TIME–SHARING POLICIES FOR CONTROLLED MARKOV CHAINS

by

Eitan Altman and Adam Shwartz

Electrical Engineering, Technion, Haifa 32000, Israel

ABSTRACT

We propose a class of non-stationary policies called *"policy time sharing"* (p.t.s.), which possess several desirable properties for problems where the criteria are of the average-cost type; an optimal policy exists within this class, the computation of optimal policies is straightforward, and the implementation of this policy is easy. While in the finite state case stationary policies are also known to share these properties, the new policies are much more flexible, in the sense that they can be applied to solve adaptive problems, and that they suggest new ways to incorporate the particular structure of the problem at hand into the derivation of optimal policies. In addition, they provide insight into the pathwise-structure of Controlled Markov chains.

To use p.t.s. policies one alternates between the use of several stationary deterministic policies, switching when reaching some pre-determined state.

In some (countable state) cases optimal solutions of the policy time sharing type are available and easy to compute, whereas optimal stationary policies are not available. Examples illustrating the last point and the usefulness of the new approach are discussed, involving constrained optimization problems with countable state space or compact action space.

Key words: Undiscounted rewards, non-stationary policies, time sharing, additional constraints, state-action frequencies.

Submitted November 1987, revised October 1988 and February 1991.

INTRODUCTION

The average cost criterion for a Markov Decision process clearly depends only on the limiting behavior of the processes. This implies considerable flexibility in the choice of control strategies, especially with regard to their finite-time behavior. A case where this is of importance is when the control serves additional purposes except optimization, such as estimation of some unknown parameters. If the values of the unknown parameters are needed for the computation of the optimal control, the estimation becomes part of the optimization procedure, and the resulting adaptive policy is necessarily non stationary.

On the other hand, it is desirable to reduce the complexity of the search for optimal policies, and to employ policies which are easy to implement. Traditionally, this is done by proving that an optimal policy exists within the class of stationary (or even stationary-deterministic) policies. In this paper we propose and analyze a class of non stationary policies, which contains optimal policies for optimization and constrained optimization problems. Over that class, the search for optimal policies is relatively simple, and it allows the degree of flexibility required for adaptive problems. Moreover, the derivation provides a sample-path approach to MDPs, in that the average cost is shown to depend only on **asymptotic, sample-path** properties of the control.

In Section 2 we introduce the non-stationary "policy time sharing" (p.t.s.) class of policies. These are obtained by alternating, when reaching some pre-determined state, between several stationary deterministic policies. The cost under such a policy is shown to be *linear* in the costs of the deterministic policies, and the coefficients are calculated. These results allow the derivation, in Section 3, of a new solution for the constrained optimization problem (with finite state and action spaces). This solution can be obtained by solving a new linear program. Conditions under which this approach is computationally attractive are illustrated via a constrained optimization problem for a queuing system. The state space for this system is countable, so that the standard linear programs (Derman 1970, Hordijk and Kallenberg 1984) do not apply, and a stationary solution is not available. Using p.t.s. policies makes it possible to take advantage of the structure of this queueing system and, using the linear representation of the costs, to derive a **finite** linear program for computing an optimal control. The resulting policy is simple to compute explicitly, and easy to implement. Finally an application to optimization under a single constraint when the action space is compact is given (Beutler and Ross (1985)); the use of time sharing results in this case in reduced computation. Using these policies it is possible to solve adaptive problems, as in Altman and Shwartz (1987, 1991a) and Ma and Makowski (1988). Fox and Rolph (1973) have used a sampling idea with a degenerate form of p.t.s policies—these sampling methods can also be combined with p.t.s policies to obtain adaptive policies—see Altman and Shwartz (1991a).

1. MODEL AND ASSUMPTIONS.

Let $\{X_t\}_{t=1}^{\infty}$ be the basic discrete time process where $\{X_t\}$ is defined on the countable *state space* $\mathbf{X}$={0,1,...}; an *action* a belongs to the countable *action space* $\mathbf{A}$, and A_t is the action taken at time t. We denote by $A(x)$ the set of actions available when in state x. $h_t := (X_1, A_1, X_2, A_2, ..., X_t, A_t)$ is the *history* of $\{X_t, A_t\}$. Denote the conditional probabilities for the controlled Markov chain by

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

A policy u in the *policy space* U is described as $u = \{u_1, u_2, ...\}$, here 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 $\mathbf{A}$. Let Ω denote the canonical sample space, i.e. $\Omega := (\mathbf{X} \times \mathbf{A})^{\infty}$, and let $\mathcal{F}$ denote the product σ-field on Ω. Then (see e.g. Schäl 1975) each policy u induces a probability measure denoted by P^u on $(\Omega, \mathcal{F})$ such that (1.1) holds, and X_t and A_t are random variables. The corresponding expectation operator is denoted by E_u.

A *Markov policy* $u \in U(M)$ is characterized by the dependence of $u_{t+1}(\cdot \mid h_t, X_{t+1})$ 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 $\mathbf{A}$, so that $p^g_{A(x)|x} = 1$; under g, $\{X_t\}$ becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} = \sum_{a \in A(x)} p^g_{a|x} P_{xay}$.

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 one point for each x.

Let $c(x, a)$, $d^k(x, a)$, $k = 1, ..., K$ be (real valued) cost functions and define the average costs of a policy u:

$$C_z(u) = \overline{\lim_{t\to\infty}} \frac{1}{t} E_u \left[\sum_{s=1}^{t} c(X_s, A_s) \mid X_1 = z \right] \quad , \quad D^k_z(u) = \overline{\lim_{t\to\infty}} \frac{1}{t} E_u \left[\sum_{s=1}^{t} d^k(X_s, A_s) \mid X_1 = z \right] \tag{1.2}$$

Introduce the constrained optimization problem **COP:**

$$\textit{minimize } C_z(u) \qquad \textit{subject to} \qquad D^k_z(u) \le V_k \quad 1 \le k \le K \tag{1.3}$$

A policy u is said to be feasible (for COP) if $D^k_z(u) \le V_k$, $1 \le k \le K$. A class of policies U' is called *sufficient* for an optimization problem if for any feasible policy u there is a feasible policy $u' \in U'$ which performs at least as well, i.e. $C_z(u') \le C_z(u)$.

Often, (see Derman (1970) and Altman and Shwartz (1987, 1991b)) the costs in (1.2) can be represented in terms of the *expected state-action frequencies*: let $\bar{f}^T_{x,u}(y, a) := \frac{1}{T} \sum_{s=1}^{T} P^u(X_s = y, A_s = a \mid X_1 = x)$. A state action frequency matrix $\bar{f}_{x,u}$ is any accumulation point of the matrix $\{\bar{f}^T_{x,u}(y, a)\}_{y,a}$ as $T \to \infty$, where convergence is pointwise, i.e. for each point (x, a). (This matrix thus describes an accumulation point in a countable-dimensional space with one coordinate for each state action pair). Let $\bar{F}_{x,u}$ denote the set of all matrices $\bar{f}_{x,u}$.

A class of policies U' determines a set of accumulation points denoted $L_x(U') := \cup_{u \in U'} \bar{F}_{x,u}$. The set of frequencies obtained from all policies is $L_x := \cup_{u \in U} \bar{F}_{x,u}$.

A class of policies U' is called complete if for each state x, policy u and frequency matrix $\bar{f}_{x,u}$ there is a policy u' in U' such that $\bar{f}_{x,u'} = \bar{f}_{x,u}$, and $\bar{F}_{x,u'}$ is a singleton. The notion of completeness will be used for establishing the sufficiency of p.t.s. policies.

The following notation is used below: $1\{A\}$ is the indicator function of the set A. coA denotes the convex hull of a set A, and $\bar{c}oA$ is the closure of coA.

2. POLICY TIME SHARING

We begin by defining a class of p.t.s. policies $U(pts)$. Fix a finite set of stationary deterministic policies $\{g_i \ , \ 1 \leq i \leq n\}$, assume:

A0. The state space forms a single positive recurrent class when using any g_i.

Fix an arbitrary state and, without loss of generality, let it be state zero. A cycle is defined as the period between consecutive visits to this state. More precisely, pose:

$\eta(1) := \inf \{t \geq 1 : X_t = 0\}$

$\eta(k+1) := \inf \{t > \eta(k) : X_t = 0\}$ with the standard convention that $\inf \emptyset := \infty$, and $\eta(k) = \infty$ implies $\eta(k+1) = \infty$. Thus $\eta(k)$ is the time when the kth cycle begins.

Roughly speaking, a policy $\hat{\alpha}$ of the p.t.s. type is specified through the vector of constants $\alpha := (\alpha_1, ..., \alpha_n) \in [0,1]^n$ satisfying $\sum_{i=1}^{n} \alpha_i = 1$. It is defined as any policy that uses during each cycle a constant policy g_i , and so that the fraction of cycles during which a policy g_i is used converges a.s. to α_i. For example, if $\alpha = (3/6, 2/6, 1/6, 0, ..., 0)$ then $\hat{\alpha}$ might be realized by using g_1 during three cycles, g_2 during two and g_3 for one cycle every 6 cycles, and never using other policies. Note that the α_j's need not be rational, and the realization is not necessarily periodic. More formally, let

$r(t) := \max \{k : \eta(k+1) < t\}$ be the total number of cycles completed prior to t.

$r_i(t) := \sum_{k=1}^{r(t)} 1\{g_i \ \textit{used during} \ [\eta(k), \eta(k+1))\}$ is the number of cycles prior to t during which g_i is used.

Definition: A policy $\hat{\alpha}$ is p.t.s. with parameter α if it uses a fixed policy during each cycle and, for each i

$$P^{\hat{\alpha}}\left[\lim_{k\to\infty}\frac{r_i(\eta(k))}{r(\eta(k))}=\alpha_i \ \mid \eta(k)<\infty \ \ for\ all\ k\geq 1\right]=1\ . \tag{2.1}$$

Note that from the definition of p.t.s policies $\sum_{i=1}^{n} r_i(t)=r(t)$, and that by A0, under any p.t.s. policy $\hat{\alpha}$, $\eta(k)$ are finite for all k $P^{\hat{\alpha}}$ a.s.

For a given parameter α there could be many corresponding p.t.s. policies which satisfy (2.1). Here is one example. For each k, define the set

$$I_k=\text{argmin}_i\ \left\{\frac{r_i(\eta(k))}{k}-\alpha_i\right\}$$

At cycle k, $k=1,2,\ldots$ use the policy g_i^* that satisfies $i^*=\min\{i:\ i\in I_k\}$. It is not difficult to establish (see appendix of Altman and Shwartz (1991a)) that

$$\max_i\left|\frac{r_i(\eta(k+1))}{k+1}-\alpha_i\right|\leq\max\left\{\frac{k}{k+1}\max_i\left|\frac{r_i(\eta((k))}{k}-\alpha_i\right|\ ,\ \frac{3n}{k+1}\right\}$$

so that this policy satisfies (2.1), and is therefore a p.t.s. policy with parameter α.

Another example for construction of p.t.s. policies used in adaptive problems is given in Altman and Shwartz (1991a) Section 3.1.

Technical Remark: Clearly, the $\eta(k)$ can be constructed solely from information about the state process X_t and are thus random variables (in fact stopping times) on the probability space described in Section 1. Moreover, it is not difficult to see that under the p.t.s. policy described above, $r(t)$ and $r_i(t)$ can be computed using the number of times the event $X_s=0$ occurred, $s\leq t$. This implies that $r(t)$ and $r_i(t)$ are random variables, and the policy described above satisfies the formal definition of a policy as given in Section 1. The same clearly holds whenever there is a pre-defined rule for choosing the policy g_i based on the number of times the event $X_s=0$ occurred.

However, suppose that at each $\eta(k)$, $k=1,\ldots$ one tosses a die with n facets, with probability α_i to obtain facet i, and then according to the outcome, use the respective policy g_i till $\eta(k+1)$. The law of large numbers guarantees that under this decision rule, (2.1) holds. However, this would not fit into the standard definition of policy in Markov Decision Problems as given in Section 1, since in this case, in order to use a fixed g_i throughout a cycle, we need to "remember" which policy was chosen at the beginning of the cycle. This information cannot, in general, be recovered from past states and actions alone.

Decision rules involving randomization arise naturally in the context of adaptive problems. In general, to deal with decision rules which use information about the chosen policy (e.g. allow randomization), one has to construct an appropriate enlarged probability space on which to define the MDP; see, e.g. Schäl

(1984). For the rest of the paper, we shall restrict to policies satisfying the formal definitions of Section 1, although all the results can be extended after the appropriate construction of an enlarged probability space.

2.1. The cost under p.t.s.

An expression for $C(\hat{\alpha})$ is now developed, under the following assumption:

A1: $E_{g_i}\left[\sum_{\eta(1)}^{\eta(2)-1}|c(X_s,A_s)|\right] < \infty$ for $1 \leq i \leq n$.

By A0, this holds trivially whenever the function c is bounded. Define recursively:

$\xi_i(0) = 0$, and for $k = 0, 1, \ldots$,

$\eta_i(k+1) := \inf\{t \geq \xi_i(k)+1 \;:\; X_t = 0$ and the controller decides to use g_i during the cycle starting at time $t\}$

$\xi_i(k+1) := \inf\{t > \eta_i(k+1) : X_t = 0\} - 1$.

Thus the cycle at which g_i is used for the kth time begins at time $\eta_i(k)$ and ends at time $\xi_i(k)$. Let

$Y_i(k) := \sum_{t=\eta_i(k)}^{\xi_i(k)} c(X_t, A_t)$ is the total cost over the cycle during which g_i is used for the kth time.

If $\eta_i(k) = \infty$ set $Y_i(k) := 0$. Let u be a policy that uses g_i during at least l cycles (P^u a.s.).

Lemma 2.1: *Assume A0-A1. Then the expected cost $W_i := E_u[Y_i(k)]$ exists, and does not depend on k, $k \leq l$.*

Proof: This is a consequence of the strong Markov property, the definition of $Y_i(k)$ and the result for the stationary case in Chung (1967) p. 84. ■

Under A0-A1 the cost (1.2) under any stationary policy is independent of the initial state, which is then omitted (see e.g. Chung 1967) . Denote the duration of the kth cycle by $\sigma(k) := \eta(k+1) - \eta(k)$, $k = 1, 2, \ldots$ and the expected duration of a cycle, when using g_i by

$$\tau_i := E_{g_i}(\sigma(1) \mid X_1 = 0) < \infty \; . \tag{2.2}$$

Note that Lemma 2.1 implies that $E_u[\xi_i(k) + 1 - \eta_i(k)] = \tau_i$ is independent of k.

Lemma 2.2: *Under A0, for a p.t.s. policy $\hat{\alpha}$*

$$\lim_{t\to\infty} \frac{r_i(t)}{t} = \frac{\alpha_i}{\sum_{j=1}^{n} \tau_j \alpha_j} \qquad P^{\hat{\alpha}} \textit{ a.s. and in } L_1 \; .$$

Proof: By A0, under any g_j, $\eta(1) < \infty$ w.p.1, and by a simple induction argument $\eta(k) < \infty$ w.p.1 under any policy $\hat{\alpha}$. This implies $\lim_{t\to\infty} r(t) = \infty$ $P^{\hat{\alpha}}$ *a.s.* Thus, since $r_i(t)$ and $r(t)$ change only at cycle completion times, (2.1) implies

$$\lim_{t\to\infty} \frac{r_i(t)}{r(t)} = \alpha_i \qquad P^{\hat{\alpha}}\ a.s. \quad , \qquad 1 \le i \le n \tag{2.3}$$

Let $b_i(t) := \sum_{s=1}^{t} 1\{g_i \text{ used at } s\}$ denote the total time prior to t during which g_i is used. Note that $\sum_{i=1}^{n} b_i(t) = t$. For the rest of the proof, fix i such that $\alpha_i \neq 0$. For large t, the ratio on the left side of (2.4) is then well defined, and (2.3) implies

$$\lim_{t\to\infty} \frac{r_j(t)}{r_i(t)} = \frac{\alpha_j}{\alpha_i} \qquad P^{\hat{\alpha}}\ a.s., \qquad 1 \le i, j \le n \tag{2.4}$$

Now for any j for which $\lim_{t\to\infty} r_j(t) = \infty$ (and in particular, for $j = i$),

$$\varliminf_{k\to\infty} \frac{b_j(\eta(k))}{r_j(\eta(k))} \le \varliminf_{t\to\infty} \frac{b_j(t)}{r_j(t)} \le \varlimsup_{t\to\infty} \frac{b_j(t)}{r_j(t)} \le \varlimsup_{k\to\infty} \frac{b_j(\eta(k))}{r_j(\eta(k-1))} = \varlimsup_{k\to\infty} \frac{b_j(\eta(k))}{r_j(\eta(k))} \tag{2.5}$$

But by the strong law of large numbers the first and last expressions in (2.5) equal τ_j, so

$$\lim_{t\to\infty} \frac{b_j(t)}{r_j(t)} = \lim_{k\to\infty} \frac{b_j(\eta(k))}{r_j(\eta(k))} = \tau_j \qquad P^{\hat{\alpha}}\ a.s. \tag{2.6}$$

whenever $\lim_{t\to\infty} r_j(t) = \infty$. On the other hand, if $\lim_{t\to\infty} r_j(t) < \infty$ on a set of positive probability, then necessarily $\lim_{t\to\infty} b_j(t) < \infty$ $P^{\hat{\alpha}}$ a.s. on that set, so that $\varlimsup_{t\to\infty} \frac{b_j(t)}{r_j(t)} < \infty$ $P^{\hat{\alpha}}$ a.s. on that set. It therefore follows from (2.4) that if $\alpha_j = 0$ then $\lim_{t\to\infty} \frac{b_j(t)}{r_i(t)} = \lim_{t\to\infty} \frac{b_j(t)}{r_j(t)} \frac{r_j(t)}{r_i(t)} = 0$. Setting $J_\alpha := \{1 \le j \le n \mid \alpha_j \neq 0\}$, since $\sum_{j=1}^{n} b_j(t) = t$ the last remark implies

$$\lim_{t\to\infty} \frac{b_i(t)}{t} = \lim_{t\to\infty} \frac{b_i(t)/r_i(t)}{\sum_{j=1}^{n} b_j(t)/r_i(t)} = \lim_{t\to\infty} \frac{b_i(t)/r_i(t)}{\sum_{j\in J_\alpha} b_j(t)/r_i(t)} = \lim_{t\to\infty} \frac{b_i(t)/r_i(t)}{\sum_{j\in J_\alpha} [b_j(t)/r_j(t)][r_j(t)/r_i(t)]} \tag{2.7}$$

since the limits of all ratios in the last equation exist $P^{\hat{\alpha}}$ a.s. Finally, using (2.4), (2.6) and (2.7)

$$\lim_{t\to\infty} \frac{b_i(t)}{t} = \frac{\alpha_i \tau_i}{\sum_{j\in J_\alpha} \tau_j \alpha_j} = \frac{\alpha_i \tau_i}{\sum_{j=1}^{n} \tau_j \alpha_j} := z_i(\alpha) \qquad P^{\hat{\alpha}}\ a.s. \tag{2.8}$$

A fortiori, (2.8) holds for $\alpha_j = 0$ as well, since $\sum_1^n z_i = 1$ and by (2.8) $\lim_{t\to\infty}\ t^{-1} \sum_{j\in J_\alpha} b_j(t) = 1$ and so for all $j \notin J_\alpha$ we have $\varlimsup_{t\to\infty} \frac{b_j(t)}{t} = 0$.

As a result of (2.6) and (2.8) we obtain for $\alpha_i \neq 0$:

$$\lim_{t\to\infty} \frac{r_i(t)}{t} = \frac{\alpha_i}{\sum_{j=1}^{n} \tau_j \alpha_j} \qquad P^{\hat{\alpha}}\ a.s. \text{ and in } L_1 \tag{2.9}$$

The convergence in L_1 follows from the a.s. convergence since $r_j(t)/t$ is bounded by 1. If $\alpha_i = 0$ (2.9) holds as well, since $t \ge r(t)$ and in that case by (2.3), $r_i(t)/r(t) \to 0$. ∎

Remark: For rational α_i's one can use a periodic realization of the p.t.s. policy. For such a realization, the calculation (2.8) follows from well known results on alternating Markov renewal processes (e.g. Karlin and Taylor 1975).

The parameters α_i will be rational, for example, if all transition probabilities, constraints and cost functions are rational numbers. This may be a reasonable assumption in the modelling of computers, where only rational numbers can be stored. Rational parameters can also serve as an approximation, provided some continuity properties are established.

However, in general there is no guarantee that the solution of problems such as COP would yield rational α_i's. Moreover, even if all α_i are rational, requiring a periodic realization may place an undesirable restriction on the policy.

Lemma 2.3: *Under A0-A1, for each j:*

$$\lim_{t\to\infty} E_{\hat\alpha} \left[\frac{1}{t}\sum_{k=1}^{r_j(t)} Y_j(k)\right] = \lim_{t\to\infty} E_{\hat\alpha} \left(\frac{r_j(t)}{t}\right) W_j$$

The proof is technical and is left to the Appendix.

Theorem 2.4: *Under assumptions A0 and A1 the cost $C_x(\hat\alpha)$ of a p.t.s. policy $\hat\alpha$ is independent of the initial state x, and is given by*

$$C(\hat\alpha) = \sum_{j=1}^{n} z_j(\alpha) C(g_j) \tag{2.10}$$

where $z_j(\alpha)$ is given by

$$z_j(\alpha) = \alpha_j \tau_j \left[\sum_{l=1}^{n} \alpha_l \tau_l\right]^{-1} \tag{2.11}$$

Proof: In calculating $C(\hat\alpha)$ we note that under A0-A1 the expected cost till $\eta(1)$ has zero contribution, and so does the expected cost between the last visit to state 0 and t (see Chung 1967 p. 89 and p. 93). This fact and the definitions imply that $C(\hat\alpha)$ is obtained by the following calculation:

$$C(\hat\alpha) = \overline{\lim_{t\to\infty}} E_{\hat\alpha}\, \frac{1}{t}\sum_{s=1}^{t} c(X_s, A_s) = \overline{\lim_{t\to\infty}}\, E_{\hat\alpha} \sum_{j=1}^{n} \left[\frac{1}{t}\sum_{k=1}^{r_j(t)} Y_j(k)\right] = \sum_{j=1}^{n} \overline{\lim_{t\to\infty}}\, E_{\hat\alpha} \left[\frac{1}{t}\sum_{k=1}^{r_j(t)} Y_j(k)\right] \tag{2.12}$$

where the change of order between the (finite) summation and the "$\overline{\lim}$" is justified since by Lemma 2.2 and Lemma 2.3 $\lim_{t\to\infty} E_{\hat\alpha} \left[\frac{1}{t}\sum_{k=1}^{r_j(t)} Y_j(k)\right]$ exists for all j. From Chung (1967 pp. 95-96), under A0-A1

$$C(g_i) = W_i/\tau_i \tag{2.13}$$

We finally obtain using Lemma 2.2, Lemma 2.3 and (2.13)

$$C(\hat{\alpha}) = \sum_{j=1}^{n} \lim_{t\to\infty} E_{\hat{\alpha}} \left(\frac{r_j(t)}{t}\right) W_j \; = \; \sum_{j=1}^{n} \frac{\alpha_j \tau_j}{\sum_{k=1}^{n} \tau_k \alpha_k} C(g_j) \; = \; \sum_{j=1}^{n} z_j(\alpha) C(g_j) \tag{2.14}$$

which establishes (2.10). ■

Remarks:

1. The theorem implies that $C(\hat{\alpha})$ depends only on α.
2. Both the theorem and the calculations that follow hold in fact not only to p.t.s. of deterministic policies, but also to time sharing between a finite number of *stationary* policies (possibly randomized).
3. When implementing p.t.s. policies, the controller uses the ratios vector α in order to satisfy (2.1). As is seen below (e.g. Theorem 3.2) the calculation of optimal policies yields the desired value of the vector z. From (2.11) it then follows that α is obtained as

$$\alpha_i(z) = \frac{z_i}{\tau_i} \left[\sum_{j=1}^{n} \frac{z_j}{\tau_j} \right]^{-1} \tag{2.15}$$

Assumption A1 may be difficult to check. It is related to finiteness of the cost;

Lemma 2.5: *Assume A0. If $C(g_i)$ is finite and $c(x,a)$ is uniformly bounded from below, then A1 holds. If A1 holds then $C(g_i)$ is finite.*

Proof: Let $c_t^M := c(X_t, A_t) 1\{c(X_t, A_t) \le M\}$ and define $Y_i^M(k)$, W_i^M and $C_x^M(u)$ as in the definition of $Y_i(k)$, W_i and $C_x(u)$, with c_t^M replacing $c(X_t, A_t)$. Then for all ω the following monotone convergence holds:

$$Y_i(1) \; = \; \sum_{s=\eta_i(1)}^{\xi_i(1)} c(X_s, A_s) \; = \; \lim_{M\to\infty} \sum_{s=\eta_i(1)}^{\xi_i(1)} c_s^M \; = \; \lim_{M\to\infty} Y_i^M(1) \tag{2.16}$$

By (2.13), $C(g_i) \ge \overline{\lim}_{M\to\infty} C^M(g_i) = \overline{\lim}_{M\to\infty} W_i^M/\tau_i = W_i/\tau_i$ by the monotone convergence theorem. The second part of the Theorem follows from (2.13) (Chung 1967). ■

2.2. The finite irreducible case: sufficiency of p.t.s.

In this section we show that p.t.s. are sufficient for a large class of optimization problems (e.g. minimizing the cost $C_x(u)$, under several constraints).

Throughout this section, assume A0 and

A2. Both state and action spaces are finite.

Let $\{g_i, 1 \leq i \leq n\}$ be the set of all stationary deterministic policies and define $U(pts)$ as the collection of p.t.s. policies obtained by using those policies. By Lemma 2.5, under A0 and A2 assumption A1 is equivalent to $C(g_i) < \infty$, so that A0 and A2 imply A1.

The reason for introducing the state-action frequencies is given in the following lemma (Derman 1970):

Lemma 2.6: *Under A0 and A2, for every policy $u \in U$, there exists some $\bar{f}_{x,u}$ for which the cost $C_x(u)$ can be represented by:*

$$C_x(u) = \sum_{\{y\in\mathbf{X}\}} \sum_{\{a\in A(y)\}} c(y,a)\bar{f}_{x,u}(y,a) \tag{2.17}$$

This representation of the cost in terms of the state-action frequencies makes it possible to transform optimization problems of $C_x(u)$ over the space of all policies, to optimization problems over the state-action frequencies. The advantage of doing so is in the linear form of the cost, when expressed in terms of state-action frequencies.

Note that $U(pts)$ does not include all stationary policies; however, this class is as rich as $U(S)$ in the following sense:

Theorem 2.7: *Under assumptions A0 and A2, U(pts) is complete.*

Proof: To establish the Theorem we need to show that for any u, the matrix $\bar{f}_{x,u}$ is in $L_x(pts)$, or that $L_x \subseteq L_x(pts)$. Derman (1970) has already shown that under A2 $L_x = L_x(S) = \bar{co}(L_x(SD))$. Therefore we only have to show that $\bar{co}(L_x(SD)) \subseteq L_x(pts)$ (because obviously, $L_x(pts) \subseteq L_x$). Since $L_x(SD)$ is closed (being finite) and since by definition $L_x(SD) \subseteq L_x(pts)$, it suffices to show that any convex combination of frequencies in $L_x(SD)$ is in $L_x(pts)$, i.e. that for any $\beta_j, 1 \leq j \leq l$, $\sum_{j=1}^{l} \beta_j = 1$ $\beta_j \geq 0$ and stationary deterministic policies g_j, there exists $\hat{\alpha} \in U(pts)$ so that:

$$\bar{f}_{\hat{\alpha}}(y,a) = \sum_{j=1}^{l} \beta_j \bar{f}_{g_j}(y,a) \tag{2.18}$$

for all y and a. To prove (2.18), fix y, a and define $c(a', x) := 1\{a' = a, x = y\}$. Then $C(g_k) = \bar{f}_{g_k}(y,a)$. Substitute β in (2.15) to obtain $\alpha = \alpha(\beta)$. Theorem 2.4 then implies that (2.18) holds for each $\{y, a\}$. ■

Corollary 2.8: *Under the assumptions A0 and A2, $U(pts)$ are sufficient for COP.*

Proof: Follows from Altman and Shwartz (1991b) Corollary 2.9. ■

3. OPTIMIZATION UNDER CONSTRAINTS

In this section we give some applications of p.t.s. policies. The results use the linearity of time average cost (1.2) (when using p.t.s.) in the time-ratio coefficients z_i's (Theorem 2.4), and the completeness of U(pts) (Theorem 2.7). These results enable the easy derivation of simple algorithms for constrained optimization problems using p.t.s. policies.

The first applications in subsection 3.1 deal with optimization under several constraints. The second application in 3.2 is the case of finite state space and a compact metric action space, with one constraint.

3.1. Optimization under several constraints.

Consider the optimization problem COP with m constraints on average cost functionals, with finite irreducible state space and finite action space. This problem is discussed in Hordijk and Kallenberg (1984) and Ross (1988), who find an optimal stationary policy. Hordijk and Kallenberg (1984) generalize this problem to multichain case, and find an optimal Markovian policy. (At the end of this subsection we compare these methods to ours).

Denote the (finite) number of stationary deterministic policies $\{g_j\}$ by l. Theorem 3.2 shows that the solution of COP can be reduced to a solution of the following linear programming problem:

P3.1: Find a vector of l real numbers: $\beta := \{\beta_1, ..., \beta_l\}$ so as to Minimize:

$$C(\hat{\beta}) = \sum_{j=1}^{l} \beta_j C(g_j) \tag{3.1a}$$

Under the constraints:

$$D^k(\hat{\beta}) = \sum_{j=1}^{l} \beta_j D^k(g_j) \leq \ V^k \qquad 1 \leq k \leq m \tag{3.1b}$$

$$\sum_{j=1}^{l} \beta_j \ = \ 1 \quad , \quad \beta_j \geq 0 \tag{3.1c}$$

Denote by U_V the class of policies that meet the constraints in COP. The following theorem states that p.t.s. policies are sufficient for problem COP, i.e. if U_V is non empty then there exists an optimal policy for COP of the p.t.s. type which can be found by solving P3.1.

Theorem 3.2: *Under A0 and A2,*

(i) U_V is nonempty if and only if there exists a vector β that meets the constraints (3.1b) and (3.1c).

(ii) If β_i is optimal for P3.1 then the policy $\hat{\alpha}$ is optimal for COP, where $\alpha_i = \frac{\beta_i}{\tau_i}\left[\sum_{j=1}^{n}\frac{\beta_j}{\tau_j}\right]^{-1}$

Proof: Follows from Corollary 2.8 and the representation of the cost through (2.10)–(2.11). ■

Remark: In general, there are methods for solving problem COP which are computationally more efficient. Using the methods of Derman (1970), Hordijk and Kallenberg (1984) and Ross (1988), if there are $|\mathbf{X}|$ states and $|A|$ actions per state, then the number of decision variables in the associated linear program is of the order of $|\mathbf{X}| \cdot |A|$, and no additional computations are required. Solving COP via P3.1 requires $|A|^{|\mathbf{X}|}$ decision variables; furthermore, all $C(g_i)$'s and τ_i's must be calculated. Thus in general solving COP through P3.1 is feasible only when the number of states (and actions) is small. However, in some cases *it is possible to determine a-priori some small subclass of policies G*, so that P3.1 may be restricted to this subclass. If this subclass is small enough, P3.1 becomes feasible, and sometimes even preferable to existing methods. The constrained optimization problem below serves to illustrate this point.

Since p.t.s. policies are defined through the limiting property (2.1), they can be used to solve adaptive problems (see Altman and Shwartz (1987, 1991a)).

A countable-state constrained optimization problem: Consider a system of K infinite-capacity queues with i.i.d. arrivals. A single server can choose to serve one queue at a time, and provides Bernoulli service with queue-dependent parameter. The cost functions are of the form $c(x,a) = \sum_{i=1}^{K} c_i X_i(t)$ where $c_i \geq 0$ (with the same structure for $d^k(x,a)$) and $X_i(t)$ is the number of units in queue i at time t.

The state space for this system is not finite, and therefore the methods of Derman, Hordijk and Kallenberg or Ross are not applicable. On the other hand, it is shown in Altman and Shwartz (1986, 1987) that in order to solve such a constrained optimization problem it suffices to consider p.t.s. policies that make use of the class G of the $K!$ strict priority policies. Thus an optimal policy is found by solving P3.1. For this system there are explicit methods to compute $C(g_j)$ and τ_j, so that P3.1 provides (the only) solution of this problem which can be computed explicitly.

Note that due to well known results in linear programming (e.g. Minoux 1983), the number of different stationary deterministic policies that are used by the optimal policy $\hat{\beta}$ (or in other words the number of nonzero β_i's) in COP need not exceed $m+1$. In that sense there is a resemblance to the stationary policies, as it is known that we can find an optimal stationary policy for problem COP that randomizes at no more than m states (Ross 1988).

3.2. One constraint with compact action space.

In this section, we apply p.t.s. for the case of one constraint with a finite state space and a compact metric action space. This case was already investigated by Beutler and Ross (1985). They showed that if there exists any feasible policy then there exists an optimal "randomized" policy obtained by randomly choosing at each decision epoch one out of two deterministic stationary policies. As we show below, it is possible to solve this problem using the p.t.s. approach, i.e. replace the randomization with time-sharing. The advantage of doing so is that the parameter α which determines the optimal p.t.s. policy has an explicit expression (3.13), whereas the parameter that determines the optimal randomized policy suggested in Beutler and Ross (1985) can be obtained only through (usually complicated) numerical approximations.

As in Beutler and Ross (1985) we assume:

A3. the state space $\mathbf{X} = \{0, 1, ..., N\}$ is finite. The action space A is a compact metric space equipped with the σ-algebra generated by its open sets.

Let $d(x, a)$ be the cost paid when at state x and when choosing action a. Assume

A4. $c(x, a)$ and $d(x, a)$ are non negative, and $c(x, \cdot)$ and $d(x, \cdot)$ are continuous on the action space A.

A5. For every $g \in U(SD)$, the state 0 is accessible from each $x \in \mathbf{X}$.

The optimization problem consists of maximizing the average reward R, subject to a constraint on the average cost K, where:

$$R_x(u) := \underline{\lim}_{t\to\infty} \frac{1}{t} E_u \left[\sum_{s=1}^{t} c(X_s, A_s) \mid X_1 = x \right] \tag{3.2}$$

$$K_x(u) := \overline{\lim}_{t\to\infty} \frac{1}{t} E_u \left[\sum_{s=1}^{t} d(X_s, A_s) \mid X_1 = x \right] \tag{3.3}$$

The constraint on the average cost is given by:

$$K_x(u) \leq V \tag{3.4}$$

Here again we define U_V as the subclass of U on which the constraint is met, and policies in U_V are called feasible. For the precise statement of the problem, define:

$$R_x = \ sup_{u \in U_V} R_x(u) \tag{3.5}$$

The problem we solve in this section is:

P3.3: Find an optimal constrained policy for $R_x(\cdot)$, i.e. a feasible policy attaining R_x for every x.

In order to solve problem P3.3, we introduce an unconstrained problem with parameter λ as in Beutler and Ross (1985). Finding optimal policies for the new problem is easier, and that will help to determine a class of optimal p.t.s. policies for problem P3.3. Define the reward of the new problem:

$$b^\lambda(x,a) = c(x,a) - \lambda d(x,a) \tag{3.6}$$

and the average reward:

$$J_x^\lambda(u) := \underline{\lim}_{t\to\infty} \frac{1}{t} E_u \left[\sum_{s=1}^{t} b^\lambda(X_s, A_s) \mid X_1 = x \right] \tag{3.7}$$

Define $J_x^\lambda = \; sup_{u\in U} J_x^\lambda(u)$. Assume P3.3 is feasible yet non trivial, i.e. U_V is non empty but if $\tilde{R}_x(u') := \sup_{u\in U} R_x(u)$ then $K_x(u') > V$. Under this assumption it was shown by Beutler and Ross (1985) that there exist a positive real number γ and two deterministic policies g_1, g_2 with the properties that $J_x^\gamma(g_1) \;=\; J_x^\gamma(g_2) \;=\; J_x^\gamma$ and $K(g_1) \le V \le K(g_2)$.

Remarks: There are no explicit methods for calculating the policies g_i although a characterization is given by Beutler and Ross (1985). Moreover, Beutler and Ross require that these policies be solutions of the optimality equation associated with the cost b^γ. This requirement is not needed here, so that adequate policies may be easier to find.

It is possible to generalize the problem discussed in this section to some systems with countable state space, under appropriate assumptions. See Altman and Shwartz (1986, 1988) and references therein. In the case treated there, g_i are easily found.

We now construct a class of optimal policies $\hat{\alpha}$ for P3.3 by time-sharing between the deterministic policies g_1 *and* g_2. Under A3 and A5, the expected cycle times τ_1 and τ_2 defined in (2.2) are finite. (Note that the fact that g_1 and g_2 are stationary deterministic allows us to use expressions (2.10)-(2.15), (2.15) for the cost, in spite of the fact that this Section deals with compact metric action space). By Theorem 2.4,

$$J_x^\gamma(\hat{\alpha}) \;=\; z_1 J_x^\gamma(g_1) + z_2 J_x^\gamma(g_2) = J_x^\gamma \tag{3.8}$$

where $z_i = \alpha_i \tau_i [\alpha_1\tau_1 \;+\; \alpha_2\tau_2]^{-1}$ and the last equality follows since $z_1 + z_2 \;=\; 1$ and $J_x^\gamma(g_1) \;=\; J_x^\gamma(g_2) \;=\; J_x^\gamma$.

$$K_x(\hat{\alpha}) \;=\; z_1 K_x(g_1) + z_2 K_x(g_2) \tag{3.9}$$

$$R_x(\hat{\alpha}) \;=\; z_1 R_x(g_1) + z_2 R_x(g_2) \tag{3.10}$$

¿From Theorem 2.4, the costs (3.2), (3.3) and (3.7) are obtained as limits, and so by (3.6),

$$J_x^\gamma(\hat{\alpha}) \;=\; R_x(\hat{\alpha}) - \gamma K_x(\hat{\alpha}) \tag{3.11}$$

Lemma 3.4: *If $K_x(\hat{\alpha}) = V$ then $\hat{\alpha}$ is an optimal policy for P3.3.*

Proof: for any u and any x, $J_x^\gamma(\hat{\alpha}) \;\geq\; J_x^\gamma(u)$ because of (3.8). Now by definition of J, R and K, $J_x^\gamma(u) \;\geq\; R_x(u) - \gamma K_x(u)$. If u is feasible, then $K_x(u) \leq V$, so we get:

$$R_x(\hat{\alpha}) \;\geq\; R_x(u) + \gamma[V - K_x(u)] \;\geq\; R_x(u) \tag{3.12}$$

and since $\hat{\alpha}$ is feasible, the result follows. ■

A class of optimal p.t.s. policies for P3.3 is now easily obtained:

Theorem 3.5: *The class of p.t.s. policies $\hat{\alpha}$ for which α is given by:*

$$\alpha_1 \;=\; \frac{\tau_2[V - K(g_2)]}{\tau_2[V - K(g_2)] - \tau_1[V - K(g_1)]} \qquad\qquad \alpha_2 = 1 - \alpha_1 \tag{3.13}$$

is optimal for problem P3.3.

Proof: According to Lemma 3.4, if $V \;=\; K_x(\hat{\alpha})$ then $\hat{\alpha}$ is optimal for P3.3. From (3.9) this implies $V \;=\; z_1 K(g_1) + z_2 K(g_2)$ and since $z_1 + z_2 = 1$, we obtain

$$z_1 = \frac{V - K(g_2)}{K(g_1) - K(g_2)} \qquad\qquad z_2 = \frac{V - K(g_1)}{K(g_2) - K(g_1)} \tag{3.14}$$

Finally, the required expression for α is obtained from equations (2.15) and (3.14). ■

Acknowledgments: We wish to thank Professor Stidham and the anonymous associate editor and reviewers for their helpful and constructive remarks.

This work was supported in part by Grant No. 85-00306 from the United States—Israel Binational Science Foundation (BSF), Jerusalem, Israel. The research of Adam Shwartz was also performed in part while visiting the Mathematical Sciences Research Center, AT&T Bell Laboratories. The support and hospitality of the Mathematics of Networks and Systems Department is gratefully acknowledged.

4. APPENDIX

Proof of Lemma 2.3. We follow a similar calculation by Chung (1967) p. 94. By Lemma 2.1 we have $E_{\hat{\alpha}}[Y_j(k)] = W_j$. The arguments in Chung (1967) p. 94 can be extended to show that $Y_j(k)$ is independent of $\{r_j(t)+1<k\}$, so:

$$\int_{\{r_j(t)+1<k\}} Y_j(k)P^{\hat{\alpha}}(d\omega) \;=\; P^{\hat{\alpha}}\{r_j(t)+1<k\}W_j \tag{4.1}$$

Using the fact that $r_j(t) \leq t$ we have:

$$E_{\hat{\alpha}}\left[\sum_{k=1}^{r_j(t)} Y_j(k)\right] \;=\; E_{\hat{\alpha}}\left[\sum_{k=1}^{t} 1_{\{r_j(t)\geq k\}}Y_j(k)\right] \;=\; \sum_{k=1}^{t}\int_{\{r_j(t)\geq k\}} Y_j(k)P^{\hat{\alpha}}(d\omega) \;= \tag{4.2}$$

$$= \sum_{k=1}^{t}\left[\int Y_j(k)P^{\hat{\alpha}}(d\omega) \;-\; \int_{\{r_j(t)+1<k\}} Y_j(k)P^{\hat{\alpha}}(d\omega) \;-\; \int_{\{r_j(t)+1=k\}} Y_j(k)P^{\hat{\alpha}}(d\omega)\right] \;=$$

$$= \sum_{k=1}^{t}\left[W_j - P^{\hat{\alpha}}\{r_j(t)+1<k\}W_j - \int_{\{r_j(t)+1=k\}} Y_j(k)P^{\hat{\alpha}}(d\omega)\right] \;=$$

$$= W_j\sum_{k=1}^{t} P^{\hat{\alpha}}\{r_j(t)+1\geq k\} - \sum_{k=1}^{t}\int_{\{r_j(t)+1=k\}} Y_j(k)P^{\hat{\alpha}}(d\omega) \;=$$

$$= W_jE_{\hat{\alpha}}[r_j(t)+1] - \sum_{k=1}^{t}\int_{\{r_j(t)+1=k\}} Y_j(k)P^{\hat{\alpha}}(d\omega)$$

In order to estimate the last term above, we write:

$$\sum_{k=1}^{t}\int_{\{r_j(t)+1=k\}} Y_j(k)P^{\hat{\alpha}}(d\omega) \;\leq\; \sum_{k=1}^{t}\int_{\{r_j(t)+1=k\}} \max_{1\leq k\leq t} Y_j(k)P^{\hat{\alpha}}(d\omega) \;=\; E_{\hat{\alpha}}\left[\max_{1\leq k\leq t} Y_j(k)\right] \tag{4.3}$$

It is shown in Chung (1967) that for any sequence of t i.i.d. random variables having finite expectation, the expectation of the maximal random variable divided by t converges to 0 as $t\to\infty$. Applying that, we obtain from (4.2)

$$\lim_{t\to\infty}\; E_{\hat{\alpha}}\;\left[\frac{1}{t}\sum_{k=1}^{r_j(t)} Y_j(k)\right] \;=\; \lim_{t\to\infty}\; E_{\hat{\alpha}}\;\left(\frac{r_j(t)}{t}\right)W_j\;. \tag{4.4}$$

■

REFERENCES

Altman E. and A. Shwartz, 1986. "Optimal priority assignment with general constraints", *Proc. of the 24th Allerton Conf. on Communications, Control and Computing,* Univ. of Illinois Urbana-Champaign.

Altman E. and A. Shwartz, 1987. "Non stationary policies for controlled Markov chains," EE Pub. 633, Technion, Israel.

Altman E. and A. Shwartz, 1989. "Optimal priority assignment: a time sharing approach", *IEEE Trans. Auto. Control* **AC-34** pp. 1098-1102.

Altman E. and A. Shwartz, 1991a. "Adaptive Control of Constrained Markov Chains", *IEEE Trans. Auto. Control*, to appear.

Altman E. and A. Shwartz, 1991b. "Markov Decision Problems and State-Action Frequencies", *SIAM J. Control and Optimization*, to appear.

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

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

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

Fox B. L. and J. E. Rolph, 1973. "Adaptive policies for Markov renewal programs", *Ann. Stat.,* Vol. 1 pp. 334-341.

Hordijk A. and L. C. M. Kallenberg, 1984. "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* Vol. 9, no. 2 pp. 276-289.

Karlin S. and H. M. Taylor, 1975. *A first course in stochastic processes,* Academic Press, 2nd edition.

Ma D.-J. and A. M. Makowski, 1988. "Steering policies for Markov decision processes under a recurrence condition", *Proceedings of the 27th IEEE Conference on Decision and Control,* Austin (TX).

Minoux M., 1983. *"Programmation mathematique theorie et algorithmes",* Dunod, Paris.

Ross K. W., 1988. "Randomized and past-dependent policies for Markov decision processes with multiple constraints", to appear in *Journal of Operational Research.*

Schäl, M., 1975. "Conditions for optimality in dynamic programming and for the limit of n-stageoptimal policies to be optimal," *Z. Wahr. verw. Gebiete* **32**, pp. 179-196.

Schäl, M., 1984. "Asymptotic results for sequential Markov Decision models under uncertainty," *Statistics and Decision* **2**, pp. 39-62.