ADAPTIVE CONTROL OF CONSTRAINED MARKOV CHAINS

Eitan Altman and Adam Shwartz

Electrical Engineering

Technion — Israel Institute of Technology

Haifa, 32000 Israel

ABSTRACT

We consider the adaptive control of finite-state Markov chains, where the optimal performance is characterized through the minimization of a long-run average cost functional, subject to constraints on several other such functionals.

In contrast with the unconstrained problem, applying the control which is optimal for the current parameter estimates ("certainty equivalence" control) may not yield optimal performance. This problem is related to the fact that a feasible set of a Linear Program is not, in general, continuous in the parameters.

Under mild structural and feasibility conditions we exhibit two explicit adaptive control policies for the case where the transition probabilities are unknown, which are optimal under the constrained optimization criterion. These policies rely on a powerful estimation scheme of "probing", which provides consistent estimators for the transition probabilities. This scheme is of independent interest, as it provides strong consistency under a large number of adaptive schemes, and is independent of any "identifiability" conditions.

As an application we derive an optimal adaptive policy for a system of K competing queues with **countable** state space, and for which the constrained criteria arise naturally in the context of communication networks.

Keywords: Controlled Markov chains, additional constraints, adaptive control.

Please address all correspondence to the second author at the above address.

Telephone: (4) 29 4743. Electronic address (bitnet): adam@techsel.

Submitted March 1989, revised February 1990.

INTRODUCTION

The problem of adaptive control of Markov chains has received considerable attention in recent years; for motivation and existing results see e.g. the survey paper by Kumar [18], Militio and Cruz [21], the book by Hernandez-Lerma [11] and references therein. In the setup considered there, the transition probabilities of a Markov chain are parameterized, and the "true" parameter value is not known. One then tries to devise a control policy which minimizes the long-run average cost, based on some estimate of the parameters. Schäl [22] introduced an asymptotic discounted cost criterion and studied the related adaptive optimal policies (see also Hernandez-Lerma and Marcus [12,16,15] and the extensions [13,14] to incomplete state information). In the constrained problem [10] the optimization criterion is the minimization of a long-run average cost (1.2a), but subject to several constraints, given also in terms of long-run average-cost functionals (1.2b).

Below we formulate and solve the problem of optimal adaptive control of finite state Markov chains, under average-cost constraints. We assume no prior information about the transition probabilities and their dependence on the control. This is formalized by taking the transition probabilities as the unknown parameters. The precise model and basic assumptions are given in Section 1.

Adaptive policies for the case of a single constraint were first introduced by Makowski and Shwartz [20,23] using the Lagrange approach of Beutler and Ross [7]. This approach is, however, limited to a single constraint. Altman and Shwartz [2] obtain an optimal adaptive policy for the finite-parameter case. The situation here is considerably more complicated than the unconstrained, single constraint or finite-parameter cases since, as Example 4.1 illustrates, the ("certainty equivalence") approach of using current estimates to compute the optimal control may fail. The difficulty is that the optimal non-adaptive control may have discontinuities as a function of the constraints, for the following reason. Even when all parameter are known, the only available approach to the computation of optimal policies is through an associated Linear Program (Derman [10], Hordijk and Kallenberg [17], Altman and Shwartz [2,3]). Thus under the "certainty equivalence" paradigm, the computation of approximate controls from parameter estimates is necessarily carried out through a Linear Program whose coefficients are estimated on-line. However, it is well-known that the feasible region of a Linear Program and hence also the optimal solution may exhibit discontinuities as a function of the coefficients. Example 4.1 illustrates the two difficulties with the "certainty equivalence" approach when utilizing a Linear Program; existence of solutions, and convergence of the controls. This makes it a non trivial task to derive conditions under which the "approximate problems" possess solutions, and these solutions converge to the optimal solution.

The "classical" Linear Program (Derman [10], Hordijk and Kallenberg [17]) and a different one (using the "Policy Time Sharing" idea of Altman and Shwartz [2,3]) associated with the non-adaptive constrained problem are recalled in Section 2. We describe two classes of policies, which are suitable for adaptive problems; Action Time Sharing (Section 2.2) and Policy Time Sharing (Section 2.1). While the first is computationally more attractive, the second can be extended in some cases to systems with countable state space (Section 3.1.3). In Section 3 we present general "probing" methods for modifying policies so as to obtain strongly consistent estimators. The advantage of this approach is that it alleviates the usual identifiability problem of adaptive control [18].

If we assume the availability of some "continuous" method to compute policies based on these estimators, then it is shown that the adaptive policy which combines probing with the substitution of current estimates into the control rule is optimal. In order to develop adaptive policies for the constrained problem it is necessary to overcome the discontinuities in the feasible region of the Linear Program. The theory of sensitivity analysis for Linear Programs [9] is used in Section 4 to resolve this issue. As a result we obtain controls which converge to the optimal policy (i.e. "self tuning" is achieved). It remains to be shown that this convergence of the controls implies optimality, i.e. that under the adaptive policy minimal cost is achieved and the constraints are met. In particular, we need to establish that probing does not increase the cost. This is shown by an application of the results of Altman and Shwartz [2,4] on "time-sharing policies" (Lemma 2.1, Lemma 3.3).

The resulting optimal adaptive policies for constrained problems are obtained under weak conditions; they possess a natural structure and can be implemented using simple standard operations. The underlying methods are quite flexible, and offer an alternative even to adaptive methods of unconstrained optimization [18,21]. The general "probing" approach presented here and in particular the adaptive algorithms are useful in other situations as well; this is illustrated through an application to a **countable** state space system of K competing queues [6]. Altman and Shwartz [3] obtain optimal controls for the constrained non-adaptive problem. In this paper we obtain optimal, adaptive, and implementable policies for the constrained optimization problem of this system.

1. MODEL AND ASSUMPTIONS.

Let $\{X_t\}_{t=1}^{\infty}$ be the state process, defined on the finite *state space* $\mathbf{X} = 0, 1, ..., N$; the action A_t taken at time t takes values in the finite *action space* $\mathbf{A}$. To simplify the notation we assume that in any state x all actions in $\mathbf{A}$ are available (although the results are independent of this assumption).

Denote by $h_t := (X_1, A_1, ..., X_t, A_t)$ the *history* of the process up to time t. If the state at time t is x and action a is applied, then the next state will be y with probability

$$P_{xay} := P(X_{t+1} = y \mid X_t = x; A_t = a) = P(X_{t+1} = y \mid h_{t-1} = h, X_t = x; A_t = a) \qquad (1.1)$$

A policy u in the *policy space* U is described as $u = \{u_0, u_1, ...\}$, where u_t is applied at time epoch t, and $u_{t+1}(\cdot \mid h_t, X_{t+1})$ is a conditional probability measure over $\mathbf{A}$. Each policy u induces a probability measure denoted by P^u on the space of paths $\Omega := (\mathbf{X} \times \mathbf{A})^\infty$ which serves as the canonical sample space. The corresponding expectation operator is denoted by E^u.

A *Markov policy* $u \in U(M)$ is characterized by the dependence of u_{t+1} on X_{t+1} only; i.e. $u_{t+1}(\cdot \mid h_t, X_{t+1}) = u_{t+1}(\cdot \mid X_{t+1})$. A *stationary policy* $g \in U(S)$ is characterized by a single conditional probability measure $p^g_{\cdot|x}$ over $\mathbf{A}$; under g, X_t becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} = \sum_{a \in \mathbf{A}} p^g_{a|x} P_{xay}$. The class of *stationary deterministic policies* $U(SD)$ is a subclass of $U(S)$, and every $g \in U(SD)$ is characterized by a mapping $g : \mathbf{X} \to \mathbf{A}$ so that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is concentrated at the point $g(x)$ for each x.

Let $\{c(x,a),\, d^k(x,a)\ ,\ k = 1, ..., K\}$ be (real valued) cost functions and define

$$C_x(u) = \overline{\lim_{t\to\infty}} \frac{1}{t} E^u \left[\sum_{s=1}^{t} c(X_s, A_s) \mid X_1 = x \right] \qquad (1.2a)$$

$$D^k_x(u) = \overline{\lim_{t\to\infty}} \frac{1}{t} E^u \left[\sum_{s=1}^{t} d^k(X_s, A_s) \mid X_1 = x \right] \qquad k = 1, ..., K \quad (1.2b)$$

Given the real numbers $\{V_k\ ,\ k = 1, ..., K\}$, define the constrained optimization problem **COP**;

$$\textit{minimize } C_x(u) \qquad \textit{subject to } D^k_x(u) \le V_k \qquad 1 \le k \le K \qquad (1.3)$$

In this paper we assume that the transition probabilities P_{xay} are unknown. We thus solve the adaptive constrained problem **ACOP**, which is to find an optimal policy for COP based on the available information (i.e. using on-line estimation to update and improve the policy).

Throughout the paper we impose the following assumptions under the true parameters:

A1: The state space forms a single positive recurrent class under any policy in $U(SD)$.

A2: COP is feasible, i.e. there exists some $u \in U$ for which (1.3) holds.

A1 is a standard assumption that holds for many queuing systems (see e.g. Section 5). The analysis can be extended to include certain transient states, at a cost of technical complications.

We use the following notation: $1\{A\}$ is the indicator function of the set A and $\delta_a(x)$ is the Kronecker delta function. $\overline{B}$ is the closure of a set B, and $|B|$ is the number of elements in the (finite) set. For sets $\{H_n\}$ in the $\mathbb{R}^l$ we use the (slightly nonstandard) notation (see e.g. [9]);

$\underline{\lim}_{n\to\infty} H_n := \{x \in \mathbb{R}^l : x = \lim_{n\to\infty} x_n, \text{ for some sequence } x_n \in H_n\}$,

$\overline{\lim}_{n\to\infty} H_n := \{x \in \mathbb{R}^l\text{: for some infinite subsequence } x_{n_r} \in H_{n_r}, x = \lim_{r\to\infty} x_{n_r}\}$.

Thus $\underline{\lim}_{n\to\infty} H_n$ contains all limits of converging sequences, whereas $\overline{\lim}_{n\to\infty} H_n$ contains all accumulation points. Note that $\overline{\lim}_{n\to\infty} H_n := \cap_{m=1}^{\infty} \overline{\cup_{n=m}^{\infty} H_n}$ (the standard definition does not include the closure). If $\underline{\lim}_{n\to\infty} H_n = \overline{\lim}_{n\to\infty} H_n$ then write $\lim_{n\to\infty} H_n = \underline{\lim}_{n\to\infty} H_n$.

2. THE NON-ADAPTIVE CASE

All the available methods for computing optimal policies for the (non-adaptive) constrained problem rely on associated Linear Programs. Two such programs are presented below.

2.1 An optimal time-sharing policy for COP.

Policy Time Sharing (PTS) policies where introduced by Altman and Shwartz [2, 3, 19]. First note that since both action space and state space are finite, we can order all stationary deterministic policies as $\{g_1, g_2, ..., g_l\}$, with $l = |U(SD)| = |\mathbf{X}| \times |\mathbf{A}| = (N+1)|\mathbf{A}|$. Define a "cycle" as the time between two consecutive visits to state 0. A PTS policy is characterized [2,3] by a l dimensional vector $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$, and the following properties;

(1) : During each cycle, a fixed stationary deterministic policy is used.

(2) : α_i is the limiting proportion of the number of cycles during which policy g_i is used.

(The formal definitions are in Section 3.1.1). Thus the α_i are necessarily positive and sum to 1. Any PTS policy with parameter α is denoted by $\hat{\alpha}$. Note that there are infinitly many PTS policies with a given parameter α, since (1)–(2) identify a whole class of policies. However, it will be seen below that the distinction between policies with the same parameter α is of no consequence in our case; in particular, by (2.1)–(2.2) these policies have the same costs.

Let τ_i be the expected cycle duration when using the deterministic policy g_i. It is shown in [2] that the cost $C_x(\hat{\alpha})$ is obtained as limit (rather than the $\overline{\lim}$ in (1.2)) and

$$C_x(\hat{\alpha}) = \sum_{i=1}^{l} z_i(\alpha) C(g_i) \tag{2.1}$$

is independent of the initial state x, where $z_j(\alpha)$ is given by

$$z_i(\alpha) = \alpha_i \tau_i \left[\sum_{j=1}^{l} \alpha_j \tau_j\right]^{-1} \tag{2.2}$$

The same linear representation holds for $D_x^k(\hat{\alpha})$. Denote by $\hat{\Gamma}$ the subset of PTS policies which are best for problem COP, i.e. feasible policies for which the cost $C(\hat{\alpha})$ is no greater than that of any other feasible PTS policy. Since obviously $\tau_i > 0$ for each i it follows from (1.2) and (2.1) that $\hat{\Gamma}$ (or the corresponding vectors α) can be obtained as follows. Solve the Linear Program:

LP_{pts} : find $z := \{z_1, ..., z_l\}$ that minimizes $\sum_{i=1}^{l} z_i C(g_i)$, subject to

$$\sum_{i=1}^{l} z_i D^k(g_i) \le V_k \quad 1 \le k \le K, \qquad \sum_{i=1}^{l} z_i = 1, \qquad z_i \ge 0 \quad \text{for} \quad 1 \le i \le l \tag{2.3}$$

If z satisfies the equality and nonnegativity constraints in (2.3) define $\gamma(z)$ by the inverse of (2.2):

$$\gamma_i = \frac{z_i}{\tau_i} \left[\sum_{j=1}^{l} \frac{z_j}{\tau_j} \right]^{-1} \tag{2.4}$$

Any z which is feasible for LP_{pts}, i.e. satisfies (2.3), defines a feasible PTS policy through $\alpha = \gamma(z)$, and vice versa. In particular, let B denote the set of optimal solutions of LP_{pts}. The set $\Gamma := \{\gamma(z) : z \in B\}$ then satisfies $\hat{\Gamma} = \{\hat{\alpha} : \alpha \in \Gamma\}$. In fact, any vector in Γ defines an optimal policy for COP (and not only best among PTS policies), as the following result shows [2];

Lemma 2.1: *Under A1 COP is feasible iff LP_{pts} is. If A2 also holds, then any PTS policy $\hat{\alpha}$ where $\alpha \in \Gamma$ is optimal for COP.*

2.2 An optimal stationary policy for COP [10] [17].

Let now z be a $|\mathbf{X}| \times |\mathbf{A}|$ dimensional real matrix and consider the following Linear Program: **LP_s:** Find z that minimizes $\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} c(y,a) z(y,a)$ subject to:

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} (\delta_x(y) - P_{yax}) z(y,a) = 0 \qquad 0 \le x \le N \tag{2.5a}$$

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} d^k(y,a) z(y,a) \le V_k \qquad 1 \le k \le K \tag{2.5b}$$

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} z(y,a) = 1 \tag{2.5c}$$

$$z(y,a) \ge 0 \quad y \in \mathbf{X}, a \in \mathbf{A} \tag{2.5d}$$

Denote the set of optimal solutions of LP_s by B. For any z that satisfies (2.5a, 2.5c, 2.5d) define the matrix $\gamma(z)$ through (2.6); Lemma 2.2 below establishes that γ is well defined.

$$\gamma_y^a(z) := \frac{z(y,a)}{\sum_{a' \in \mathbf{A}} z(y,a')} \tag{2.6}$$

Each matrix γ defines a stationary policy, say ν through $p^{\nu}_{a|y} = \gamma^a_y$. Denote $\Gamma := \{\gamma(z) : \ z \in B\}$. The Linear Program LP_s is related to the COP in the following way [17].

Lemma 2.2: *Under A1, LP_s is feasible iff A2 holds. If z satisfies (2.5a, 2.5c, 2.5d) then $\gamma^a_y(z)$ is well defined. If a stationary policy g is defined by $p^g_{a|y} = \gamma^a_y(z)$, then $\sum_a z(y,a) = \pi^g(y)$. Conversely, for each stationary g, $z(y,a) := \pi^g(y) \cdot p^g_{a|y}$ satisfies (2.5a, 2.5c, 2.5d). In particular, for each $\beta \in B$ the stationary policy ν defined by $p^{\nu}_{a|y} = \gamma^a_y(\beta)$ is optimal for COP, and $\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} d^k(y,a)\beta(y,a) = D^k(\nu), \quad 1 \le k \le K$ and $\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}} c(y,a)\beta(y,a) = C(\nu)$.*

Some important properties of LP_s which are needed later are collected below.

Lemma 2.3 [10 p. 80]: *Under A1, (i) there is at least one z that satisfies constraints (2.5a), (2.5c) and (2.5d) and (ii) any such z has, for each state y, at least one non-zero component $z(y,a)$.*

Lemma 2.4: *Under A1 there are exactly N+1 linearly independent equations among (2.5a), (2.5c).*

Proof: Summing over $x = 0, 1, 2, ..., N$ in (2.5a) shows that these equations are linearly dependent. Suppose there are $L \le N$ independent equations among (2.5a) and (2.5c). Consider the feasible region defined by (2.5a), (2.5c) and (2.5d). ¿From standard results on Linear Programs, there exist some feasible z with at most L non-zero components. But this contradicts Lemma 2.3 (ii). $\square$

Conclusion: In both the PTS and the stationary cases one obtains a Linear Program of the form:

LP1: Find z that minimizes $c \cdot z$, subject to $\quad g(z) = 0$, and

$$z \ge 0 \tag{2.7a}$$

$$d^k \cdot z \le V_k \qquad 1 \le k \le K \tag{2.7b}$$

where g is a vector of affine functions (and the scalar product is the summation over all common indices). It will sometimes be convenient to use generic notation for the inequality constraints, and we shall, without further mention, replace (2.7a-2.7b) with the single inequality $f(z) \le 0$. If the stationary method is used then we omit one (redundant) equality constraints in (2.5a) so that by Lemma 2.4 in either method all the equality constraints are linearly independent.

Denote by B the set of z's which are optimal for $LP1$. Note that B is closed and convex. If B is a singleton then we denote its single element by β. The set Γ as defined below (2.4) or (2.6) is clearly also closed, but not necessarily convex.

3. THE ADAPTIVE CASE: PROBING APPROACH.

Below we show that the solution of ACOP can be decomposed into two phases: the estimation, and the control. In this section we obtain strongly consistent estimators, assuming only A1. After each step, the new information obtained dictates an update of the estimation and hence of the policy in use. One standard way to do so is to solve the non-adaptive problem COP through LP1 while substituting the estimates instead of the real (unknown) coefficients in the Linear Program. The resulting policy is then used till the next estimate is obtained; this is known as the "certainty equivalence" principle. There are two difficulties with this approach. The first is a continuity problem; in Section 4 we give conditions under which, if the estimation is consistent, then the policy converges ("self tuning") and the policy is optimal. The second difficulty is that under a Certainty Equivalence policy it is not generally possible to construct consistent estimators. In this section we shall assume that for each possible value of the parameter, we are already given some control laws for which convergence does occur; this is formalized below. Under this additional condition and A2 we construct consistent estimators using a probing approach, and show that the costs obtained when all parameters are known are also achieved in the adaptive case. Hence the estimation does not affect the cost, even though probing controls, which are not close to to the optimal control, are used to enhance the estimation.

3.1 The Time Sharing approach

3.1.1 Estimation of the costs

The first method involves the estimation of $C(g_i)$ and $D^k(g_i)$ for all l deterministic policies g_i and $1 \leq k \leq K$. Throughout subsection 3.1 we use PTS policies only. Denote

$\sigma(1) := \min\{t > 0, X_t = 0\}$, $\sigma(j+1) := \min\{t > \sigma(j), X_t = 0\}$.

The n^{th} cycle is thus the period $[\sigma(n), \sigma(n+1))$, and during each cycle, by property (1) of PTS policies a fixed deterministic policy is used. We impose property (1) also in $[1, \sigma(1))$. Let

$\sigma_i(1) := \min\{t > 0, X_t = 0$ and policy g_i is used during the cycle beginning at time $t\}$.

$\sigma_i(j+1) := \min\{t > \sigma_i(j), X_t = 0$ and policy g_i is used during the cycle beginning at time $t\}$.

$b_i(n)$:= the total time during which g_i was used during the first n cycles.

$m_i(n)$:= the number of cycles during which g_i was used during the first n cycles.

Property (2) of a PTS policy $\hat{\alpha}$ thus requires that for $i = 1, ..., l$, $\lim_{n\to\infty} n^{-1} m_i(n) = \alpha_i$ with probability one (under A1, in any realization of any PTS policy the number of cycles becomes infinite as $t \to \infty$ with probability one [2] and hence the definitions above are valid).

The costs under policy g_i are estimated as

$$\hat{C}_n(g_i) := [b_i(n)]^{-1} \sum_{j=1}^{n} \sum_{t=\sigma(j)}^{\sigma(j+1)-1} c(X_t, A_t) \cdot 1\{g_i \text{ is used during cylce } j\} \tag{3.1a}$$

$$\hat{D}_n^k(g_i) := [b_i(n)]^{-1} \sum_{j=1}^{n} \sum_{t=\sigma(j)}^{\sigma(j+1)-1} d^k(X_t, A_t) \cdot 1\{g_i \text{ is used during cylce } j\},\ 1 \le k \le K \tag{3.1b}$$

To obtain consistent estimators, i.e. that for a PTS policy u we have for all i

$$\lim_{n\to\infty} \hat{C}_n(g_i) = C(g_i) \quad and \quad \lim_{n\to\infty} \hat{D}_n^k(g_i) = D^k(g_i) \quad P^u - a.s.$$

it suffices to use every g_i during infinitely many cycles. This is shown as follows; let

$$Y_i(j) := \sum_{t=\sigma_i(j)}^{\sigma_i(j+1)-1} c(X_t, A_t) \cdot 1\{during\ the\ j^{th}\ cycle\ policy\ g_i\ is\ used\}$$

If u is any PTS policy that uses g_i infinitely often, then $Y_i(n),\ n = 1, 2, \ldots$ are i.i.d. Define $E^u Y_i(1) := W_i$ and recall $Eb_i(n) = \tau_i$. By the Strong Law of Large Numbers

$$\lim_{n\to\infty} \hat{C}_n(g_i) = \lim_{n\to\infty} \frac{\sum_{j=1}^{n} Y_i(j)}{b_i(n)} = \frac{\lim_{n\to\infty} n^{-1}\sum_{j=1}^{n} Y_i(j)}{\lim_{n\to\infty} n^{-1} b_i(n)} = \frac{W_i}{\tau_i} = C(g_i) \quad P^u\ a.s. \tag{3.2}$$

where the last equality follows from Chung [8, pp. 95-96]. A similar result holds for $D^k(g_i)$.

3.1.2 The optimal control.

Recall that under a PTS policy, some fixed deterministic policy is used during each cycle. Denote by $\hat{C}_n$ and $\hat{D}_n$ the set of estimates $\{\hat{C}_n(g_i),\ \hat{D}_n^k(g_i),\ 1 \le k \le K,\ 1 \le i \le l\}$. Suppose we are given a sequence $\pi := \{\pi^r,\ r = 1, 2, \ldots\}$ of approximations of an optimal $\gamma \in \Gamma$ (defined below (2.4)) which are parameterized by the estimators of the costs. Let $d(z, z') := \max_{1\le i\le l}\{|z(i) - z'(i)|\}$ and fix a decreasing sequence $\epsilon_r \downarrow 0$. In the following Algorithm 3.1 the r^{th} estimate of C and D is substituted in π^r to yield a series $\gamma(r)$ of approximations of some optimal γ. These in turn define a sequence of policies. The performance of the Algorithm is obtained in Theorem 3.2 below.

Algorithm 3.1:

1. Set $N_1 = 0$ and $r = 1$.
2. Use g_i during the $N_r + i^{th}$ cycle, $1 \le i \le l$.

3. Calculate $\hat{C}_{N_r+l}$ and $\hat{D}^k_{N_r+l}$ to obtain $\gamma(r) := \pi^r(\hat{C}_{N_r+l}, \hat{D}_{N_r+l})$.

4. Starting from cycle N_r+l+1 choose at the n^{th} cycle g_i that satisfies: $i \in \operatorname{argmin}\{\frac{m_i(n)}{n} - \gamma_i(r)\}$.

5. Continue until cycle $N_{r+1} := \min\{n > N_r + l + 1 : d[\frac{m(n)}{n}, \gamma(r)] < \epsilon_r\}$. Increase r by one and repeat from step 2.

Denote the policy resulting from algorithm 3.1 by $\mu = \mu(\pi, \hat{C}, \hat{D})$

Theorem 3.2: *Under A1 (i) $N_r < \infty$ w.p.1 all r, (ii) $\hat{C}_n(g_i) \to C(g_i)$ and $\hat{D}^k_n(g_i) \to D^k(g_i)$ w.p.1. If moreover A2 holds and the parameterized approximations π are continuous, i.e. if (ii) implies the convergence $\gamma(r) \to \gamma$, then Algorithm 3.1 yields an optimal PTS policy for ACOP.*

Proof: The first claim is established by induction. By definition, $N_1 = 0 < \infty$. Assume that $N_j < \infty,\ \ j \le r$. We shall show that N_{r+1} is finite as well. Clearly for $N_r \le n < N_{r+1}$,

$$d\left[\frac{m(n+1)}{n+1}, \gamma(r)\right] \le d\left[\frac{m(n)}{n}, \gamma(r)\right] + \frac{1}{n} \tag{3.3}$$

Now if the $n+1^{st}$ cycle is not a probing cycle, then it is shown in the appendix that

$$d\left[\frac{m(n+1)}{n+1}, \gamma(r)\right] \le \max\left\{\frac{n}{n+1} d\left[\frac{m(n)}{n}, \gamma(r)\right], \frac{3l}{n+1}\right\} \tag{3.4}$$

Since there are exactly l probing cycles between N_r and N_{r+1} and since $d[\frac{m(N_r)}{N_r}, \gamma(r)] < \infty$, (3.3) and (3.4) imply that step 5 is concluded in a finite time, and *(i)* is established. It immediately follows that probing cycles are used infinitely often, and hence by the argument following (3.1) *(ii)* holds. To establish the last claim note that by hypothesis, *(ii)* implies $\gamma(r) \to \gamma$ so step 5 of Algorithm 3.1 implies that $\lim_{r\to\infty} {N_r}^{-1} m_i(N_r) = \gamma_i$ for $1 \le i \le l$. As there are exactly l probing cycles among the cycles $N_r \le n < N_{r+1}$, it follows from this and (3.3) that $\lim_{n\to\infty} n^{-1} m_i(n) = \gamma_i$ for $1 \le i \le l$. But this establishes that the adaptive policy is a PTS policy with parameter γ. By Lemma 2.1 this policy is optimal, and the last claim is established. □

Remarks: Theorem 3.2 shows that the "probing cycles" have no effect on the cost; the reason is that they occur less and less frequently, so that they do not have any impact on the last limit. The decreasing frequency of probing is intuitively clear from the fact that ϵ_r is decreasing, so that the number of cycles it takes to "correct" the bias due to probing is increasing.

The information obtained during the non-probing cycles could also be used to improve the estimation, and decrease the size of the probing periods. This will increase the rate of convergence of both estimators and controls. The proofs are similar, but we shall not pursue these issues here.

3.1.3 The countable case

When the state and action spaces are very large the Linear Program required by either methods described in Section 2 becomes impossible to solve, and approximation schemes are required. Yet in some non-adaptive applications PTS policies have been successfully used to obtain optimal policies using finite Linear Programs (e.g. [3]) even when the state space is countably infinite. Such a result holds when the following condition can be verified:

A3a: There is a finite set $G \subset U(SD)$, so that an optimal policy exists among PTS policies that are obtained by switching between policies in G only.

A3b: Under every $g_i \in G$, $E_{g_i}\left[\sum_{s=\sigma(1)}^{\sigma(2)-1} |c(X_s, A_s)|\right] < \infty$, and similarly for d^k.

When A3 holds, the optimal policy is obtained using exactly the same method as in section 2.1, where l stands for the number of stationary deterministic policies in the subset G. It is easy to check that the adaptive scheme described in Section 3.1.1-3.1.2 carries over as well. In section 5 an example of a queueing network is given where this method is applied.

3.2 Generalizing the stationary policies

3.2.1 Estimation of the transition probabilities.

The second method involves a direct estimation of the transition probabilities. Define

$$\hat{P}^t_{xay} := \frac{\sum_{s=2}^{t} 1\{X_{s-1} = x, A_{s-1} = a, X_s = y\}}{\sum_{s=2}^{t} 1\{X_{s-1} = x, A_{s-1} = a\}}$$

If the denominator is zero then $\hat{P}^t$ is chosen arbitrarily, but such that for every x and a, $\hat{P}^t_{xay}$ is a probability measure. For this estimator to be consistent under some policy u, i.e.

$$\lim_{t\to\infty} \hat{P}^t_{xay} = P_{xay} \quad P^u \ a.s. \tag{3.5}$$

for all states $x, y \in \mathbf{X}$ and $a \in \mathbf{A}$, it is sufficient that each state-action pair is visited infinitely often. If this holds, then the type of renewal argument leading to (3.2) also establishes (3.5) (see [5] (7.12)-(7.13)). It turns out ([2] Lemma 4.1) that under *any policy* u each state x is visited infinitely often P^u a.s. By an appropriate "probing" (described below) we shall guarantee that $\sum_{s=2}^{\infty} 1\{X_{s-1} = x, A_{s-1} = a\} = \infty \quad P^u \ a.s.$ for all $x \in \mathbf{X}$, $a \in \mathbf{A}$, implying consistent estimation.

3.2.2 The optimal control.

To obtain an optimal policy for ACOP, the estimator $\hat{P}^t$ of the transitions P_{xay} is combined with an updating rule for the control. Clearly, the resulting policy cannot be stationary; nevertheless, we would like to apply the Linear Program LP_s of section 2.2 as a tool for computing optimal policies. Altman and Shwartz [2] construct "Action Time Sharing" (abbreviated ATS) policies, which generalize the stationary policies. If $\sum_{s=1}^t \{X_s = y\} = 0$ set $f^t(a|y) := 0$. Otherwise define

$$f^t(a|y) := \frac{\sum_{s=1}^t 1\{X_s = y, A_s = a\}}{\sum_{s=1}^t 1\{X_s = y\}}$$

A policy u is an ATS policy with parameter $\alpha := \{\alpha_y^a : y \in \mathbf{X}, a \in \mathbf{A}\}$ if $\lim_{t\to\infty} f^t(a|y) = \alpha_y^a \; P^u$ a.s. As in the case of PTS, the parameter α does not determine an ATS policy uniquely, and we denote be $\hat{\alpha}$ any such policy. All these policies achieve the same cost, as we show in Lemma 3.3 below, and hence the notation $\hat{\alpha}$ does not cause difficulties.

Lemma 3.3 [2] Theorem 4.2: *The costs under a stationary policy ν are also achieved by any ATS policy $\hat{\alpha}$ for which $\alpha_y^a = p_{a|y}^\nu \; P^{\hat{\alpha}}$ a.s. Consequently, any policy for which such a limit α exists and satisfies $\alpha_y^a = \gamma_y^a$ for some $\gamma \in \Gamma$ (given in (2.6)) is optimal for COP.*

As in Section 3.1.2 assume we are given a sequence $\pi := \{\pi^r,\ r = 1, 2, ...\}$ of approximations of an optimal $\gamma \in \Gamma$ (defined below (2.6)) which are parameterized by the estimators $\hat{P}_{xay},\ x, y \in \mathbf{X},\ a \in \mathbf{A}$ of the transition probabilities. In the following Algorithm 3.4 the r^{th} estimate of the transition probabilities P is substituted in π^r to yield a series $\gamma(r)$ of approximations of an optimal γ (i.e. $\gamma \in \Gamma$). Let now $d(z, z') := \max\{|z(y, a) - z'(y, a)| : y \in \mathbf{X}, a \in \mathbf{A})\}$. Fix a decreasing sequence $\epsilon_r \downarrow 0$ and define the sequences of (random) stopping times $\{T_i\}_{i=1}^\infty$ and $\{S_i\}_{i=1}^\infty$ as follows. Set $T_1 := 0$.

For $r \geq 1$ define $S_r := \inf\{t > T_r$; each state has been visited at least $|\mathbf{A}|$ times after $T_r\}$.

$T_{r+1} := \min\{t > S_r : d[f^t(\cdot|\cdot), \gamma(r)] < \epsilon_r\}$, where $\gamma(r)$ is obtained at time S_r using π.

Algorithm 3.4:

1. Set $r = 1$ and choose $\gamma(0)$ arbitrarily.

2. Let $\mathbf{A} = \{a_1, ..., a_{|\mathbf{A}|}\}$. Probe by choosing action a_i at the ith visit to state y after T_r.

3. Use $a' \in \operatorname{argmin}_{a\in\mathbf{A}}\{f^t(a|y) - \gamma_y^a(r-1)\}$ whenever $X_t = y$ except when probing. Continue until S_r.

4. Calculate $\hat{P}^{S_r}$ to obtain $\gamma(r) = \pi^r(\hat{P}^{S_r})$. If $X_t = y$ then $a' \in \operatorname{argmin}_{a\in\mathbf{A}}\{f^t(a|y) - \gamma_y^a(r)\}$. Continue until T_r, increase r by 1 and repeat from step 2.

Denote the policy resulting from algorithm 3.4 by $\mu = \mu(\pi, \hat{P})$.

Remark: As Theorem 3.5 shows, the controls used at "probing times" have no effect on the cost. Since they occur less and less frequently they do not influence the conditional frequencies $f^t(a|y)$ as $t \to \infty$ (see the remark following Theorem 3.2).

Theorem 3.5: *Under A1, (i) $T_r < \infty$ and $S_r < \infty$ for all r P^μ a.s. and (ii) $\hat{P}^{T_r} \to P \quad P^\mu$ a.s. Assume moreover A2 and that the sequence π is continuous in the sense that (ii) implies that $\gamma(r)$ converges to some optimal γ. Then Algorithm 3.4 yields an optimal ATS policy for ACOP.*

Proof: (i) and (ii) are proved in the same way as in Theorem 3.2. By (ii), the definition of T_r and the hypothesis, under Algorithm 3.4 $\lim_{r\to\infty} f^{T_r}(a|y) = \gamma_y^a$ for all $a \in \mathbf{A}$, $y \in \mathbf{X}$. Using arguments as in the proof of Theorem 3.2 it can be shown that in fact $\lim_{t\to\infty} f^t(a|y) = \gamma_y^a$ for all $a \in \mathbf{A}$, $y \in \mathbf{X}$. Hence the adaptive policy is in fact an ATS policy with parameter γ. by Lemma 3.3 and from LP_s (Section 2), this policy is optimal. The details are omitted. $\square$

3.3: Comparison of the two methods

The number of estimated parameters and the number of coefficients in the associated Linear Program are of the order of $|\mathbf{A}|^{|\mathbf{X}|}$ under the PTS method, and $|\mathbf{X}|^2 \cdot |\mathbf{A}|$ under the second method. Since both methods require solving Linear Programs (for the control law), it follows that the second method is much more efficient than the first. However, the PTS method has several advantages. First, it enables to solve specific problems involving a countable state space (Section 5). Second, it easily generalizes to problems where the costs $c(x, a)$ and $d^k(x, a)$ are not known a-priori and can only be measured with some noise, for then the estimator (3.1) is consistent.

4. THE CONTROL LAWS

In order to construct an optimal policy, we needed a sequence π that enables us to obtain a "converging" sequence of approximations $\beta(n)$ using estimates of some coefficients of LP1 (Section 2). If LP1 has a unique solution β for the true parameters then $\beta(n)$ are estimates of that β and the convergence implies, through (2.6) or (2.4), that $\gamma(n) \to \gamma$ as required in Theorem 3.2 (3.5). Below we implement the "certainty equivalence" approach of substituting the estimates in $LP1$, and illustrate via Example 4.1 the difficulties with this approach. Denote by B the set of optimal solutions of LP1 (under the true parameters). We consider the case where the parameter estimates converge (this holds, for example, under the probing schemes of section 3). In Theorem 4.6 we provide conditions under which if $\beta(n)$ are a sequence of solutions for the "approximate" LP1, then any limit is in B, i.e. $\overline{\lim}_{n\to\infty}\{\beta(n)\} \subset B$. This implies $\overline{\lim}_{n\to\infty}\{\gamma(n)\} \subset \Gamma$. In Section 5 we combine the results of Sections 3-4 to obtain optimal adaptive controls for several systems.

Roughly speaking, the "certainty equivalence" π_{ce} control is obtained by substituting the estimates generated by Algorithm 3.1 (3.4) into LP1, and defining π^r in Step 3 of Algorithm 3.1 (Step 4 of Algorithm 3.4) as the solution of $LP1_n$. For both methods, the Linear Program takes the form (cf. LP1 (2.7))

$\mathbf{LP1_n}$: Find z^n that minimizes $c_n \cdot z$ subject to $g^n(z) = 0$, $z \geq 0$ and

$$d_n^k \cdot z \leq V_k \qquad 1 \leq k \leq K$$

where the coefficients in $LP1_n$ are obtained through the estimation scheme. Choose $\beta(0)$ arbitrarily but such that $\beta_i(0) \geq 0 \quad 1 \leq i \leq l$ and $\sum_{i=1}^{l} \beta_i(0) = 1$. Let $B(n)$ be the set of solutions of $LP1_n$. If $B(n)$ is empty, i.e. $LP1_n$ is not feasible then set $\beta(n) := \beta(n-1)$. Otherwise, pick any one element of $B(n)$ and denote it by $\beta(n)$. This defines a control $\gamma(n)$ through either (2.6) or (2.4).

The certainty equivalence approach was used successfully in many unconstrained problems [18]. In the constrained adaptive case some serious difficulties arise. For simplicity, these are illustrated through a parameterized ACOP, although we may anticipate similar problems in our case.

Example 4.1: Let $\mathbf{X} := \{1, 2\}$, $\mathbf{A} := \{p, q\}$, $V_1 = -V_2 = V_3 = -V_4 = 1$, with transitions

$$P_{y,p,1} = P_{y,q,2} = 1 - P_{y,p,2} = 1 - P_{y,q,1} = \zeta, \qquad y = 1, 2$$

The costs are given by $c(1,p) = c(1,q) = c(2,p) = 0$ and $c(2,q) = 1$,

$$d^1(x,a) = 1\{x = 1, a = p\} \cdot b^{-1} \qquad d^2(x,a) = -1\{x = 1, a = p\} \cdot b^{-1}$$

$$d^3(x,a) = 1\{x = 1, a = q\} \cdot b^{-1} \qquad d^4(x,a) = -1\{x = 1, a = q\} \cdot b^{-1}$$

where a and b are parameters. Note that the four inequality constraints impose two equality constraints. We shall illustrate the continuity problems that arise in the corresponding Linear Program when ζ is replaced by a series $\zeta_n \to \zeta$ and when b is replaced by a series $b_n \to b$.

The Linear Program corresponding to COP using stationary policies takes the form

LP: Find z that minimize $\sum_{a,y} z(a,y)c(a,y) = z(2,q)$ subject to:

$$\sum_{y=1,2} \sum_{a=p,q} (\delta_1(y) - P_{ya1})z(y,a) = (1-\zeta)z(1,p) - \zeta z(2,p) + \zeta z(1,q) + (\zeta - 1)z(2,q) = 0 \quad (4.1a)$$

$$b^{-1}z(1,p) = b^{-1}z(1,q) = 1 \tag{4.1b}$$

$$\sum_{y,a} z(y,a) = 1 \qquad z(y,a) \geq 0 \tag{4.1c}$$

Where one redundant equality constraint was omitted, as indicated below LP1. Let β be any optimal solution of the LP. For any $\zeta \neq 0.5$ the feasible region contains at most the single point

$$z(1,p) = z(1,q) = b,\ z(2,p) = (-3b + 2\zeta + 1 - \zeta)[1-2\zeta]^{-1},\ z(2,q) = (b - \zeta + 2\zeta)[1-2\zeta]^{-1}$$

for all b such that (4.1c) holds. In particular, for $b = 0.25$ all $z(\cdot,\cdot)$ are equal to 0.25.

For $\zeta = 0.5$ the feasible region is nonempty for $b = 0.25$ only, and then it consists of a hyperplane

$$z(2,p) + z(2,q) = 0.5\ , \quad z(1,p) = z(1,q) = 0.25$$

Conclusion: we can anticipate two problems when using the certainty equivalence control law:

(i) $LP1_n$ is not feasible for some (perhaps infinitely many) n even though $LP1$ is feasible. In our example, this is the case when $\zeta = 0.5$ is fixed and b is replaced by estimates different than 0.25.

(ii) If $b = 0.25$ is fixed and $\zeta = 0.5$ is unknown and is replaced by some estimates, the respective γ which determines the optimal policy for ACOP is given by $\gamma_1^p = \gamma_1^q = 0.5,\ \gamma_2^p = 1,\ \gamma_2^q = 0$ whereas whenever the estimate of ζ is not equal 0.5, the approximating control is $\gamma_1^p = \gamma_1^q = \gamma_2^p = \gamma_2^q = 0.5$. Thus we can expect suboptimal behavior.

In order to understand and overcome these problems, we use the well known theory of sensitivity analysis for Linear Programs. We begin with a lemma of Dantzig, Folkman and Shapiro [9] which gives conditions for the convergence of $\beta(n)$ to B — the solution set of LP1. Let $\{f, f^n\}$, $\{g, g^n\}$ be affine functions with domain $\mathbb{R}^l$ and ranges in $\mathbb{R}^m$ and $\mathbb{R}^{m'}$ respectively, such that $f^n \to f$, and $g^n \to g$ pointwise. Given a Linear Program (e.g. LP1) with l decision variables, $f = (f_1, f_2, ..., f_m)$ (f^n) represents the m inequality constraints and g (g^n) represents the m' equality constraints. Given any function σ from $\mathbb{R}^l$ to $\mathbb{R}$ and a set $H \in \mathbb{R}^l$, define $M(\sigma|H)$ to be the subset of H where σ achieves its minimum, i.e. $M(\sigma|H) := \{x \in H : \sigma(x) = \inf\{\sigma(y)|y \in H\}\}$. Denote $H(f,g) := \{x \in \mathbb{R}^l \mid f(x) \leq 0,\ \text{and}\ g(x) = 0\}$; this is the feasible set of the Linear Program associated with f, g.

Lemma 4.2 [9] Theorem I.2.2: *Assume that $lim_{n\to\infty} H(f^n, g^n) = H$ for some set H. Let σ and $\{\sigma_n\}$ be linear functions such that $\sigma_n \to \sigma$ pointwise. Then*

$$\varliminf_{n\to\infty} M(\sigma_n|H(f^n,g^n)) \subset M(\sigma|H) \tag{4.2}$$

Thus convergence of sets of optimal solutions of LP's depends on convergence of feasible sets. Denote

$I := \{i \mid 1 \le i \le m, f_i(x) = 0 \text{ for all } x \in H(f,g)\}$, $f_I := \{f_i,\ i \in I\}$ and denote the rank of the matrix whose $|I| + m'$ rows are the coefficiences of the linear functions $f_i, i \in I$, and $g_j, 1 \le j \le m'$ by $\mathrm{rank}(f_I, g)$.

Lemma 4.3 [9] Cor. II.3.4: *Assume that* $\overline{\lim}_{n\to\infty} \mathrm{rank}(\mathrm{f}_\mathrm{I}^\mathrm{n}, \mathrm{g}^\mathrm{n}) \le \mathrm{rank}(\mathrm{f}_\mathrm{I}, \mathrm{g})$ *and that* $H(f,g)$ *is non empty. Then either* $\lim_{n\to\infty} H(f^n, g^n) = H(f,g)$ *or* $H(f^n, g^n)$ *is empty for infinitely many* n*.*

In order to apply the previous Lemmas to $LP1_n$ we need the following hypothesis.

A2$'$: COP is feasible when replacing the inequalities in (1.3) by strict inequalities, i.e. by

$$D_z^k(u) < V_k \qquad 1 \le k \le K$$

Note that under A2$'$ there also exist a stationary policy g and a PTS policy $\hat{\alpha}$ such that $D_z^k(g) < V_k$ and $D_z^k(\hat{\alpha}) < V_k \quad 1 \le k \le K$. For $K = 0$ this is immediate due to Lemma 2.1 and Lemma 2.2. For $K > 0$, let v be any policy that satisfies $D_z^k(v) < V_k \quad 1 \le k \le K$. Consider

COP': find a policy u that minimizes $D_z^K(u)$, such that $D_z^k(u) < D_z^k(v) \quad 1 \le k \le K-1$.

It is feasible since v satisfies the constraint. Hence according to Lemmas 2.1 and 2.2 there exist an optimal PTS policy $\hat{\alpha}$ and an optimal stationary policy g, from which the claim follows.

We show that A1 and A2$'$ imply the hypotheses of Lemma 4.3 and that under A1–A2$'$ the feasible sets of $LP1_n$ are empty only a finite number of times. We then apply Lemma 4.2 to obtain the convergence of the set of optimal solutions of $LP1_n$.

Lemma 4.4: *Under A1,* A2$'$ *is equivalent to* $H(f,g) \neq \emptyset$ *together with* $I = \emptyset$*, and implies* $\mathrm{rank}(\mathrm{f}_\mathrm{I}^\mathrm{n}, \mathrm{g}^\mathrm{n}) \le \mathrm{rank}(\mathrm{f}_\mathrm{I}, \mathrm{g})$ *for all* n*.*

Proof: Since g is full rank (see Conclusion in Section 2) $I = \emptyset$ implies $\mathrm{rank}(\mathrm{f}_\mathrm{I}^\mathrm{n}, \mathrm{g}^\mathrm{n}) \le \mathrm{rank}(\mathrm{f}_\mathrm{I}, \mathrm{g})$ for all n. Assume $H(f,g) \neq \emptyset$. Since the feasible set of LP1 is convex, $I = \emptyset$ is equivalent to the requirement that there exists some z that satisfies $g(z) = 0$ and $f(z) < 0$. ¿From Lemma 2.2 (stationary case) or from the representation (2.1, 2.2) (PTS), it follows that $I = \emptyset$ implies A2$'$. On the other hand, we conclude from Lemma 2.2 or (2.1, 2.2) that A2$'$ implies the existence of some stationary (or PTS) policy v whose corresponding z (in the representation of LP1) satisfies

$$d^k \cdot z < V_k \qquad 1 \le k \le K \qquad g(z) = 0 \tag{4.3}$$

We shall construct some z' satisfying both (4.3) and $z' > 0$, which corresponds to some other stationary or PTS policy. Consider first the case that LP1 is obtained from the PTS approach. From (2.1,2.2) it follows that $D^k(v)$ is continuous in z. Hence by picking some z' that satisfies $z' > 0$ and $\sum_i z'_i = 1$ but close enough to z, (4.3) still holds. In the stationary case, by construction, z satisfies (4.3) and $g(z) = 0$ (z is obtained by $z(y,a) = \pi^v(y) \cdot p^v_{a|y}$). Pick any other stationary policy w that satisfies $p^w_{a|y} > 0,\ a \in \mathbf{A}, y \in \mathbf{X}$. Let $z''(y,a) = \pi^w(y) \cdot p^w_{a|y}$. By assumption A1 and Lemma 2.2, z'' satisfies $z'' > 0$ and $g(z'') = 0$. Let $z' := (1-\delta)z + \delta z''$ where δ is some positive constant. By choosing δ small enough, $z' > 0$ and clearly satisfies (4.3). □

Lemma 4.5: *Assume A1,A2'. Assume that for all but a finite number of n, there exists some point z_n that satisfies $g^n(z_n) = 0$ and $z_n \geq 0$. Then $H(f^n, g^n)$ is at most finitely often empty.*

Proof: Consider the auxiliary Linear Program

$\mathbf{LP2_n}$: Find z and η that minimize η, subject to $g^n(z) = 0$, $z \geq 0$ and

$$d^j_n \cdot z \leq V^j + \eta \tag{4.4}$$

Denote the optimal η by η^n. Let H'_n denote the set of (z, η) satisfying the constraines in $LP2_n$. Let $LP2_\infty$ denote the Linear Program obtained by replacing g^n by g and d^j_n with d^j in $LP2_n$. It follows from A2$'$ that for both the PTS and the stationary cases, the solution η^∞ of $LP2_\infty$ satisfies $\eta^\infty < 0$. Since the state and action spaces are finite, there exists some positive constant L such that $D^k_x(u) \leq L$ for all policies u and all k. Thus by choosing $\eta' = L + \max_k\{\mid V^k \mid\}$ we have $(z_n, \eta') \in H'_n$, so that by hypotheses H'_n is nonempty except for a finite number of times. Define $\tilde{f}^n$ as the respective inequality constraints in $LP2_n$. Note that g^n is exactly the function defining $LP1_n$.

¿From Lemma 4.4 and (4.4) it easily follows that

$$I := \{i \mid 1 \leq i \leq m,\ \tilde{f}_i(x) = 0 \text{ for all } x \in H'_\infty(\tilde{f}, g)\} = \emptyset$$

As in Lemma 4.4 we conclude $\mathrm{rank}(\tilde{\mathrm{f}}^{\mathrm{n}}_{\mathrm{I}}, \mathrm{g}^{\mathrm{n}}) \leq \mathrm{rank}(\tilde{\mathrm{f}}_{\mathrm{I}}, \mathrm{g})$ for all n. It then follows from Lemma 4.3 that the feasible sets H'_n converge to H'_∞. We claim that $\eta^n \leq 0$ for all large enough n. To prove the claim, assume the converse. Then, since H'_n is bounded and nonempty for all large n, there exists a subsequence n_i such that $0 \leq \eta^{n_i} \leq \eta'$ and $(z^*_n, \eta^{n_i}) \to (z^*, \eta)$, where $\eta \geq 0$. But by Lemma 4.2 applied to the subsequence n_i, $\eta^{n_i} \to \eta^\infty < 0$, a contradiction, and the claim follows. Thus $LP2_n$ is feasible for all n large, so that $H(f^n, g^n)$ is at most finitely often empty. □

Using the previous Lemmas, the following Theorem gives conditions for $\overline{\lim}_{n\to\infty}\{\beta(n)\} \subset B$, and for the convergence of $\beta(n)$ to β, given that $B = \{\beta\}$. Using Theorems 3.2 and 3.5, the last convergence implies optimality of the certainty equivalence approach. Recall that B (B_n) is the set of optimal solutions of $LP1$ ($LP1_n$). Define f, g (f^n, g^n) as in LP1 ($LP1_n$ respectively).

Theorem 4.6: *Consider $LP1_n$. Under A1,A2, (i) $\underline{\lim}_{n\to\infty} B(n) \subset B$, (ii) $\overline{\lim}_{n\to\infty}\{\beta(n)\} \subset B$ and (iii) if moreover LP1 corresponding to the true parameters has only one solution, i.e. $B = \{\beta\}$, then $\lim_{n\to\infty}\beta(n) = \beta$.*

Proof: For either PTS or stationary case, there is a point z_n such that $g^n(z_n) = 0$ and $z \geq 0$ for all n. By Lemma 4.4, H_n is at most finitely often empty, and Lemma 4.2 implies (i). Pick any sequence $z_n \in B(n)$ and let n_i be a subsequence along which z_{n_i} converges. Then (ii) follows by applying (i) to that subsequence. (iii) is immediate from the last argument and (ii). $\square$

5. APPLICATIONS.

By way of conclusion, we give some specific results on the ACOP problem.

5.1 The finite ACOP

Let us introduce an assumption, whose significance we discuss below.

A4 At the true parameter values, LP1 possesses a unique solution β.

Theorem 5.1: *Assume A1, A2′ and A4. Then Algorithm 3.1 (3.4 respectively), using the sequence π_{ce}, provides an optimal PTS (resp. ATS) policy for ACOP.*

Proof: By Theorem 4.6 the sequence π_{ce} possesses the continuity property requeired in Theorem 3.2 (3.5), from which the result now follows. $\square$

The value of this result is limited mainly by assumption A4, since A1 is a weak assumption (which can also be slightly relaxed), while A2′ is close to the necessary assumption A2. The difficulty with A4 is that it *does not* hold for all values of the parameters, while on the other hand one clearly cannot check this assumption at the unknown value of the parameters. It is possible to obtain a result such as Theorem 5.1 without assuming A4. For this we need to extend Theorem 3.2 (3.5), and change the Certainty Equivalence control. This change is necessary since a policy which uses alternately one of two optimal actions may not be optimal (this is in contrast with controls obtained through Dynamic Programming for the unconstrained case). We shall not pursue these matters here, since the technical complications are considerable. Note that assumptions such as

A4 or other continuity assumptions are common in the litterature of adaptive problems (see e.g. Theorem 7.2 p.347 or the example in page 349 of [18]).

However, note that the set of parameters for which A4 does not hold is small in the following sense. The feasible set of LP1 is a polyhedron, and the minimization in LP1 is equivalent to finding the shortest vector (whose direction is determined by c) from the hyperplane through the origin which is perpendicular to c, to the boundary of the polyhedron. If there are two directions for which this vector has the same length, then the slightest change in the appropriate parameter will change this, and we are back in the situation where the optimal solution is unique, i.e. A4 holds. Formally, if we consider all parameters as beloging to some Euclidean space $\mathbb{R}^n$, then the Lebesgue measure of the set of parameters for which A4 does not hold is zero.

This fact can be utilized to "force" uniqueness as follows. It suffices to add to the cost c a small randomized perturbation, were the randomiztion possesses a density, then it can easily be seen that with probability one (with respect to the randomization) A4 holds, while the costs can be kept arbitrarily close to optimal by choosing small perturbation.

Thus assumption A4, can be overcome using extensions of the theory presented here, yielding ϵ-optimal policies.

Parametrized models: In many applications (see [18]) it is natural to have a parameterization of the unknowns in the system, and these need not include all parameters of $LP1_n$. It is obvious that any parameter that depends continuously on the parameters estimated in Algorithm 3.1 (3.4) can also be estimated consistently. Moreover, under such parameterization, if the optimal control as a function of the parameter is continuous, then Theorem 3 holds.

5.2 Application to a queuing system.

In this section we apply Algorithm 3.1 of section 3.1 to solve an adaptive problem for the following discrete-time queuing model with *countable state space.*

At time t, M_t^j customers arrive to queue j, $1 \leq j \leq J$. Each input stream is received in an infinite capacity buffer. Arrivals are independent from slot to slot, and the arrival process $M_t = \{M_t^1, ..., M_t^J\}$ forms a renewal sequence with finite means λ_j. During a time slot (t,t+1) a customer from any class j, $1 \leq j \leq J$ may be served, according to some policy, which is a prespecified dynamic priority assignment. If served, with probability μ_j it completes its service and leaves the system; otherwise it remains in its queue. The state $X_t = \{X_t^1, X_t^2, ..., X_t^J\}$ represents a J dimensional vector of the different queues' size at time t. Altman and Shwartz [3,4] solve the non-adaptive problem with a countable state space and with constraints on the average size of several queues. They considered the following linear cost functions: $c(X_t, A_t) = \sum_{j=1}^{J} c_j X_t^j$ and $d^k(X_t, A_t) = \sum_{j=1}^{J} d_j^k X_t^j$ for $1 \le k \le K$, where c_j and d_j^k are non-negative constants.

We denote COP_{queues} and $ACOP_{queues}$ the problems COP and ACOP with the above dynamic and cost structure. Assume the standard stability condition on the traffic intensity $\rho := \sum_{j=1}^{J} \frac{\lambda_j}{\mu_j} < 1$. This is a sufficient condition for A1 (see [6]). Altman and Shwartz used PTS policies to solve this problem in [3,4]. They show that in fact one can restrict to the finite class of PTS policies obtained by switching only between the $l = J!$ different priority policies g_i, each time the queues are empty. In fact, Assumption A3 (introduced in 3.1.3) holds [3] when the second moment of the number of arrivals per time-slot is bounded. As indicated in 3.1.3, this implies that an optimal policy for COP_{queues} is obtained as in Section 2 (see [3]). Moreover, an optimal policy for $ACOP_{queues}$ is obtained following the probing approach of Sections 3.1.1-3.1.2.

In this queueing example LP_{pts} and (2.1-2.2) can be simplified. In fact [3] the τ_i's are equal for all the priority policies g_i, hence instead of (2.1-2.2) the cost can be expressed as $C_x(\hat{\alpha}) = \sum_{i=1}^{l} \alpha_i C(g_i)$ and LP_{pts} reduces to

LP_{queues}: find α that minimizes $\sum_{i=1}^{l} \alpha_i C(g_i)$, subject to

$$\sum_{i=1}^{l} \alpha_i D^k(g_i) \le V_k \quad 1 \le k \le K, \qquad \sum_{i=1}^{l} \alpha_i = 1, \qquad \alpha_i \ge 0 \quad for \quad 1 \le i \le l$$

Denote B the set of optimal solutions of LP_{queues}. Similarly to Lemma 2.1, we have [3]

Lemma 5.2: *Under A1, COP_{queues} is feasible iff LP_{queues} is. If A2 holds, then any PTS policy $\hat{\alpha}$ that satisfies $\alpha \in B$ is optimal for COP_{queues}.*

Thus the simple form that LP_{pts} and (2.1-2.2) take simplifies the calculations. Except for that, ACOP is solved following Section 3.2.1-3.2.2.

Theorem 5.3: *Assume A1, A2$'$ and assume that LP_{queues} has a single optimal solution $B = \{\beta\}$. Then Algorithm 3.1 yields an optimal policy for $ACOP_{queues}$, when using the control law π_{ce} (introduced in Section 4)*

Proof: By Lemma 4.4, Assumption A2$'$ holds. Hence by Theorem 4.5 the control law π_{ce} yields $\beta(n) \to \beta$. According to Section 3.1.3 we can use Theorem 3.2 since Assumption A3 hold, which establishes the Theorem. $\square$

APPENDIX

Proof of (3.4): For $n > N_r + l$, let $J_n := \{i : \frac{m_i(n)}{n} > \gamma_i(r)\}$, and $d_n^* := \sum_{i=1}^{l} \mid m_i(n) - n\gamma_i(r) \mid$. Note that since $\sum_i \frac{m_i(n)}{n} = 1 = \sum_i \gamma_i$, $d_n^* = 2\sum_{i\in J_n}[m_i(n) - n\gamma_i(r)]$. We now show that $d_{n+1}^* \leq \max\{d_n^*, 3l\}$. If $d_n^* \leq 2l$ then clearly by (3.3) $d_{n+1}^* \leq 3l$. Now suppose that $d_n^* > 2l$. According to Algorithm 3.1, at the $n+1^{st}$ cycle some g_j is chosen such that $n\gamma_j(r) - m_j(n) > 1$. Thus $m_i(n+1) = m_i(n) + 1\{i = j\}$, so that $J_{n+1} \subset J_n$. Since $j \notin J_{n+1}$, we obtain:

$$\frac{1}{2}d_{n+1}^* = \sum_{i\in J_{n+1}} [m_i(n+1) - (n+1)\gamma_i(r)] = \sum_{i\in J_{n+1}} [m_i(n) - n\gamma_i(r)] - \sum_{i\in J_{n+1}} \gamma_i(r)$$

$$\leq \sum_{i\in J_n} [m_i(n) - n\gamma_i(r)] = \frac{1}{2}d_n^*$$

Hence $\frac{d_{n+1}^*}{n+1} \leq \max\{\frac{n}{n+1}\cdot\frac{d_n^*}{n}\ ,\ \frac{3l}{n+1}\}$. Since $d\left[\frac{m_i(n)}{n}, \gamma(r)\right] \leq \frac{d_n^*}{n}$, (3.4) follows.

REFERENCES

[1] Agrawal R., D. Teneketzis and V. Anantharam, "Asymptotically Efficient Adaptive Allocation Schemes for Controlled Markov Chains: Finite Parameter Space", Communications and Signal Processing Laboratory, University of Michigan, *Technical Report n. 254.*

[2] — , "Non-stationary policies for controlled Markov Chains", June 1987, submitted.

[3] — , "Optimal priority assignment: a time sharing approach", *IEEE Trans. on Automatic Control* Vol. AC-34 No. 9, 1989.

[4] — , "Markov decision problems and state-action frequencies", EE. PUB. No. 693, Technion, June 1988, submitted.

[5] — , "Adaptive Control of constrained Markov chains: Criteria and Policies", submitted to Annals of Oper. Res., 1989.

[6] Baras, J. S., D. -J. Ma, and A. M. Makowski, "K competing queues with geometric service requirements and linear costs: the μc rule is always optimal," *Systems and Control Letters,* Vol. 6 No. 3 pp. 173-180, August 1985.

[7] Beutler F. J. and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint", *Math. Anal. Appl.* Vol. 112, pp. 236-252, 1985.

[8] Chung K. L., *Markov chains with stationary transition probabilities*, 2nd edition, Springer Verlag, New York, 1967.

[9] Dantzig G. B., J. Folkman and N. Shapiro, "On the continuity of the minimum set of a continuous function", *J. Math. Anal. and Applications,* Vol. 17, pp. 519-548, 1967.

[10] Derman C., *Finite State Markovian Decision Processes,* Academic Press, New York, 1970.

[11] Hernandez-Lerma O., *Adaptive Markov Control Processes*, Springer-Verlag, 1989.

[12] Hernandez-Lerma O. and S. I. Marcus, "Adaptive Control of Discounted Markov Decision Chains", *J. Opt. Theory Appl.* Vol. 46 No. 2, pp. 27-235, June 1985.

[13] — , "Adaptive control of Markov processes with incomplete state information and unknown parameters", *J. of Opt. Theo. Appl.* Vol. 52 No. 2, pp. 227-241, February 1987.

[14] — , "Nonparametric adaptive control of discrete-time partially observable stochastic systems", *J. Math. Analysis of Appl.,* to appear.

[15] Hernandez-Lerma O. and S. I. Marcus, "Discretization procedures for adaptive Markov control processes", *J. Math. Anal . Appl.,* to appear.

[16] Hernandez-Lerma O. and S. I. Marcus, "Adaptive policies for discrete-time stochastic systems with unknown disturbance distribution" *System Control Letters* Vol. 9, pp. 307-315, North Holland, 1987.

[17] Hordijk A. and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* Vol. 9, No. 2, May 1984.

[18] Kumar P. R., "A survey of some results in stochastic adaptive control", *SIAM J. Control Opt.*, Vol. 23 No. 3, pp. 329-380, May 1985.

[19] Ma D. -J., A. M. Makowski and A. Shwartz, "Estimation and optimal control for constrained Markov chains", *Proc. 25th CDC,* Athens, Vol.2, pp. 994-999, 1986.

[20] Makowski A. M. and A. Shwartz, "Implementation issues for Markov decision processes", *Stochastic Differential Systems, Stochastic Control Theory and Applications,* Springer Verlag — IMA Volumes in Math. and Appl. No 10, Edited by W. Fleming and P.-L. Lions, 1988.

[21] Militio R. A. and J. B. Cruz, Jr., "An optimization-oriented approach to the adaptive control of Markov chains", *IEEE Trans. Automatic Control,* Vol. 32 No. 9, pp. 754-762, Sept. 1987.

[22] Schäl M., "Estimation and control in discounted dynamic programming," *Stochastics* **20**, pp. 51-71, 1987.

[23] Shwartz A. and A. M. Makowski, "An Optimal Adaptive Scheme for Two Competing Queues with Constraints", *Analysis and Opt. of Systems,* edited by A. Bensoussan and J. L. Lions, Springer Verlag lecture notes in Control and Info. Scie. No. 83, 1986.