CONSTRAINED DISCOUNTED DYNAMIC PROGRAMMING

EUGENE A. FEINBERG

SUNY at Stony Brook

ADAM SHWARTZ

Technion—Israel Institute of Technology

August 1994; Revised: September 1995. To appear, MOR

Abstract

This paper deals with constrained optimization of Markov Decision Processes with a countable state space, compact action sets, continuous transition probabilities, and upper semi-continuous reward functions. The objective is to maximize the expected total discounted reward for one reward function, under several inequality constraints on similar criteria with other reward functions.

Suppose a feasible policy exists for a problem with M constraints. We prove two results on the existence and structure of optimal policies. First, we show that there exists a randomized stationary optimal policy which requires at most M actions more than a nonrandomized stationary one. This result is known for several particular cases. Second, we prove that there exists an optimal policy which is (i) stationary (non-randomized) from some step onward, (ii) randomized Markov before this step, but the total number of actions which are added by randomization is at most M, (iii) the total number of actions that are added by nonstationarity is at most M.

We also establish Pareto optimality of policies from the two classes described above for multi-criteria problems. We describe an algorithm to compute optimal policies with properties (i)–(iii) for constrained problems.

The policies that satisfy properties (i)–(iii) have the pleasing aesthetic property that the amount of randomization they require over any trajectory is restricted by the number of constraints. In contrast, a randomized stationary policy may require an infinite number of randomizations over time.

AMS 1980 subject classification: Primary: 90C40.
IAOR 1973 subject classification: Main: Programming, Markov Decision.
OR/MS Index 1978 subject classification: Primary: 119 Dynamic Programming/Markov
Key words: Dynamic programming, discounted Markov decision processes, additional constraints.

1. Introduction.

This paper deals with discrete time Markov Decision Processes (MDPs) with the expected total discounted criteria (2.1). The goal is to maximize the expected total discounted reward (2.2) for one reward function under M inequality constraints (2.3) for expected total discounted rewards, where M is a finite positive integer. We consider countable state and compact action sets, continuous transition probabilities, and upper-semicontinuous rewards.

Example 1 in Frid (1972) indicates that randomized policies may be significantly better than nonrandomized ones for discounted problems with constraints. In the case of finite state and action sets (Kallenberg 1983, Heyman and Sobel 1984), there exist randomized stationary optimal policies. Borkar (1988) showed that, for discounted countable MDPs, for any policy π there exists a randomized stationary policy ϕ such that, for any reward function, the expected total discounted rewards for ϕ and π are equal.

It is known that the number of states in which a randomized stationary optimal policy should select decisions randomly depends on M. We say that a randomized stationary policy is m-randomized if it uses at most m actions more than a (nonrandomized) stationary policy; for precise definitions see Section 2. By Theorem 4.2 in Feinberg and Shwartz (1995), for finite state and action sets with M constraints, there exist M-randomized stationary optimal policies. For countable state models satisfying some additional assumptions, similar results were proved by Sennott (1991) for a model with one constraint and with compact action sets and by Altman (1994), Remark (3), for a model with finite action sets. Theorem 2.1(i) below establishes the existence of M-randomized stationary optimal policies for countable state MDPs with compact action sets and with M constraints.

To the best of our knowledge, all studies of constrained discounted MDPs, except Feinberg and Shwartz (1995), deal with randomized stationary policies. Feinberg and Shwartz (1995) considered MDPs with finite state and action sets, and with weighted discounted rewards (a sum of several standard discounted criteria, each with a different discount factor). Under this criterion, good randomized stationary policies may not exist even for unconstrained problems; see Example 1.1 in Feinberg and Shwartz (1994). The class of (m, N)-policies was introduced in Feinberg and Shwartz (1995). Policies in this class are characterized by the following two properties: they are (nonrandomized) stationary from epoch N onwards; they are Markov at steps $0, \ldots, N-1$, but they use at most m actions more than a (nonrandomized) Markov policy; see (2.5). For a weighted discounted MDP with finite state and action sets and with M constraints, the existence of optimal

(M, N)-policies for some $N < \infty$ was established in Feinberg and Shwartz (1995), Theorem 6.8.

Note that (M, N)-policies have the pleasing aesthetic property that the total number of randomizations they require over any trajectory is restricted by M. By contrast, a randomized stationary policy may require an infinite number of randomizations over time – every time when one of the states where randomization is required is reached; see Example 2.2 below. As noted in Feinberg and Shwartz (1995), the existence of optimal (M, N)-policies was a new result for constrained discounted MDPs with finite state and action sets, even in the case of a single discount factor. But in fact, a stronger result holds in this case.

We say that a policy π is a strong (m, N)-policy if it is an (m, N)-policy and there exists an m-randomized stationary policy ϕ such that for each state the set of actions used by π over the entire horizon is limited to the set of actions used by ϕ; for a precise definition see Section 2. The results of Feinberg and Shwartz (1995) imply that, for constrained discounted MDPs with finite state and action sets, there exist optimal strong (M, N)-policies. To see this, select an M-randomized stationary optimal policy ϕ using Theorem 4.2 in Feinberg and Shwartz (1995). Next consider the MDP where the feasible action sets are replaced by the action sets used by ϕ. Now apply Theorem 6.8 in Feinberg and Shwartz (1995) on the existence of optimal (M, N)-policies to the new MDP. This policy is strong (M, N)-optimal for the original MDP.

In this paper we prove that, for a constrained discounted MDP with countable state and compact action sets, for some $N < \infty$ there exists an optimal strong (M, N)-policy; Theorem 2.1(ii). We also provide an algorithm to compute such policies for models with finite state and action sets. We remark that optimal (M, N)-policies may not exist for a weighted discounted MDP when either the state or the action set is infinite; see Examples 3.16 and 3.17 in Feinberg and Shwartz (1994).

In Section 2 we formulate the problem and state the main results. In Section 3 we formulate geometrical lemmas which are used later when we study the geometric properties of the performance set. Unlike the sets of occupation measures studied by Borkar (1988), Altman and Shwartz (1991), and Altman (1993, 1994), the performance set has finite dimension $(M + 1)$, and is hence a more convenient object of study.

The main results of the paper are proved in Section 4. Most of that section deals with the structure of Pareto optimal policies in the set of all policies. We establish that any Pareto optimal point can be attained by an M-randomized stationary policy, and also by a strong (M, N)-policy; Theorem 4.1. This leads to the main result of the paper, the existence of optimal M-randomized and strong (M, N)-policies; Theorem 2.1. In Section 5, we prove the existence of M-mixed stationary optimal policies. In Section 6, we describe an algorithm to compute optimal strong (M,N)-policies in models with finite state and action sets.

2. Problem Formulation and Main Results.

Let $\mathbb{N}_0 = \{0,1,\ldots\}$, $\mathbb{N} = \{1,2,\ldots\}$, and fix $M \in \mathbb{N}_0$. We consider a discrete-time MDP with a countable state space $\mathbf{X}$, action space $\mathbf{A}$, sets of actions $\mathbf{A}(x) \subseteq \mathbf{A}$ available at $x \in \mathbf{X}$, transition probabilities $\{p(y \mid x,a)\}$, and reward functions $r_m(x,a)$, where $m = 0,\ldots,M$, $x,y \in \mathbf{X}$ and $a \in \mathbf{A}(x)$, and discount factor $\beta \in [0,1)$.

Throughout the paper, the following assumptions hold.

Assumptions. The set $\mathbf{A}$ is a metric space, $\mathbf{A}(x)$ are compact sets, the functions $p(y|x,a)$ are continuous in a and $\sum_{y\in\mathbf{X}} p(y \mid x,a) = 1$, the functions $r_m(x,a)$ are upper semicontinuous in a and uniformly bounded above, where $m = 0,\ldots,M$, $x \in \mathbf{X}$, and $a \in \mathbf{A}(x)$.

Let $H_n = \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^n$ be the space of histories up to the epoch $n \in \mathbb{N}_0$ or $n = \infty$. Let $H = \bigcup_{0\le n<\infty} H_n$ be the space of all finite histories. The spaces H_n, $n \in \mathbb{N}_0$ or $N = \infty$, and H are endowed with σ-fields $\mathcal{B}_n$ and $\mathcal{F}$ generated by $2^{\mathbf{X}}$ and by the Borel σ-field on $\mathbf{A}$. A policy π is a function that assigns to each prehistory $h_n = x_0a_0x_1\ldots x_n \in H_n$, $n = 0,1,\ldots$, a probability distribution $\pi(\cdot \mid h_n)$ on $\mathbf{A}$ satisfying the conditions that $\pi(\mathbf{A}(x_n) \mid h_n) = 1$ and $\pi(E|h)$ is a measurable function on H for any (Borel) measurable subset E of $\mathbf{A}$. We denote by Π the set of all policies.

Using standard notation and construction, each policy π and initial state x induce a probability measure $\mathbb{P}_x^\pi$ on H_∞. We denote the corresponding expectation operator by $\mathbb{E}_x^\pi$.

For $m = 0,\ldots,M$, we define the expected total discounted rewards

$$V_m(x,\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta^n r_m(x_n,a_n), \qquad m = 0,\ldots,M. \tag{2.1}$$

The $(M+1)$-dimensional vector $V(x,\pi) = (V_0(x,\pi), V_1(x,\pi),\ldots,V_M(x,\pi))$ characterizes the performance of a policy $\pi \in \Pi$ when the initial state is $x \in \mathbf{X}$. We are interested in the following constrained optimization problem: given $x \in \mathbf{X}$ and the numbers $c_1,\ldots,c_M$

$$\text{maximize} \quad V_0(x,\pi) \tag{2.2}$$

$$\text{subject to} \quad V_m(x,\pi) \ge c_m, \quad m = 1,\ldots,M, \tag{2.3}$$

over $\pi \in \Pi$. If π satisfies (2.3) then it is called feasible, and if it also achieves the (constrained) maximum then it is called optimal at x, or in short (since x is fixed throughout) optimal. We remark than a more general case, when an initial probability distribution μ on $\mathbf{X}$ is fixed, can be reduced to the situation, when the initial state is fixed, by introducing an additional state, from which the process moves according to the distribution μ, by starting the decision process one time earlier, and by multiplying all the constraints by β.

A policy π is called *randomized Markov* if, for each $n = 0, 1, \dots$ and each $y \in \mathbf{X}$, there exists a distribution $\pi_n(\cdot \mid y)$ such that $\pi(\cdot \mid h_n) = \pi_n(\cdot \mid x_n)$ for any $h_n \in H$. Let Π_{RM} be the set of randomized Markov policies. Given an initial state $x \in \mathbf{X}$, for any policy π there exists a randomized Markov policy σ such that $V_m(x, \sigma) = V_m(x, \pi)$ for criteria (2.1), $m = 0, \dots, M$; see Derman and Strauch (1966), Hordijk (1974) Theorem 13.2, or Feinberg and Shwartz (1995) for details. Therefore, without loss of generality, the set of all policies Π may be narrowed to Π_{RM} in problem (2.2)–(2.3).

A randomized Markov policy is called *randomized stationary* if $\pi_n(\cdot \mid y) = \pi_0(\cdot \mid y)$ for any $n = \mathbb{N}$ and any $y \in \mathbf{X}$. A *Markov* policy ϕ is a sequence of mappings $\phi_n : \mathbf{X} \to \mathbf{A}$ such that $\phi_n(y) \in \mathbf{A}(y)$ for any $y \in \mathbf{X}$. A Markov policy is called *stationary* if $\phi_n(y) = \phi_0(y)$ for any $n = \mathbb{N}$ and any $y \in \mathbf{X}$. Given $0 \leq N < \infty$, a Markov policy π is called (N, ∞)-*stationary* if there exists a stationary policy ϕ such that $\pi_n(x) = \phi(x)$ for any $x \in \mathbf{X}$ and for any $n = N, N+1, \dots$. Stationary policies are $(0, \infty)$-stationary and vice versa.

For each $y \in \mathbf{X}$ fix a nonempty compact subset $B(y)$ of $\mathbf{A}(y)$. An MDP with the action sets $B(y)$, $y \in \mathbf{X}$, is called a submodel of the original MDP. We observe that any function $B : \mathbf{X} \to 2^{\mathbf{A}}$ with compact values $B(y) \subseteq \mathbf{A}(y)$, $y \in \mathbf{X}$, defines a submodel. For brevity, we refer to the submodel with action sets $B(y)$, $y \in \mathbf{X}$, as submodel B. We denote by Π^B the set of all policies in submodel B. If the sets $B(y)$ have finite cardinality for all $y \in \mathbf{X}$, then the submodel is called *finite*. For any set E, we denote by $|E|$ its cardinality (which is the number of elements in E if E is finite). For a submodel B, we call $\sum_{y \in \mathbf{X}} (|B(y)| - 1)$ the index of the submodel. If the index of a submodel is finite then the submodel is finite. A finite submodel has index 0 if and only if all sets $B(y)$ are singletons, $y \in \mathbf{X}$.

Note that a randomized Markov policy π belongs to submodel B if and only if $\pi_n(B(y)|y) = 1$ for all $y \in \mathbf{X}$ and $n \in \mathbb{N}$. We say that a randomized Markov policy has index $m \in \mathbb{N}$ if it belongs to a submodel with index m. Since any randomized stationary policy is randomized Markov, indexes are defined for randomized stationary policies as well. A randomized stationary policy is called m-*randomized* if it has index m. The notions of 0-randomized stationary and stationary policies coincide. If a randomized Markov policy π has index m, then it uses at most m actions more than a stationary policy.

For an integer m, we say that a Markov policy π is a *randomized Markov policy of order* m if

$$\sum_{n=0}^{\infty} \sum_{y \in \mathbf{X}} (|\{a \in \mathbf{A}(y) : \ \phi_n(a|y) > 0\}| - 1) \leq m. \tag{2.5}$$

In words, a randomized Markov policy is randomized Markov of order m if, throughout all state-time pairs (x_n, n), this policy uses at most m actions more than a (nonrandomized) Markov policy. We note that the notions of Markov and randomized Markov policy of order 0 coincide.

A policy π will be called a (m, N)*-policy*, where $m, N \in \mathbb{N}_0$, if π is a randomized Markov policy of order m and, in addition, $\pi_n(\phi(x)|x) = 1$ for any $x \in \mathbf{X}$, for some stationary policy ϕ, and for any $n \geq N$. In words, a policy is an (m, N)-policy if on steps $0, \ldots, N-1$ it coincides with a randomized Markov policy of order m, and on steps $N, N+1, \ldots$ it coincides with a stationary policy. We note that the notions of a $(0, N)$-policy and (N, ∞)-stationary policy coincide.

We say that π is a strong (m, N)-policy if it is an (m, N)-policy and has index m. In words, this policy is a randomized Markov policy of order m, has an index m, and is stationary from epoch N onward. In this paper, we prove the following result.

Theorem 2.1. *If problem (2.2)–(2.3) is feasible then*

(i) *there exist an optimal M-randomized stationary policy,*

(ii) *for some $N < \infty$ there exists an optimal strong (M, N)-policy.*

As discussed in the introduction, statement (i) generalizes or strengthens several known results. Statement (ii) describes a new class of optimal policies. Implementation of strong (M, N)-policies can be more natural for many situations since these policies require at most M randomized decisions for all possible state-time pairs and, in addition, they eventually become (nonrandomized) stationary. In contrast, M-randomized stationary policies may require an infinite number of randomized decisions. The following simple example illustrates the difference between these optimal policies.

Example 2.2. Let $\mathbf{X} = \{1\}$, $\mathbf{A} = \mathbf{A}(1) = \{b, d\}$, $p(1|1, b) = p(1|1, d) = 1$, $M = 1$, $r_0(1, b) = r_1(1, d) = 1$, $r_0(1, d) = r_1(1, b) = 0$, and $\beta = .5$. If $c_1 = 1$ then a randomized stationary optimal policy ϕ is unique and $\phi(c|1) = \phi(d|1) = .5$. In this case, the Markov policy σ with $\sigma_0(1) = d$ and $\sigma_n(1) = b$ for $n \geq 1$ is an optimal strong $(1, 1)$-policy. The Markov policy σ' with $\sigma'_0(1) = b$ and $\sigma'_n(1) = d$ for $n \geq 1$ is optimal as well.

Now let $c_1 = 1\frac{1}{3}$. Then ψ, with $\psi(b|1) = \frac{1}{3}$ and $\psi(d|1) = \frac{2}{3}$ is the unique randomized stationary optimal policy. Define π by $\pi_0(b|1) = \frac{2}{3}$, $\pi_0(d|1) = \frac{1}{3}$ and $\pi_n(d|1) = 1$ for $n \geq 1$. Then π is an optimal strong $(1, 1)$-policy that uses a randomized procedure at step 0.

The randomized stationary optimal policies ϕ and ψ require randomization at each step. Policy σ is nonrandomized and has a simple structure. Policy π has a simple structure and requires a randomization procedure just once.

3. Extreme Subsets and Pareto Optimality.

In this section we describe properties of Pareto optimal subsets of convex sets which will be used in the proof of Theorem 2.1. The main result of this section is Lemma 3.2. Since these are only tools for the main results of Section 4, we relegate the proofs to the Appendix.

Let $\mathbb{R}^{M+1}$ be the $(M + 1)$-dimensional Euclidean space. We say that $u \in \mathbb{R}^{M+1}$ dominates $v \in \mathbb{R}^{M+1}$ if $(u - v)$ has nonnegative coordinates. Given a set $E \subseteq \mathbb{R}^{M+1}$, a point $u \in E$ is called Pareto optimal in E if $v = u$ for any point $v \in E$ which dominates u.

We recall some definitions and facts from convex analysis; see Stoer and Witzgall (1970). A convex subset D of a convex set E is called extreme if any representation $u_3 = \lambda u_1 + (1 - \lambda)u_2$, $0 < \lambda < 1$, with $u_1, u_2 \in E$ of a point $u_3 \in D$ is possible only if $u_1, u_2 \in D$. A subset D of E is called exposed if there is a hyperplane H supporting E such that $D = H \cap E$. Extreme and exposed subsets other than E are called proper. Any exposed subset of a convex set is extreme; Stoer and Witzgall (1970) p. 43, but the converse may not hold.

Intersection of any number of extreme (exposed) subsets of a convex set is also an extreme (exposed) subset; Stoer and Witzgall (1970), Proposition (3.6.2). For every point u in a convex set E we denote by $G(u)$ the minimal extreme subset that contains u. This subset is the intersection of all extreme subsets that contain u.

We say that a subset E of $\mathbb{R}^{M+1}$ is bounded above if there is an element of $\mathbb{R}^{M+1}$ which dominates any element of E. For a set E in $\mathbb{R}^{M+1}$ we denote by $\mathrm{Par}(E)$ the set of Pareto optimal elements of E. We denote by e_i the unit vector in direction i, i.e. e_i is the element of $\mathbb{R}^{M+1}$ whose i^{th} coordinate is 1 and all other coordinates are zeros.

Lemma 3.1. *Let E be a convex set in $\mathbb{R}^{M+1}$. Then $G(u) \subseteq \mathrm{Par}(E)$ for any $u \in \mathrm{Par}(E)$.*

Lemma 3.2. *Let $E \subseteq \mathbb{R}^{M+1}$ be a convex set and let $u = (u_0, \ldots, u_M) \in \mathrm{Par}(E)$. Then for some* *finite $1 \le k \le M+1$ there exist hyperplanes*

$$P_j = \left\{ (v_0, \ldots, v_M) \in \mathbb{R}^{M+1} : \sum_{i=0}^{M} b_{ji} v_i = d_j \right\}, \quad j = 1, \ldots, k,$$

with the following properties:

(i) $G(u) = E \cap \left(\bigcap_{j=1}^{k} P_j \right)$,

(ii) $\sum_{i=0}^{M} b_{1i} v_i \le d_1$ *for all* $(v_0, \ldots, v_M) \in E$,

(iii) $\sum_{i=0}^{M} b_{ji} v_i \le d_j$ *for all* $(v_0, \ldots, v_M) \in E_j = E \cap \left(\bigcap_{l=1}^{j-1} P_l \right)$, $j = 2, \ldots, k$,

(iv) $b_{ji} \ge 0$ *for all* $j = 1, \ldots, k-1$, $i = 0, \ldots, M$,

(v) $b_{ki} > 0$ *for all* $i = 0, \ldots, M$.

We remark that (i)–(iii) imply that P_1 supports E at u and P_j support E_j at u, $j = 1, \ldots, k$. In our case, (iv) is particularly important since linear combinations of the reward functions with negative coefficients are not upper semicontinuous. Note that this lemma provides an algorithmic characterization of minimal extreme subsets containing Pareto optimal points.

Corollary 3.3. *If $E \subset \mathbb{R}^{M+1}$ is a closed convex set which is bounded above then $G(u)$ is convex and compact for each $u \in \mathrm{Par}(E)$.*

Lemma 3.4. *Let $E \subset \mathbb{R}^{M+1}$ be a closed convex set which is bounded above and let $u \in \mathrm{Par}(E)$. Then either $G(u) = \{u\}$ or u is a relatively internal point of $G(u)$.*

Remark 3.5. Consider the space $\tilde{R}^{M+1}$ of vectors $(u_0, \ldots, u_M)$ where $-\infty \le u_m < \infty$, that is, we allow the coordinates to take the value $(-\infty)$. For any subset $E \subseteq \tilde{R}^{M+1}$ denote

$$E_{>} = \{u = (u_0, \ldots, u_M) : u \in E \text{ and } u_m > -\infty, m = 0, \ldots, M\}.$$

If E is convex then $E_>$ is convex. Suppose E is convex and $u \in E_> \cap \mathrm{Par}(E)$. If $G(u)$ is the minimal extreme subset of E which contains u, then $G(u) \subseteq E_>$. This follows since if D is an extreme subset of a convex set E then $D_>$ is an extreme subset. We conclude that Lemmas 3.1, 3.2, 3.4 and Corollary 3.3 hold in the expanded space $\tilde{R}^{M+1}$ under the additional condition that all coordinates of the point u are finite.

4. Pareto Optimal Policies.

The main goal of this section is to prove Theorem 2.1. To prove this theorem on constrained optimization, we develop a theory, of independent interest, of vector optimization for discounted MDPs. This topic has been studied in the literature; see Furukawa (1980), Henig (1983, 1985), and references therein.

Furukawa (1980) discussed vector optimality with respect to a cone (a point u in a set $E \in \mathbb{R}^{M+1}$ is optimal if there is no point v in E such that $v - u$ is in the cone). This concept generalizes Pareto optimality as well as lexicographic optimality. He considered the performance set over *nonrandomized stationary policies* only and obtained a characterization of the optimal policies, as well as a "policy improvement" algorithm. Henig (1983) considered a more general dynamic programming model, and with a notion of "optimality" that is more general than in Furukawa (1980). He considered history dependent, but nonrandomized policies. Nonrandomized stationary policies were then proved to be "optimal," and every "optimal" point in the performance set is shown to be a convex combination of points attained by (nonrandomized) stationary policies. Henig (1985) generalized this model to the case where returns are in some partially ordered set (again restricting to nonrandomized policies), proved the validity of Bellman's optimality principle, showed that the set of "optimal points" is nonempty and provided some approximation algorithms.

We consider the performance set $V(x, \Pi)$ over *all policies* and describe the structure of Pareto optimal policies in Π for the standard coordinate-wise partial order in $\mathbb{R}^{M+1}$. Theorem 4.1 describes the structure of Pareto optimal policies. This result implies Theorem 2.1. The proof of Theorem 4.1 combines geometrical results from Section 3 and properties of MDPs. In particular, we use the fact that the set of optimal policies for standard discounted problems forms a submodel of the original MDP; Theorem 4.2. For the case of MDPs with finite state and action sets and with weighted discounted criteria, the structure of these sets, so-called funnels, was established in Feinberg and Shwartz (1995). The situation considered here is simpler than that of Feinberg and Shwartz (1995) in that we do not have to consider funnels, but more complicated in that the performance set may be non-compact.

We remark that there are some connections between vector optimization of stochastic processes and Statistical Decision Theory; see Wald (1950), Blackwell and Girshick (1954), Le Cam (1955), and Ferguson (1967). For example, a notion of a Pareto optimal policy is related to a notion of an admissible decision rule; linear combinations of criteria with nonnegative weights considered in the paper are related to the concept of Bayes optimization.

For a subset of policies $\Delta \in \Pi$, we denote

$$V(x, \Delta) = \{u \in \mathbb{R}^{M+1} : u = V(x, \pi) \text{ for some } \pi \in \Delta\}.$$

Then $V(x, \Pi)$ is the "performance set." A policy π is called Pareto optimal if $V(x, \pi)$ is Pareto optimal in $V(x, \Pi)$. We say that a policy π dominates a policy σ at x if $V(x, \pi)$ dominates $V(x, \sigma)$. Policies π and σ are called equivalent at x if $V(x, \pi) = V(x, \sigma)$.

The goal of this section is to prove Theorem 4.1. Theorems 4.1, 4.5 and Lemma 4.4 imply Theorem 2.1. The proof of Theorem 4.1 uses the results of Section 3.

Theorem 4.1. *If π is a Pareto optimal policy then*

(i) *there exists an equivalent M-randomized stationary policy,*

(ii) *for some $N < \infty$ there exists an equivalent strong (M, N)-policy.*

We consider a single-criterion problem first. For this problem, we write r instead of r_0 and, since $V(x, \pi) = V_0(x, \pi)$, we write V instead of V_0. Let $\mathbf{V}(x) = \sup_{\pi \in \Pi} V(x, \pi)$. In the single-criterion case, a policy π is called optimal if $V(x, \pi) = \mathbf{V}(x)$ for all $x \in \mathbf{X}$. For functions $f : \mathbf{X} \to [-\infty, \infty)$ which are bounded above define

$$P^a f(x) = \sum_{y \in \mathbf{X}} p(y|x, a) f(y), \qquad a \in \mathbf{A}(x),$$

and $T^a f(x) = r(x, a) + \beta P^a f(x)$. We define the optimality operator $Tf(x) = \sup_{a \in \mathbf{A}(x)} T^a f(x)$.

It is well-known that $\mathbf{V}$ satisfies the optimality equation $\mathbf{V} = T\mathbf{V}$; see Hinderer (1970), Theorem 6.7, Dynkin and Yushkevich (1979), Section 6.1, or van der Wal (1981), Theorem 2.1. For any $x \in \mathbf{X}$ we define the set of conserving actions $\mathbf{A}^c(x) = \{a \in \mathbf{A}(x) : \ T^a\mathbf{V}(x) = \mathbf{V}(x)\}$.

Theorem 4.2. *(i) For any $x \in \mathbf{X}$ the set $\mathbf{A}^c(x)$ is nonempty and compact. (ii) A randomized stationary policy π is optimal if and only if $\pi(\mathbf{A}^c(x)|x) = 1$ for all $x \in \mathbf{X}$. (iii) For an arbitrary policy π, $V(x, \pi) = \mathbf{V}(x)$ if and only if $a_n \in \mathbf{A}^c(x_n)$ ($\mathbb{P}_x^\pi$-a.s.).*

Proof. (i) follows from Schäl (1975), Lemma 3.4. However, since our proof is short and direct, we include it for completeness. We show first that $P^a\mathbf{V}(x)$ is an upper-semicontinuous function on $\mathbf{A}(x)$ for every $x \in \mathbf{X}$. Let $\mathbf{V}^N(x) = \max\{\mathbf{V}(x), -N\}$, $N \in \mathbb{N}$. The functions $\mathbf{V}^N$ are bounded (and continuous in the discrete topology on $\mathbf{X}$). Since the probability distributions $p(\cdot|x, a)$ are weakly continuous in $a \in \mathbf{A}(x)$, the functions $P^a\mathbf{V}^N(x)$ are continuous in $a \in \mathbf{A}(x)$. Since $\mathbf{V}^N(x) \searrow \mathbf{V}(x)$, by the monotone convergence theorem $P^a\mathbf{V}^N(x) \searrow P^a\mathbf{V}(x)$. Therefore, for each $x \in \mathbf{X}$ the function $T^a\mathbf{V}(x)$ is upper-semicontinuous in a; Royden (1968), p. 49. It therefore achieves its maximum $\mathbf{V}(x)$, and moreover the nonempty set $\mathbf{A}^c(x)$ is closed. Finally, it is compact since it is a closed subset of the compact set $\mathbf{A}(x)$.

Once (i) is established, propositions (ii) and (iii) follow from the fact that for any policy the expected discounted rewards at epoch n are bounded above by $\beta^n K$ where $r(y,a) \le K < \infty$ for all $y \in \mathbf{X}$ and for all $a \in \mathbf{A}(x)$. This is a particular case of summable above models considered in Hinderer (1970) (cf. Theorems 3.9 and 17.1) and in Dynkin and Yushkevich (1979) where statements similar to (ii) and (iii) are proved. Note that the proof of (ii) and (iii) is similar to the proof of Lemma 3.1 in Feinberg and Shwartz (1994). ∎

Consider now an $(M+1)$-criterion problem. For $C \subseteq \mathbb{R}^{M+1}$ denote by $\overline{C}$ the closure of C.

Lemma 4.3. *For each $u \in \overline{V(x,\Pi)}$ there exists $v \in V(x,\Pi)$ which dominates u.*

Proof. For each criterion V_m, $m = 0,\dots,M$, the MDP satisfies conditions (C), (S), and (W) from Schäl (1975); see also Balder (1989) for more general conditions. Therefore, one can define a topology on Π such that Π is compact and $V_m(x,\pi)$ are upper-semicontinuous in $\pi \in \Pi$. Let $u = (u_0,\dots,u_M) \in \overline{V(x,\Pi)}$. Then there exists a sequence of policies $\{\pi_n\}_{n=1}^{\infty}$ such that $V(x,\pi_n) \to u$ as $n \to \infty$. Since Π is compact, we can select a convergent subsequence $\{\pi_{n_k}\}$ with limit $\pi^* \in \Pi$. Then, by the semi-continuity of $V_m(x,\pi)$, we have that $V_m(x,\pi^*) \ge u_m$ for all $m = 0,\dots,M$. ∎

If some of the functions r_m are not continuous in a, then $V(x,\Pi)$ may not be closed. However, Theorem 4.5 below shows that the Pareto optimal points and the minimal extreme subsets containing them are the same for $V(x,\Pi)$ and its closure.

Lemma 4.4 (Corollary 3.4 in Feinberg and Shwartz 1995; see also Section 3.5 in Dynkin and Yushkevich 1979 and Theorem 7.11 in Schäl 1979). *$V(x,\Pi)$ is convex.*

Let $\Pi_> = \{\pi : V_m(\pi) > -\infty \text{ for all } m = 0,\dots,M\}$. For a Pareto optimal policy $\pi \in \Pi_>$, we denote by $G(x,\pi)$ the minimal extreme subset of $\overline{V(x,\Pi)}$ containing $V(x,\pi)$, and by $\tilde{G}(x,\pi)$ the minimal extreme subset of $V(x,\Pi)$ containing $V(x,\pi)$. To be consistent with the notation of Section 3, we note that if $E = \overline{V(x,\Pi)}$ then $G(x,\pi) = G(V(x,\pi))$ and if $E = V(x,\Pi)$ then $\tilde{G}(x,\pi) = G(V(x,\pi))$.

Theorem 4.5.

(i) $\mathrm{Par}\left(\overline{V(x,\Pi)}\right) = \mathrm{Par}(V(x,\Pi))$.

(ii) If $\pi \in \Pi_>$ is a Pareto optimal policy, then $G(x,\pi) = \tilde{G}(x,\pi)$.

Proof. (i) Let $w \in V(x,\Pi)$ be a Pareto optimal point in $V(x,\Pi)$. If there exists $u \in \overline{V(x,\Pi)}$ such that $u \neq w$ and u dominates w then, by Lemma 4.3, there exists $v \in V(x,\Pi)$ such that v dominates u. Therefore, v dominates w and $v \neq w$. This contradicts the Pareto optimality of w in $V(x,\Pi)$. So, if u is Pareto optimal in $V(x,\Pi)$, it is Pareto optimal in $\overline{V(x,\Pi)}$.

Let $u \in \overline{V(x,\Pi)}$ be Pareto optimal in $\overline{V(x,\Pi)}$. By Lemma 4.3, there exists $v \in V(x,\Pi)$ such that v dominates u. Since $V(x,\Pi) \subseteq \overline{V(x,\Pi)}$, $v = u$. Therefore, u is a Pareto optimal element of $V(x,\Pi)$, and (i) is established.

(ii) By Lemma 3.1 and Theorem 4.5 (i), $G(x,\pi) \subseteq \mathrm{Par}\left(\overline{V(x,\Pi)}\right) = \mathrm{Par}(V(x,\Pi)) \subseteq V(x,\Pi)$. Therefore, $G(x,\pi)$ is an extreme subset of $V(x,\Pi)$. This implies that $\tilde{G}(x,\pi) \subseteq G(x,\pi)$.

If $\tilde{G}(x,\pi) \neq G(x,\pi)$ then $\tilde{G}(x,\pi)$ is not an extreme subset of $\overline{V(x,\Pi)}$. Therefore there exists a $u_3 \in \tilde{G}(x,\pi)$ and $u_1, u_2 \in \overline{V(x,\Pi)}$ such that $u_3 = \alpha u_1 + (1-\alpha)u_2$ for some $0 < \alpha < 1$ and either $u_1 \notin \tilde{G}(x,\pi)$ or $u_2 \notin \tilde{G}(x,\pi)$. By our previous argument $u_3 \in \mathrm{Par}\left(\overline{V(x,\Pi)}\right)$ and so, since u_3 is a convex combination of u_1 and u_2 which are in $\overline{V(x,\Pi)}$, we have that $u_1, u_2 \in \mathrm{Par}\left(\overline{V(x,\Pi)}\right)$. By Theorem 4.5(i), this implies that u_1 and u_2 are in $V(x,\Pi)$. However, this is impossible because $\tilde{G}(x,\pi)$ is an extreme subset of $V(x,\Pi)$. ■

Due to Theorem 4.5 (ii) we shall not use the notation $\tilde{G}(x,\pi)$ any more. Recall that Π^B denotes the set of policies in a submodel B.

Lemma 4.6. *Let $\pi \in \Pi_>$ be a Pareto optimal policy. Let $E = V(x,\Pi)$ and consider the hyperplanes $P_1,\ldots,P_k$ whose existence is established in Lemma 3.2. Then $V(x,\Pi) \cap P_1 = V(x,\Pi^B)$ for some submodel B.*

Proof. Consider a one-criterion problem with the objective function $\sum_{i=0}^{M} b_{1i}V_i(x,\pi)$. This criterion is the expected total discounted reward for the reward function $\sum_{i=0}^{M} b_{1i}r_i(y,a)$, where $y \in \mathbf{X}$ and $a \in \mathbf{A}(x)$. Since $b_{1i} \geq 0$ for all $i = 0,\ldots M$, this reward function is bounded above and upper-semicontinuous in a for each y. By theorem 4.2, a policy σ is optimal at the initial state x for this criterion if and only if $a_n \in B(x_n)$ $\mathbb{P}_x^\sigma$-a.s. where $B(y) = \mathbf{A}^c(y) \neq \emptyset$, $y \in \mathbf{X}$. Since $\sup_\sigma \sum_{i=0}^{M} b_{1i}V_i(x,\sigma) = d_1$, we have that $V(x,\Pi^B) = V(x,\Pi) \cap P_1$. ■

Theorem 4.7. *For any Pareto optimal policy $\pi \in \Pi_>$, there exists a submodel B such that $G(x,\pi) = V(x,\Pi^B)$.*

Proof. Consider the hyperplanes $P_1, \ldots, P_k$ defined in Lemma 3.2 for $E = V(x, \Pi)$ and $u = V(x, \pi)$. By Lemma 4.6, there exists a submodel B_1 such that $V(x, \Pi^{B_1}) = V(x, \Pi) \cap P_1$. If $k > 1$, we consider the submodel B_1 instead of the original model. Then Lemma 4.6 implies that there exists a submodel B_2 such that $V(x, \Pi^{B_2}) = V(x, \Pi) \cap P_1 \cap P_2$. Repeat this argument k times and define $B = B_k$. Lemma 3.2 implies that $G(x, \pi) = V(x, \Pi^B)$. ■

Corollary 4.8. *Let $\pi \in \Pi_>$ be a Pareto optimal policy. If $V(x, \pi)$ is an extreme point of $\overline{V(x, \Pi)}$ then*

(i) $\{V(x, \pi)\} = V(x, \Pi^B)$ *for some submodel B,*

(ii) $V(x, \pi) = V(x, \phi)$ *for some stationary policy ϕ.*

Proof. In this case, $G(x, \pi) = \{V(x, \pi)\}$. Therefore, (i) follows from Theorem 4.7. Take any stationary policy $\phi \in \Pi^B$ to obtain (ii). ■

Corollary 4.9. *Let $\pi \in \Pi_>$ be a Pareto optimal policy. If u is an extreme point of $G(x, \pi)$ then there exists a stationary policy ϕ such that $V(x, \phi) = u$.*

Proof. By Theorem 4.7, $G(x, \pi) = V(x, \Pi^B)$ for some submodel B. We replace the original model with submodel B and apply Corollary 4.8 (ii). ■

Recall that the sets H_n, $n \in \mathbb{N}_0$ or $n = \infty$ and H are endowed with σ-fields $\mathcal{B}_n$ and $\mathcal{F}$. Let $\mathcal{F}_n$ be the minimal σ-field on H_∞ generated by the cylinder sets with bases in $\mathcal{B}_n$, $n \in \mathbb{N}_0$. Let τ be a stopping time with respect to the σ-fields $\mathcal{F}_n$, where $n \in \mathbb{N}$ and $\tau \le \infty$. For a policy π, $m = 0, \ldots, M$, and bounded above functions g_m on $\mathbf{X}$, we define

$$V_m(x, \pi, \tau, g_m) = \mathbb{E}_x^\pi \left\{ \sum_{j=0}^{\tau-1} \beta^j r_m(x_j, a_j) + \beta^\tau g_m(x_\tau) \right\}$$

and $V_m(x, \pi, \tau) = V_m(x, \pi, \tau, g_m)$ when $g_m(y) = 0$ for all $y \in \mathbf{X}$.

Lemma 4.10 (Feinberg and Sonin 1983, Lemma 3.5). *Let τ_n, $n \in \mathbb{N}_0$ be stopping times and $\tau_n \nearrow \infty$ as $n \to \infty$ for any trajectory $h_\infty \in H_\infty$. Then for $m = 0, \ldots, M$,*

$$V_m(x, \pi) = \lim_{n \to \infty} V_m(x, \pi, \tau_n).$$

For $Y \subseteq \mathbf{X}$, define the stopping time τ^Y as the first exit time from Y, i.e. $\tau^Y(h_\infty) = \min\{n \in \mathbb{N}_0 : x_n \notin Y\}$. Let $\tau_n^Y(h_\infty) = \min\{\tau^Y(h_\infty), n\}$.

Lemma 4.11. *Fix $l \in \mathbb{N}$ and let $\phi^1, \ldots, \phi^l$ be stationary policies with $\phi^i \in \Pi_>$ for all $i = 1, \ldots, l$. Let g_m be bounded above functions on $\mathbf{X}$, $m = 0, \ldots, M$. Then for any $\epsilon > 0$ there exist a number $N \in \mathbb{N}_0$ and a finite subset Y of $\mathbf{X}$ such that*

$$V_m(x, \phi^i) \geq V_m(x, \phi^i, \tau_N^Y, g_m) - \epsilon$$

for all $i = 1, \ldots, l$ and for all $m = 0, \ldots, M$.

Proof. Let $0 < K < \infty$ be an upper bound for g_m for all m, i.e. $g_m(x) \leq K$ for all $x \in \mathbf{X}$ and $m = 0, \ldots, M$. Let Y_n, $n \in \mathbb{N}_0$ be a sequence of finite subsets of $\mathbf{X}$ such that the starting point $x \in Y_0$, $Y_n \subset Y_{n+1}$, and $\mathbf{X} = \bigcup_{n=0}^{\infty} Y_n$. We denote $\tau_n = \tau_n^{Y_n}$. We observe that $\tau_n \nearrow \infty$ as $n \to \infty$. By the monotone convergence theorem, $\lim_{n\to\infty} \mathbb{E}_x^\pi \beta^{\tau_n} = 0$ for any policy π. From this equality and from Lemma 4.10 we have that for any $\epsilon > 0$ there exists a number $N \in \mathbb{N}_0$ such that

$$\mathbb{E}_x^{\phi^i} \beta^{\tau_N} g(x_{\tau_N}) \leq K\, \mathbb{E}_x^{\phi^i} \beta^{\tau_N} \leq \epsilon/2$$

and

$$V_m(x, \phi^i) \geq V_m(x, \phi^i, \tau_N) - \epsilon/2.$$

Therefore

$$V_m(x, \phi^i) \geq V_m(x, \phi^i, \tau_N) - \epsilon/2 \geq V_m(x, \phi^i, \tau_N) + \mathbb{E}_x^{\phi^i} \beta^{\tau_N} g(x_{\tau_N}) - \epsilon = V_m(x, \phi^i, \tau, g_m) - \epsilon.$$

■

For stationary policies ϕ, ψ and for a stopping time τ, we define a nonrandomized policy $\sigma = \sigma[\phi, \tau, \psi]$ by

$$\sigma(h_n) = \begin{cases} \phi(x_n), & \text{if } n < \tau; \\ \psi(x_n), & \text{if } n \geq \tau; \end{cases}$$

where $n \in \mathbb{N}_0$ and $h_n = x_0 a_0 \ldots x_n$. The policy σ prescribes the use of ϕ until epoch τ and the use of ψ thereon. We observe that $V_m(x, \sigma[\phi, \tau, \psi]) = V_m(x, \phi, \tau, V_m^\psi)$, where $V_m^\psi(y) = V_m(y, \psi)$ for all $y \in \mathbf{X}$ and all $m = 0, \ldots, M$.

Lemma 4.12. *Let $\pi \in \Pi_>$ be a Pareto optimal policy. Consider the submodel B of Lemma 4.7. Let ψ be a stationary policy in Π^B and $l \in \mathbb{N}$. Then for any stationary policies $\phi^1, \ldots, \phi^l$ in Π^B and for any $\epsilon > 0$ there exist a number $N \in \mathbb{N}_0$ and a finite subset Y of $\mathbf{X}$ such that*

$$||V(x, \phi^i) - V(x, \sigma[\phi^i, \tau_N^Y, \psi])|| \leq \epsilon, \qquad i = 1, \ldots, l.$$

Proof. By Lemma 3.2 (v) and Theorem 4.7, there exist positive numbers $b_0, \ldots, b_M$ such that

$$\sum_{m=0}^{M} b_m V_m(x,\pi) = \sum_{m=0}^{M} b_m V_m(x,\sigma) \tag{4.1}$$

for any $\pi, \sigma \in \Pi^B$. Let $\delta = \epsilon \cdot \min\{b_0, \ldots, b_K\}/(M \max\{b_0, \ldots, b_M\})$. By Lemma 4.11 there exists a finite subset Y of $\mathbf{X}$ and an integer N such that

$$V_K(x, \phi^i) \geq V_m(x, \sigma[\phi^i, \tau, \psi]) - \delta$$

for all $m = 0, \ldots, M$ and all $i = 1, \ldots, l$. This inequality and (4.1) imply Lemma 4.12. ■

Lemma 4.13 (Derman and Strauch 1966, Hordijk 1974, Theorem 13.2, van der Wal 1981, Lemma 2.1). *Let $\pi^1, \pi^2, \ldots$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, \ldots$ a sequence of nonnegative real numbers with $\sum\limits_{i=1}^{\infty} \lambda_i = 1$. Let σ be a randomized Markov policy defined by*

$$\sigma_n(C|y) = \frac{\sum\limits_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}(x_n = y, a_n \in C)}{\sum\limits_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}(x_n = y)}\,, \qquad n \in \mathbb{N}_0,\ y \in \mathbf{X},$$

for Borel subsets C of $\mathbf{A}(y)$, provided the denominator is positive. We let $\sigma_n(\cdot|y)$ be an arbitrary distribution on $\mathbf{A}(y)$ when the denominator is 0. Then, for any $n \in \mathbb{N}_0$, any $y \in \mathbf{X}$, and any Borel subset C of $\mathbf{A}$,

$$\mathbb{P}_x^{\sigma}(x_n = y, a_n \in C) = \sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}(x_n = y, a_n \in C),$$

and, therefore, $V_m(x,\sigma) = \sum\limits_{i=1}^{\infty} \lambda_i V_m(x, \pi^i)$ for $m = 0, \ldots, M$.

Lemma 4.14. *Let π be a Pareto optimal policy. Then there exists a submodel B such that for any stationary policy ψ in B there is a randomized Markov policy σ in B with the following properties:*

(i) $V(x,\sigma) = V(x,\pi)$,

(ii) *there exist an integer $N \in \mathbb{N}_0$ and a finite subset Y of $\mathbf{X}$, such that $\sigma_n(\{\psi(y)\}|y) = 1$ if either $n \geq N$ or $y \notin Y$,*

(iii) *for each $y \in \mathbf{X}$ and $n \in \mathbb{N}_0$, the distribution $\sigma_n(\cdot|y)$ is atomic and the set $\mathbf{A}_n^{\sigma}(y) = \{a \in \mathbf{A}(y) : \sigma_n(a|y) > 0\}$ is finite.*

Proof. If $V_m(x,\pi) = -\infty$ for all $m = 0, \ldots, M$ then all criteria equal $-\infty$ for any policy. So, any Markov policy σ which satisfies (ii) and (iii), satisfies the conditions of the lemma (in particular, we can take $\sigma = \psi$). If $V_m(x,\pi) = -\infty$ for some $m = 0,\ldots,M$ then $V_m(x,\pi') \geq V_m(x,\pi)$ for any policy π'. So, if we find σ such that (ii), (iii) hold and the vector relation (i) holds except for coordinate m, then by the Pareto optimality it must hold for m. Therefore, in this case the m^{th} criterion may be dropped. So, we assume without loss of generality that $\pi \in \Pi_>$.

Suppose $G(x,\pi)$ is a singleton, i.e. $G(x,\pi) = \{V(x,\pi)\}$. Then, by Corollary 4.8, there exists a submodel B so that $V(x,\pi) = V(x,\psi)$ for any stationary policy ψ in B, and so (i)–(iii) hold.

If $G(x,\pi) \neq \{V(x,\pi)\}$ then, by Lemma 3.4, $V(x,\pi)$ is a (relatively) internal point of $G(x,\pi)$. By Corollary 3.3, and Theorem 4.5 (i), $G(x,\pi)$ is convex and compact. Let $\dim G(x,\pi) = l \leq M$. Then there exists a polytope with dimension l such that $V(x,\pi)$ is an internal point of this polytope and its vertices $u^1,\ldots,u^k$ are extreme points of $G(x,\pi)$. By Theorem 4.7, $G(x,\pi) = V(x,\Pi^B)$ for some submodel B. By Corollary 4.9, there exist stationary policies ϕ^i in submodel B such that $u^i = V(x,\phi^i)$, $i = 1,\ldots,k$. Since $V(x,\pi)$ is an internal point of the polytope, there exists $\epsilon > 0$ such that if, for each i, we select any point $v^i \in G(x,\pi)$ in an ϵ-neighborhood of u^i then $V(x,\pi)$ belongs to the polytope with vertices $v^1,\ldots,v^k$. This implies that

$$V(x,\pi) = \sum_{i=1}^{k} \lambda_i v^i \tag{4.2}$$

for some nonnegative $\lambda_1,\ldots,\lambda_k$ with

$$\sum_{i=1}^{k} \lambda_i = 1. \tag{4.3}$$

Let ψ be a stationary policy in submodel B. By Lemma 4.12, there exist a finite set Y and $n \in \mathbb{N}_0$ such that $V(x,\sigma[\phi^i,\tau_N^Y,\psi])$ is in the ϵ-neighborhood of $V(x,\phi^i)$. Since $\sigma[\phi^i,\tau_N^Y,\psi] \in \Pi^B$, we have that $V(x,\sigma[\phi^i,\tau_N^Y,\psi]) \in G(x,\pi)$. So, from (4.2) we have that

$$V(x,\pi) = \sum_{i=1}^{k} \lambda_i V(x,\sigma[\phi^i,\tau_N^Y,\psi])$$

with nonnegative $\lambda_1,\ldots,\lambda_k$ satisfying (4.3). Let $\pi^i = \sigma[\phi^i,\tau_N^Y,\psi]$, $i = 1,\ldots,k$. Then the randomized Markov policy σ defined in Lemma 4.13 satisfies (i)–(iii). ■

Corollary 4.15. *Let π be a Pareto optimal policy. Then there exists a finite subset Y of $\mathbf{X}$ and a submodel B^* such that:*

(i) $B^*(y)$ is a singleton for any $y \in \mathbf{X} \setminus Y$,

(ii) $B^*(y)$ *is finite for any* $y \in Y$,

(iii) $V(x,\sigma) = V(x,\pi)$ *for some* $\sigma \in \Pi^{B^*}$.

Proof. Consider the policy σ constructed in Lemma 4.14. Define

$$B^*(y) = \{\psi(y)\} \bigcup \left(\bigcup_{n=1}^{N-1} \mathbf{A}_n^\sigma(y) \right), \quad y \in \mathbf{X},$$

where the sets $\mathbf{A}_n^\sigma(y)$ are defined in Lemma 4.14(iii). ■

For $Y \subseteq \mathbf{X}$, denote by $\xi(h_n)$ the amount of time spent in Y up to epoch n (not including epoch 0):

$$\xi(h_n) = \sum_{i=1}^{n} \mathbf{1}\{x_i \in Y\}, \quad h_n \in H_n,\ n \in \mathbb{N}.$$

Let τ^i be the epoch when the process x_n is in Y for the i^{th} time (if $x \in Y$, epoch 0 is not counted):

$$\tau^0(h_\infty) = 0,\ \tau^{i+1}(h_\infty) = \min\{n > \tau^i :\ x_n \in Y\} = \min\{n > 0 : \xi(h_n) = i+1\},\ h_\infty = h_n \ldots \in H_\infty.$$

Let $\theta(h_n) = n - \tau^{\xi(h_n)}(h_\infty[h_n])$ be the time elapsed since the last visit to Y, where $h_\infty[h_n]$ is any infinite trajectory that coincides with h_n up to epoch n.

A policy is called *randomized* Y*-embedded* if all decisions depend on three factors: (i) the current state x_n, (ii) the amount of time spent in Y prior to the current epoch, (iii) amount of time elapsed since the last visit to Y. In other words, a policy π is randomized Y-embedded if for any $y \in \mathbf{X}$ and $m, k \in \mathbb{N}_0$ there exists a distribution $\pi(\cdot|x,m,k)$ on $\mathbf{A}$ such that $\pi(\mathbf{A}(y)|x,m,k) = 1$ and $\pi(\cdot|h_n) = \pi(\cdot|x_n, \xi(h_n), \theta(h_n))$ for any $h_n = x_0 a_0 \ldots x_n \in H$; see Feinberg (1986) for details.

We observe that $\{x_n = y, \xi(h_n) = i, \theta(h_n) = j \text{ for some } n \in N_0\} = \{x_{\tau^i + j} = y, \tau^{i+1} > \tau^i + j\}$. For any policy π, define a randomized Y-embedded policy σ as follows: for any $y \in \mathbf{X}$, $i, j \in N_0$, and a Borel set $C \in \mathbf{A}$,

$$\sigma(C|y,i,j) = \begin{cases} \mathbb{P}_x^\pi(a_{\tau^i+j} \in C | x_{\tau^i+j} = y, \tau^{i+1} > \tau^i + j), & \text{if } \mathbb{P}_x^\pi(x_{\tau^i+j} = y, \tau^{i+1} > \tau^i + j) > 0; \\ \text{arbitrary probability distribution on } \mathbf{A}(y), & \text{otherwise.} \end{cases} \tag{4.4}$$

Lemma 4.16 (Feinberg 1986, Theorem 4.1). *Let* π *be an arbitrary policy,* Y *be a subset of* $\mathbf{X}$, *and* σ *be the randomized* Y*-embedded policy of (4.4). Then*

$$\mathbb{P}_x^\sigma(x_{\tau^i+j} = y, a_{\tau^i+j} \in C, \tau^{i+1} > \tau^i + j) = \mathbb{P}_x^\pi(x_{\tau^i+j} = y, a_{\tau^i+j} \in C, \tau^{i+1} > \tau^i + j),$$

for any $y \in \mathbf{X}$, $i, j \in \mathbb{N}_0$, *and for any Borel subset* C *of* $\mathbf{A}$, *and therefore* $V_m(x, \sigma) = V_m(x, \pi)$ *for all* $m = 0, \ldots, M$.

We remark that Lemma 4.16 is a generalization of Lemma 4.13. In the particular case $Y = \mathbf{X}$, Lemma 4.16 implies Lemma 4.13 for $\lambda_1 = 1$. The extension of Lemma 4.16 to a similar statement with arbitrary nonnegative λ_i whose sum is 1 is routine. For instance, we may take a policy π such that

$$\mathbb{P}_x^{\pi} = \sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i};$$

see Dynkin and Yushkevich (1979), Section 3.6. Then we apply Lemma 4.16 to the policy π.

Lemma 4.17. *If the sets* $\mathbf{X}$ *and* $\mathbf{A}$ *are finite then for any Pareto optimal policy* π *there exists an equivalent* M*-randomized stationary policy.*

Proof. We set $c_m = V_m(x, \pi)$, $m = 1, \ldots, M$, and consider problem (2.2)–(2.3). Then π is a feasible policy. By Theorem 4.2 in Feinberg and Shwartz (1995), there exists an M-randomized stationary optimal policy ϕ for this problem. Since $V_m(x, \phi) \geq V_m(x, \pi)$ for all $m = 0, \ldots, M$ and π is Pareto optimal, $V(x, \phi) = V(x, \pi)$ so that π and ϕ are equivalent. ■

Remark 4.18. Theorem 4.2 in Feinberg and Shwartz (1995) states the existence of M-randomized stationary optimal policies for discounted MDPs with finite state and action sets and with M constraints. Actually, the theorem states that the linear programming algorithm (Kallenberg 1983, Heyman and Sobel 1984) leads to an M-randomized optimal policy. A similar result holds for Semi-Markov Decision Processes (SMDPs). SMDPs are a generalization of MDPs when sojourn times are arbitrary nonnegative random variables with positive expectations; see Ross (1970), Kallenberg (1983), Heyman and Sobel (1984), or Feinberg (1994) for precise definitions.

It is well known that a discounted SMDP may be reduced to a discrete time MDP with discount factors depending on states and actions, $\beta = \beta(y, a)$. If the state and action sets are finite then $0 \leq \beta(y, a) \leq \beta^* $ for all $y \in \mathbf{X}$ and $a \in \mathbf{A}$ and for some $\beta^* < 1$; see Heyman and Sobel (1984) p. 202. In the original SMDP, a controller may use information about actual sojourn times. This information is not available in the equivalent MDP, but Theorem 4.3 and Corollary 4.4 in Feinberg (1994) show that for any policy π in the SMDP there exists a policy σ in the equivalent MDP such that $V_m(x, \sigma) = V_m(x, \pi)$ for all $m = 0, \ldots, M$. Therefore, a controller cannot benefit from the additional information available in SMDPs and SMDPs are indeed equivalent to MDPs with discount factors depending on states and actions.

MDPs with discount factors $\beta(y,a)$ depending on states and actions and with a bound on the discounts $\beta(y,a) \le \beta^* < 1$ for all $y \in \mathbf{X}$ and $a \in \mathbf{A}$ may be reduced to standard MDPs. There are two ways to verify this statement. One way is to check that all proofs for standard MDPs may be expanded to MDPs with variable discount factors. Another way is to embed the model in a larger one, as follows. Add an additional point x^* to the state space with a single control a^*, $\mathbf{A}(x^*) = \{a^*\}$ and define $r_m(x^*, a^*) = 0$ for all $m = 0, \ldots, M$. The transition probabilities p^* are now defined by $p^*(x^*|x^*, a^*) = 1$ and $p^*(y|z,a) = \beta(y,a)p(y|z,a)/\beta^*$ for $y \in \mathbf{X}$ and $p^*(x^*|z,a) = 1 - \sum_{y \in \mathbf{X}} p^*(y|z,a)$. So, we have transformed a SMDP to a standard discounted MDP with the discount factor β^* and with finite state and action sets. Therefore, if there exists a feasible policy for a SMDP with finite state and action sets, then there exists an M-randomized stationary optimal policy. The Linear Program (4.8)–(4.11) in Feinberg and Shwartz (1995), with β replaced with $\beta(u,a)$ and placed inside the double summation in formula (4.9) there, provides an LP for the computation of an optimal M-randomized stationary policy for a SMDP with finite state and action sets.

Remark 4.19. Lemma 4.17 and remark 4.18 imply that if the sets $\mathbf{X}$ and $\mathbf{A}$ are finite in a discounted SMDP with $M+1$ criteria then for any Pareto optimal policy there exists an equivalent M-randomized stationary policy.

Proof of of Theorem 4.1(i). First replace the original model with submodel B^* from Corollary 4.15. Next, we replace this MDP with a SMDP with the finite state space Y. For $z \in Y$, the set of available actions is $B^*(z)$, which is finite for all $z \in Y$. A transition probability from $z \in Y$ to $y \in Y$, if an action a is chosen at z in the SMDP is the joint probability that the process will hit Y again and that y is the point of new entrance (it is possible that this happens in the next step; it is also possible that the sum of these transition probabilities over all y in Y is less than 1 in this SMDPs, since sojourn times may be infinite with positive probability). Rewards are expected total discounted rewards until the next entrance to Y. More precisely, at $y \in \mathbf{X} \setminus Y$, the process is controlled by a stationary policy ϕ, where $\{\phi(y)\} = B^*(y)$ for $y \in \mathbf{X} \setminus Y$. If an action $a \in B^*(z)$ is chosen at $z \in Y$, the one-step rewards are $\mathbb{E}_z^\psi \sum_{n=0}^{\tau-1} \beta^n r_m(x_n, a_n)$, where ψ is any stationary policy such that $\psi(z) = a$ and $\psi(y) = \phi(y)$ for $y \in \mathbf{X} \setminus Y$, $\tau(h_\infty) = \min\{n \in \mathbb{N} : x_n \in Y\}$, and $m = 0, \ldots, M$. The transition probabilities are $\mathbb{P}_z^\psi(\tau < \infty,\ x_\tau = y)$, $y \in Y$.

Lemma 4.16 implies that, for any policy in submodel B^*, there exists an equivalent Y-embedded policy. From the definitions it follows that there is a one-to-one correspondence between randomized Y-embedded policies in the submodel B^* and randomized Markov policies in the SMDP with state space Y. Since all actions $B^*(y)$ are singletons for $y \in \mathbf{X} \setminus Y$, this correspondence provides a one-to-one correspondence between m-randomized stationary policies in submodel B^* and m-randomized stationary policies in the defined SMDP, for all $m \in \mathbb{N}$. If the problem is feasible, then by Remark 4.19 there exists an M-randomized stationary policy ϕ in the SMDP and such that $V(x, \phi) = V(x, \pi)$. Therefore, if we expand this policy from Y to $\mathbf{X}$ by setting $\phi(B^*(y)|y) = 1$ for $y \in \mathbf{X} \setminus Y$, we have that $\phi \in \Pi^{B^*}$ is equivalent to π. This policy ϕ is an M-randomized stationary policy in the original MDP. ■

Lemma 4.20. *Let π be a Pareto optimal policy. Then for some $N \in \mathbb{N}_0$ there exists an equivalent (M, N)-policy.*

Proof. Let σ be the policy of Lemma 4.14. We consider an MDP with the state space $(\mathbf{X} \times \{0, \ldots, N-1\}) \cup \mathbf{X}$. The action set is $\mathbf{A}$ and the initial state is $(x, 0)$. For states in $\mathbf{X}$, the sets of available actions, transition functions, and rewards coincide with the same objects in the original MDP. We also set $\mathbf{A}(y, n) = \mathbf{A}_n^\sigma(y)$ (see Lemma 4.14(iii)) and

$$p(z|y, n, a) = \begin{cases} p(z'|y, a), & \text{if } n < N-1 \text{ and } z = (z', n+1); \\ p(z|y, a), & \text{if } n = N-1 \text{ and } z \in \mathbf{X}; \\ 0, & \text{otherwise}; \end{cases}$$

and $r_m(y, n, a) = r_m(y, a)$.

Now σ defines a randomized stationary policy in the new model. Also, any m-randomized stationary policy in the new model corresponds to a (m, N)-policy in the original model and vice versa.

We apply Theorem 4.1(i) to the new model. Then in the new model there exists an M-randomized stationary policy ψ which is equivalent to σ and, therefore, to π. The policy that corresponds to ψ in the original model is an (M, N)-policy which is equivalent to π. ■

Proof of of Theorem 4.1(ii). By Theorem 4.1(i), there exists an M-randomized stationary policy ϕ such that $V(x, \phi) = V(x, \pi)$. Let $B(y) = \{a \in \mathbf{A}(y) : \phi(a|y) > 0\}$, $y \in \mathbf{X}$. We observe that ϕ is a Pareto optimal policy in submodel B. This submodel is an MDP with an index M or less. We apply Lemma 4.20 to the Pareto optimal policy ϕ in submodel B. Since the index of B is not greater than M, the equivalent (M, N)-policy is a strong (M, N)-policy. ■

Proof of Theorem 2.1. Let problem (2.2)–(2.3) be feasible. We show that there exists a Pareto optimal policy which is optimal for this problem. This fact and Theorem 4.1 imply the theorem.

Let $c_0 = \sup\{V_0(x,\pi) : \pi \in \Pi, V_m(x,\pi) \ge c_m, m = 1,\ldots,M\}$ and $V' = \overline{V(x,\Pi)} \cap \{u \in \mathbb{R}^{M+1} : u_m \ge c_m, m = 0,\ldots,M\}$. Since V' is closed and bounded above, $\mathrm{Par}(V') \ne \emptyset$. Since any point from V' cannot be dominated by any element of $\overline{V(x,\Pi)} \setminus \{u \in \mathbb{R}^{M+1} : u_m \ge c_m, m = 0,\ldots,M\}$, we have that $\mathrm{Par}(V') \subseteq \mathrm{Par}(\overline{V(x,\Pi)}) = \mathrm{Par}(V(x,\Pi))$, where the equality is Theorem 4.5(i). So there exists a Pareto optimal policy which is optimal for problem (2.2)–(2.3). ■

5. Optimality of M-Mixed Stationary Policies.

Let $m \in \mathbb{N}_0$. A policy σ is called m-mixed stationary if

$$\mathbb{P}_y^\sigma = \sum_{i=0}^{m} \lambda_i \mathbb{P}_y^{\phi^i}, \qquad y \in \mathbf{X}, \tag{5.1}$$

for some stationary policies $\phi^0,\ldots,\phi^m$ and for some nonnegative numbers $\lambda_0,\ldots,\lambda_m$ whose sum equals one.

For any policies $\phi^0,\ldots,\phi^m$ and any nonnegative coefficients $\lambda_0,\ldots,\lambda_m$ whose sum equals one, there exists a policy σ such that (5.1) holds; see Chapter 3.5 in Dynkin and Yushkevich (1979) or Theorem 7.11 in Schäl (1979). Therefore, the definition of m-mixed stationary policies is correct; see Feinberg (1994) for details.

In this section, we show that for any m-randomized stationary policy there exists an equivalent m-mixed stationary policy (Theorem 5.1). This result, combined with Theorem 4.1(i) and Theorem 2.1(i) respectively, implies the existence of M-mixed stationary policies which are Pareto optimal (Corollary 5.2) and optimal for problem (2.2)–(2.3) (Corollary 5.2).

The existence of 1-mixed stationary policies for a problem with one constraint was studieded by Sennott (1991; they were called Bernoulli modulated there). Makowski and Shwartz (1991) studied the existence and properties of optimal 1-mixed stationary policies for discounted and undiscounted MDPs with one constraint and with a special structure. When the action space is finite (and under recurrence assumptions), Theorem 5.1 in Altman and Shwartz (1991) establishes that the sets of state-action frequencies coincide for randomized stationary, general, and mixed stationary policies (in the case of average rewards per unit time, state action frequencies are the analog of the occupation measures for the total discounted reward). Feinberg (1994) proved the existence of M-mixed stationary policies in undiscounted SMDPs with finite state and action sets and with M constraints.

For a policy π we define occupation measures on $\mathbf{X} \times \mathbf{A}$:

$$Q_x^\pi(y, E) = \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \beta^t \mathbf{1}\{x_t = y, a_t \in E\},$$

where $y \in \mathbf{X}$ and E is a Borel subset of $\mathbf{A}$. Then for $m = 0, \ldots, M$

$$V_m(x, \pi) = \sum_{y \in \mathbf{X}} \int_{\mathbf{A}} r_m(y, a) Q_x^\pi(y, da).$$

When r_m are bounded, this equality follows from Fubini's theorem. For r_m bounded above, it follows from Fubini's theorem and monotone convergence; see the proof of Theorem 4.2 for details.

Theorem 5.1. *Let $m \in \mathbb{N}_0$. For any m-randomized stationary policy π there exists an m-mixed stationary policy σ such that $Q_x^\sigma(y, E) = Q_x^\pi(y, E)$ for any $y \in \mathbf{X}$ and for any measurable subset E of $\mathbf{A}$ and, therefore, the policies π and σ are equivalent.*

Proof. Suppose we show that

$$Q_x^\pi = \alpha Q_x^\phi + (1 - \alpha) Q_x^{\pi^*}, \tag{5.2}$$

where $\alpha \in [0, 1]$, ϕ is a stationary policy, and π^* is an $(m-1)$-randomized stationary policy. Then, by induction,

$$Q_x^\pi = \sum_{i=0}^{m} \lambda_i Q_x^{\phi^i},$$

where $\lambda_0 = \alpha$, $\phi^0 = \phi$, the policies $\phi^0, \ldots, \phi^m$ are stationary, and the constants $\lambda_0, \ldots, \lambda_m$ are nonnegative and sum to one. Then the m-mixed stationary policy σ defined by (5.1) satisfies $Q_x^\sigma = Q_x^\pi$.

So, it remains to prove (5.2). Let $\mathbf{A}^\pi(y) = \{a \in \mathbf{A}(y) : \ \pi(a|y) > 0\}$, $y \in \mathbf{X}$, and Y be a subset of $\mathbf{X}$ such that $\mathbf{A}^\pi(y)$ is not a singleton. This set is finite, and π is stationary outside of Y. Consider the stopping time $\tau = \min\{n > 0 : x_n \in Y\}$ which is the first time that the process enters Y after epoch 0. Given the initial state $y \in Y$, define

$$\mu(z, E|y, a) = \mathbb{E}_y^{\psi[a]} \sum_{n=1}^{\tau-1} \beta^n \mathbf{1}\{x_n = z, a_n \in E\}, \tag{5.3}$$

where $a \in \mathbf{A}(y)$, $z \in \mathbf{X} \setminus Y$, E is a measurable subset of $\mathbf{A}$, and $\psi = \psi[a]$ any stationary policy that prescribes action a at y, while outside Y it prescribes the same actions as the policy π. We also set for $y \notin Y$

$$\mu(z,E|y) = \mathbb{E}_y^{\psi} \sum_{n=0}^{\tau-1} \beta^n \mathbf{1}\{x_n = z, a_n \in E\}. \tag{5.4}$$

We observe that for any randomized stationary policy π' in the submodel $\mathbf{A}^\pi$, for any $z \in \mathbf{X}\setminus Y$, and for any measurable subset E of $\mathbf{A}$,

$$Q_x^{\pi'}(z,E) = \mathbf{1}\{x \notin Y\}\mu(z,E|x) + \sum_{y\in Y}\sum_{a\in\mathbf{A}^\pi(y)} \mu(z,E|y,a)Q_x^{\pi'}(y,a). \tag{5.5}$$

In order to prove (5.5), we set $\tau_0 = \min\{n \ge 0 : x_n \in Y\}$ and $\tau_{k+1} = \min\{n > \tau_k : x_n \in Y\}$, $k \in \mathbb{N}_0$. We also fix $z \in \mathbf{X}\setminus Y$, a measurable subset E of $\mathbf{A}$, and a randomized stationary policy π' in the submodel $\mathbf{A}^\pi$. Then

$$\begin{aligned}
&Q_x^{\pi'}(z,E)\\
&= \mathbb{E}_x^{\pi'} \sum_{n=0}^{\infty} \beta^n \mathbf{1}\{x_n = z, a_n = E\}\\
&= \mathbb{E}_x^{\pi'} \sum_{n=0}^{\tau_0-1} \beta^n \mathbf{1}\{x_n = z, a_n = E\} + \mathbb{E}_x^{\pi'} \sum_{k=0}^{\infty}\sum_{n=\tau_k}^{\tau_{k+1}-1} \beta^n \mathbf{1}\{x_n = z, a_n = E\} \qquad (5.6)\\
&= \mathbf{1}\{x \notin Y\}\,\mathbb{E}_x^{\psi} \sum_{n=0}^{\tau-1} \beta^n \mathbf{1}\{x_n = z, a_n \in E\} + \mathbb{E}_x^{\pi'} \sum_{k=0}^{\infty} \beta^{\tau_k} \sum_{n=\tau_k+1}^{\tau_{k+1}-1} \beta^{n-\tau_k} \mathbf{1}\{x_n = z, a_n = E\}\\
&= \mathbf{1}\{x \notin Y\}\mu(z,E|x) + \sum_{k=0}^{\infty} \mathbb{E}_x^{\pi'} \beta^{\tau_k} \sum_{n=\tau_k+1}^{\tau_{k+1}-1} \beta^{n-\tau_k} \mathbf{1}\{x_n = z, a_n = E\}.
\end{aligned}$$

The second equality follows from $x_{\tau_k} \in Y$ and $z \notin Y$. The last equality follows from (5.4). For every $k \in \mathbb{N}_0$

$$\begin{aligned}
&\mathbb{E}_x^{\pi'} \beta^{\tau_k} \sum_{n=\tau_k+1}^{\tau_{k+1}-1} \beta^{n-\tau_k} \mathbf{1}\{x_n = z, a_n = E\}\\
&\qquad = \mathbb{E}_x^{\pi'} \beta^{\tau_k}\, \mathbb{E}_{x_{\tau_k}}^{\psi[a_{\tau_k}]} \sum_{n=1}^{\tau-1} \beta^n \mathbf{1}\{x_n = z, a_n \in E\}\\
&\qquad = \mathbb{E}_x^{\pi'} \beta^{\tau_k} \mu(z,E|x_{\tau_k}, a_{\tau_k}) \qquad (5.7)\\
&\qquad = \mathbb{E}_x^{\pi'} \sum_{y\in Y}\sum_{a\in\mathbf{A}^\pi(y)} \beta^{\tau_k} \mathbf{1}\{x_{\tau_k} = y, a_{\tau_k} = a\}\mu(z,E|y,a)\\
&\qquad = \sum_{y\in Y}\sum_{a\in\mathbf{A}^\pi(y)} \mu(z,E|y,a)\,\mathbb{E}_x^{\pi'} \beta^{\tau_k} \mathbf{1}\{x_{\tau_k} = y, a_{\tau_k} = a\}.
\end{aligned}$$

The first equality in (5.7) follows from the fact that (x_n, a_n) is a homogeneous Markov chain under any randomized stationary policy. The second equality follows from (5.3). We observe that

$$\sum_{k=0}^{\infty} \sum_{y\in Y} \sum_{a\in \mathbf{A}^\pi(y)} \mu(z,E|y,a)\, \mathbb{E}_x^{\pi'} \beta^{\tau_k} \mathbf{1}\{x_{\tau_k} = y, a_{\tau_k} = a\}$$

$$= \sum_{y\in Y} \sum_{a\in \mathbf{A}^\pi(y)} \mu(z,E|y,a)\, \mathbb{E}_x^{\pi'} \sum_{k=0}^{\infty} \mathbf{1}\{x_{\tau_k} = y, a_{\tau_k} = a\} \tag{5.8}$$

$$= \sum_{y\in Y} \sum_{a\in \mathbf{A}^\pi(y)} \mu(z,E|y,a) Q_x^{\pi'}(y,a).$$

Formulas (5.6)–(5.8) imply (5.5).

Fix an arbitrary stationary policy ϕ such that $\phi(y) \in \mathbf{A}^\pi(y)$ for all $y \in \mathbf{X}$. The measures Q_x^π and Q_x^ϕ are atomic and are concentrated on the set $D = \{(y,a) \in \mathbf{X}\times\mathbf{A} :\ y \in \mathbf{X},\ a \in \mathbf{A}^\pi(y)\}$. If $Q_x^\phi = Q_x^\pi$ then (5.2) holds for any $\alpha \in [0,1]$ with $\pi^* = \phi$.

So, we consider the case $Q_x^\phi \neq Q_x^\pi$. We replace our model with the submodel $\mathbf{A}^\pi$. Given an initial state x, a nonnegative measure Q on $\mathbf{X}\times\mathbf{A}$ is an occupation measure for some randomized stationary policy π^* in this submodel if and only if $Q(D) = (1-\beta)^{-1}$ and for all $y \in \mathbf{X}$

$$\sum_{a\in \mathbf{A}^\pi(y)} Q(y,a) - \beta \sum_{z\in\mathbf{X}} \sum_{a\in \mathbf{A}^\pi(z)} p(y|z,a) Q(z,a) = \mathbf{1}\{y = x\}; \tag{5.9}$$

see Altman (1994), Theorem 5.1(1.1), (2), and (2.1). In addition, if Q is a nonnegative measure with $Q(D) = (1-\beta)^{-1}$ which satisfies (5.9) then $Q = Q_x^{\pi^*}$ where

$$\pi^*(a|y) = \begin{cases} 1, & \text{if } \mathbf{A}^\pi(y) = \{a\}; \\ Q(y,a)\left[\sum_{a'\in\mathbf{A}^\pi(y)} Q(y,a')\right]^{-1}, & \text{if the denominator is positive;} \\ \text{arbitrary distribution on } \mathbf{A}^\pi(y), & \text{otherwise.} \end{cases} \tag{5.10}$$

Let $\alpha = \sup\{c : cQ_x^\phi(y,a) \le Q_x^\pi(y,a) \text{ for all } y \in \mathbf{X},\ a \in \mathbf{A}^\pi(y)\}$. Because of (5.5),

$$\alpha = \sup\{c : cQ_x^\phi(y,a) \le Q_x^\pi(y,a) \text{ for all } y \in Y,\ a \in \mathbf{A}^\pi(y)\}.$$

We observe that $Q_x^\phi \neq Q_x^\pi$, and that if $Q_x^\pi(y,a) = 0$ for some y and a then $Q_x^\phi(y,a) = 0$. Since the sets Y and $\mathbf{A}^\pi(y)$, $y \in Y$ are finite, the supremum α is positive and is attained at some y and a. Since $Q_x^\phi(\mathbf{X}\times\mathbf{A}) = Q_x^\pi(\mathbf{X}\times\mathbf{A}) = (1-\beta)^{-1}$, we have $\alpha \in (0,1)$. Let $Q = (Q_x^\pi - \alpha Q_x^\phi)/(1-\alpha)$. The definition of α implies that if $Q_x^\pi(y,a)=0$ then $Q(y,a)=0$ and, since the supremum is attained, $Q(y,a)=0$ for some $x \in Y$ and $a \in \mathbf{A}^\pi(y)$ such that $Q_x^\pi(y,a)>0$. We observe that Q satisfies (5.9). Therefore, $Q = Q_x^{\pi^*}$ for the $(k-1)$-randomized stationary policy π^* defined by (5.10) and (5.2) follows. ■

Theorem 4.1(i) implies Corollary 5.2 and Theorem 2.1(i) implies Corollary 5.3 below.

Corollary 5.2. *For any Pareto optimal policy there exists an equivalent M-mixed stationary policy.*

Corollary 5.3. *If problem (2.2)–(2.3) is feasible then there exists an optimal M-mixed stationary policy.*

The methods developed in the proof of Theorem 5.1 allow us to prove the following result.

Corollary 5.4. *Let π be an m-randomized stationary Pareto optimal policy for some $m \in \mathbb{N}_0$. Then any policy $\sigma \in \Pi^{\mathbf{A}^\pi}$ is Pareto optimal.*

Proof. By Lemma 3.1 in Borkar (1988), for any policy σ there exists an equivalent randomized stationary policy. We apply this to submodel $\mathbf{A}^\pi$ and to a policy σ in this submodel. So, without loss of generality we assume that σ is a randomized stationary policy.

The randomized stationary policies π and σ define homogeneous Markov chains on $\mathbf{X}$ with the same initial state x. In addition, if the one-step transition probability from $z \in \mathbf{X}$ to $y \in \mathbf{X}$ is positive for the Markov chain defined by σ, then it is positive for the Markov chain defined by π. This follows since the sets $\mathbf{A}^\pi(y)$, $y \in \mathbf{X}$ are discrete (in fact, finite) and $\pi(a|y)>0$ for any $a \in \mathbf{A}^\pi(y)$. This implies that if $Q_x^\sigma(y,a)>0$ for some $y \in \mathbf{X}$, $a \in \mathbf{A}^\pi(y)$, then $Q_x^\pi(y,a)>0$.

The last implication and (5.5) applied to $\pi' = \pi$ and to $\pi' = \sigma$, and to the finite set $Y = \{y \in \mathbf{X} : |\mathbf{A}^\pi(y)| > 1\}$ implies that there exists α, $0<\alpha<1$, such that $Q_x^\pi(y,C) \geq \alpha Q_x^\sigma(y,C)$ for any $y \in \mathbf{X}$ and for any Borel subset C of $\mathbf{A}$. Set $Q_x(\cdot) = (1-\alpha)^{-1}\left(Q_x^\pi(\cdot) - \alpha Q_x^\sigma(\cdot)\right)$. Then Q_x is a nonnegative measure on the set D defined below (5.5), and $Q_x(D) = (1-\beta)$. Equation (5.9) and the results from Altman (1994) quoted below (5.9) imply that there exists a randomized stationary policy σ' such that $Q_x^{\sigma'} = Q_x$.

We have that $Q_x^\pi = \alpha Q_x^\sigma + (1-\alpha) Q_x^{\sigma'}$. This implies that $V_m(x,\pi) = \alpha V_m(x,\sigma) + (1-\alpha) V_m(x,\sigma')$ for $m = 0,\ldots,M$. Since π is Pareto optimal, σ is Pareto optimal. ■

6. Computation of Optimal Strong (M, N)-Policies.

When the state and action sets are finite, optimal M-randomized stationary policies can be computed by a linear programming (LP) algorithm described in Kallenberg (1983), Heyman and Sobel (1984), and Feinberg and Shwartz (1995). Once such a policy is found, the proof of Lemma 6.2 provides an algorithm that computes an optimal M-mixed stationary policy. In this section, we formulate an algorithm for computing optimal strong (M, N)-policies for an MDP with finite state and action sets. The algorithm consists of three major steps and one intermediate step.

Step 1. Compute an M-randomized stationary optimal policy.

Remark 6.1. Such a policy exists if and only if problem (2.2)–(2.3) is feasible; Theorem 2.1(i). When $\mathbf{X}$ and $\mathbf{A}$ are finite, the LP algorithm that realizes Step 1 was proposed by Kallenberg (1983); see also Heyman and Sobel (1984), and Feinberg and Shwartz (1995). In order to find such a policy, one solves the LP described in Kallenberg (1983), Heyman and Sobel (1984), and Feinberg and Shwartz (1995). The solution of this LP defines an optimal M-randomized policy ψ and the value of this LP is $V_0(x, \psi)$; see Theorem 4.2 in Feinberg and Shwartz (1995) for details.

Step 2. Set $c_0 = V_0(x, \psi)$, where ψ is an optimal M-randomized stationary policy found in Step 1 (it is also the value of problem (2.2)–(2.3) and the value of the corresponding LP that should be solved at Step 1). Replace the action sets $\mathbf{A}(y)$ with the subsets $\mathbf{A}^{\psi}(y)$ of actions that the policy ψ uses at $y \in \mathbf{X}$. Use the LP algorithm described in Theorem 4.2 in Feinberg and Shwartz (1995) and compute an optimal randomized stationary policy σ for the following problem in submodel $\mathbf{A}^{\psi}$:

$$\text{maximize} \quad \sum_{m=0}^{M} \alpha_m V_m(x, \pi) \tag{6.1}$$

$$\text{subject to} \quad V_m(x, \pi) \geq c_m, \qquad m = 0, \ldots, M, \tag{6.2}$$

where α_m, $m = 0, \ldots, M$ are positive constants; for instance $\alpha_m = 1$, $m = 0, \ldots, M$.

Remark 6.2. Problem (6.1)–(6.2) is an optimization problem of discounted rewards under $M + 1$ constraints. Therefore, the LP algorithm mentioned in Step 1 is also applicable to Step 2. An optimal policy for problem (6.1)–(6.2) is Pareto optimal in submodel $\mathbf{A}^{\psi}$. The constraints c_m, $k = 1, \ldots, M$, in (6.2) may be substituted with $V_m(x, \psi)$. Let σ be a randomized stationary policy computed at Step 2.

Intermediate Step. If $V(x, \sigma) \neq V(x, \psi)$ then redefine ψ to be $\psi = \sigma$. Select any stationary policy ϕ such that $\phi(y) \in \mathbf{A}^{\psi}(y)$ for all $y \in \mathbf{X}$, where $\mathbf{A}^{\psi}(y) = \{a \in \mathbf{A}(y) : \ \psi(a|y) > 0\}$, $y \in \mathbf{X}$. Select any $N \in \mathbb{N}_0$.

Remark 6.3. As noted above, the LP algorithm that finds an optimal policy for discounted constrained problems (Kallenberg 1983, Heyman and Sobel 1984, Feinberg and Shwartz 1995) computes $V_m(x,\sigma)$, $m = 0,\ldots,M$, first and then it computes σ. Therefore, if $V(x,\sigma) = V(x,\psi)$, there is no need to compute σ.

Remark 6.4. Policy ψ is either an M-randomized policy computed at Step 1 or $V(x,\psi) \neq V(x,\sigma)$ and $\psi = \sigma$, where σ is a randomized stationary optimal policy computed at Step 3. If $V(x,\psi) \neq V(x,\sigma)$ then one of the constraints in (6.2) is not tight. This implies that the resulting policy σ is M-randomized stationary. Therefore, ψ is an M-randomized stationary policy after the Intermediate Step.

Step 3. Compute an optimal Markov policy π^* of order M for the following finite horizon problem in the submodel $\mathbf{A}^\psi$:

$$\text{maximize} \quad \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{N-1} \beta^n r_0(x_n,a_n) + \beta^N V_0(x_N,\phi) \right\} \tag{6.3}$$

$$\text{subject to} \quad \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{N-1} \beta^n r_m(x_n,a_n) + \beta^N V_m(x_N,\phi) \right\} \geq c_m, \quad m = 1,\ldots,M. \tag{6.4}$$

Then the objective function is $V_0(x,\pi^*)$, where the policy π^* at epoch $n = 0,\ldots,N-1$ is computed at Step 3 and $\pi_n^*(\phi(y)|y) = 1$ for $n \geq N$ and $y \in \mathbf{X}$. If $V_0(x,\pi^*) = V_0(x,\psi)$ then π^* is an optimal strong (M,N)-policy. If $V_0(x,\pi^*) < V_0(x,\psi)$ then increase N and repeat Step 3.

Remark 6.5. Algorithm 4.3 in Feinberg and Shwartz (1995) computes an optimal randomized Markov policy of order M for finite horizon models with M constraints (similar algorithms were formulated in Derman and Klein 1965 and Kallenberg 1981). Similar to infinite horizon discounted rewards, this algorithm computes $V_0(x,\pi^*)$ first and then computes the policy π^* at epochs $0,\ldots,N-1$. Therefore, if $V_0(x,\pi^*) < V_0(x,\psi)$ then there is no need to compute π^*. Step 3 computes an optimal policy for N sufficiently large. If we do not replace the original MDP with submodel $\mathbf{A}^\psi$ at steps 2 and 3, then the dimension of the corresponding LPs a solution of (2.1) and (2.2) which is a Pareto optimal at steps 2 and 3 may increase significantly. If we do not replace the original MDP with submodel $\mathbf{A}^\psi$ at step 2, the algorithm computes a Pareto optimal policy.

Remark 6.6. The algorithm converges for the following reasons. A policy which is optimal in submodel $\mathbf{A}^\psi$ is optimal, where $\mathbf{A}^\psi$ is the randomized optimal policy found in Step 1. After the intermediate step, ψ is optimal for problem (2.2)–(2.3) and Pareto optimal in submodel $\mathbf{A}^\psi$. Theorem 5.4 implies that ϕ is Pareto optimal in submodel $\mathbf{A}^\psi$ and arguments as in the proof of

Lemma 4.14 imply that, for N large enough, there exists an optimal (N, M)-policy with "tail" ϕ. Step 3 computes this policy. Theorem 5.4 implies that an optimal solution found by this algorithm is Pareto optimal in submodel $\mathbf{A}^{\psi}$ for the $(M+1)$-dimensional problem.

Remark 6.7. The algorithm computes an optimal strong (M, N)-policy in a finite number of steps. An algorithm that computes an optimal (M, N)-policy for a problem with weighted discounted rewards is not available. Algorithm 8.1 in Feinberg and Shwartz (1995) computes just an approximate solution for this problem. The difficulty in computing an optimal (M, N)-policy for a weighted discounted problem is that it is not clear how to find a "tail" stationary policy ϕ. In the case of standard discounting, Steps 1, 2 and the Intermediate Step find this policy.

The algorithm described above provides a general method for the computation of optimal strong (M, N)-policies for models with countable state spaces. However, in this case, the related LPs at steps $1-3$ become infinite dimensional. Under some additional assumptions, approximation methods for the solution of such LPs were studied by Altman (1993, 1994).

APPENDIX

In this section we provide proofs for the results of Section 3.

Proof of Lemma 3.1. Fix u in E and let

$$D(u) = \{v \in E : \alpha v + (1-\alpha)w = u \text{ for some } w \in E, \alpha \in (0,1)\}.$$

We notice that $u \in D(u)$. We show that $D(u)$ is an extreme subset of E.

Let $v \in D(u)$. We need to show that if $v = \alpha_1 v^1 + (1-\alpha_1)v^2$ for some $v^1, v^2 \in E$ and $\alpha_1 \in (0,1)$ then $v^1, v^2 \in D(u)$. Since $v \in D(u)$, we have $u = \alpha v + (1-\alpha)w$ for some $w \in E$, $\alpha \in (0,1)$. Thus

$$u = \alpha\alpha_1 v^1 + \alpha(1-\alpha_1)v^2 + (1-\alpha)w = \alpha\alpha_1 v^1 + (1-\alpha\alpha_1)\tilde{w},$$

where $\tilde{w} = \lambda v^2 + (1-\lambda)w$, $\lambda = \alpha(1-\alpha_1)/(1-\alpha\alpha_1)$, and $0 < \alpha\alpha_1 < 1$, $0 < \lambda < 1$. Since E is convex, $\tilde{w} \in E$ so that $v^1 \in D(u)$. Similarly, $v^2 \in D(u)$, and we have established that $D(u)$ is an extreme subset that contains u. By definition, $G(u) \subseteq D(u)$ (in fact, it is possible to show that $D(u) = G(u)$ but we do not need the equality here).

Fix a Pareto optimal u in E and any v in $D(u)$. Then $u = \alpha v + (1-\alpha)w$ for some w in E. If $v' \in E$ dominates v then $u' = \alpha v' + (1-\alpha)w \in E$ dominates u, which is a contradiction. Therefore $v \in \mathrm{Par}(E)$ and $G(u) \subseteq D(u) \subseteq \mathrm{Par}(E)$. ■

Proof of Lemma 3.2. Let $E_1 \subseteq \mathbb{R}^{M+1}$ be the set of points dominated by elements of E,

$$E_1 = \left\{w = (w_0, \ldots, w_M) \in \mathbb{R}^{M+1} : w_i \le v_i,\ i = 0, \ldots, M \text{ for some } v = (v_0, \ldots, v_M) \in E\right\}.$$

Then $\text{Par}(E_1) = \text{Par}(E)$ and $G(u)$ is also the minimal extreme subset of E_1 that contains u. Let w be a nonzero vector with nonnegative coordinates. If $v \in G(u)$ then $(v - w) \in E_1$ and $(v - w) \notin \text{Par}(E_1)$. By Lemma 3.1, $(v - w) \notin G(u)$. Therefore, $G(u)$ is a proper subset of E_1. By Proposition 3.6.5 in Stoer and Witzgall (1970), the proper extreme subset $G(u)$ of E_1 is contained in a proper exposed subset $\tilde{E}$ of E_1. Let $\tilde{E} = E_1 \cap P_1$, where P_1 is a supporting hyperplane of E_1, defined through

$$P_1 = \left\{ (v_0, \ldots, v_M) : \sum_{i=0}^{M} b_{1i} v_i = d_1 \right\} \text{ and } \sum_{i=0}^{M} b_{1i} v_i \leq d_1 \text{ for all } (v_0, \ldots, v_M) \in E_1.$$

Since $G(u) \in P_1$, we have that $\sum\limits_{i=0}^{M} b_{1i} u_i = d_1$. Since $E \subseteq E_1$, the hyperplane P_1 supports E at u. Note that $\dim E_1 = M + 1$ and $\dim \tilde{E} \leq M$.

Since $G(u) \subseteq E \subseteq E_1$, it follows that if $G(u) = E_1 \cap \left(\bigcap\limits_{j=1}^{k} P_j \right)$ then (i) holds. Let the hyperplanes $P_1, \ldots, P_j$ be given for some $j \in \mathbb{N}$. Since $u \in E$ and $E \subseteq E_1$, if a hyperplane P_j supports $E_1 \cap (\bigcap_{l=1}^{j-1} P_l)$ at u then it supports E_j at u. Therefore, if we construct $P_1, \ldots, P_k$ so that (i)–(v) hold when E is replaced with E_1 in (i)–(iii), this will prove the lemma. So, without loss of generality we prove Lemma 3.2 for a particular case when $E = E_1$, so that $\tilde{E} = E_2$ as defined in (iii).

The plane P_1 is already constructed. We observe that $b_{1i} \geq 0$ for all $i = 0, \ldots, M$. Indeed, suppose $b_{1i} < 0$ for some $i = 0, \ldots, M$. Then for $u' = u - e_i \in E_1$ we have $\sum\limits_{l=0}^{M} b_{1l} u'_l = d_1 - b_i > d_1$, in contradiction to the definition of P_1.

We continue by induction. Suppose that, for $1 \leq n \leq M$, the hyperplanes $P_1, \ldots, P_n$ are given, properties (ii)–(iv) hold for $k = n$ and, in addition, the following two properties hold for $k = n$:

(vi) $G(u) \subseteq E_1 \cap \left(\bigcap\limits_{j=1}^{k} P_j \right)$,

(vii) $\dim E_k \leq M - k + 2$, where the set E_k is defined in (iii).

Recall that (ii)–(iv) and (vi)–(vii) were established for $k = 1$. We show that either $E_n = G(u)$ or there exists a hyperplane P_n such that properties (ii)–(iv), (vi), and (vii) hold for $k = n + 1$. If $E_n = G(u)$ then properties (i)–(iv) hold.

If $E_n \neq G(u)$ then $G(u)$ is a proper extreme subset of E_n. Therefore, there exists a proper exposed subset E_{n+1} of E_n that contains $G(u)$. Let $E_{n+1} = E_n \cap P_n$, where P_n is a hyperplane that supports E_n, with

$$P_n = \left\{ (v_0, \ldots, v_M) : \sum_{i=0}^{M} b_{ni} v_i = d_n \right\} \text{ and } \sum_{i=0}^{M} b_{ni} v_i \le d_n \text{ for all } (v_0, \ldots, v_M) \in E_n.$$

Since $G(u) \in P_n$, we have that $\sum_{i=0}^{M} b_{ni} u_i = d_n$. By Proposition 3.6.3 in Stoer and Witzgall (1970), $\dim E_{n+1} < \dim E_n$. So, $\dim E_{n+1} \le M - n + 1$. Thus properties (iii), (vi), and (vii) hold when $k = n+1$.

We show that P_n can be selected in a way that property (iv) holds for $k = n+1$, i.e. $b_{ni} \ge 0$ for all $i = 0, \ldots, M$. Define

$$I_+ = \{ i : \ b_{ji} > 0 \text{ for some } j = 0, \ldots, n-1 \}$$

$$I_0 = \{ i : \ b_{ji} = 0 \text{ for all } j = 0, \ldots, n-1 \}.$$

Since (iv) holds for $k = n$, we have $I_+ \cup I_0 = \{0, \ldots, M\}$. Note that $b_{ni} \ge 0$ for any $i \in I_0$. Indeed, let $b_{ni} < 0$ for some $i \in I_0$, and consider $u' = u - e_i$. Since $\sum_{i=0}^{M} b_{ji} u_i = d_j$ and $b_{ji} = 0$ for all $j = 1, \ldots, n-1$, we have $u' \in E_n$. However, $\sum_{i=0}^{M} b_{ni} u'_i = d_n - b_{ni} > d_n$. This contradicts the definition of P_n.

Next we show that P_n can be selected so that $b_{ni} > 0$ for all $i \in I_+$. If $b_{ni} \le 0$ for some $i \in I_+$, we set, for $\epsilon > 0$ and $l = 0, \ldots, M$,

$$b'_{nl} = \sum_{j=1}^{n-1} b_{jl} + \epsilon b_{nl}, \quad d'_n = \sum_{j=1}^{n-1} d_j + \epsilon d_n$$

and $b'_{ni} = \epsilon b_{ni} \ge 0$ for $i \in I_0$. Choose $\epsilon > 0$ small enough and then $b'_{nl} > 0$ for all $l \in I_+$. Let

$$P'_n = \left\{ (v_0, \ldots, v_M) : \sum_{i=0}^{M} b'_{ni} v_i = d'_n \right\}.$$

Then by definition of b_{ni} and P_n

$$\sum_{i=0}^{M} b'_{ni} v_i \le d'_i \text{ for } (v_0, \ldots, v_M) \in E_n \quad \text{and} \quad \bigcap_{j=0}^{n} P_j = \left(\bigcap_{j=0}^{n-1} P_j \right) \cap P'_n.$$

Therefore, P_n' supports E_n and $E_{n+1} = E_{n+1} \cap P_n'$. Replacing P_n with P_n' we see that properties (iii), (iv), (vi), and (vii) hold for $j = k+1$, and the induction step is complete. The convexity of E_k together with (vii) imply that $E_k = G(u)$ for some $1 \le k \le M+1$.

Finally, to establish (v) suppose $b_{ki} = 0$ for some i and note that by our construction of P_n', this implies that $b_{ji} = 0$ for all $j = 1, \ldots, k$. But then $u' = u - e_i \in E_k = G(u)$, so that u' is not Pareto optimal. This contradicts Lemma 3.1. ■

Proof of Corollary 3.3. Convexity and closeness of $G(u)$ follows from Lemma 3.2 (i). Let v be an element of $\mathbb{R}^{M+1}$ that dominates all elements of E. The set D of all points in $\mathbb{R}^{M+1}$ dominated by v is a cone and $E \subseteq D$. Lemma 3.2 (v) implies that $D \cap P_k$ is a bounded set. Since $G(u) \subseteq E \cap P_k \subseteq D \cap P_k$, we have that $G(u)$ is bounded. ■

Proof of Lemma 3.4. Suppose $G(u) \neq \{u\}$ and u belongs to the boundary of $G(u)$. Since by Corollary 3.3 $G(u)$ is convex and compact, u belongs to a proper extreme subset D of $G(u)$; Proposition 3.6.7 in Stoer and Witzgall (1970). The existence of D contradicts the definition of $G(u)$ as a minimal extreme subset of E which contains u. ■

Acknowledgments. The authors are indebted to an anonymous referee and to Professor William Sudderth for enlightening comments and in particular for pointing out some connections to Statistical Decision Theory.

Research of the first author was partially supported by NSF Grant DMI-9500746.

Research of the second author was supported in part by the *Fund for promotion of research at the Technion*, and in part by the *Israel Science Foundation*, administered by the Israel Academy of Sciences and Humanities.

References

Altman, E. (1993). Asymptotic properties of constrained Markov decision processes. *ZOR—Methods and Models of Oper. Res.* **37** 151–170.

______ (1994). Denumerable constrained Markov decision problems and finite approximations. *Math. Oper. Res.* **19** 169–191.

______ and Shwartz, A. (1991). Markov decision models and state-action frequencies. *SIAM J. Control Optimization* **29** 786–809.

Balder, E.J. (1989). On compactness of the space of policies in stochastic dynamic programming. *Stochastic Processes Appl.* **32** 141–150.

Blackwell, D. and Girshick, M.A. (1954). *Theory of Games and Statistical Decisions.* Wiley, NY.

Borkar, V.S. (1988). A convex analytic approach to Markov decision processes. *Probab. Th. Rel. Fields* **78** 583–602.

Derman, C. and Klein, M. (1965). Some remarks on finite horizon Markovian decision models. *Oper. Res.* **13** 272–278.

Derman, C. and Strauch, R.E. (1966). A note on memoryless rules for controlling sequential processes. *Ann. Math. Stat.* **37** 276–278.

Dynkin, A.A. and Yushkevich, A.A. (1979). *Controlled Markov Processes.* Springer-Verlag, NY.

Feinberg, E.A. (1986). Sufficient classes of strategies in discrete dynamic programming I: decomposition of randomized strategies and embedded models. *Theory Prob. Appl.* **31** 658–668.

______ (1994). Constrained semi-Markov decision processes with average rewards. *ZOR—Methods and Models of Oper. Res.* **39** 257–288.

______ and Shwartz, A. (1994). Markov decision models with weighted discounted criteria. *Math. Oper. Res.* **19** 152–168.

______ and ______ (1995). Constrained Markov decision models with weighted discounted rewards. *Math. Oper. Res.* **20** 302–320.

______ and Sonin, I.M. (1983). Stationary and Markov policies in countable state dynamic programming. In *Lecture Notes in Math.* **1021** 111–129, Springer Verlag, NY.

Ferguson, T. S. (1967). *Mathematical Statistics.* Academic Press, New York.

Frid, E.B. (1972). On optimal strategies in control problems with constraints. *Theory Prob. Appl.* **17** 188–192.

Furukawa, M. (1980). Characterization of optimal policies in vector-valued Markovian decision processes. *Math. Oper. Res.* **5** 271–279.

Henig, M.I. (1983). Vector-valued dynamic programming. *SIAM J. Control Opt.* **21** 490–499.

Henig, M.I. (1985). The principle of optimality in dynamic programming with returns in partially ordered sets. *Math. Oper. Res.* **10** 462–471.

Heyman, D.P. and Sobel, M.J. (1984). *Stochastic Models in Operations Research. Volume II: Stochastic Optimization.* McGraw-Hill, NY.

Hinderer, K. (1970). *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter.* Springer-Verlag, Berlin.

Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tracts **51**, Mathematisch Centrum, Amsterdam.

Kallenberg, L.C.M. (1981). *Unconstrained and constrained dynamic programming over a finite horizon.* Report 81-46, Inst. Appl. Math. and Computer Sci., University of Leiden, The Netherlands.

______ (1983). *Linear Programming and Finite Markovian Control Problems.* Mathematical Centre Tracts **148**, Mathematisch Centrum, Amsterdam.

Le Cam , L. (1955). An extension of Wald's theory of statistical decision functions. *Ann. Math. Stat.* **26** 69–81.

Makowski, A.M. and Shwartz, A. (1993). On constrained optimization of the Klimov network and related Markov decision processes. *IEEE Trans. Auto. Control* **38** 354–359.

Royden, H.L. (1968). *Real Analysis.* Second Edition, Macmillan, NY.

Ross, S.M. (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, S.F.

Sennott, L.I. (1991). Constrained discounted Markov decision chains. *Probability in the Engineering and Informational Sciences* **5** 463–476.

Schäl, M. (1975). On dynamic programming: compactness of the space of policies. *Stochastic Processes Appl.* **3** 345–364.

Schäl, M. (1979). On dynamic programming and statistical decision theory. *Ann. Stat.* **7** 432–445.

Stoer, J. and Witzgall, C. (1970). *Convexity and Optimization in Finite Dimensions I,* Springer-Verlag, New York.

van der Wal, J. (1981). *Stochastic Dynamic Programming.* Mathematical Centre Tracts **139**, Mathematisch Centrum, Amsterdam.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

FEINBERG: 303 HARRIMAN HALL, SUNY AT STONY BROOK, STONY BROOK, NY 11794-3775

SHWARTZ: DEPARTMENT OF ELECTRICAL ENGINEERING, TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL