde{g}_i^k(u, \tilde{\theta}), \quad \forall u \in \Re^K, \tilde{\theta} \in \Theta,$$

from which the result follows ■

Later in Section IV we show an example where Assumption 4 can be verified with the help of this result on a relevant application.

III. Adaptive Optimization

We introduce a simple modification to the "batch" algorithm of [1]. The modified algorithm achieves the purpose of optimizing the average reward of the system, while discovering the value θ^* of the unknown parameters. The procedure starts with an initial estimate $\tilde{\theta}_0$ of the unknown parameter θ^*, an initial guess of the control parameters u_0, and an estimate $\tilde{\lambda}_0$ of $\lambda(u, \theta)$. Starting at state $i^* \in S$, a special state recurrent under all u, simulate the process under parameter vector u_0 until we return to state i^*, known as a regenerative cycle. Let t_m be the time of the m-th return to the special state i^*. At those times, new values for u_m, $\tilde{\lambda}_m$, and $\tilde{\theta}_m$ are calculated with the recursive formula

$$u_{m+1} = u_m + \gamma_m F(u_m, \tilde{\lambda}_m, \tilde{\theta}_m), \tag{4}$$

$$\tilde{\lambda}_{m+1} = \tilde{\lambda}_m + \eta\gamma_m \sum_{n=t_m}^{t_{m+1}-1} (g_{i_n}(u_m, \tilde{\theta}_m) - \tilde{\lambda}_m), \tag{5}$$

$$\tilde{\theta}_{m+1} = \Pi_\Theta[\tilde{\theta}_m + \mu\gamma_m \mathcal{E}(u_m, \tilde{\theta}_m)], \tag{6}$$

where $\eta > 0$, and $\mu > 0$ are stepsize factors, the function

$$\begin{aligned}F(u_m, \tilde{\lambda}_m, \tilde{\theta}_m) = \sum_{n=t_m}^{t_{m+1}-1} &[\tilde{v}_{i_n}(u_m, \tilde{\theta}_m) L_{i_{n-1}i_n}(u_m, \tilde{\theta}_m) \\ &+ \nabla g_{i_n}(u_m, \tilde{\theta}_m)],\end{aligned}$$

and $\tilde{v}_{i_n}(u_m, \tilde{\theta}_m)$ is an approximation to the cost-to-go function to return to state i^*, under parameter values u_m. Specifically

$$\tilde{v}_{i_n}(u_m, \tilde{\theta}_m) = \sum_{k=n}^{t_{m+1}-1} (g_{i_k}(u_m, \tilde{\theta}_m) - \tilde{\lambda}_m).$$

Even though the chain is different to the one being simulated (under the correct values θ^*), it will be shown that the difference is asymptotically negligible under Assumptions 1-3, in the sense that the stability of the algorithm is not compromised.

This recursive procedure coincides with the algorithm introduced by Marbach and Tsitsiklis in [1], differing only in our use of estimates $\tilde{\theta}_m$ of the unknown values θ^* (instead of assuming perfect knowledge of θ^* as in theirs), and the introduction of the estimation recursion of Eqn. 6.

A. Convergence of the Algorithm

In this section, it will be shown that under Assumptions 1-3 the average behavior of the procedure converges asymptotically to the one with full knowledge of the parameters, and the error obtained in such process is asymptotically negligible. To set the stage for the discussion, let $r_m = (u_m, \tilde{\lambda}_m, \tilde{\theta}_m)$, and let

$$H(r_m) = \left[\begin{array}{c} F(u_m, \tilde{\lambda}_m, \tilde{\theta}_m) \\ \eta \sum_{n=t_m}^{t_{m+1}-1} (g_{i_n}(u_m, \tilde{\theta}_m) - \tilde{\lambda}_m) \\ \mu \mathcal{E}(u_m, \tilde{\theta}_m) \end{array} \right], \qquad (7)$$

with which the algorithm can be rewritten as:

$$r_{m+1} = r_m + \gamma_m H(r_m),$$

without consideration of the projection. Furthermore, let $h(r_m) = E[H(u_m, \tilde{\lambda}_m, \theta^*) \mid \mathcal{F}_m]$ with respect to the filtration $\mathcal{F}_m = \sigma(u_0, \tilde{\lambda}_0, \tilde{\theta}_0, i_0, i_1, \ldots, i_{t_m})$. Notice that the expectation is taken assuming knowledge of θ^*. Using the standard ODE approach, rewrite

$$r_{m+1} = r_m + \gamma_m h(r_m) + \varepsilon_m, \qquad (8)$$

where $\varepsilon_m = \gamma_m(H(r_m) - h(r_m))$.

Lemma 3: $H(u, \tilde{\lambda}, \tilde{\theta})$ is differentiable with respect to $\theta, \forall \theta \in \Theta$, and there exists a constant C such that $E[\|\nabla H(u, \tilde{\lambda}, \tilde{\theta})\|] \leq C$ for all u, $\tilde{\lambda}$, and $\tilde{\theta}$.

Proof: The first component of $H(u, \tilde{\lambda}, \tilde{\theta})$, $F(u, \tilde{\lambda}, \tilde{\theta})$, is a sum of differentiable functions g, ∇g, and $L_{ij}(u, \theta)$. For the latter, Assumption 3 guarantees that the function will be differentiable, since is the quotient of differentiable function, while the function in the denominator is always different than zero. Therefore, $F(u, \tilde{\lambda}, \tilde{\lambda})$ is differentiable.

Notice that the derivative of $L_{ij}(u, \theta)$ is also bounded as a consequence that $p_{ij}(u, \theta) > \epsilon$. A consequence of positive recurrence, is that $E[T] < \infty$. Therefore the expected value of the gradient of $F(u, \tilde{\lambda}, \tilde{\theta})$ is a finite sum of bounded terms, which guarantees is bounded. Similar arguments can be made for the other components of $H(u_m, \tilde{\lambda}, \tilde{\theta})$. ■

Since function H is differentiable, we can rewrite

$$H(r_m) = H(u_m, \tilde{\lambda}_m, \theta^*) + \nabla_{\theta^*} H(u_m, \tilde{\lambda}, \theta') \cdot (\tilde{\theta}_m - \theta^*),$$

for some $\theta' \in \Theta$, by the Mean Value Theorem. This leads to the following result.

Lemma 4: Under Assumptions 4-5,

$$\sum_m \|\varepsilon_m\| < \infty,$$

with probability one.

Proof: By Lemma 3, an

$$\begin{aligned} E[\|\varepsilon_m\|] &= E[\|\gamma_m(H(u_m, \tilde{\lambda}_m, \theta^*) \\ &\quad + \nabla_\theta H(u_m, \tilde{\lambda}, \theta') \cdot (\tilde{\theta}_m - \theta^*) - h(r_m))\|] \\ &= E[\|\gamma \nabla_\theta H(u_m, \tilde{\lambda}, \theta') \cdot (\tilde{\theta}_m - \theta^*)\|] \\ &\leq \gamma_m C \|\tilde{\theta}_m - \theta^*\|. \end{aligned}$$

Therefore, by Corollary 2, we have that $\sum_m E[\|\varepsilon_m\|] < \infty$, which along with the summability of the second moment, implies the desired result. ■

Proposition 1: The recursive procedure described in 4-6 converges with probability one to $(u, \tilde{\lambda}, \theta^*)$, where $\nabla_u \lambda(u) = 0$, and $\tilde{\lambda} = \lambda(u)$, under Assumptions 1-5.

Proof: Notice that by rewriting the recursive procedure as 8, the driving direction of the process is the same as if the parameter was known. Therefore, the stability analysis of [1] holds. To see this clearly, it is possible to rewrite 8 as

$$r_{m+1} = r_m + \gamma_m h(r_m) + \varepsilon'_m + \varepsilon''_m,$$

where $\varepsilon'_m = \gamma_m(H(r_m) - H(u_m, \tilde{\lambda}_m, \tilde{\theta}_m)$, and $\varepsilon''_m = \gamma_m(H(u_m, \tilde{\lambda}_m, \tilde{\theta}_m) - h(r_m))$, and notice that by omitting ε'_m the recursion

$$r_{m+1} = r_m + \gamma_m h(r_m) + \varepsilon''_m,$$

which is equal to the one in [1]. In other terms, the systems being analyzed is a perturbed version of the algorithm where all parameters are known, but this perturbation is asymptotically negligible ■

IV. Pricing of Multi-class Loss Networks: an Illustration

As an application, we study a problem of optimal pricing services in a multi-class loss network. Consider a multi-class loss system with a set N of resources with capacity C_n, $n = 1, 2, \ldots, N$ shared by users of K classes of calls or services. Each service class $k \in \{1, 2, \ldots, K\}$ can be described by a Poisson arrival rate function $\alpha_k(u_k, \theta^*)$, which we assume to be a function of a price $u_k \in (0, \bar{u}_k)$ [1], and a set of unknown but fixed parameters θ^* that belongs to Θ, a compact convex set. We assume $\alpha_k(u_k, \theta)$ is a bounded function of both u_k and θ, with bounded first and second derivatives with respect to both u_k, and θ, for all $\theta \in \Theta$. We assume that requests for calls constitute a Poisson arrival process with rate $\alpha(u_k, \theta^*)$ whenever price u_k is used.

We assume that $\alpha_k(\bar{u}_k, \theta^*) > 0$ for all $k = \{1, 2, \ldots, K\}$. Let $\mathcal{U} = \prod_{k=1}^K (0, \bar{u}_k)$ denote the set of all feasible price combinations. The duration of a call of class k is assumed to be exponentially distributed with mean $1/\beta_k(\theta^*)$, during which the network reserves $\mathbf{m}_k = (m_{k,1}, m_{k,2}, \ldots, m_{k,N})$ out of the capacity of the corresponding resources $C_1, C_2, \ldots, C_N$. Let M be the matrix of resources used by each type of call, i.e. a matrix whose column k is vector $\mathbf{m}_k$. We assume that $\beta(\theta)$ is bounded, with bounded first and second derivatives with respect to

[1] Note that an open interval can be continuously mapped to the real line without sacrificing convergence.

$\theta \in \Theta$. At a time t, the state of the system can be described as $i(t) = (i_1(t), i_2(t), \ldots, i_K(t))$, a vector that represents the number of ongoing calls of class k present in the system. We use I to denote the set of all feasible usage profiles $(i_!, i_2, \ldots, i_K)$.

Finally, we assume that the system implements a strict admission control rule. Suppose that at time t a request for a call of type k is received. The call will be admitted if $M \cdot i(t) + \mathbf{m}_k \leq C$, where matrix M and vector $\mathbf{m}_k$ are defined as above. To simplify notation, define $A(i) = \{k | \, M \cdot i + \mathbf{m}_k \leq C\}$ to be the set of calls that satisfy the admission control rule when the system is in state $i \in I$. We also define $a_k(i) = \mathbf{1}_{k \in A(i)}$, to indicate if class k is admissible at state i.

A. Formulation as a Markov Control Process

We focus on the so-called static pricing policies, which have recently been shown to be asymptotically optimal for scaled versions of this system (see [32]–[34]). In the static pricing formulation, the same class-dependent price u_k always applies regardless of the state of the process. The revenue maximization problem then can be expressed as

$$\max_{u \in \mathcal{U}} \lambda(u) = \lim_{t \to \infty} \frac{1}{T} E \left[\int_0^T \sum_{k \in A(i(t))} \alpha_k(u_k, \theta^*) u_k dt \right]. \tag{9}$$

In order to apply the algorithms of Section III we transform the continuous-time process into a discrete one through the process of uniformization, which is described in [35]. At a particular state $i = (i_1, i_2, \ldots, i_K)$ and under prices $u = (u_1, u_2, \ldots, u_K)$ and θ^*, the transition rate out of state i is given by

$$\nu_i(u, \theta^*) = \sum_{k \in A(i)} \alpha_k(u_k, \theta^*) + \sum_{k=1}^{K} i_k \beta_k(\theta^*).$$

In order to apply the process of uniformization, we need the following assumption.

Assumption 6: The functions $\alpha_k(\cdot, \cdot)$ and $\beta(\cdot)$ are such that there exists a constant ν^* such that

$$\nu^* \geq \nu_i(u, \tilde{\theta}), \quad \forall i \in S, \tilde{\theta} \in \Theta, u \in \Re^K.$$

Through uniformization, we obtain the transition probabilities

$$p_{ij}(u) = \begin{cases} \frac{i_k \beta_k(\theta^*)}{\nu^*} & \text{if } j = i - e_k, k = 1, ..., n, \\ \frac{\alpha_k(u_k, \theta^*)}{\nu^*} & \text{if } j = i + e_k, k = 1, ..., n, \\ 1 - \frac{\nu_i(u, \theta^*)}{\nu^*} & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Also, the expected reward per stage under parameter u when visiting state i is

$$g_i(u, \theta^*) = \frac{1}{\nu^*} \sum_{k \in A(i)} \alpha_k(u_k, \theta^*) u_k.$$

Assumption 7: Suppose

$$\theta^* = (\alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*, \ldots, \alpha_K^*, \beta_K^*).$$

Furthermore, suppose that the arrival rate functions $\alpha_k(u_k, \theta^*) = \alpha_k(u_k, \alpha_k^*)$, service rates $\beta_k(\beta_k^*)$ for class k are monotonic and concave functions of α_k^*, and β_k^*, respectively.

Proposition 2: Under Assumptions 6 and 7 the iterative procedure described in Eqns. 4-6 converges to the true values θ^* and to a local maximizer of $\lambda(u)$.

Proof: First, notice that when the system is at a state i, the only possible events are arrivals of all classes $k \in A(i)$, and departures of users currently in the system. According to Eqn. 3, we have that $E[\mathcal{L}(u, \tilde{\theta})] =$

$$\begin{aligned} &= \sum_{i \in S} \pi_i(u) \sum_{j \in S} p_{ij}(u, \theta^*) \frac{\nabla_{\tilde{\theta}} p_{ij}(u, \tilde{\theta})}{p_{ij}(u, \tilde{\sum})} \\ &= \sum_{i \in S} \pi_i(u) \times \\ &\times \begin{bmatrix} a_1(i) \frac{\partial \alpha_1(u_1, \tilde{\alpha}_1)}{\partial \tilde{\alpha}_1} \left(\frac{\alpha_1(u_1, \alpha_1^*)}{\alpha_1(u_1, \tilde{\alpha}_1)} - \frac{\nu^* - \nu_i(u, \theta^*)}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ i_1 \frac{\partial \beta_1(\tilde{\beta}_1)}{\partial \tilde{\beta}_1} \left(\frac{\beta_1(\beta_1^*)}{\beta_1(\tilde{\beta}_1)} - \frac{\nu^* - \nu_i(u, \theta^*)}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ \vdots \\ a_K(i) \frac{\partial \alpha_K(u_K, \tilde{\alpha}_K)}{\partial \tilde{\alpha}_K} \left(\frac{\alpha_K(u_K, \alpha_K^*)}{\alpha_K(u_K, \tilde{\alpha}_K)} - \frac{\nu^* - \nu_i(u, \theta^*)}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ i_K \frac{\partial \beta_K(\tilde{\beta}_K)}{\partial \tilde{\beta}_K} \left(\frac{\beta_K(\beta_K^*)}{\beta_K(\tilde{\beta}_K)} - \frac{\nu^* - \nu_i(u, \theta^*)}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \end{bmatrix} \end{aligned}$$

By defining $\Delta \nu_i(u, \tilde{\theta}) = \nu_i(u, \theta^*) - \nu_i(u, \tilde{\theta})$, we can rewrite $E[\mathcal{L}(u, \tilde{\theta})] =$

$$\begin{aligned} &= \sum_{i \in S} \pi_i(u) \times \\ &\times \begin{bmatrix} a_1(i) \frac{\partial \alpha_1(u_1, \tilde{\alpha}_1)}{\partial \tilde{\alpha}_1} \left(\frac{\alpha_1(u_1, \alpha_1^*)}{\alpha_1(u_1, \tilde{\alpha}_1)} - 1 + \frac{\Delta \nu_i(u, \tilde{\theta})}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ i_1 \frac{\partial \beta_1(\tilde{\beta}_1)}{\partial \tilde{\beta}_1} \left(\frac{\beta_1(\beta_1^*)}{\beta_1(\tilde{\beta}_1)} - 1 + \frac{\Delta \nu_i(u, \tilde{\theta})}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ \vdots \\ a_K(i) \frac{\partial \alpha_K(u_K, \tilde{\alpha}_K)}{\partial \tilde{\alpha}_K} \left(\frac{\alpha_K(u_K, \alpha_K^*)}{\alpha_K(u_K, \tilde{\alpha}_K)} - 1 + \frac{\Delta \nu_i(u, \tilde{\theta})}{\nu^* - \nu_i(u, \tilde{\theta})} \right) \\ i_K \frac{\partial \beta_K(\tilde{\beta}_K)}{\partial \tilde{\beta}_K} \left(\frac{\beta_K(\beta_K^*)}{\beta_K(\tilde{\beta}_K)} - 1 + \frac{\Delta \nu_i(u, \tilde{\theta})}{\theta^* - \nu_i(u, \tilde{\theta})} \right) \end{bmatrix} \end{aligned}$$

Noting that as a consequence of Assumption 7, we have that for any $\tilde{\theta} \in \Theta$, and $k = 1, 2, \ldots, K$

$$\frac{\partial \alpha_k(u_k, \tilde{\alpha}_k)}{\partial \tilde{\alpha}_k} \left(\frac{\alpha_k(u_k, \alpha_k^*)}{\alpha_k(u_k, \tilde{\alpha}_k)} - 1 \right) (\alpha_k^* - \tilde{\alpha}_k) \geq 0,$$
$$\frac{\partial \beta_k(\tilde{\beta}_k)}{\partial \tilde{\beta}_k} \left(\frac{\beta_k(\beta_k^*)}{\beta_k(\tilde{\beta}_k)} - 1 \right) (\beta_k^* - \tilde{\beta}_k) \geq 0.$$

Therefore $E[\mathcal{L}(u,\tilde{\theta})] \cdot (\theta^* - \tilde{\theta}) =$

$$\sum_{i \in S} \pi_i(u)[\sum_{k \in A(i)} \frac{\partial \alpha_k(u_k, \tilde{\alpha}_k)}{\partial \tilde{\alpha}_k} \left(\frac{\alpha_k(u_k, \alpha_k^*)}{\alpha_k(u_k, \tilde{\alpha}_k)} - 1 \right) (\alpha_k^* - \tilde{\alpha}_k)$$

$$+ \sum_{k=1}^{K} i_k \frac{\partial \beta_k(\tilde{\beta}_k)}{\partial \tilde{\beta}_k} \left(\frac{\beta_k(\beta_k^*)}{\beta_k(\tilde{\beta}_k)} - 1 \right) (\beta_k^* - \tilde{\beta}_k)$$

$$+ \frac{\Delta \nu_i(u, \tilde{\theta})}{\nu^* - \nu_i(u, \tilde{\theta})} \{ \sum_{k \in A(i)} \frac{\partial \alpha_k(u_k, \tilde{\alpha}_k)}{\partial \tilde{\alpha}_k} (\alpha_k^* - \tilde{\alpha}_k)$$

$$+ \sum_{k=1}^{K} i_k \frac{\partial \beta_k(\tilde{\beta}_k)}{\partial \tilde{\beta}_k} (\beta_k^* - \tilde{\beta}_k) \}]$$

By concavity we have that

$$\begin{aligned} \frac{\partial \alpha_k(u_k, \tilde{\alpha}_k)}{\partial \tilde{\alpha}_k} (\alpha_k^* - \tilde{\alpha}_k) &\geq \frac{\partial \alpha_k(u_k, \alpha_k')}{\partial \tilde{\alpha}_k} (\alpha_k^* - \tilde{\alpha}_k), \forall \alpha_k' \\ \frac{\partial \beta_k(\tilde{\beta}_k)}{\partial \tilde{\beta}_k} (\beta_k^* - \tilde{\beta}_k) &\geq \frac{\partial \beta_k(\beta_k')}{\partial \tilde{\beta}_k} (\beta_k^* - \tilde{\beta}_k), \forall \beta_k'. \end{aligned}$$

In particular by the mean value theorem, there are α_k'', and β_k'' such that

$$\begin{aligned} \frac{\partial \alpha_k(u_k, \alpha_k'')}{\partial \tilde{\alpha}_k} (\alpha_k^* - \tilde{\alpha}_k) &= \alpha_k(u_k, \alpha_k^*) - \alpha_k(u_k, \tilde{\alpha}_k) \\ \frac{\partial \beta_k(\beta_k'')}{\partial \tilde{\beta}_k} (\beta_k^* - \tilde{\beta}_k) &= \beta_k(\beta_k^*) - \beta(\tilde{\beta}_k) \end{aligned}$$

Therefore $E[\mathcal{L}(u,\tilde{\theta})] \cdot (\theta^* - \tilde{\theta}) \geq$

$$\sum_{i \in S} \pi_i(u)[\sum_{k \in A(i)} \frac{\partial \alpha_k(u_k, \tilde{\alpha}_k)}{\partial \tilde{\alpha}_k} \left(\frac{\alpha_k(u_k, \alpha_k^*)}{\alpha_k(u_k, \tilde{\alpha}_k)} - 1 \right) (\alpha_k^* - \tilde{\alpha}_k)$$

$$+ \sum_{k=1}^{K} i_k \frac{\partial \beta_k(\tilde{\beta}_k)}{\partial \tilde{\beta}_k} \left(\frac{\beta_k(\beta_k^*)}{\beta_k(\tilde{\beta}_k)} - 1 \right) (\beta_k^* - \tilde{\beta}_k)$$

$$+ \frac{\Delta \nu_i(u, \tilde{\theta})^2}{\nu^* - \nu_i(u, \tilde{\theta})}] \geq 0$$

Notice that the equality will only hold when $\tilde{\theta} = \theta^*$. Therefore, the function $\mathcal{L}(u, \tilde{\theta})$ is asymptotically stable around θ^*, since for the Lyapunov function $V(\tilde{\theta}) = \|\theta^* - \tilde{\theta}\|^2$, we have that

$$E[\mathcal{L}(u, \tilde{\theta})] \cdot \nabla_{\tilde{\theta}} V(\tilde{\theta}) = E[\mathcal{L}(u, \tilde{\theta})] \cdot 2(\tilde{\theta} - \theta^*) < 0,$$

for all $\tilde{\theta} \in \Theta, \tilde{\theta} \neq \Theta$, and $V(\tilde{\theta}) = 0$ if $\tilde{\theta} = \theta^*$ ■

The result shows that the estimator $\mathcal{L}(u, \tilde{\theta})$ is indeed asymptotically stable around θ^*, and hence the results of Proposition 1 apply.

B. Example: Adapting to Unknown Service Rates

Suppose that the unknown parameters in this case are the service rates $\beta_1, \ldots, \beta_K$. In this case, we define $\theta^* = \beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_K^*)$, and $\beta_k(\theta^*) = \beta_k^*$, for all $k = 1, 2, \ldots, K$. In this case, we will use the estimates $\tilde{\theta} = \tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_K)$, and there is no dependence of the arrival rates functions on $\tilde{\theta}_k$, and the functions satisfy Assumption 7.

To illustrate the procedure numerically, the Algorithm of Eqns. 4-6 is applied to estimate class services rates, using the estimation function $\mathcal{L}(u, \tilde{\theta})$. In the example, three classes of traffic share a small network of three links with equal capacities $C_n = 10$, for $n = 1, 2, 3$. The three classes of traffic have identical arrival rate functions and service time parameters, and differ only in the routes they use. Each class uses a unique two link path from the network and requires 1 unit of the capacity of the link. The route parameters are described at the bottom of Table I. The service rates are $\beta_1 = \beta_2 = \beta_3 = 5$ calls per second, and the arrival functions are defined by $\alpha_k(u_k) = \bar{\alpha}_k(1 - u_k/\breve{u}_k)_+$. The value of these parameters are defined in Table I and are time-invariant.

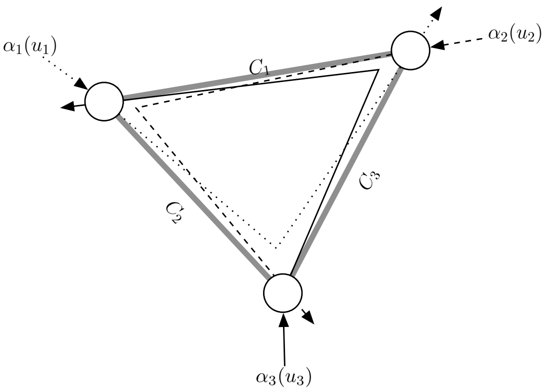

Fig. 1. Topology for the numerical example: three classes with identical arrival rate function $\alpha_k(u_k)$ and service parameters $\beta_k(\theta^*) = \beta_k^*$, for $k = 1, 2, 3$.

TABLE I
PARAMETERS FOR EXAMPLE IV-B. ($C_1 = C_2 = C_3 = 10$).

Parameter	Class 1	Class 2	Class 3
$\bar{\alpha}_k$	50.0	50.0	50.0
$\breve{u}_k$	1.0	1.0	1.0
$\bar{u}_k$	.95	.95	.95
Requirements			
Link 1	1	1	0
Link 2	1	0	1
Link 3	0	1	1
β_k	5	5	5
$\tilde{\theta} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$	7.5	5	2.5

Given the symmetry of the traffic classes in Example IV-B, it is clear that any optimal solution has to be of the form $u_1 = u_2 = u_3$, since any solution where the equality does not hold will lead to lower utilization of some of the links. For example, if $u_1 > u_2, u_3$ then on average link 3 will be saturated, while the other two will be under utilized. As Figures 2 and 3 show, the prices for the three classes achieve a similar value (approximately $u_1 = u_2 = u_3 = .59$. Notice that the estimation process converges quickly to the correct values, as shown by Figure 4. The values of β_i are restricted to the ranges $[1, 10]$.

V. CONCLUSIONS

We have presented an algorithm for selecting optimal parameters for controllable Markov chains. The algorithm is robust to parametric uncertainty, in the sense that 1)

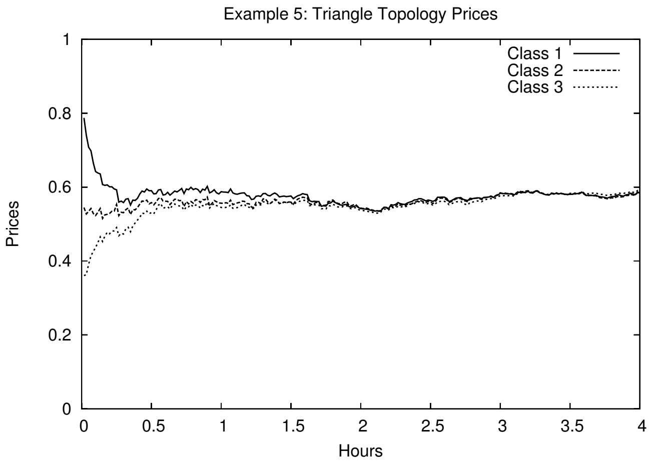

Fig. 2. Evolution of the prices for Example IV-B.

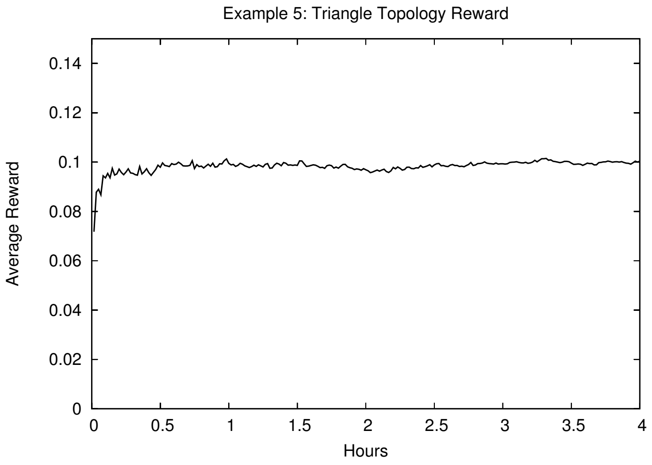

Fig. 3. Evolution of average reward for Example IV-B.

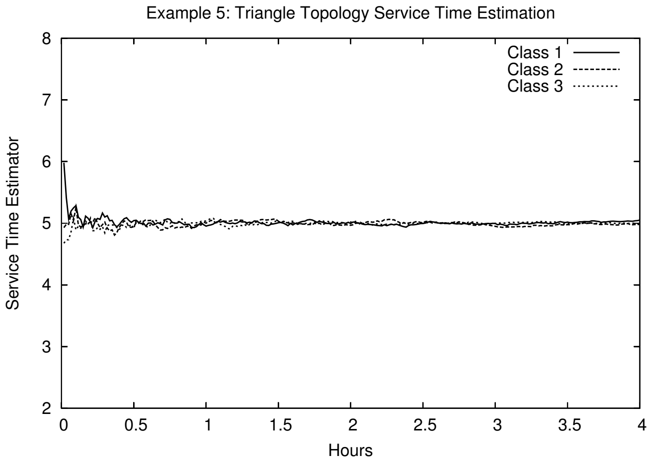

Fig. 4. Evolution of the estimates $\tilde{\theta} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$ for Example IV-B.

computes estimates of unknown but fixed parameters, and 2) maximizes the average reward by moving the tunable parameters in the approximate direction of the gradient of the objective function. We provide sufficient conditions for the convergence of such an algorithm. As was shown through numerical examples in [36], the algorithm can be used with constant stepsizes to track slow changing parameters, although we lose asymptotic stability, i.e. when the fixed but unknown parameters θ^* are replaced with a slowly changing set of parameters $\theta^*(t)$.

References

[1] P. Marbach and J. N. Tsitsiklis. Simulation-Based Optimization of Markov Reward Processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, February 2001.

[2] D. Bertsekas and J. S. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[3] A. G. Barto, S. J. Bradtke, and S. P. Singh. Neuron-like Elements that Can Solve Difficult Learning Control Problems. *IEEE Trans. on Ssytems, Man, and Cybernetics*, 13:835–846, 1983.

[4] A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to Act Using Real-Time Dynamic Programming. *Artificial Intelligence*, Special Volume: Computational Research on Interaction and Agency(72):81–138, 1995.

[5] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[6] J. N. Tsitsiklis and B. Van Roy. Average Cost Temporal-Difference Learning. *Automatica*, 35(11):1799–1808, November 1999.

[7] V. R. Konda and J. N. Tsitsiklis. Actor-Critic Algorithms. *submitted to SIAM Journal on Control and Optimization*, February 2001.

[8] M. C. Fu and Y. C. Ho. Using Perturbation Analysis for Gradient Estimation, Averaging, and Updating in a Stochastic Approximation Algorithm. *Proceedings of the 1988 Winter Simulation Conference*, pages 509–517, 1988.

[9] M. Fu and J. Q. Hu. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, 1997.

[10] F. J. Vazquez-Abad. Strong Points of Weak Convergence: A Study of Using RPA Gradient Estimation for Automatic Learning. *Automatica*, 35:1255–1274, 1999.

[11] P. Marbach and J. N. Tsitsiklis. Gradient-Based Optimization of Markov Reward Processes: Practical Variants. Technical report, Laboratory for Information and Decision Systems / MIT, March 2000.

[12] P. Marbach and J. N. Tsitsiklis. Approximate Gradient Methods in Policy-Space Optimization of Markov Reward Processes. *Journal of Discrete Event Dynamical Systems*, 13:111–148, 2003.

[13] X. Cao and H. Chen. Perturbation Realization, Potentials, and Sensitivity Analysis of Markov Processes. *IEEE Transactions on Automatic Control*, 42(10):1382–1393, 1997.

[14] X. Cao and Y. Wan. Algorithms for Sensitivity Analysis of Markov Systems Through Potentials and Perturbation Realization. *IEEE Transactions on Control Systems Technology*, 6(4):482–494, 1998.

[15] H. T. Fang, H. F. Chen, and X. R. Cao. Recursive Approaches for Single Sample Path Based Markov Reward Processes. *Asian Journal of Control*, 3(1):21–26, 2001.

[16] X.R. Cao. From Perturbation Analysis to Markov Decision Processes and Reinforcement Learning. *Discrete Event Dynamic Systems: Theory and Applications*, (13):9–39, 2003.

[17] H.T. Fang and X.R. Cao. Potential-Based On-line Policy Iteration Algorithms for Markov Decision Processes. *IEEE Transactions on Automatic Control*, 2004.

[18] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[19] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

[20] P. Glasserman and P. W. Glynn. Gradient Estimation for Regenerative Processes. *Proceedings of the Winter Simulation Conference*, 1992.

[21] P. Mandl. Estimation and Control in Markov Chains. *Advances in Applied Probability*, 6:40–60, 1974.

[22] V. S. Borkar and P. Varaiya. Adaptive Control of Markov Chains, I: Finite Parameter Set. *IEEE Transactions on Automatic Control*, AC(24):953–957, 1979.

[23] P. R. Kumar and A. Becker. A New Family of Optimal Adaptive Controllers for Markov Chains. *IEEE Transactions on Automatic Control*, AC-27(1), February 1982.

[24] B. Doshi. Strong Consistency of a Modified Maximum Likelihood Estimator for Controlled Markov Chains. *J. Appl. Prob.*, 17:726–734, 1980.

[25] Y. M. El-Fattah. Gradient approach for recursive estimation and control in finite Markov chains. *Advances in Applied Probability*, 13:778–803, 1981.

[26] V. S. Borkar and P. Varaiya. Identification and Adaptive Control of Markov Chains. *SIAM J. Control and Optimization*, 20(4):470–89, July 1982.

[27] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, 1989.

[28] Z. Ren and B. H. Krogh. Adaptive Control of Markov Chains with Average Cost. *IEEE Transactions on Automatic Control*, 46(4), April 2001.

[29] G. Santharam and P. S. Sastry. A reinforcement learning neural network for adaptive control of markov chains. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(5), September 1997.

[30] W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 1964.

[31] D. J. Ma, A. M. Makowski, and A. Shwartz. Stochastic Approximations for Finite-State Markov Chains. *Stochastic Processes and their Applications*, 35:27–45, 1990.

[32] X. Lin and N. B. Shroff. Simplification of Network Dynamics in Large Systems. Technical report, Purdue University, 2001.

[33] I. Ch. Paschalidis and Y. Liu. Pricing in Multiservice Loss Networks: Static Pricing, Asymptotic Optimality, and Demand Substitution Effects. *IEEE/ACM Transactions on Networking*, 10(3):425–438, 2002.

[34] I. Ch. Paschalidis and J. N. Tsitsiklis. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, 2000.

[35] D. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 1995.

[36] E. Campos-Nanez and S. D. Patek. On-Line Pricing of Network Resources. *Proceedings of IEEE INFOCOM*, 2003.