SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 786–809, July 1991

003

MARKOV DECISION PROBLEMS AND STATE-ACTION FREQUENCIES*

EITAN ALTMAN† AND ADAM SHWARTZ†‡

Abstract. Consider a controlled Markov chain with countable state and action spaces. Basic quantities that determine the values of average cost functionals are identified. Under some regularity conditions, these turn out to be a collection of numbers, one for each state-action pair, describing for each state the relative number of uses of each action. These "conditional frequencies," which are defined *pathwise*, are shown to determine the "state-action frequencies" that, in the finite case, are known to determine the costs. This is extended to the countable case, allowing for unbounded costs. The space of frequencies is shown to be compact and convex, and the extreme points are identified with stationary deterministic policies.

Conditions under which the search for optimality in several optimization problems may be restricted to stationary policies are given. These problems include the standard Markov decision process, as well as constrained optimization (both in terms of average cost functionals) and variability-sensitive optimization. An application to a queueing problem is given, where these results imply the existence and explicit computation of optimal policies in constrained optimization problems.

The pathwise definition of the conditional frequencies implies that their values can be controlled directly; moreover, they depend only on the limiting behavior of the control. This has immediate application to adaptive control of Markov chains, including adaptive control under constraints.

Key words. Markov decision process, average cost, constrained optimization, state-action frequencies, nonstationary control

AMS(MOS) subject classifications. 90B50, 60G17, 60J10, 93E20

Introduction. An important issue in optimization problems is the reduction of the space of policies over which we optimize. This is motivated by the need to reduce the complexity of the search for optimal policies, and by the desire to restrict attention to those policies that are easy to implement. Indeed, in many optimization problems we first show that it is possible to restrict the optimization to the class of stationary policies. This simplifies the search, since many computational methods are available in the stationary case. Furthermore, the implementation requires relatively little memory. Conditions that ensure that we may indeed restrict the search of optimal policies for Markov decision processes to stationary policies (or even to deterministic stationary policies) are an active area of research (see, e.g., Borkar [9], [10], Cavazos-Cadena [12], Dekker and Hordijk [14], Sennott [28], [29] and references therein).

On the other hand, it is of interest to know how flexible we can be in the choice of policies, in a way that does not change the values of average cost criteria. This is the case in adaptive optimization, where we often use on-line estimation schemes to generate an approximation of the optimal control (the certainty equivalence approach). The goal in this case is to achieve the same performance as in the case of full information.

These two issues are treated in this paper in the framework of the following question. For a given policy, what are the quantities that determine the values of average cost functionals? Fix a state x and an action a. For each t, consider the (random) number of times the process visited state x and action a was used by time t. It turns out that in many cases average costs are determined by the limits (in time) of the expectations of such "state-action frequencies." For each time t, consider the (random) ratio of the number of uses of action a while in state x, to the number of

* Received by the editors November 14, 1988; accepted for publication (in revised form) June 15, 1990.
† Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000 Israel.
‡ The research of this author was performed in part while he was visiting the Mathematical Sciences Research Center, AT&T Bell Laboratories.

visits to state x. Below we show that the pathwise limits of these "conditional frequencies" are the more basic quantity, in that they determine the expected state action frequencies.

We deal with countable state and action spaces, and obtain classes of policies that achieve every possible state action frequency; we term such classes "complete." In the finite case, some questions of completeness are investigated in [2], [15], and [22]; Derman [15] gives conditions for the completeness of Markov policies. Hordijk and Kallenberg [22] strengthen this result to Markov policies having just one accumulation point of the "matrix" of frequencies. Derman [15] and, later, Hordijk and Kallenberg [22] give conditions for the completeness of stationary policies. Two time sharing policies were introduced by Altman and Shwartz [1]–[3], who show that under the conditions of Derman [15], completeness depends on pathwise limit properties only, and in particular may be achieved using deterministic (but nonstationary) policies. In this paper we show that in the countable case the space of achievable frequencies is a compact convex set whose extreme points are frequencies obtained by deterministic (stationary) policies. This extends the geometric characterization given in the finite case by Derman [15], Hordijk and Kallenberg [22], and Altman and Shwartz [2].

We give conditions under which some classes of policies (such as the stationaries) are "sufficient" in the countable case for several optimization problems, including optimization under several constraints. These results allow the use of steady-state analysis of systems, which simplifies the search for optimal policies considerably. It becomes possible to translate results on performance, which in many cases deal with "steady state," into results concerning optimization (see, e.g., Altman and Shwartz [1], [3]). Previous results on the sufficiency of stationary policies in the case of countable state space dealt only with the minimization (or maximization) of a single criterion.

Then, we introduce a larger family of "sufficient" policies—the action time sharing (*ats*) policies—which is characterized by the existence of a with probability one limit to the conditional frequencies. In contrast with the standard "small" classes of policies such as the stationary policies, these policies are flexible enough to be useful for adaptive problems, as they have the following important property: the expected frequencies (and thus the cost) achieved by any policy depends only on the (pathwise) limiting behavior of the control mechanism. More precisely, it depends only on the limit of the conditional frequencies, described above. Therefore it is possible to use nonstationary algorithms based on real-time estimation of unknown parameters, and still obtain optimality. Moreover, whereas existing results on adaptive control of Markov chains consider only the optimization of a single criterion, the present results can be used to obtain adaptive controls under more general criteria, such as constrained optimization. An application of these ideas in the case of *finite* state and action spaces is given in Altman and Shwartz [2, § 5], [4]. The computation of optimal policies of the *ats* type is equivalent to the computation over the more restricted class of stationary policies, and the implementation is just as simple.

After introducing the model and some notation, § 1 provides the basic motivation by introducing the standard Markov decision problem and a constrained optimization problem. In § 2 we derive conditions under which the frequencies determine the value of optimization criteria, and under which stationary policies or other complete classes of policies are sufficient for the two optimization problems. In § 3 the basic results concerning the completeness of the stationary policies and the role of the conditional frequencies in determining the behavior of the process are derived. Since the state and action spaces are countable (and thus not compact), a tightness condition is used. The literature concerning the tightness problem is extensive; in § 4 we adapt some applicable

results. The case where tightness does not hold is treated by imposing conditions on the cost, under which tight policies are "better" than nontight ones. In § 5 it is shown that the space of frequencies is compact, and has the geometric characterization as the convex hull of the frequencies of stationary deterministic policies. This has implications to the existence of optimal policies in constrained optimization problems. Finally, we apply and extend the results of the previous sections. In § 6 we treat a queueing network, and in § 7 an equivalence between the constrained optimization problem and an associated linear program (which is well known in the finite case [15], [22]) is extended to the countable case. Section 8 treats some lesser known optimization problems involving variance.

1. The model and the problems. Let $\{X_t\}_{t=1}^{\infty}$ be a discrete time process defined on the countable *state space* $\mathbf{X}=\{0, 1, \cdots\}$. At time t an action A_t from the countable *action space* $\mathbf{A}$ is taken. Denote by $A(x)$ the set of actions available when in state x. $h_n := (X_1, A_1, X_2, A_2, \cdots, X_n, A_n)$ is the *history* of $\{X_t\}$. Denote the transition probabilities for the controlled Markov chain by

$$(1.1)\quad P_{xay} := P(X_{n+1}=y \mid X_n=x;\ A_n=a) = P(X_{n+1}=y \mid h_{n-1}=h,\ X_n=x;\ A_n=a).$$

A policy u in the *policy space* U is defined by $u=\{u_1, u_2, \cdots\}$, where u_t is applied at time epoch t, and $u_t(\cdot \mid h_{t-1}, X_t)$ is a conditional probability measure over $\mathbf{A}$. Each policy u induces a probability measure P^u on the space of paths (which serves as the canonical sample space Ω). The corresponding expectation operator is denoted by E_u.

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

Let $c(x, a)$ be a real valued function on $\mathbf{X}\times\mathbf{A}$, possibly unbounded, and let

$$C^t_x(u) = \frac{1}{t} E_u\left[\sum_{s=1}^{t} c(X_s, A_s) \mid X_1 = x\right].$$

We assume throughout that for each u, the cost $C^t_x(u)$ is well defined. This will usually follow from uniform integrability assumptions on $c(X_t, A_t)$, or from a condition that $c(\cdot, \cdot)$ is bounded below. The optimization problem OP involves the minimization of average cost functionals:

$$(1.2a)\qquad \bar{C}_x(u) = \overline{\lim_{t\to\infty}}\ C^t_x(u),$$

$$(1.2b)\qquad \underline{C}_x(u) = \underline{\lim_{t\to\infty}}\ C^t_x(u).$$

These include the standard "positive" and "negative" Markov decision problems. Given the constants V_k, $1 \leqq k \leqq K$, the constrained optimization problem COP is defined as

$$(1.3a)\qquad \text{minimize}\quad \bar{C}_x(u) \qquad \text{subject to}\quad \bar{D}^k_x(u) \leqq V_k,\quad 1 \leqq k \leqq K,$$

$$(1.3b)\qquad \text{minimize}\quad \underline{C}_x(u) \qquad \text{subject to}\quad \bar{D}^k_x(u) \leqq V_k,\quad 1 \leqq k \leqq K,$$

where $\bar{D}^k_x(u)$ is defined similarly to $\bar{C}_x(u)$ with $c(x, a)$ replaced by $d^k(x, a)$, and both $c(x, a)$ and $d^k(x, a)$ may be unbounded. For finite state and action spaces, a solution

to the constrained optimization problem based on linear programming was already obtained by Derman [15] and Hordijk and Kallenberg [22], and some variables of this linear program are limits of the state-action frequencies (1.4).

These *expected state-action frequencies* (Derman [15]) and *expected state frequencies*

$$\bar{f}^T_{x,u}(y,a) := \frac{1}{T}\sum_{s=1}^{T} P^u(X_s = y, A_s = a \,|\, X_1 = x),$$
$$\bar{f}^t_{x,u}(y) := \frac{1}{t}\sum_{s=1}^{t} P^u(X_s = y \,|\, X_1 = x) \tag{1.4}$$

are key quantities in the analysis below. Let the "matrix" $\{\bar{f}_{x,u}(y,a)\}_{y,a}$ denote a generic accumulation point of the infinite "matrix" $\{\bar{f}^T_{x,u}(y,a)\}_{y,a}$ as $T\to\infty$ (i.e., an accumulation point in a countable-dimensional space with one coordinate for each state action pair), and let $\{\bar{f}_{x,u}(y)\}_y$ denote any accumulation point of the infinite vector $\{\bar{f}^t_{x,u}(y)\}_y$. Let $\bar{F}_{x,u}$ denote the set of all limit matrices $\bar{f}_{x,u}(\cdot,\cdot)$. Any class of policies U' determines a set of accumulation points $L_x(U') := \bigcup_{u\in U'} \bar{F}_{x,u}$ and the set of all such limits is denoted by $L_x := \bigcup_{u\in U} \bar{F}_{x,u}$. We use the abbreviations $L_x(S) := L_x(U(S))$ and $L_x(D) := L_x(U(SD))$.

The following definitions are useful for the sample-path analysis of § 3. Let $f^T(y,a) := (1/T)\sum_{s=1}^{T} 1\{X_s = y, A_s = a\}$ denote the sample-path frequency at which the event of being at state y and choosing action a occurs till time T. The expectation of the random variable $f^T(y,a)$ under u starting at x is thus $\bar{f}^T_{x,u}(y,a)$. The frequency at which the event of being at state y occurs till time T is denoted by $f^T(y) = (1/T)\sum_{s=1}^{T} 1\{X_s = y\}$. Finally, $f^T(a\,|\,y) := f^T(y,a)[f^T(y)]^{-1}$ is the frequency at which action a is chosen conditioned on being in state y, until time T. If $f^T(y) = 0$ define $f^T(a\,|\,y) := 0$. Denote by $f(y,a)$ (respectively, $f(y)$) any accumulation point of $f^T(y,a)$ (respectively, $f^T(y)$) as $T\to\infty$.

Let g be a stationary policy. The following standard result will be frequently used.

LEMMA 1.1. *Assume that under g the process* $\{X_t\}_{t=1}^{\infty}$ *has one positive recurrent class, and that from any transient state, absorption into the recurrent class occurs in finite expected time. Then*

$$\bar{f}_{x,g}(y,a) = \pi^g(y)p^g_{a|y} = \lim_{t\to\infty} f^t(y,a) \quad P^g \text{ a.s.}$$

For the last equality to hold, it suffices that absorption occurs with probability one.

A class of policies U' is called *complete* if $L_x = \bigcup\{\bar{F}_{x,u'}: u'\in U'$ and $\bar{F}_{x,u'}$ *is a singleton*$\}$. U' is called *weakly complete* if

$$L_x \cap \left\{\zeta: \sum_{y,a}\zeta(y,a) = 1\right\} \subset \bigcup\{\bar{F}_{x,u'}: u'\in U' \text{ and } \bar{F}_{x,u'} \text{ is a singleton}\}.$$

Note that for each t, $\bar{f}^t_{x,u}(y,a)$ can be considered as a probability measure over $\mathbf{X}\times\mathbf{A}$. The condition $\sum_{y,a}\bar{f}_{x,u}(y,a) = 1$ for every limit point $\bar{f}_{x,u}(y,a)$ of a subsequence $\{\bar{f}^{t_n}_{x,u}(y,a)\}_n$ is equivalent to tightness of this set of measures [8]. Thus weak completeness considers only tight frequencies.

A class of policies U' is called *sufficient* for an optimization problem if for any policy u there is a policy $u'\in U'$ that performs at least as well. The motivation for studying questions of completeness and the spaces of frequencies is provided in § 2 below, where the connection between completeness and sufficiency is established. Note that sufficiency does not imply existence of an optimal policy, but rather that the search for "good" policies can be restricted to any subclass that is sufficient.

The following assumptions are used frequently in the paper:

(A0) At each state x, the set of available actions $A(x)$ is finite.

(A1) Under any policy $g \in U(S)$ the state space contains a single *positive* recurrent class, and absorption into the positive recurrent class takes place in finite expected time.

It follows from Fisher [18] that under (A0) and (A1), if there are no transient states under any policy in $U(SD)$ then the chain is ergodic under each policy in $U(S)$ if and only if it is ergodic under each policy in $U(SD)$ (see also § 5, Corollary 5.3).

(A2(u)) Given a policy u, the expected frequencies $\{\bar{f}^t_{x,u}(y, a)\}_t$ are tight.

(A2) Assumption (A2(u)) holds for all policies $u \in U$.

(A2*) The family of stationary probabilities corresponding to policies in $U(SD)$ is tight.

Remarks. (i) The issue of tightness is treated in § 4. In Lemma 4.1 we show that under the appropriate conditions, (A2) is equivalent to (A2*). We give simple verifiable sufficient conditions for (A2*), and develop some methods for the nontight case.

(ii) Assumption (A2(u)) depends on the initial state x, even when (A1) holds. For example, let u' be a policy that violates (A2(u')) (e.g., the policy constructed in [17]). Let g be a policy for which (A2(g)) holds (under (A1), this holds for any stationary policy). If u equals u' whenever $X_0 = x$ and otherwise uses g, then clearly (A2(u)) holds for all initial states except x. Throughout the paper, reference to (A2(u)) will implicitly assume a fixed initial state, which is omitted from the notation.

To make the discussion more concrete, we cite Theorem 3.2, whose proof is given in § 3.

THEOREM 3.2. *Under* (A1) *the class of stationary policies is weakly complete.*

As will become clear in § 3, the property of completeness does not depend on stationarity; it is more naturally defined through conditional frequencies. This will be seen to provide a large degree of flexibility, which can be applied in a straightforward manner to adaptive optimization problems [4].

2. Sufficiency and completeness. The aim of this section is to establish the relation between optimization problems and state-action frequencies, and in particular between sufficiency and completeness. In the case of finite state and action spaces it is known that the time average expected cost has a representation as a linear function of the expected state-action frequencies (e.g., [15]). We extend this result to the countable case, and establish sufficient conditions under which the costs (1.2a) and (1.2b) can be represented as linear functionals (2.4) of the frequencies. The advantage of this approach is that it deals directly with the cost functionals, and therefore applies to many classes of optimization problems. In the following sections we investigate the optimization problems OP and COP, and show the connection between completeness and sufficiency. In particular, we present conditions under which the search for solutions of OP and COP can be restricted to those policies for which the costs have the linear representation (2.4) in terms of the frequencies. Similar results are obtained in § 8 for other optimization problems. These results motivate the further investigation of the achievable frequencies under various classes of policies, which is carried out in § 5. We will be especially interested in finding out which classes of policies are complete. This will indicate when a class of policies is sufficient for the optimization problems OP and COP, or, in other words, whether we may restrict the search for optimal policies to smaller classes of policies. Moreover, as will become clear is § 3, this approach identifies the key quantities that determine the costs, and allows for a flexible choice of controls while keeping the cost fixed.

The results of this section concerning optimization problems are given under condition (A2), which is a rather strong "uniform stability" assumption. In § 4 we provide natural sufficient conditions for (A2), and also show how the results can be extended when tightness does not hold.

2.1. Representation of costs through frequencies. Unlike the case of finite state and action spaces, the time average expected cost in the countable case need not even have a representation as a function of the expected state-action frequencies.

Counterexample 2.1. The deterministic case. Let $P_{xy} = 1\{x+1 = y\}$ and $c(x, a) = 1$ for all states and actions. The action is thus irrelevant to both the dynamics and cost and we may assume that there is just one possible action. Under any policy u we clearly have $\bar{f}_{x,u}(y, a) = 0$, while $C_x(u) = 1$.

In this example (A1) does not hold. Counterexample 3.5 in § 3 presents a case where (A1) holds but (A2) does not, and which exhibits similar behavior.

Lemmas 2.2 and 2.3 provide conditions under which a linear representation (2.4) holds. Fix an initial state x and a policy u. Since by assumption $C_x^t(u)$ is well defined, the definitions imply

$$\bar{C}_x(u) = \overline{\lim_{t\to\infty}} \sum_{y,a} \bar{f}^t_{x,u}(y, a)c(y, a). \tag{2.1}$$

Let $\{s_n\}_n$ be a subsequence along which the limit is obtained, i.e.,

$$\bar{C}_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{s_n}_{x,u}(y, a)c(y, a). \tag{2.2}$$

Using diagonalization, choose a further subsequence $\{t_n\}_n$ so that $\bar{f}^{t_n}_{x,u}(y, a) \to \bar{f}_{x,u}(y, a)$, for all y and a.

LEMMA 2.2. *Assume* (A1) *and let* $\{c(X_s, A_s)\}_s$ *be uniformly integrable under* P^u. *If* (A2(u)) *holds then the costs* (1.2) *are a function* (2.4) *of the* $\bar{f}_{x,u}$ *defined above. If* v *is a policy such that* $\bar{F}_{x,v} = \{\bar{f}_{x,u}\}$ *and* $\{c(X_s, A_s)\}_s$ *are also uniformly integrable under* P^v, *then* $\bar{C}_x(u) = \bar{C}_x(v)$ *and* $\underline{C}_x(u) = \underline{C}_x(v)$.

Proof. Consider first the cost function defined through (1.2a). Note that for each t, $\bar{f}^t_{x,u}(\cdot, \cdot)$ can be considered as a probability measure over $\mathbf{X} \times \mathbf{A}$, and the cost $c(\cdot, \cdot)$ can then be viewed as a random variable over $\mathbf{X} \times \mathbf{A}$. The convergence $\bar{f}^{t_n}_{x,u}(y, a) \to \bar{f}_{x,u}(y, a)$ for all y and a thus translates under (A2(u)) into weak convergence of probability measures along t_n. As $\{c(X_s, A_s)\}_s$ is uniformly integrable with respect to P^u, $c(\cdot, \cdot)$ is also uniformly integrable with respect to $\{\bar{f}^t_{x,u}\}_t$; this follows from the fact that for every function h,

$$\sum_{y,a} \bar{f}^t_{x,u}(y, a)h[c(y, a)] = E_u \frac{1}{t} \sum_{s=1}^{t} [h[c(X_s, A_s)] \,|\, X_1 = x]. \tag{2.3}$$

This weak convergence of $\bar{f}^t_{x,u}$ implies the convergence of $c(\cdot, \cdot)$ in distribution, and combining this to the uniform integrability of $c(\cdot, \cdot)$ we finally obtain [8]

$$\bar{C}_x(u) = \sum_{y,a} \bar{f}_{x,u}(y, a)c(y, a). \tag{2.4}$$

The argument for (1.2b) is identical. The last claim is now immediate since $\bar{f}_{x,u} = \bar{f}_{x,v}$. □

It is not difficult to establish the following converse to Lemma 2.2. If (2.4) holds for each limit point $\bar{f}_{x,u}$ in $\bar{F}_{x,u}$ and c satisfies $\infty > \bar{c} > c(y, a) > \varepsilon > 0$, for all y, a, then (A2(u)) holds. But for an arbitrary c (2.4) may not imply (A2(u)) (for example, $c = 0$ provides no information).

Next, we discuss the representation (2.4) for stationary policies. Fix $g \in U(S)$ and an initial state x, and let $0 \in \mathbf{X}$ be recurrent under g. With the standard convention that $\inf \varnothing := \infty$, define $\eta(1) := \inf\{t \geqq 1: X_t = 0\}$, $\eta(k+1) := \inf\{t > \eta(k): X_t = 0\}$, where $\eta(k) = \infty$ implies $\eta(k+1) = \infty$.

LEMMA 2.3. *Assume* (A1) *and let* $g \in U(S)$. *Then* (A2(g)) *holds, and the representation* (2.4) *for the costs* (1.2) *holds whenever one of the following is true*: (i) $\{c(X_s, A_s)\}_s$ *are uniformly integrable with respect to* P^g; (ii) c *is bounded from below and* (A3(g)) *holds*; (iii) c *is bounded from above and* (A3(g)) *holds, where*

(A3(g))
$$E_g\left[\sum_{s=1}^{\eta(1)-1} |c(X_s, A_s)| \,\middle|\, X_1 = x\right] = \infty$$
$$\textit{implies } E_g\left[\sum_{\eta(1)}^{\eta(2)-1} |c(X_s, A_s)| \,\middle|\, X_1 = x\right] = \infty.$$

Note that if, under g, x belongs to the recurrent class then (A3(g)) always holds (see the proof below). In particular, when there are no transient states under g, (A3(g)) holds.

Proof. The first claim follows from Lemma 1.1, and (i) is obtained by combining this with Lemma 2.2.

To prove (ii), consider first a cost of the form $c(x, a) = c(x)$. Recall that the initial condition is fixed, and is omitted from the notation below. Denote by $\tau_k := E_g[\eta(k+1) - \eta(k)]$ the expected time between consecutive visits to state zero under g. (We call such a period a cycle.) From Chung [13], under (A1), $\tau := \tau_k$ is independent of k. Denote the (sample) cost over the kth cycle by $Y(k) := \sum_{t=\eta(k)}^{\eta(k+1)-1} c(X_t)$. If $\eta(k) = \infty$ set $Y(k) := 0$. Assume that

(*)
$$E_g\left[\sum_{\eta(1)}^{\eta(2)-1} |c(X_s)|\right] < \infty.$$

Denote by $W := E_g[Y(1)]$ the total expected cost during the first cycle. Since c is bounded below, (A1) implies that (*) is equivalent to W being finite. From Chung [13] it follows that under (A1), (*), and (A3(g)),

$$\bar{C}_x(g) = \underline{C}_x(g) = C(g) = \frac{W}{\tau} = \sum_{y \in \mathbf{X}} \pi_y^g c(y). \tag{2.5}$$

But as g is stationary, $\bar{f}_{x,g}(y) = \pi_y^g$ whereas the tightness impliess $\sum_{a \in \mathbf{A}} \bar{f}_{x,g}(y, a) = \bar{f}_{x,g}(y)$. Hence

$$C(g) = \sum_{y \in \mathbf{X}} \bar{f}_{x,g}(y) c(y) = \sum_{y,a} \bar{f}_{x,u}(y, a) c(y). \tag{2.6}$$

Next, if (*) does not hold then $W = \infty$. Denote $c^M(x) := c(x) 1\{c(x) \leqq M\}$, and define W^M, $Y^M(k)$, and $C_x^M(u)$ as before, but with $c^M(x)$ replacing $c(x)$. The following monotone convergence holds pathwise:

$$Y(1) = \sum_{s=\eta(1)}^{\eta(2)-1} c(X_s) = \lim_{M \to \infty} \sum_{s=\eta(1)}^{\eta(2)-1} c^M(X_s) = \lim_{M \to \infty} Y^M(1).$$

Clearly, $C(g) \geqq \overline{\lim}_{M \to \infty} C_x^M(g) = \overline{\lim}_{M \to \infty} W^M/\tau = W/\tau$ by (2.5) and the monotone convergence theorem. Thus $C(g) = W/\tau$, i.e., all but the last equality in (2.5) hold independently of (*). It then follows from (2.5) that

$$\infty = \lim_{M \to \infty} \frac{W^M}{\tau} = \lim_{M \to \infty} \sum_{y \in \mathbf{X}} \pi_y^g c^M(y) = \sum_{y \in \mathbf{X}} \pi_y^g c(y) \tag{2.7}$$

using the monotone convergence theorem. The argument leading to (2.6) now implies (2.4).

Finally, we allow the cost to depend on the action. Let $\hat{c}(x) := \sum_{a \in A(x)} p^g_{a|x} c(a, x)$. Note that

$$C_x(g) = \overline{\lim_{t\to\infty}} \frac{1}{t} E_u\left[\sum_{s=1}^{t} \hat{c}(X_s) \,|\, X_1 = x\right] = \sum_{y\in\mathbf{X}} \bar{f}_{x,g}(y)\hat{c}(y)$$
$$= \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}} \bar{f}_{x,g}(y) p^g_{a|y} c(y, a) = \sum_{y,a} \bar{f}_{x,g}(y, a) c(y, a),$$

since $\bar{f}_{x,g}(y)p^g_{a|y} = \bar{f}_{x,g}(y, a)$, and where changes of order of summation are justified since c is bounded below. The proof in the case the cost bounded from above is identical. □

That (A3(g)) is necessary can be seen through the following example. Let $\mathbf{X}$ be the positive integers and let $\mathbf{A} := \{a\}$. Let $P_{0a0} = 1$, and $P_{xa0} = 0.5$, $P_{xa(x+1)} = 0.5$ for $x > 0$. With $c(y, a) = b^y$, (A3(g)) is violated and (2.4) fails to hold whenever $b \geqq 2$.

2.2. Optimization and frequencies. Using the previous lemmas we next discuss optimization under the expected average cost criterion.

LEMMA 2.4. *Assume $c(x, a)$ is bounded below, and let $C_x(u)$ denote either of the costs* (1.2). *If*

$$C_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,u}(y, a) c(y, a) \tag{2.8}$$

for some u and sequence $\{t_n\}$, then for any accumulation point $\bar{f}_{x,u}$ of $\{\bar{f}^{t_n}_{x,u}\}$, $C_x(u) \geqq \sum_{y,a} \bar{f}_{x,u}(y, a)c(y, a)$.

Proof. Assume first that c is positive and let $\{s_n\}$ be a subsequence of $\{t_n\}$ such that $\bar{f}^{s_n}_{x,u} \to \bar{f}_{x,u}$ for all (a, y). An application of Fatou's lemma, where c is considered as a σ-finite measure yields

$$C_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{s_n}_{x,u}(y, a)c(y, a) \geqq \sum_{y,a} \varliminf_{n\to\infty} \bar{f}^{s_n}_{x,u}(y, a)c(y, a) = \sum_{y,a} \bar{f}_{x,u}(y, a)c(y, a). \tag{2.9}$$

In general, shifting c to obtain such a measure, the same argument applies. □

COROLLARY 2.5. *Let u and v be two policies such that $\bar{F}_{x,v} = \{\bar{f}_{x,v}\}$, and $\bar{f}_{x,u} = \bar{f}_{x,v}$, for some accumulation point $\bar{f}_{x,u} \in \bar{F}_{x,u}$. Assume that under v the representation* (2.4) *holds and $c(x, a)$ is bounded below. Then $C_x(v) \leqq C_x(u)$, where $C_x(u)$ stands for either of the costs* (1.2).

The following theorem gives conditions under which the search for optimal (or ε optimal) policies can be restricted to a subclass of policies.

THEOREM 2.6. *Consider the problem of minimizing $\bar{C}_x(u)$ (or minimizing $\underline{C}_x(u)$). Assume* (A1) and (A2) *and let U' be complete. Then U' is sufficient if one of the following assumptions holds*:

(i) *$\{c(X_s, A_s)\}_s$ is uniformly integrable with respect to P^u for each $u \in U$.*

(ii) *For each $u' \in U'$* (2.4) *holds and $c(x, a)$ is bounded below.*

Proof. (i). For any $u \in U$, there exists a $v \in U'$ satisfying the hypotheses of Lemma 2.2, and sufficiency follows. The proof of (ii) follows from Lemma 2.4 and Corollary 2.5. □

Note that the question of existence of optimal policies is not raised here.

2.3. Constrained optimization. The reason for restricting problem COP to cost functionals D^k defined through (1.2a) is that, when the constraints are defined through (1.2b), a complete class of policies may not be sufficient, even if the state and action spaces are finite. For example:

Counterexample 2.7. Optimization under a constraint. Let $\mathbf{X}=\{x\}$, and $\mathbf{A}=\{p, q\}$. Set $c(x,p)=d(x,q)=-1$, $d(x,p)=c(x,q)=0$. Define $\underline{C}$ and $\underline{D}$ through (1.2b). The objective is to minimize $\underline{C}_x(u)$ under the constraint $\underline{D}_x(u)\leqq -0.5$.

In the finite, single class case, it is well known that the class of stationary policies achieves all possible frequencies. It is easy to see that the best stationary policy g chooses p or q with equal probability, and $\underline{C}_x(g)=-0.5=\underline{D}_x(g)$. Consider the policy u that uses p at times $(2n)^2\leqq t<(2n+1)^2$ $n=1,2,\cdots$ and action q at the remaining epochs. Then $\underline{C}_x(u)=\underline{D}_x(u)=-1$, and we conclude that there is no stationary optimal policy.

THEOREM 2.8. *Consider problem* COP (1.3a) *and* (1.3b). *Under* (A1) *and* (A2) *the stationary policies are sufficient if one of the following holds*;

(i) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1\leqq k\leqq K$ *are uniformly integrable with respect to* P^u *for each* u, *or*

(ii) $c(x,a)$ *and* $d^k(x,a)$, $1\leqq k\leqq K$ *are bounded below and* (A3(g)) *holds for all* $g\in U(S)$, *with respect to* c *and* d^k, $1\leqq k\leqq K$.

Remark. It clearly suffices to check (A3(g)) for those policies that satisfy the constraints (see also § 4).

Proof. Consider first (1.3a). Fix an arbitrary policy u. Let t_n be a subsequence such that $\bar{C}_x(u)=\lim_{n\to\infty}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)c(y,a)$, and such that the limits $\bar{f}^{t_n}_{x,u}(y,a)\to\bar{f}_{x,u}(y,a)$ for all y and a, and $\lim_{n\to\infty}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)d^k(y,a)$ exist. According to Corollary 3.3 the class of stationary policies $U(S)$ is complete, hence there exists a stationary policy g such that $\bar{f}_{x,g}=\bar{f}_{x,u}$. Under assumption (i), Lemma 2.3 implies that (A2(g)) holds, and so Lemma 2.2 implies $\bar{C}_x(g)=\bar{C}_x(u)$. Finally,

$$\bar{D}^k_x(u)=\overline{\lim_{t\to\infty}}\sum_{y,a}\bar{f}^{t}_{x,u}(y,a)d^k(y,a)\geqq\overline{\lim_{n\to\infty}}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)d^k(y,a)$$
$$=\sum_{y,a}\bar{f}_{x,u}(y,a)d^k(y,a)=\bar{D}^k_x(g),$$

where the next to last equality is validated by the arguments of the proof of Lemma 2.2. This proves the theorem under assumption (i).

Under (ii), since (A1) implies that $\bar{f}_{x,g}$ is a singleton, Lemma 2.3 and Corollary 2.5 can be invoked to conclude $\bar{C}_x(g)\leqq\bar{C}_x(u)$ and $\bar{D}^k_x(g)\leqq\bar{D}^k_x(u)$ for each k, and the proof for (1.3a) is concluded. The proof for (1.3b) is identical. □

Finally, we consider an arbitrary complete class.

COROLLARY 2.9. *Assume* (A1) *and* (A2) *and consider* COP (1.3). *Let* U' *be any complete class of policies. Assume* (2.4) *holds for all* $u'\in U'$ *and for* c *and* d^k. *Then* U' *is sufficient if one of the following holds*;

(i) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1\leqq k\leqq K$ *are uniformly integrable with respect to* P^u *for each* u, *or*

(ii) $C(x,a)$ *and* $d^k(x,a)$, $1\leqq k\leqq K$ *are bounded below.*

Proof. Stationarity is used in the proof of Theorem 2.8 solely to guarantee that $\bar{F}$ is a singleton and the representation (2.4) holds. □

3. Completeness: action time sharing. In Theorem 3.2 we prove that the class of stationary policies is weakly complete under (A1), and complete (Corollary 3.3) under the additional assumption (A2). This and the results of § 2 allow us to recover and extend classical results, concerning optimality of stationary policies.

The classical approach to Markov optimization problems relies on the specific class of stationary policies, and on their statistical properties. In contrast, the point of view taken here is to find weak sufficient conditions for a class of policies to be complete. The class of "action time sharing" policies introduced below includes the

stationary policies. However, the novelty of this approach is expressed in Theorem 3.6, which states that the frequencies achieved by "*ats*" policies depend only on their **pathwise** conditional frequencies. This implies (Theorem 3.7) that completeness can be achieved within subclasses other than stationaries: for example, using deterministic, nonstationary policies.

Fix $\alpha := \{\alpha_y^a, y \in \mathbf{X}, a \in A(y)\}$, where $\alpha_y^a \geqq 0$, and $\sum_{a \in A(y)} \alpha_y^a = 1$ for each $y \in \mathbf{X}$.

DEFINITION. A policy u is an "action time sharing" (*ats*) policy with parameter α, and is denoted as $\hat{\alpha}$ if for every state y that is visited infinitely often P^u almost surely and any action a,

$$f^T(a \mid y) \to \alpha_y^a \quad \text{as } T \to \infty \quad P^u \text{ a.s.}$$

Thus an *ats* policy $\hat{\alpha}$ alternates between several actions at each state so as to achieve a prescribed (state dependent) limiting relative frequency for each action. There are no restrictions as to how the limiting frequencies $f^T(a \mid y)$, are achieved, and there are many ways such a policy can be realized.

A realization of an *ats* policy with parameter α is a mapping h from the space S_α of all possible collections α to the space of all policies U. Given such a mapping h, let $U^h(ats) := h(S_\alpha)$ denote the subclass of *ats* policies of the form $\hat{\alpha} = h(\alpha)$ for some $\alpha \in S_\alpha$. For example, setting $p_{a|y} = \alpha_y^a$ defines a stationary policy, where $p_{a|y}$ are the conditional distributions. We thus recover the class of stationary policies, where the realization is by randomly choosing the actions using unfair dice. Another possible realization of *ats* policies is through the use of a counter for each $y \in \mathbf{X}$, $a \in A(y)$. We then choose in a *deterministic way* which action to use for every state, so that the appropriate conditional frequencies are achieved.

The main result of this section, Theorem 3.7, states that under (A1) and (A2), $U^h(ats)$ is complete for any h (see the definition of completeness in § 1). Moreover, the frequencies $\bar{f}_{x,\hat{\alpha}}(y, a)$ depend only on the parameter α, and not on the realization $\hat{\alpha} = h(\alpha)$ (Theorem 3.6). We proceed to investigate the completeness of stationary policies, and will then turn to *ats* policies. But first we need a technical lemma.

LEMMA 3.1. *Under* (A1) *for any transition matrix* P^g, $g \in U(S)$ *there exists a measure* π *such that*

$$\pi(y) \geqq \sum_{z \in \mathbf{X}} \pi(z)[P^g]_{zy}. \tag{3.1}$$

The measure π *is finite, is unique up to a multiplicative constant, and in fact* $\pi(y) = \sum_{z \in \mathbf{X}} \pi(z)[P^g]_{zy}$.

Remark. This result is well known when there are no transient states (see, e.g., [27, Thm. 1.10, p. 67]).

Proof. Existence. Let R and T denote the recurrent and transient classes under g. By Theorem 1.10 of [27, p. 67], there exists a finite measure $\tilde{\pi}$, unique up to a multiplicative constant, such that

$$\tilde{\pi}(y) = \sum_{z \in R} \tilde{\pi}(z)[P^g]_{zy}.$$

Define the measure π on $\mathbf{X}$ by $\pi(y) = \tilde{\pi}(y)$ for $y \in R$ and $\pi(y) = 0$ otherwise. Then it is easy to check that π solves (3.1), in fact with equality. To prove uniqueness, let π be a solution of (3.1). Iterating (3.1), we obtain for every $n > 0$

$$\pi(y) \geqq \sum_{z \in \mathbf{X}} \pi(z)[(P^g)^n]_{zy} \geqq \sum_{z \in R} \pi(z)[(P^g)^n]_{zy}. \tag{3.2}$$

Again, by Theorem 1.10 of [27, p. 67], there exists $\{\pi(y), y \in R\}$, unique up to a multiplicative constant and independent of n, such that for all $y \in R$, $\pi(y) = \sum_{z \in R} \pi(z)$ $[(P^g)^n]_{zy}$. But this and (3.2) imply that π satisfies $\sum_{z \in T} \pi(z)[(P^g)^n]_{zy} = 0$ for all $y \in R$ and all n. Fix some $y \in R$; by (A1), for each $z \in T$ there exists a finite n such that $[(P^g)^n]_{zy} > 0$. Thus we conclude $\pi(z) = 0$ and the uniqueness is established. □

THEOREM 3.2. *Under* (A1) *the class of stationary policies is weakly complete.*

Proof. First note that $\bar{F}_{x,g}$ is a singleton for any stationary policy g. This follows from the existence of a unique stationary probability, under (A1). Pick any frequency matrix $\zeta \in L_x$ that is achieved, say, by a policy $u \in U$. To establish the theorem we need to show that whenever $\{\bar{f}^t_{x,u}\}_t$ is tight, there exists a policy $g \in U(S)$ so that $\bar{f}_{x,g} = \zeta$. Thus let u be a policy such that $\{\bar{f}^t_{x,u}\}_t$ is tight. Let $\{t_n\}_n$ be an increasing sequence of times (chosen by diagonalization) along which $\lim_{n\to\infty} \bar{f}^{t_n}_{x,u}(y, a) := \bar{f}_{x,u}(y, a) = \zeta(y, a)$ and $\lim_{n\to\infty} \bar{f}^{t_n}_{x,u}(y) := \bar{f}_{x,u}(y)$ exist for all y and $a \in A(y)$. Note that for each y,

$$\bar{f}_{x,u}(y) = \lim_{n\to\infty} \bar{f}^{t_n}_{x,u}(y) = \lim_{n\to\infty} \sum_{a \in A(y)} \bar{f}^{t_n}_{x,u}(y, a) \tag{3.3}$$

and by tightness and the convergence $\bar{f}^{t_n}_{x,u}(y, a) \to \bar{f}_{x,u}(y, a)$, the probability measures $\bar{f}^{t_n}_{x,u}$, $n = 1, 2, \cdots$ converge weakly (see the Portmanteau theorem in [8]), so that

$$\bar{f}_{x,u}(y) = \lim_{n\to\infty} \sum_{a \in A(y)} \bar{f}^{t_n}_{x,u}(y, a) = \sum_{a \in A(y)} \lim_{n\to\infty} \bar{f}^{t_n}_{x,u}(y, a) = \sum_{a \in A(y)} \bar{f}_{x,u}(y, a). \tag{3.4}$$

Set $\beta^a_y := \bar{f}_{x,u}(y, a)[\bar{f}_{x,u}(y)]^{-1}$ whenever $\bar{f}_{x,u}(y) \neq 0$. If $\bar{f}_{x,u}(y) = 0$ then the β^a_y are chosen arbitrarily but such that $0 \leqq \beta^a_y \leqq 1$ for all a, and $\sum_{a \in A(y)} \beta^a_y = 1$. By (3.4), $\sum_{a \in A(y)} \beta^a_y = 1$ for every y. Define the stationary policy g by $p^g_{a|y} = \beta^a_y$. Then

$$P^g_{xy} = \sum_{a \in A(x)} \beta^a_x P_{xay}. \tag{3.5}$$

Since for every $s > 1$ we have $P^u_x\{X_s = y\} = \sum_{z,a} P^u_x\{X_{s-1} = z, A_{s-1} = a\} P_{zay}$, we get after some algebra

$$\bar{f}^t_{x,u}(y) - \frac{1}{t} P^u_x\{X_1 = y\} = \sum_{z,a} \bar{f}^t_{x,u}(z, a) P_{zay} - \frac{1}{t} \sum_{z,a} P^u_x\{X_t = z, A_t = a\} P_{zay}. \tag{3.6}$$

Since the left side of (3.6) converges along the sequence t_n to $\bar{f}_{x,u}(y)$, so does the right. Fix y and consider P_{zay} as a σ-finite measure on $\mathbf{X} \times \mathbf{A}$. Applying Fatou's lemma we obtain using (3.6)

$$\bar{f}_{x,u}(y) = \lim_{n\to\infty} \sum_{z,a} \bar{f}^{t_n}_{x,u}(z, a) P_{zay} \geqq \sum_{z,a} \bar{f}_{x,u}(z, a) P_{zay} \tag{3.7}$$

since the last term in (3.6) is bounded by t^{-1}. From (3.5), (3.7) and from the definition of β^a_y we obtain

$$\bar{f}_{x,u}(y) \geqq \sum_z \bar{f}_{x,u}(z) \cdot P^g_{zy}. \tag{3.8}$$

From (3.8) we conclude that $\bar{f}_{x,u}(\cdot)$ is an excessive measure with respect to the transition matrix P^g. It follows from (3.4) that $\{\bar{f}^{t_n}_{x,u}(\cdot)\}_{t_n}$ are tight, and hence $\bar{f}_{x,u}(\cdot)$ is in fact a probability measure. But under (A1), Lemma 3.1 implies that $\bar{f}_{x,u} = \pi^g$. Using the definition of β and g this finally implies that

$$\zeta(y, a) = \bar{f}_{x,u}(y, a) = \bar{f}_{x,u}(y) \cdot \beta^a_y = \pi^g(y) p^g_{a|y} = \bar{f}_{x,g}(y, a) \tag{3.9}$$

by Lemma 1.1. □

From Theorem 3.2 we immediately obtain Corollary 3.3.

COROLLARY 3.3. *Under* (A1) *and* (A2) *the class of stationary policies is complete.*

Combining this with the theorems of § 2 we thus conclude that, under the relevant assumptions, the stationary policies are sufficient for problems OP and COP.

Assumption (A2) is used to guarantee that $\sum_{y,a} \bar{f}_{x,u}(y, a) = 1$ and that $\sum_{a \in A(y)} \beta_y^a = 1$ for every y. Assumption (A0) guarantees the latter; note that it does not imply that $\mathbf{A}$ is finite.

COROLLARY 3.4. *Under* (A1) *and* (A0), *for every policy* u *and a matrix* $\bar{f}_{x,u}(\cdot, \cdot)$, *there exists a stationary policy* g *and a constant* $0 \leqq \delta \leqq 1$ *such that* $\bar{f}_{x,u}(y, a) = \delta \cdot \bar{f}_{x,g}(y, a)$, $y \in \mathbf{X}$, $a \in A(y)$. *Under* (A2), $\delta = 1$.

Proof. Following the proof of Theorem 3.2, observe that $\bar{f}_{x,u}(\cdot)$ is an excessive measure due to (3.8) and is thus, by Lemma 3.1, proportional to π^g. But $\bar{f}_{x,u}(\cdot, \cdot)$ is clearly a subprobability measure, i.e., $\sum_{y,a} \bar{f}_{x,u}(y, a) \leqq 1$. Thus by the argument of (3.9), $\bar{f}_{x,u}(y, a) = \delta \cdot \bar{f}_{x,g}(y, a)$, $y \in \mathbf{X}$, $a \in A(y)$. If (A2) holds then it is a probability measure, and $\delta = 1$ by Theorem 3.2. □

Remark. If under every $u \in U(SD)$ there are no transient states then δ in Corollary 3.4 is always strictly positive; moreover, $\bar{f}_{x,u}(y) > \varepsilon(y)$ uniformly in $u \in U$ (see, e.g., [18]).

Before we show that $\bar{f}_{x,\hat{\alpha}}$ depends only on α, we present a simple example that demonstrates the importance of (A2), and shows that a condition such as (A0) is necessary for (A2).

Counterexample 3.5. Countable action space. Consider problem OP with $\mathbf{X} = \{x\}$ and $\mathbf{A} = \{1, 2, \cdots\}$, and let $c(x, a) = 1 + a^{-1}$. Clearly, $X_t = x$ for all t. Under any stationary policy g, $\bar{C}_x(g) > 1$ and $\sum_{a \in A(x)} \bar{f}_{x,g}(x, a) = \bar{f}_{x,g}(x) = 1$. Let u be the nonstationary policy that chooses action $a = t$ at time t. Clearly, we have $\bar{C}_x(u) = 1$, $\bar{f}_{x,u}(x, a) = 0$, $\bar{f}_{x,u}(x) = 1$.

This example demonstrates that even under the unichain assumption, the expected state-action frequencies may not be tight while the expected state frequencies are, and the average expected cost is not necessarily a function of the expected state action frequencies. Moreover, the stationary policies are not complete, and due to the noncompactness of the action space, the cost achieved by some nonstationary policy can be strictly smaller than the cost of any stationary policy. This is in contrast with the case of finite state and action spaces (see Derman [15]).

A counterexample where both (A1) and (A0) hold yet (A2) is not satisfied is presented in Fisher and Ross [17]. They show that indeed without (A2) the stationary policies may be incomplete.

THEOREM 3.6. *Under* (A1) *and* (A2), $\bar{F}_{x,\hat{\alpha}} = \{\bar{f}_{x,\hat{\alpha}}\}$ *is a singleton. Moreover,* $\bar{f}_{x,\hat{\alpha}}$ *depends only on* α *and is independent of the realization* h.

Proof. Let $v = \hat{\alpha} = h(\alpha)$ be some *ats* policy with parameter α. Define the stationary policy g by $p^g_{a|y} = \alpha^a_y$. By the strong law of large numbers, g is also an *ats* policy with parameter α. The proof is concluded by showing that $\bar{f}_{x,v}(y, a) = \bar{f}_{x,g}(y, a)$. Since the initial state is fixed, we suppress it in the notation of P and E. Let

$$M_t := \sum_{s=2}^{t} 1\{X_s = y\} - \sum_{s=2}^{t} \sum_{x,a} 1\{X_{s-1} = x, A_{s-1} = a\} P_{xay}.$$

Then for any u, M_t is a P^u martingale and by the stability theorem (e.g., [20, Thm. 2.22])

$$\lim_{t \to \infty} \frac{1}{t} \left[\sum_{s=2}^{t} 1\{X_s = y\} - \sum_{s=2}^{t} \sum_{x,a} 1\{X_{s-1} = x, A_{s-1} = a\} P_{xay} \right] = 0 \quad P^u \text{ a.s.} \tag{3.10}$$

Let N be the P^v-null set of ω for which either (3.10) or the convergence in the definition of the *ats* policy v do not hold. Fix $\omega \in \Omega - N$ and an arbitrary increasing sequence of times t_n. Using diagonalization, construct a subsequence s_n to t_n along which for all y and a, $f^{s_n}(y, a)$, $f^{s_n}(y)$ and $f^{s_n}(a \mid y)$ converge to some limits $f(y, a)$, $f(y)$, and

$f(a|y)$, respectively. Note that from the definition of the *ats* policy v it follows that $f(y)=\sum_a f(y,a)\ P^v$ almost surely. For that ω we have from (3.10) for all $y\in\mathbf{X}$:

$$f(y)=\lim_{n\to\infty}\sum_{x,a} f^{(s_n-1)}(x,a)P_{xay}. \tag{3.11}$$

An argument as in (3.7) and (3.8) implies

$$f(y)\geqq\sum_{x,a} f(x,a)P_{xay}. \tag{3.12}$$

From the definition of the *ats* policy v it is easy to see that

$$\alpha_x^a f(x)=f(a|x)f(x)=f(x,a). \tag{3.13}$$

From (3.12) and (3.13) we obtain, since all terms are nonnegative

$$f(y)\geqq\sum_{x\in\mathbf{X}} f(x)\left[\sum_{a\in A(x)}\alpha_x^a P_{xay}\right]=\sum_{x\in\mathbf{X}} f(x)P_{xy}^g. \tag{3.14}$$

Using the same argument that followed the proof of Corollary 3.4 we obtain for all $y\in\mathbf{X}$:

$$f(y)=\delta(\omega,\{s_n\})\cdot\pi_y^g \tag{3.15}$$

for some constant δ satisfying $0\leqq\delta\leqq 1$. Thus, for all y, z in $\mathbf{X}$,

$$\lim_n\left[f^{s_n}(y)\pi^g(z)-f^{s_n}(z)\pi^g(y)\right]=0.$$

Since the sequence $\{t_n\}_n$ was arbitrary, we conclude that in fact

$$\lim_t\left[f^t(y)\pi^g(z)-f^t(z)\pi^g(y)\right]=0$$

and this holds for P^v almost surely. But by the bounded convergence theorem,

$$\lim_t\left[E_v f^t(y)\pi^g(z)-E_v f^t(z)\pi^g(y)\right]=\lim_t\left[\bar f_{x,v}^t(y)\pi^g(z)-\bar f_{x,v}^t(z)\pi^g(y)\right]=0.$$

By assumption $\{\bar f_{x,v}^t\}_t$ is tight. Fix any subsequence $\{r_n\}_n$ such that $\bar f_{x,v}^{r_n}\to\bar f_{x,v}$. Then $\bar f_{x,v}(y)\pi^g(z)=\bar f_{x,v}(z)\pi^g(y)$. However, the only probability measure that solves this equation is $\bar f_{x,v}=\pi^g$, and we conclude that $\bar f_{x,v}^t\to\pi^g$. From the definition of the *ats* policy v and the bounded convergence theorem, we have

$$\overline{\lim_{t\to\infty}} E_v|\alpha_y^a-f^t(y,a)[f^t(y)]^{-1}|=0.$$

Thus, using the bounded convergence theorem and the tightness (A2),

$$\lim_{t\to\infty}\bar f_{x,v}^t(y,a)=\lim_{t\to\infty}E_v f^t(y)\cdot\frac{f^t(y,a)}{f^t(y)}=\lim_{t\to\infty}\bar f_{x,v}^t(y)\cdot\alpha_y^a=\alpha_y^a\pi^g(y) \tag{3.16}$$

for all a, y. Finally, Lemma 1.1 implies $\bar f_{x,g}(y,a)=\alpha_y^a\cdot\pi^g(y)=\bar f_{x,v}(y,a)$. Since π^g depends only on α and not on $\hat\alpha$ this concludes the proof. □

Combining Theorem 3.2 and Theorem 3.6 it follows that the completeness is determined by the α only, so that complete classes can be easily generated.

THEOREM 3.7. *Under* (A1) *and* (A2), *for any realization* $h: S_\alpha\to U$, $U^h(ats)$ *is complete.*

4. Tightness. The issue of tightness for Markov decision processes has been investigated extensively. It is easy to see that, in general, unless the sets $A(x)$ are finite (compact), (A2) need not hold. In the compact case, Lemma 4.1 provides a simple sufficient condition for (A2). We describe briefly several approaches that provide sufficient conditions for tightness in this compact case (i.e., under (A0)).

If compactness of the actions is not assumed, we can usually construct a policy u for which (A2(u)) does not hold, so that (A2) will not hold. However, since the tightness appears in connection with the optimization problems, we derive conditions on the cost functions that guarantee that the search for optimal policy can be restricted to policies satisfying (A2(u)). This extends the results of §§ 2.2 and 2.3 to cases where (A2) does not hold.

LEMMA 4.1. *Under* (A0) *and* (A1), (A2) *implies* (A2*). *If in addition there are no transient states, then* A2 *is equivalent to* (A2*).

Proof. It is shown in the proof of Lemma 7.3 of [10] that, under (A0) and (A1) and when there are no transient states, (A2*) implies that for each state x and policy u the expected frequencies $\{\bar{f}^t_{x,u}(y)\}_t$ are tight. To see that the converse holds, assume (A0) and (A1) and let g_i be a sequence of policies in $U(SD)$ such that the sequence of corresponding invariant distributions π_i is not tight. Clearly, $\bar{f}^t_{x,g_i}(y) \to \pi_i(y)$ for all x, y. Thus we can find an increasing sequence $\{t_i\}$ and construct a policy u where $u_t(\cdot \mid H_{t-1}, X_t) = g_i(x_t)$ for $t_i < t \leqq t_{i+1}$ such that $\bar{f}^{t_i}_{x,u} \to \bar{f}_{x,u}$, and $\sum_y \bar{f}_{x,u}(y) < 1$. Thus it suffices to show that $\{\bar{f}^t_{x,u}(y)\}_t$ is tight if and only if $\{\bar{f}^t_{x,u}(y, a)\}_t$ is tight.

By definition, $\{\bar{f}^t_{x,u}(y)\}_t$ is tight if and only if for any $\varepsilon > 0$ there exists a compact (finite) set $K(\varepsilon) \subset \mathbf{X}$ such that $\sum_{y \in K(\varepsilon)} \bar{f}^t_{x,u}(y) > 1 - \varepsilon$, and similarly for $\{\bar{f}^t_{x,u}(y, a)\}_t$. Given $K(\varepsilon)$, let $K'(\varepsilon) := \{(y, a)\colon y \in K(\varepsilon), a \in A(y)\}$. Then $K'(\varepsilon)$ is compact and since

$$\bar{f}^t_{x,u}(y) = \sum_{a \in A(y)} \bar{f}^t_{x,u}(y, a) \tag{4.1}$$

we have $\sum_{(y,a) \in K'(\varepsilon)} \bar{f}^t_{x,u}(y, a) = \sum_{(y) \in K(\varepsilon)} \bar{f}^t_{x,u}(y) > 1 - \varepsilon$. To prove the converse, given $K'(\varepsilon) \subset \mathbf{X} \times \mathbf{A}$ let $K(\varepsilon) := \{y\colon (y, a) \in K'(\varepsilon)$ for *some* $a \in A(y)\}$. The same argument now concludes the proof. □

Assumption (A2*) is quite common in the literature on controlled Markov chains with a countable state space, and sufficient conditions are available. Borkar [10, § III] shows that (A2*) is equivalent to the time between visits to some recurrent state being uniformly integrable under all $u \in U(SD)$. The whole § IX in [10] is then devoted to different sufficient conditions for that uniform integrability. Hordijk [21] presents several sufficient conditions for (A2*), in terms of the measures $P^g_{x,K} := \sum_{y \in K} P^g_{xy}$;

(i) The set of probability measures $\{P(X_2 = \cdot \mid X_1 = x)\colon x \in \mathbf{X}, g \in U(S)\}$ is tight [21, Lemma 10.3, § 10].

(ii) Given any $\varepsilon > 0$ there exist a finite set $K(\varepsilon)$ and an integer $N(\varepsilon)$ such that for all $x \in \mathbf{X}$ and $g \in U(S)$,

$$[(P^g)^{N(\varepsilon)}]_{x,K(\varepsilon)} \geqq 1 - \varepsilon.$$

(iii) The simultaneous Doeblin condition. There exist a finite set K, a positive integer n, and a positive real number c such that $[P^g]^n_{x,K} \geqq c$ for all $x \in \mathbf{X}$ and all $g \in U(S)$ [21, § 11.1].

Two other assumptions that are equivalent to (iii) above (and are thus sufficient for (A2*)) are presented in Theorem 11.3 of [21]. To formulate these conditions, denote

$$_AP^{g,t}_{x,B} := P^g\{X_t \in B, X_s \notin A, 1 < s < t \mid X_1 = x\}, \qquad m_g(x, A) := \sum_{t=2}^{\infty} {}_AP^{g,t}_{x,\mathbf{X}}.$$

(iv) There exist a finite set K, $c > 0$, and n such that $\sum_{s=2}^{n} {}_KP^{g,s}_{x,K} \geqq c$ for all $x \in \mathbf{X}$ and $g \in U(S)$.

(v) There exist a finite set K and a real number b such that for all $x \in \mathbf{X}$ and $g \in U(S)$, $m_g(x, K) \leqq b$.

In the absence of tightness, it may be possible to restrict the optimization problem to a subclass of policies under which tightness holds, if the structure of the costs makes it unprofitable to use nontight policies (see also Borkar [10]).

LEMMA 4.2. *Assume there exists a sequence of increasing compact (i.e., finite) subsets* K_i *of* $\mathbf{X}\times\mathbf{A}$ *such that* $\cup_i K_i = \mathbf{X}\times\mathbf{A}$, *and such that the cost function* $c(y, a)$ *satisfies*

$$\liminf_{i\to\infty}\{c(y, a);\ (y, a)\notin K_i\} = \infty. \tag{4.2}$$

Then for any policy u *such that* $\bar{C}_x(u) < \infty$ (*or* $\underline{C}_x(u) < \infty$), *the frequencies* $\{\bar{f}^t_{x,u}(\cdot,\cdot)\}_t$ *are tight.*

Proof. By (4.2), $c(x, a)$ is bounded below, say by B. Assume $\{\bar{f}^t_{x,u}(\cdot,\cdot)\}$ is not tight. Then there exists some $\varepsilon > 0$ and an increasing sequence $\{t_l\}$ such that $\sum_{(y,a)\notin K_l} \bar{f}^{t_l}_{x,u}(y, a) > \varepsilon$. Let $c_j := \inf\{c(y, a);\ (y, a)\notin K_j\}$. Clearly, $\bar{C}^{t_l}_x(u) \geqq c_j\varepsilon - |B|$. But by (4.2) $\lim_{j\to\infty} c_j = \infty$, and hence $\bar{C}_x(u) = \infty$, contradicting the hypotheses. The proof using $\underline{C}$ is identical. □

A complete class of policies (or even a weakly complete class of policies) may thus be sufficient even when the tightness assumption (A2) does not hold.

THEOREM 4.3. *Assume* (A1) *and consider the problem of minimizing* $\bar{C}_x(u)$. *Let* U' *be a weakly complete class such that* (2.4) *holds for each* $v\in U'$. *If* $c(\cdot,\cdot)$ *satisfies the conditions of Lemma* 4.2, *then* U' *is sufficient for* OP.

Note that the stationary policies are in fact weakly complete (Corollary 3.4) and, under (A3(g)), satisfy (2.4).

Proof. From Lemma 4.2 we conclude that if $\{\bar{f}^t_{x,u}\}_t$ is not tight (so that for some limit point $\sum_{y,a} \bar{f}_{x,u}(y, a) < 1$), then necessarily $C_x(u) = \infty$. Thus we may limit the optimization to policies u for which $\{\bar{f}^t_{x,u}\}_t$ is tight, so that $\sum_{y,a} \bar{f}_{x,u}(y, a) = 1$. By (4.2) $c(\cdot,\cdot)$ is bounded from below, and hence by Corollary 2.5 U' is sufficient. □

Similarly for the constrained problem COP, we may relax (A2) in Theorem 2.8.

THEOREM 4.4. *Assume* (A1) *and consider problem* COP. *Let* U' *be a weakly complete class such that* (2.4) *holds for each* $v\in U'$. *If either* $c(\cdot,\cdot)$ *or* $d^k(\cdot,\cdot)$, *some* k *satisfies the conditions of Lemma* 4.2, *then* U' *is sufficient for* COP.

Proof. The proof is similar to that of Theorem 4.3. □

Next we present another method that provides conditions for sufficiency in cases that the tightness does not hold. It is a generalization of conditions that Borkar [9] introduced for the case of instantaneous cost that depends only on the state. Following [9], $c(\cdot,\cdot)$ is said to be "V-almost monotone" if there exists a collection of compact sets $\{K_i\}_i$ as in Lemma 4.2 such that $\liminf_{i\to\infty}\{c(y, a);\ (y, a)\notin K_i\} \geqq V$.

LEMMA 4.5. *Assume* (A0) *and* (A1) *and let* U' *be a weakly complete class of policies such that every* $u\in U'$ *satisfies* (2.4). *If* $c(\cdot,\cdot)$ *is* V*-almost monotone and* $C_x(u') \leqq V$, *some* $u'\in U'$, *then* U' *is sufficient for* OP.

Proof. Note that $c(\cdot,\cdot)$ is bounded below. Consider first the minimization of $\bar{C}_x$, fix an arbitrary v, and note that if $C_x(v) \geqq V$ then we are done. Thus assume $C_x(v) < V$ and let t_n be a subsequence such that $C_x(v) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y, a)c(y, a)$ and $\bar{f}^{t_n}_{x,v}$ converges to some $\bar{f}_{x,v}$. By Corollary 3.4, there exists a $g\in U(S)$ such that $\delta\bar{f}_{x,g} = \bar{f}_{x,v}$ for some $0 \leqq \delta \leqq 1$. By completeness there exists a $u\in U'$ such that $\delta\bar{f}_{x,u} = \bar{f}_{x,v}$. Let ε_i be such that $\inf\{c(y, a);\ (y, a)\notin K_i\} \geqq V - \varepsilon_i$. For every i we have

$$\begin{aligned} C_x(v) &= \lim_{n\to\infty}\left[\sum_{(y,a)\in K_i} \bar{f}^{t_n}_{x,v}(y, a)c(y, a) + \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y, a)c(y, a)\right] \\ &\geqq \sum_{(y,a)\in K_i} \bar{f}_{x,v}(y, a)c(y, a) + \varliminf_{n\to\infty} \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y, a)c(y, a), \end{aligned}$$

$$\underline{\lim}_{n\to\infty} \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y,a)c(y,a) \geqq (V-\varepsilon_i)\, \underline{\lim}_{n\to\infty} \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y,a)$$

$$= (V-\varepsilon_i)\, \underline{\lim}_{n\to\infty} \left[1 - \sum_{(y,a)\in K_i} \bar{f}^{t_n}_{x,v}(y,a)\right].$$

Using the fact that each term on the right converges to $\bar{f}_{x,v}(y,a)$ and that $\bar{f}_{x,v} = \delta \bar{f}_{x,u}$, we get

$$C_x(v) \geqq \delta \sum_{(y,a)\in K_i} \bar{f}_{x,u}(y,a)c(y,a) + (V-\varepsilon_i)\left[1 - \delta \sum_{(y,a)\in K_i} \bar{f}_{x,u}(y,a)\right].$$

By taking $i\to\infty$ we obtain $C_x(v) \geqq \delta C_x(u) + V(1-\delta)$ or $\delta C_x(v) \geqq \delta C_x(u) + (1-\delta)\times (V - C_x(v))$. Since the last term is positive, $C_x(v) \geqq C_x(u)$ that establishes the proof. The proof for $\underline{C}$ is identical. □

In the following lemma we apply the method of Lemma 4.5 in order to generalize Theorem 4.4.

LEMMA 4.6. *Assume* (A0) *and* (A1). *Let U' be a weakly complete class of policies such that every $u \in U'$ satisfies* (2.4) *for c and d^k, $1 \leqq k \leqq K$. Assume $c(\cdot,\cdot)$ is V_0-monotone and $d^k(\cdot,\cdot)$ is V_k-monotone, $1 \leqq k \leqq K$. If there exists a policy $u' \in U'$ such that $C_x(u') \leqq V_0$ and $D^k_x(u') \leqq V_k$, $1 \leqq k \leqq K$, then U' is sufficient for* COP.

Proof. The proof is the same for both $\bar{C}$ and $\underline{C}$. Note that without loss of generality, we may take the sets K_i to be the same for all costs c and d^k. Fix $v \in U'$ and note that if $C_x(v) \geqq V_0$ or for some k $D^k_x(v) > V_k$ then we are done. Thus assume $D^k_x(v) \leqq V_k$ for $k = 1, \cdots, K$ and $C_x(v) \leqq V_0$. Choose a subsequence $\{t_n\}_n$ such that $C_x(v) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)c(y,a)$ and such that $\lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)d^k(y,a)$, $1 \leqq k \leqq K$, and $\lim_{n\to\infty} \bar{f}^{t_n}_{x,v}$ exist. Then for $1 \leqq k \leqq K$, $D^k_x(v) \geqq \overline{\lim}_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)d^k(y,a)$ and by choosing u as in Lemma 4.5 we obtain by the same argument $\delta C_x(v) \geqq \delta C_x(u) + (1-\delta)(V_0 - C_x(v))$ and $\delta D^k_x(v) \geqq \delta D^k_x(u) + (1-\delta)(V_k - D^k_x(v))$. Hence $C_x(v) \geqq C_x(u)$ and $V_k \geqq D^k_x(v) \geqq D^k_x(u)$ that concludes the proof. □

5. Achievable frequencies. In this section we describe the geometry of L_x. For the case of finite state and action spaces Derman [15] has shown that under (A1), L_x is closed and convex, and its extreme points correspond to policies in $U(SD)$. In Theorem 5.1 we extend this result to the countable space. Let co B denote the convex hull of the set B, and $\overline{\text{co}}\ B$, its closed convex hull. Let η be a function from the integers onto all pairs (x,a) and fix $0 < \mu < 1$. Define a metric d on the set of subprobability measures on $\mathbf{X}\times\mathbf{A}$.

$$d(\zeta_1,\zeta_2) := \sum_{j=1}^{\infty} |\zeta_1(\eta[j]) - \zeta_2(\eta[j])|\mu^j. \tag{5.1}$$

We will use henceforth the product topology induced by this metric. Throughout this section we assume that

(A1′) The state space forms a single *positive* recurrent class under any policy $g \in U(S)$.

To prove Theorem 5.1, we need to introduce PTS ("policy time sharing") policies [2]. A PTS policy is specified through the stationary policies u_i, $i = 1, 2, \cdots, l$, a state z, and by an l-dimensional vector parameter $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_l\}$, where $\alpha_i \geqq 0$ and $\sum_i \alpha_i = 1$. Fix a state z that, due to (A1′) is positive recurrent under each u_i. Call the period between two consecutive visits to states z a cycle. A PTS policy v with parameter α is any policy that uses a fixed u_i during each cycle, and for which the relative number

of cycles during which u_i is used converges to α_i, P^v almost surely, $i=1, 2, \cdots, l$. Such a policy is denoted $\hat{\alpha}$. It follows immediately from the results of [1] and [3] that for any initial state x, $\bar{F}_{x,\hat{\alpha}}$ is a singleton, and

$$\bar{f}_{x,\hat{\alpha}} = \sum_{i=1}^{l} \gamma_i(\alpha) \bar{f}_{x,u_i}, \tag{5.2}$$

where $\gamma_i = \alpha_i \tau_i [\sum_{j=1}^{l} \alpha_j \tau_j]^{-1}$ and τ_i is the mean recurrence time of state z under u_i.

THEOREM 5.1. *Under* (A1′) *and* (A2), $L_x = L_x(S)$ *is compact. Moreover,* $L_x = \text{co}\{L_x(D)\} = \overline{\text{co}}\{L_x(D)\}$.

Proof. By Corollary 3.3, $L_x(S) = L_x$. To prove compactness, let $\{\xi_n\}_n \subset L_x$. Using diagonalization, choose a subsequence $\{\xi_{n_i}\}_i$ that (for notational convenience) is denoted $\{\zeta_n\}_n$, such that $\zeta_n(x, a) \to \zeta(x, a)$ for some ζ, for all x, a. ζ_n may all be considered measures over $\mathbf{X} \times \mathbf{A}$, and $\zeta_n(x, a) \to \zeta(x, a)$, where ζ is (possibly subprobability) measure. Our aim is to find a policy u that achieves ζ. By (A2) this implies that ζ is a probability measure.

By Corollary 3.3 there exists a stationary policy g_i that achieves ζ_i. Let $\varepsilon_i := d(\zeta, \zeta_i)$, so that $\lim_{n\to\infty} \varepsilon_n = 0$. Consider the nonstationary policy u, that uses g_1 until the time $t_1 := \min\{t: d(\zeta_1, \bar{f}^t_{u,x}) \leqq \varepsilon_1\}$, and uses g_i until between t_{i-1} and $t_i := \min\{t > t_{i-1}: d(\zeta_i, \bar{f}^t_{u,x}) \leqq \varepsilon_i\}$. The fact that $t_n < \infty$ can be proved by induction using the following fact. Suppose the policy u uses g_n from time s onward, and let $\chi_s(z) = P^u(X_s = z \mid X_1 = x)$. Then

$$\bar{f}^t_{x,u}(y, a) = \frac{s}{t} \bar{f}^s_{x,u}(y, a) + p^{g_n}_{a|y} \frac{t-s}{t} \sum_{z \in \mathbf{X}} \chi_s(z) \left(\sum_{r=1}^{t-s} [P^{g_n}]^r_{zy} \right),$$

where P^{g_n} is the transition matrix under g_n. It then follows easily that $\lim_{t\to\infty} \bar{f}^t_{x,u}(y, a) = \zeta_n$.

Thus $d(\zeta, \bar{f}^{t_n}_{x,u}) \leqq d(\zeta, \zeta_n) + d(\zeta_n, \bar{f}^{t_n}_{x,u}) \leqq 2\varepsilon_n$ and we obtain along the subsequence $\{t_n\}_n$, $\bar{f}_{x,u} = \zeta$. By (A2) ζ is a probability measure, so that L_x is closed and sequentially compact, hence compact.

To prove the convexity, recall (the first part of the proof) that $L_x = L_x(S)$. Suppose $\zeta = \beta \bar{f}_{x,u_1} + (1-\beta) \bar{f}_{x,u_2}$ for $u_1, u_2 \in U(S)$. A PTS policy u such that $\bar{f}_{x,u} = \zeta$ is obtained by setting $\alpha_1 := (\beta/\tau_1)(\beta/\tau_1 + (1-\beta)/\tau_2)^{-1}$, and $\alpha_2 = 1 - \alpha_1$ (this follows from (5.2)).

Since L_x is compact and convex in $\mathbb{R}^\infty$, by the Krein-Milman theorem it is the convex hull of its extreme points. Next we show that all extreme points of L_x correspond to deterministic stationary policies. Let g be a stationary nondeterministic policy. Then for some state $z \in \mathbf{X}$ and actions a_1 and a_2 in $A(z)$, the probability α_i to use action a_i under the policy g is strictly positive. Consider the stationary policies u_i that coincide with g in all states except for state z. In state z policy u_i uses action a_i with probability $\alpha_1 + \alpha_2$. Then according to (5.2), the PTS policy $\hat{\alpha}$ that switches in state z between u_1 and u_2 achieves $\bar{f}_{x,g} = \gamma \bar{f}_{x,u_1} + (1-\gamma) \bar{f}_{x,u_2}$. Therefore $\bar{f}_{x,g}$ is not an extreme point in L_x, and since for every policy u there is a $g \in U(S)$ with $\bar{f}_{x,u} = \bar{f}_{x,g}$ this concludes the proof. □

Theorem 5.1 enables us to strengthen theorem 2.6 as follows.

COROLLARY 5.2. *Assume* (A1′) *and* (A2) *under the uniform integrability assumption, or under the assumption that c is bounded from below, the class of deterministic policies is sufficient for problem* OP (*with C defined through either* (1.2a) *or* (1.2b)).

Proof. By Lemma 2.3, the cost of a stationary policies has the representation (2.4). An argument as in the proof of Theorem 5.1 then shows that the cost of any nondeterministic policy is a convex combination of the costs of two other stationary policies. □

Another conclusion from Theorem 5.1 is that under (A1′) and (A2) the state frequencies are bounded from below by a positive (state dependent) constant, uniformly in the policy.

COROLLARY 5.3. *Under* (A1′) *and* (A2), *for each* $y \in \mathbf{X}$ *there exists a constant* $\Delta(y) > 0$ *such that* $\bar{f}_{x,u}(y) > \Delta(y)$ *for all policies* $u \in U$.

Proof. Suppose the claim does not hold. Then there exists a sequence g_n of stationary policies and some state z such that $\lim_{n\to\infty} \bar{f}_{x,g_n}(z) = 0$. We can then construct a subsequence n_k along which $\lim_{n\to\infty} \bar{f}_{x,g_n}(y)$ exists for all $y \in \mathbf{X}$. Using Theorem 5.1 and (3.3), (3.4) there exists some stationary policy g achieving this limit, hence $\bar{f}_{x,g}(z) = 0$, which contradicts (A1). □

Remark. Fisher [18] showed that if the state space forms a single *positive* recurrent class when using any deterministic policy $g \in U(SD)$ then (A1′) holds. He then obtained the same result as in Corollary 5.3 using only the weaker condition (A0) instead of (A2).

Finally, we use Theorem 5.1 to strengthen Theorem 2.8. Theorem 2.8 states that we may restrict our search for optimal policy for COP to the stationary policies. But it does not say that an optimal policy exists. We show that this is indeed the case.

COROLLARY 5.4. *Consider problem* COP. *Assume* (A1′) *and* (A2) *and either*

(i) *Both* $c(y, a)$ *and* $d^k(y, a)$ *are bounded from below*; *or*

(ii) *Both* $c(y, a)$ *and* $d^k(y, a)$ *are uniformly integrable with respect to* $\{\bar{f}_{x,g}\}$, $g \in U(S)$.

If there is any feasible policy then there exists an optimal stationary policy.

Proof. According to Theorem 2.8 we may restrict to the stationary policies in searching for optimality. We first show that $C_x(\cdot)$ and $D_x^k(\cdot)$ are lower semicontinuous functions of the frequencies $\bar{f}_{x,g}$. Let $\{\zeta_n\}$ be a sequence of frequencies, achieved, say, by the stationary policies $\{g_n\}$ (i.e., $\zeta_n(\cdot,\cdot) = \bar{f}_{x,g_n}(\cdot,\cdot)$) converging to ζ. According to Theorem 5.1 there exists a stationary policy g such that $\zeta = \bar{f}_{x,g}$. Under (i) this implies by Fatou's lemma (using $c(\cdot,\cdot)$ as a measure) that the cost function $C_x(g)$ satisfies

$$\begin{aligned} C_x(g) &= \sum_{y,a} \zeta(y, a)c(x, a) = \sum_{y,a} \lim_{n\to\infty} \zeta_n(y, a)c(y, a) \\ &\leqq \lim_{n\to\infty} \sum_{y,a} \zeta_n(y, a)c(y, a) = \lim_{n\to\infty} C_x(g_n) \end{aligned} \tag{5.3}$$

and similarly

$$D_x^k(g) = \sum_{y,a} \lim_{n\to\infty} \zeta_n(y, a)d^k(y, a) \leqq \lim_{n\to\infty} D_x^k(g_n), \tag{5.4}$$

which establishes the lower semicontinuity for the case (i). If (ii) is assumed, then in fact we have continuity. To see that, note that the compactness of $L_x(S)$ (Theorem 5.1) implies by Prohorov's theorem that $\{\zeta_n\}$ are tight, hence converge weakly. As in the proof of Lemma 2.2, consider now $c(y, a)$ and $d^k(y, a)$ as "random variables" on the space $\mathbf{X} \times \mathbf{A}$. The weak convergence of $\{\zeta_n\}$ implies the convergence of $c(\cdot,\cdot)$ and $d^k(\cdot,\cdot)$ in distribution, and combining it with (ii) we obtain $C_x(g) = \lim_{n\to\infty} C_x(g_n)$ and $D_x^k(g) = \lim_{n\to\infty} C_x^k(g_n)$.

We thus have lower semicontinuity under either (i) or (ii). This implies that the set $\Pi_V := \{\mu: \mu \in L(S), D_x^k(\mu) \leqq V^k, 1 \leqq k \leqq K\}$ is compact, since it is obtained as the intersection of the compact set $L(S)$ and the inverse map of the closed sets $(-\infty, V_k]$. Finally, by the lower semicontinuity of $C_x(\cdot)$ on Π_V we conclude that $C_x(\cdot)$ achieves its minimum on Π_V, i.e. there exists an optimal stationary policy for COP. □

6. Application to a queueing system. In this section we apply Theorems 2.6 and 3.2 to investigate a constrained problem in the following discrete-time queueing model. At time t, M_t^k customers arrive to queue k, $1 \leqq k \leqq K$. Each input stream is received

in an infinite capacity buffer. Arrival vectors $M_t = \{M_t^1, \cdots, M_t^K\}$ are independent from slot to slot, form a renewal sequence with finite means λ_k. During a time slot $(t, t+1)$ a customer from any class k, $1 \leqq k \leqq K$ may be served, according to some policy, which is a prespecified dynamic priority assignment. If served, with probability μ_k it completes its service and leaves the system; otherwise it remains in its queue. A generic element of the state is given by $x = \{x^1, x^2, \cdots, x^K\}$ and it represents a K-dimensional vector of the different queue sizes. Altman and Shwartz [1], [3] solve a problem with constraints on the average sizes of several queues. They find an optimal nonstationary time sharing policy, using a linear program. The recurrence properties of this system as well as bounds and representations for average cost functionals for general cost functions are obtained in Makowski and Shwartz [25].

Below we present conditions for the completeness of stationary policies , and the existence of optimal stationary policies for COP with several constraints. Sufficiency is proved for costs that are nonlinear in the queue sizes. We then solve the general constrained problem with linear costs (generalizing [1], [3], [26]). Throughout we restrict to nonidling policies; using coupling (as in [11]) it can be shown that when the costs are positive and increasing (in the number of customers), idling leads to no improvement.

6.1. Completeness and sufficiency of stationary policies. We first show that (A1) and (A2) hold. We assume the standard stability condition on the traffic intensity $\rho := \sum_{k=1}^K \lambda_k/\mu_k < 1$. This is a sufficient condition for (A1) (see [25] or [26]). In order to show that (A2) is satisfied, we use Lemma 4.2. Let $c(x, a) = \sum_{k=1}^K x^k/\mu_k$. The average cost is then finite and does not depend on the policy (this follows from the μc rule [5], [6], [11]). Therefore (with the obvious choice of K_i) all conditions of Lemma 4.2 are satisfied, and (A2) holds. Hence we obtain, using Theorem 2.8, Corollary 3.3, and Corollary 6.1.

COROLLARY 6.1. *Under the foregoing assumptions on the queueing model, the stationary policies are complete. If $c(x, a)$ and $d^k(x, a)$ are bounded below then the stationary policies are sufficient for* COP.

If $c(x, a)$ and $d^k(x, a)$ are not bounded from below then the stationary policies are still sufficient for COP, provided $\{c(X_s, A_s)\}_s$ and $\{d^k(X_s, A_s)\}_s$ are uniformly integrable with respect to P^u for each policy u (Theorem 2.8). In [25], Makowski and Shwartz give the following sufficient conditions (P1) and (P2) for the uniform integrability. For any K-dimensional vector x let $|x|$ denote $\sum_{k=1}^K |x^k|$:

(P1) There exists an integer $\gamma > 1$ such that $E[|M_t|^\gamma] < \infty$ and $E[|X_1|^\gamma] < \infty$.

Note that both expectations are independent of the policy.

(P2) There exist $0 < \delta < \gamma - 1$ and L such that $|c(x, a)| \leqq L(1 + |x|^\delta)$.

These results establish that the search for optimal, or ε-optimal policies may be restricted to the stationary policies. This allows the application of steady-state analysis, of the type used in queueing theory, to problems OP and COP.

6.2. Solving COP with linear cost functions. Consider the linear cost function $c(x, a) := \sum_{k=1}^K c_k x^k$ and $d^i(x, a) = \sum_{k=1}^K d_k^i x^k$ for $1 \leqq i \leqq M$, where c_k and d_k^i are nonnegative constants. This COP problem was solved in [1] and [3] for the case $M = 1$, and for the case $d_k^i = \delta_i(k)$ and $M < K$ using "PTS" policies over the set of $l = K!$ priority policies g_i. It is shown [1] and [3] that under the condition $\rho < 1$ and $E|X_1| < \infty$, all the τ_i (defined below (5.2)) are equal, and the cost under $\hat{\alpha}$ is given by

$$C_x(\hat{\alpha}) = \sum_{i=1}^{l} \alpha_i C(g_i) \tag{6.1}$$

with a similar linear expression for $D_x^k(\hat{\alpha})$. Denote by $\hat{\beta}$ the optimal policy among all PTS policies for problem COP. From (6.1) it follows that $\hat{\beta}$ can be obtained by solving the following linear program:

(6.2) (LP) Find α that minimizes $\sum_{i=1}^{l} \alpha_i C(g_i)$

subject to $\sum_{i=1}^{l} \alpha_i D^k(g_i) \leqq V_k, \quad 1 \leqq k \leqq K, \quad \sum_{i=1}^{l} \alpha_i = 1, \quad \alpha_i \geqq 0$

for $1 \leqq i \leqq l$.

Based on Theorem 2.8(ii) we show in the following theorem that $\hat{\beta}$ is in fact overall optimal.

THEOREM 6.2. *The* PTS *policy obtained by solving* LP *is optimal for* COP.

Proof. Following [3] we define the average size of queue k by

$$\bar{X}_x^k(u) := \overline{\lim_{t\to\infty}} \frac{1}{t} E_u\left[\sum_{s=1}^{t} X_t^k \mid X_1 = x\right]. \tag{6.3}$$

Consider the class U' of all policies satisfying

$$C_x(u) = \sum_{k=1}^{K} c_k \bar{X}_x^k(u) \quad \text{and} \quad D_x^j(u) = \sum_{k=1}^{K} d_k^j \bar{X}_x^k(u), \quad 1 \leqq j \leqq M. \tag{6.4}$$

Note that $U(S) \subset U'$ and $U(PTS) \subset U'$ (this is obtained by applying Lemma 2.3 to compute $\bar{X}_x^k(u)$). According to Theorem 2.8(ii) $U(S)$ is sufficient hence U' must be sufficient. Reference [3] shows that PTS policies are "Pareto optimal" in the following sense. For any policy u there exists a PTS policy w such that $\bar{X}_x^k(w) \leqq \bar{X}_x^k(u)$, $1 \leqq k \leqq K$. This implies that $\hat{\beta}$ is optimal over U', and since U' is sufficient, this implies that $\hat{\beta}$ is overall optimal. □

This result illustrates the usefulness of the present approach. There are several results reducing optimization problems for queues to computable problems (such as linear programs). However, the optimization is usually carried out over a class of policies that is smaller than U' above (e.g., in [19] the optimization is carried out over the class of "steady state" policies). Results on sufficiency then allow to conclude optimality over the class U of *all* policies.

7. Second application: a linear program formulation for COP. Below we present a linear program that we show to be equivalent to COP. Such linear programs have been introduced for the case of finite state and action spaces (e.g., Derman [15] and Hordijk and Kallenberg [22]). In the finite case these are the most important method to compute optimal policies for COP (an alternative linear program is described in [2]). We use a different approach by which we obtain a similar linear program for the countable case. Naturally, we cannot expect to find explicit solutions for COP using an infinite-dimensional linear program, but this approach can be used to shed some light on the structure of optimal solutions for COP. Consider the LP.

Find $\{z^*(y, a)\}_{y,a}$ that minimizes $C(z) := \sum_{y,a} c(y, a) z(y, a)$ subject to

$$\sum_{y,a} z(y, a) P_{yav} \leqq \sum_{a} z(v, a), \qquad v \in \mathbf{X}, \tag{7.1a}$$

$$\sum_{y,a} d^k(y, a) z(y, a) \leqq V_k, \qquad 1 \leqq k \leqq K, \tag{7.1b}$$

$$\sum_{y,a} z(y, a) = 1, \qquad z(y, a) \geqq 0. \tag{7.1c}$$

THEOREM 7.1. *Assume* (A1) *and* (A2) *and assume either that* (i) c *and* d^k *are bounded below, and* (A3(g)) *holds for all stationary* g, *with respect to both* c *and* d^k, $1 \leqq k \leqq K$, *or that* (ii) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1 \leqq k \leqq K$ *are uniformly integrable under* P^u *for all* $u \in U$.

(i) *If the stationary policy* w *is feasible for* COP, *then* $\{z(y, a)\}$ *satisfies* (7.1), *where*

$$z(y, a) = \pi_y^w \cdot p_{a|y}^w. \tag{7.2}$$

(ii) *If* g *is an optimal stationary policy for* COP *then there exists an optimal solution for* LP *satisfying*

$$z^*(y, a) = \pi_y^g \cdot p_{a|y}^g. \tag{7.3}$$

(iii) *Conversely, let* $\{z(y, a)\}$ *satisfy* (7.1). *Then the policy* w *is feasible for* COP, *where*

$$p_{a|y}^w = \frac{z(y, a)}{\sum_{a \in A(y)} z(y, a)} \tag{7.4}$$

whenever the denominator is nonzero, and otherwise $p_{a|y}^w$ *are chosen arbitrarily but such that* $p_{\cdot|y}^w$ *is a probability measure.*

(iv) *If* z^* *solves* LP, *then the stationary policy* g *is optimal for* COP, *where*

$$p_{a|y}^g = \frac{z^*(y, a)}{\sum_{a \in A(y)} z^*(y, a)} \tag{7.5}$$

whenever the denominator is nonzero, and otherwise $p_{a|y}^g$ *are chosen arbitrarily but such that* $p_{\cdot|y}^g$ *is a probability measure.*

Proof. To prove (i) we note that $z(y, a)$ as defined in (7.2) satisfies (7.1c) since π^w and $p_{\cdot|y}^w$ are probability measures. Next we note that $z(a, y) = \bar{f}_{x,w}(a, y)$, thus (7.1b) is satisfied since its left side is equal to $D^k(g)$ by Lemma 2.3. Similarly, (7.1a) is satisfied since by definition π_y^w is invariant under the transition $P_{yv}^w = \sum_a P_{yav} p_{a|y}^w$.

To prove (iii), let $z(y) := \sum_{a \in A(y)} z(y, a)$ and substitute (7.4) in (7.1a) to obtain

$$z(y) \geqq \sum_{v \in \mathbf{X}} z(v) P_{vy}^w. \tag{7.6}$$

Following Lemma 3.1 and using (7.1c) we obtain $z(y) = \pi^w(y) = \bar{f}_{x,w}(y)$. By (7.4) and by the fact that $\bar{f}_{x,w}(y, a) = \bar{f}_{x,w}(y) \cdot p_{a|y}^w$ we obtain $z(y, a) = \pi^w(y) p_{a|y}^w = \bar{f}_{x,w}(y, a)$. It then follows by Lemma 2.3 and (7.1b) that $D^k(w) \leqq V_k$, $1 \leqq k \leqq K$, and therefore w is feasible for COP.

Parts (ii) and (iv) follow from the fact established above that (7.1) and (7.4) define a one-to-one correspondence between the z's that are feasible to LP and the stationary policies w's that are feasible to COP. Moreover, under this correspondence, it follows from Lemma 2.3 that the value $C(z)$ of LP is equal to $C_x(w)$, which establishes the proof. □

8. Extensions. In this section we outline some applications of our methods to lesser-known optimization criteria, involving variance minimization.

8.1. Variability sensitive optimization. The variability sensitive optimization problem VSOP was studied in the finite case by Filar, Kallenberg, and Lee [16] and later by Bayal-Gursoy and Ross [7];

$$\text{Maximize } R_x(u) := \lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} E_u[r(c(X_s, A_s), C_x^t(u))], \tag{8.1}$$

where $r(\cdot, \cdot)$ is called the variability function. Taking $r(x, y) = x - \lambda(x-y)^2$ the VSOP obtains the interpretation of finding a policy u that has high expected average reward but low expected variance. Other variability criteria and other variability functions are treated in the finite state-action space in [7] and [16].

In Theorem 8.1 we present conditions that ensure the sufficiency of classes of policies for problem VSOP. We will use $r(x, y) = x - \lambda(x-y)^2$. Note that when $\lambda = 0$ this reduces to problem OP.

THEOREM 8.1. *Consider problem* VSOP. *Assume* (A1) *and* (A2) *and let* U' *be complete. If* $\{c^2(X_s, A_s)\}_s$ *is uniformly integrable with respect to* P^u *for each* u, *then* U' *is sufficient.*

Proof. First note that $R_x(u)$ is equal to

$$\underline{\lim}_{t\to\infty}\left[\sum_{y,a}\bar{f}^t_{x,u}(y,a)[c(y,a)-\lambda c^2(y,a)]+\lambda\left[\sum_{y,a}\bar{f}^t_{x,u}(y,a)c(y,a)\right]^2\right]. \tag{8.2}$$

Let t_n be any subsequence of t that achieves the $\underline{\lim}$ in the expression above, and along which $\bar{f}^{t_n}_{x,u}(y,a)\to\bar{f}_{x,u}(y,a)$ for all y and a. Following the same weak convergence arguments that were used in § 2.1, we obtain from the uniform integrability

$$R_x(u)=\sum_{y,a}\bar{f}_{x,u}(y,a)[c(y,a)-\lambda c^2(y,a)]+\lambda\left[\sum_{y,a}\bar{f}_{x,u}(y,a)c(y,a)\right]^2. \tag{8.3}$$

Thus $R_x(u)$ can be represented as a function of the expected state-action frequency, so completeness implies sufficiency. □

As a simple corollary, for bounded cost completeness implies sufficiency.

8.2. The problem with constraints. VSOP can also be considered in the framework of optimization under constraints. Kawai [23] introduced the problem of minimizing the variance of some cost subject to a single constraint on the expected average cost. He treats the case of finite state and action spaces, and restricting to the stationary policies he finds an optimal solution. Kurano [24] finds a policy that is optimal among the stationary deterministic policies for the same problem as Kawai yet with general state and action spaces.

Using similar arguments as above, we show below that any complete family of policies (e.g., the stationary policies) is sufficient for the problem of Kawai; hence the solution that Kawai finds is overall optimal. Moreover, using the same kind of assumptions as in Theorem 8.1 we show (using arguments as in the proof of Theorem 2.8) that these are sufficient for the case of countable state and action spaces, and for more than one constraint on expected average cost functionals.

Denote the variance under a policy u with initial state x by $R_x(u)$ through (8.1) with $r(x, y) := (x-y)^2$. Given K real numbers $V_1, \cdots, V_K$, define the following constrained problem:

(CVSOP) minimize $R_x(u)$

subject to $\bar{D}^k_x(u) \leqq V_k, \qquad 1 \leqq k \leqq K.$

References [23] and [24] consider the case $V = V_1$ that is (ε) close to the supremum of the optimal expected average cost. The meaning of CVSOP is then to find a policy that minimizes the variance among all policies that are ε-optimal for OP.

THEOREM 8.2. *Consider problem* CVSOP. *Assume* (A1) *and* (A2) *and let* U' *be complete. If* $\{c^2(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$ $1 \leqq k \leqq K$ *are uniformly integrable with respect to* P^u *for each* u, *then* U' *is sufficient.*

Proof. The variance is given by

$$R_x(u) = \overline{\lim_{t\to\infty}} \left(\left[\sum_{y,a} \bar{f}^t_{x,u}(y,a)c(y,a) \right]^2 - \sum_{y,a} \bar{f}^t_{x,u}(y,a)c^2(y,a) \right).$$

By diagonalization, there exists some subsequence t_n along which $\lim_{n\to\infty} \bar{f}^{t_n}_{x,u}(y,a) = \bar{f}_{x,u}(y,a)$ for all y and a, such that

$$R_x(u) = \left[\sum_{y,a} \bar{f}_{x,u}(y,a)c(y,a) \right]^2 - \sum_{y,a} \bar{f}_{x,u}(y,a)c^2(y,a).$$

(similarly to the way (8.3) is obtained).

The rest of the proof now follows the same lines as the proof of Theorem 2.8. □

Acknowledgements. The authors wish to thank F. Spieksma and A. Hordijk for extensive discussions concerning this paper, and an anonymous reviewer for many constructive comments, which led to significant improvements in many respects.

The support and hospitality of the Mathematics of Networks and Systems Department, AT&T Bell Laboratories, Murray Hill, NJ, is gratefully acknowledged.

REFERENCES

[1] E. ALTMAN AND A. SHWARTZ, *Optimal priority assignment with general constraints*, in Proc. 24th Allerton Conference, University of Illinois, Urbana-Champaign, IL, October 1986.

[2] ———, *Non-stationary policies for controlled Markov chains*, EE Pub. 633, Technion-Israel Institute of Technology, Haifa, Israel, June 1987, submitted.

[3] ———, *Optimal priority assignment: a time sharing approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 1098–1102.

[4] ———, *Adaptive control of constrained Markov chains*, IEEE Trans. Automat. Control, to appear (1991).

[5] J. S. BARAS, A. J. DORSEY, AND A. M. MAKOWSKI, *Discrete time competing queues with geometric service requirements: stability, parameter estimation and adaptive control*, SIAM J. Control Optim., submitted.

[6] J. S. BARAS, D.-J. MA, AND A. M. MAKOWSKI, *K competing queues with geometric service requirements and linear costs: the μc rule is always optimal*, Systems Control Lett., 6 (1985), pp. 173–180.

[7] M. BAYAL-GURSOY AND K. W. ROSS, *Variability sensitive Markov decision processes*, Math. Oper. Res., to appear.

[8] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[9] V. BORKAR, *On minimum cost per unit time control of Markov Chains*, SIAM J. Control Optim., 22 (1983), pp. 965–978.

[10] ———, *Control of Markov chains with long-run average cost criterion*, in Proc. Stochastic Differential Systems, W. Fleming and P.-L. Lions, eds., Springer-Verlag, Berlin, New York, 1986, pp. 57–77.

[11] C. BUYYUKKOC, P. VARAIYA, AND J. WALRAND, *The cμ rule revisited*, Adv. Appl. Probab., 17 (1985), pp. 237–238.

[12] R. CAVAZOS-CADENA, *Existence of optimal stationary policies in average-reward Markov decision processes with a recurrent state*, Appl. Math. Optim., submitted.

[13] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer-Verlag, New York, 1967.

[14] R. DEKKER AND A. HORDIJK, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., 13 (1988), pp. 395–421.

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

[16] J. A. FILAR, L. C. M. KALLLENBERG, AND H. M. LEE, *Variance penalized Markov decision processes*, Math. Oper. Res., 14 (1989), pp. 147–161.

[17] L. FISHER AND S. M. ROSS, *An example in denumerable decision processes*, Ann. Math. Statist., 39 (1968), pp. 674–675.

[18] L. FISHER, *On recurrent denumerable decision processes*, Ann. Math. Statist., 39 (1968), 424–434.

[19] E. GELENBE AND I. MITRANI, *Analysis and Synthesis of Computer Systems*, Academic Press, London, 1980.

[20] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and Its Applications*, John Wiley, New York, 1980.
[21] A. HORDIJK, *Dynamic Programming and Markov Potential Theory*, Mathematical Center Tracts, no. 51, Amsterdam, the Netherlands, 1974.
[22] A. HORDIJK AND L. C. M. KALLENBERG, *Constrained undiscounted stochastic dynamic programming*, Math. Oper. Res., 9 (1984), pp. 276–289.
[23] H. KAWAI, *A variance minimization problem for a Markov Decision process*, European J. Oper. Res., 31 (1987), pp. 140–145.
[24] M. KURANO, *Markov decision processes with a minimum-variance criterion*, J. Optim. Theory Anal., 123 (1987), pp. 572–583.
[25] A. M. MAKOWSKI AND A. SHWARTZ, *Recurrence properties of a system of competing queues, with applications*, EE Pub. 627, Technion-Israel Institute of Technology, Haifa, Israel, 1987.
[26] P. NAIN AND K. W. ROSS, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control, 31 (1986), pp. 883-888.
[27] D. REVUZ, *Markov Chains*, North-Holland, Amsterdam, the Netherlands, 1975.
[28] L. I. SENNOTT, *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Oper. Res. Lett., 5 (1986), pp. 17–23.
[29] ———, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Oper. Res., 37 (1989), pp. 626–633.