CONSTRAINED MARKOV DECISION MODELS WITH WEIGHTED DISCOUNTED REWARDS

EUGENE A. FEINBERG

SUNY at Stony Brook

ADAM SHWARTZ

Technion—Israel Institute of Technology

December 1992

Revised: August 1993

Abstract

This paper deals with constrained optimization of Markov Decision Processes. Both objective function and constraints are sums of standard discounted rewards, but each with a different discount factor. Such models arise, e.g. in production and in applications involving multiple time scales. We prove that if a feasible policy exists, then there exists an optimal policy which is (i) stationary (nonrandomized) from some step onward, (ii) randomized Markov before this step, but the total number of actions which are added by randomization is bounded by the number of constraints. Optimality of such policies for multi-criteria problems is also established.

These new policies have the pleasing aesthetic property that the amount of randomization they require over any trajectory is restricted by the number of constraints. This result is new even for constrained optimization with a single discount factor, where the optimality of randomized stationary policies is known. However, a randomized stationary policy may require an infinite number of randomizations over time.

We also formulate a linear programming algorithm for approximate solutions of constrained weighted discounted models.

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: Markov decision processes, additional constraints, several discount factors.

1. Introduction.

The paper deals with discrete time Markov Decision Processes (MDP) with finite state and action sets, and with $(M+1)$ criteria. Each criterion is a sum of standard expected discounted total rewards over infinite horizon with different discount factors. We consider the problem of optimizing one criterion, under inequality constraints on the M other criteria. We prove that, given an initial state, if a feasible policy exists, then there exists an optimal Markov policy satisfying the following two properties:

(i) for some integer $N < \infty$, this policy is (nonrandomized) stationary from epoch N onward,

(ii) at epochs $0, \ldots, N-1$ this policy uses at most M actions more than a (nonrandomized) Markov policy would use at these steps.

A policy that satisfies (i) and (ii) will be called an (M, N)-policy. We formulate a linear programming algorithm for the approximate solution of constrained weighted discounted MDPs.

For the multiple criteria problem with $(M+1)$ criteria, we show that any point on the boundary of the performance set can be reached by a (M, N)-policy, for some $N < \infty$. Since any Pareto optimal point belongs to the boundary, it follows that the performance of any Pareto optimal policy can be attained by an equivalent (M, N)-policy. We also show that, given any initial state and policy, there exists an equivalent $(M+1, N)$-policy.

We remark that the existence of optimal (M, N)-policies is a new result even for constrained MDPs with one discount factor; Frid (1972), Kallenberg (1983), Heyman and Sobel (1984), Altman and Shwartz (1991, 1991a), Sennott (1991), Tanaka (1991), Altman (1993, 1991), Makowski and Shwartz (1993). The existence of optimal randomized stationary policies for constrained discounted MDPs with finite state and action sets is known; Kallenberg (1983), Heyman and Sobel (1984). The same arguments, as in Ross (1989), imply that an optimal randomized stationary policy may be chosen among policies which use, at each epoch, at most M actions more than a (non-randomized) stationary policy. But any randomized stationary policy may perform these randomizations infinitely many times over the time horizon. In contrast, the advantage of (M, N)-policies is that they perform at most M randomization procedures over the time horizon.

The first results on (unconstrained) weighted criteria were obtained by Feinberg (1981) as an application of methods developed in that paper. Filar and Vrieze (1992) considered a sum of one average and one discounted criterion, or two discounted criteria with different discount factors, in the context of a two-person zero-sum stochastic game. They proved the existence of an ϵ-optimal policy which is stationary from some stage onward. Krass (1989) and Krass, Filar and Sinha (1992) considered a sum of one average and one discounted criterion for a finite state, finite action MDP and obtained ϵ-optimal policies. Similar results for controlled diffusions and countable models are obtained by Ghosh and Marcus (1991) and by Fernandez-Gaucherand, Ghosh, and Marcus (1990). Feinberg and Shwartz (1991) developed the weighted discounted case. They considered a finite sum of standard discounted criteria, each with a different discount factor. They showed that optimal (or even ϵ-optimal) (randomized) stationary policies may fail to exist, but there exist optimal Markov (non-randomized) policies. In the case of finite state and action spaces they proved the existence of an optimal Markov policy which is stationary from some stage N onward. Moreover, they derive a necessary and sufficient condition for a Markov policy to be optimal. An effective finite algorithm for computation of optimal policies for unconstrained problems is formulated in Feinberg and Shwartz (1991).

Several applications of MDPs in finance, project management, budget allocation, and production lead to criteria which are linear combinations of objective functions of different types, for example, average and total discounted rewards or several total discounted rewards with different discount factors. Sobel (1991) describes general preference axioms leading to discounted and weighted discounted criteria. Various applications of weighted criteria were discussed in Krass (1989), Krass, Filar, and Sinha (1992), and Feinberg and Shwartz (1991). Some of these applications lead to multiple objective problems and, in particular, to constrained optimization problems. Here we describe two applications to production systems. The first example deals with the implementation of new technologies. The second example deals with a simple model of a multicomponent unreliable system.

Example 1.1. A well-known effect of learning is that, when new technologies are implemented for a production system, the productivity increases and the cost of a production of a unit decreases over time. We consider a production system. Let a new technology be implemented at epoch 0. Let $r(x,a,t)$ be a net value created at epoch $t = 0, 1, \ldots$, where x is a state of a production system, and a is a production decision, f.i. the capacity utilization, production volume, production schedule for a given epoch, and so on. The natural form of the rewards is

$$r(x,a,t) = r_1(x,a) - l(t)c(x,a),$$

where c represents transient costs, which are expected to decrease to zero as technology is improved and production methods are perfected, $r_1(x,a)$ reflects the maximal possible production efficiency for state x and decision a. The graph of l is related to a so-called learning curve. Let $l(t) = \alpha^t$, where $0 < \alpha < 1$. Let x_t and a_t be states and decisions at epochs $t = 0, 1, \ldots$. The standard discounted criterion with discount factor β and with the immediate cost r leads to a total discounted cost of the form

$$\sum_{t=0}^{\infty} \left\{\beta^t r_1(x_t, a_t) - (\beta\alpha)^t c(x_t, a_t)\right\}, \tag{1.1}$$

which is a sum of two objective functions with different discount factors. There may be some additional costs, for example, setup costs or holding costs. A multiple-criteria problem arises, for example, when we consider the vector consisting of expected discounted total production rewards as one coordinate, and expected discounted holding costs as the other coordinate. A constrained optimization problem arises, for example, if it is desired that each of these characteristics lies below or above certain given levels, while the expected total discounted reward is to be maximized.

In different applications, the function l may take different forms. A general function $l(t)$ may be approximated (according to the Stone–Weierstrass theorem) by $\sum_{k=1}^{K} d_k \alpha_k^t$, where K is some integer, d_k and α_l are constants, and $0 < \alpha_k \le 1$, $k = 1, \dots$. Then (1.1) becomes

$$\sum_{t=0}^{\infty} \left\{\beta^t r_1(x_t, a_t) - \sum_{k=1}^{K} d_k (\beta\alpha_k)^t c(x_t, a_t)\right\}$$

and we obtain a multiple criteria problem where the criteria are linear combinations of discounted rewards with different discount factors.

Example 1.2. Consider an unreliable production system consisting of two units, say 1 and 2. Unit k can fail at each epoch with probability p_k under the condition that it has been operating before. The system operates if at least one of the units operates. Let $r_k(x, a)$, $k = 1, 2$, be an operating cost for unit k, if its state is x and decision a is chosen. Let β be the discount factor. Then the total discounted reward for unit k generated by the sequences $x_t,\ a_t,\ t = 0, 1, \dots$ is

$$\sum_{t=0}^{\infty} \beta^t (1 - p_k)^t r_k(x_t, a_t).$$

The problem of minimization of the total discounted costs under constraints on the corresponding costs for each unit is a constrained weighted discounted problem.

The proofs in this paper rely on the existence results for the finite-horizon problem (section 4, see also Derman and Klein (1965), Kallenberg (1981)), on the theory of unconstrained weighted discounted criteria (Feinberg and Shwartz 1991), and on finite-dimensional convex analysis (Stoer and Witzgall 1970). A precise formulation of the problem of interest is given in section 2, followed by the details of the structure of the paper.

2. The model and overview of the results.

Let $\mathbb{N}_0 = \{0, 1, \ldots\}$, $\mathbb{N} = \{1, 2, \ldots\}$, and fix $M \in \mathbb{N}_0$. Let $\mathbb{R}^{M+1}$ be the $(M+1)$-dimensional Euclidean space, and let

$$\mathbb{R}_0^{M+1} = \left\{u = (u_0, \ldots u_M) \in \mathbb{R}^{M+1};\ u_i \geq 0,\ i = 0, \ldots, M\right\}$$

be the non-negative orthant.

Consider a discrete-time controlled Markov chain with a finite state space $\mathbf{X}$, finite action space $\mathbf{A}$, sets of actions $\mathbf{A}(x) \subseteq \mathbf{A}$ available at $x \in \mathbf{X}$, and transition probabilities $\{p(y \mid x, a)\}$. For each $x, y \in \mathbf{X}$ and $a \in \mathbf{A}(x)$, we have $p(y \mid x, a) \geq 0$ and $\sum_{y \in \mathbf{X}} p(y \mid x, a) = 1$.

Let $H_n = \mathbf{X} \times (\mathbf{A} \times \mathbf{X})^n$ be the space of histories up to the time $n = 0, 1, \ldots, \infty$. Let $H = \bigcup\limits_{0 \leq n < \infty} H_n$ be the space of all finite histories. The spaces H_n and H are endowed with σ-fields generated by $2^{\mathbf{X}}$ and $2^{\mathbf{A}}$. A policy π is a function that assigns to each prehistory $h_n = x_0 a_0 x_1 \ldots x_n \in H_n$, $n = 0, 1, \ldots$, a probability distribution $\pi(\cdot \mid h_n)$ on $\mathbf{A}$ satisfying the condition $\pi(\mathbf{A}(x_n) \mid h_n) = 1$.

A policy π is called *randomized Markov* if for each $n = 0, 1, \ldots$ and each $x \in \mathbf{X}$ there exists a distribution $\pi_n(\cdot \mid x)$ such that $\pi(\cdot \mid h_n) = \pi_n(\cdot \mid x_n)$ for any $h_n \in H$. We denote by Π the sets of all policies. In section 3 we show that, without loss of generality, this set may be narrowed to to the set of randomized Markov policies. Therefore, in sections 3 – 8, Π denotes the set of all randomized Markov policies. A randomized Markov policy is called *randomized stationary* if $\pi_n(\cdot \mid x) = \pi_0(\cdot \mid x)$ for any $n = 0, 1, \ldots$ and any $x \in \mathbf{X}$. A *Markov* policy ϕ is a sequence of mappings $\phi_n : \mathbf{X} \to \mathbf{A}$ such that $\phi_n(x) \in \mathbf{A}(x)$ for any $x \in \mathbf{X}$. A Markov policy is called *stationary* if $\phi_n(x) = \phi_0(x)$ for any $n = 0, 1, \ldots$ and any $x \in \mathbf{X}$. Given $N = 0, 1, \ldots$, a Markov policy π is called (N, ∞)-*stationary* if there exists a stationary policy ϕ such that for any $x \in \mathbf{X}$

$$\pi_n(x) = \phi(x) \qquad \text{for} \qquad n = N, N+1, \ldots$$

for any $h_n \in H_n$. Stationary policies are $(0, \infty)$-stationary and vice versa.

For a finite set B, we denote by $|B|$ the number of elements in B. For an integer m, we say that a Markov policy π is a *randomized Markov policy of order* m if

$$\sum_{(x,n) \in B} \sum_{a \in \mathbf{A}(x)} \mathbf{1}\left\{\phi_n(a|x) > 0\right\} \leq |B| + m$$

for any finite subset $B \subset \mathbf{X} \times \mathbb{N}_0$. In other words, a randomized Markov policy is randomized Markov of order m, if 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 other words, a policy is a (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 a randomized stationary policy ϕ is m-randomized stationary for some $m \in \mathbb{N}_0$, if

$$\sum_{(x,a)\in\mathbf{X}\times\mathbf{A}} \mathbf{1}\{\phi(a|x) > 0\} \leq |X| + m.$$

Note that an m-randomized stationary policy with $m \geq 1$ may randomize over time an infinite number of times; this in contrast with a randomized Markov policy of order m.

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 $\mathbb{E}_x^\pi$.

We say a point u dominates v if $(u - v) \in \mathbb{R}_0^{M+1}$. Given a set $U \subseteq \mathbb{R}_0^{M+1}$, a point $u \in U$ is called Pareto optimal in U if there is no $v \in U$ which dominates u.

Let a $(M+1)$-dimensional vector $V(x, \pi) = (V_0(x, \pi), V_1(x, \pi), \ldots, V_M(x, \pi))$ characterize the performance of a policy $\pi \in \Pi$ under an initial state $x \in \mathbf{X}$ according to $M+1$ given criteria, $M \in \mathbb{N}_0$. We denote by $U(x) = \{V(x, \pi);\ \pi \in \Pi\}$ the "performance space." A policy π is called Pareto optimal if $V(x, \pi)$ is Pareto optimal in $U(x)$. 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 $V(x, \pi) = V(x, \sigma)$.

We are interested in solutions of constrained optimization problems: given the numbers $c_1, \ldots, c_M$ and given $x \in \mathbf{X}$, for $\pi \in \Pi$ consider

$$\text{maximize} \quad V_0(x, \pi) \tag{2.1}$$

$$\text{subject to} \quad V_m(x, \pi) \geq c_k, \quad m = 1, \ldots, M. \tag{2.2}$$

For each $m = 0, \ldots, M$, let R_m be a given real-valued function (reward) defined on $\mathbf{X} \times \mathbb{N}_0 \times \mathbf{A}$. These functions are assumed to be bounded above. We consider a situation when each $V_m(x, \pi)$, $m = 0, 1, \ldots, M$, is an expected total reward criterion

$$V_m(x, \pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} R_m(x_n, n, a_n), \tag{2.3}$$

with the conventions $(-\infty) + (+\infty) = -\infty$ and $0 \times \infty = 0$. We shall follow these conventions throughout the paper. Our main interest is a particular case of expected total discounted rewards or linear combinations of expected total discounted rewards, when

$$R_m(x,n,a) = \sum_{k=1}^{K} (\beta_{mk})^n r_{mk}(x,a), \tag{2.4}$$

where r_{mk} are finite and $0 \le \beta_{mk} < 1$, $m = 1, \dots, M$, $k = 1, \dots, K$, and $K \in \mathbb{N}$. Without loss of generality (by setting some of the $r_{mk} \equiv 0$, increasing K and renumbering) we can assume that $\beta_{mk} = \beta_{m'k} = \beta_k$ is independent of m. In this case (2.3) transforms into

$$V_m(x,\pi) = \sum_{k=1}^{K} D_{mk}(x,\pi), \tag{2.5}$$

where

$$D_{mk}(x,\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta_k^n r_{mk}(x_n,a_n) \tag{2.6}$$

are the expected total discounted rewards for the discount factor β_k and reward function r_{mk}, $m = 0, \dots, M$, $k = 1, \dots, K$. We remark that for different criteria, the number of actual summands in (2.5) may be different, because it is possible that $r_{mk} \equiv 0$ for some m and k.

For an unconstrained problem, $M = 0$. In this case, $V(x,\pi) = V_0(x,\pi)$ and we use the index k instead of the double index $0k$. For an unconstrained case, our notation coincides with that of Feinberg and Shwartz (1991), except that in Feinberg and Shwartz (1991), the standard discounted rewards D_k were denoted by V_k, $k = 1, \dots, K$.

Another important subclass of models with the expected total reward criteria, which we shall require, are finite horizon models. In this case there exists $N \in \mathbb{N}_0$ such that $R.(\cdot, n, \cdot) = 0$ for $n \ge N$. For these models

$$V_m(x,\pi) = \mathbb{E}_x^\pi \sum_{n=0}^{N-1} R_m(x_n,n,a_n), \tag{2.7}$$

and we will define policies for finite horizon models only up to the finite moment of time $N - 1$. In this case, if $\mathbf{X}$ and $\mathbf{A}$ are finite then the set of Markov policies is finite.

This paper studies constrained problem (2.1)–(2.2) with weighted discounted rewards V_k defined by (2.5)–(2.6). The main result of the paper (Theorem 6.8) states that if this problem has a feasible solutions then for some $N < \infty$ there exists an optimal (M,N)-policy. As was mentioned in the introduction, this result is new even for standard constrained discounted problems. It has an advantage with respect to the known result on the existence of optimal randomized stationary policies for standard discounted models, since (M, N)-policies require at most M randomizations over time. We note that, for weighted constrained problems, this class of policies is the simplest possible, for the following reason. Randomized stationary policies may not be optimal for weighted discounted criteria, even without constraints; Feinberg and Shwartz (1991), Example 1.1. Therefore, unlike the standard discounted dynamic programming, randomized stationary policies may not be optimal in constrained problems with different discount factors.

Sections 3–5 of the paper contain the material which we use in the proof of Theorem 6.8. In section 3, we show that the sets $U(x)$ are convex and compact. In section 4, we consider a finite horizon problem, establish the existence of an optimal randomized Markov policy of order M, and formulate an LP algorithm computing this policy. The results of section 4 are similar to the known results by Derman and Klein (1965) and Kallenberg (1981), but we formulate a different LP and use a different method of proof, and show that the total number of additional actions is indeed at most M. In section 5, we describe some properties of unconstrained problems. We introduce the notion of a funnel. For subsets $\mathbf{A}_n(z) \subseteq \mathbf{A}(z)$ and a number $N < \infty$, with the property $\mathbf{A}_n(z) = \mathbf{A}_N(z)$ for all $n \geq N$ and for all $z \in \mathbf{X}$, a funnel Θ is the set of all randomized Markov policies π such that $\pi_n(\mathbf{A}_n(z)|z) = 1$, $n = 0, 1, ..., z \in \mathbf{X}$.

The notion of a funnel is natural and useful, for the following reasons. Lemma 5.5 shows that, in fact, for an unconstrained problem with a weighted discounted criterion, the set of optimal policies is a funnel. From a geometric point of view, this funnel defines an exposed subset of $U(x)$. In addition, given any funnel, one may define an MDP with finite state and action sets, and such that the set of policies for the new MDP coincides with the given funnel (see proof of Lemma 5.5). This implies that, if the set of feasible policies is restricted by a funnel, the set of optimal randomized Markov policies coincides, in fact, with another funnel which is a subset of the first one (Lemma 6.1). This in turn implies that any exposed or proper extreme subset of $U(x)$ may be represented as a set of vectors $\{V(x, \pi),\ \pi \in \Theta\}$ where Θ is a funnel (Corollary 6.2 and Lemma 6.3).

The central point in the proof of Theorem 6.8 is Theorem 6.6 which states that, for any vector u on the boundary of $U(x)$, there exists a policy π which is stationary after some finite epoch N such that $V(x, \pi) = u$. This theorem reduces an infinite horizon problem to a finite horizon one.

In section 7, we consider a multi-criteria problem with $(M + 1)$ weighted discounted criteria. We show that, for any boundary vector u of $U(x)$, there exists a (M, N)-policy whose performance vector equals u (Theorem 7.2). This result implies that for any Pareto optimal policy there exists an equivalent (M, N)-policy (Corollary 7.3). We also show that for any policy there exists an equivalent $(M+1, N)$-policy (Theorem 7.5).

In section 8 we discuss the computation of optimal policies for constrained problems with weighted rewards.

3. Convexity and compactness of $U(x)$.

The results of this sections hold without the finiteness assumptions on the state and action sets. Therefore, in this section we assume that the state space $\mathbf{X}$ is countable, the action set $\mathbf{A}$ is arbitrary, and the standard measurability conditions hold; see e.g. van der Wal (1981). In particular, we assume that $\mathbf{A}$ is endowed with a σ-field $\mathcal{A}$, the sets $\mathbf{A}(y)$ belong to $\mathcal{A}$ for all $y \in \mathbf{X}$, all single-point subsets belong to $\mathcal{A}$, and reward functions and transitional probabilities are measurable in a.

Lemma 3.1 (Hordijk (1974), Theorem 13.2, Derman and Strauch (1966)). *Let $\{\pi^i\}_{i=1}^{\infty}$ be an arbitrary sequence of policies and let $\{\alpha_i\}_{i=1}^{\infty}$ be a sequence of nonnegative real numbers with $\sum_{i=1}^{\infty} \alpha_i = 1$. Given $x \in \mathbf{X}$ let σ be a randomized Markov policy defined by*

$$\sigma_n(A \mid y) = \frac{\sum_{i=1}^{\infty} \alpha_i \, \mathbb{P}_x^{\pi^i}(x_n = y, a_n \in A)}{\sum_{i=1}^{\infty} \alpha_i \, \mathbb{P}_x^{\pi^i}(x_n = y)} \tag{3.1}$$

for all $y \in \mathbf{X}$, for all $n \in \mathbb{N}_0$, and all $A \in \mathcal{A}$, whenever the denominator is nonzero, and $\sigma_n(\cdot \mid y)$ is arbitrary when the denominator is zero . Then

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

for all $y \in \mathbf{X}$, $A \in \mathcal{A}$, and $n \in \mathbb{N}_0$.

Corollary 3.2. *Let V_m, $m = 1, 2, \ldots, M$, be expected total reward criteria defined by (2.3). For any $x \in \mathbf{X}$ and for any policy π there exists a randomized Markov policy σ such that σ is equivalent to π at x. Such a policy σ is defined by (3.1) with $\pi^1 = \pi$ and $\alpha_1 = 1$.*

In fact, this equivalence holds for any criterion which depends only on the distributions of the pairs $\{x_n, a_n\}$. Since for any policy there exists an equivalent randomized Markov policy, there is no need to consider any policies except Markov ones. Therefore, in the rest of the paper, we consider only randomized Markov policies. Consequently, "policy" will mean "randomized Markov policy". In the rest of the paper, Π denotes the set of all randomized Markov policies.

Corollary 3.3. *Let V_m, $m = 1, 2, \ldots, M$, be expected total reward criteria defined by (2.3) and let σ be a randomized Markov policy defined by (3.1). Then*

$$V_m(x, \sigma) = \sum_{i=1}^{\infty} \alpha_i V_m(x, \pi^i).$$

Corollary 3.4. *In models with expected total reward criteria (2.3), the sets $U(x)$, $x \in \mathbf{X}$, are convex.*

Lemma 3.5. *Let V_m, $m = 0, \ldots, M$, be linear combinations of expected total discounted rewards defined by (2.4)–(2.5). Assume that $\mathbf{A}(x)$ are compact subsets of a Borel space. If the functions $r_{mk}(x, a)$ and $p(y|x, a)$ are continuous in a, and if $|r_{mk}(x, a)| \le D$ for some $D < \infty$ and for any $x, y \in \mathbf{X}$, $m = 0, \ldots, M$ and $k = 1, \ldots, K$, then the sets $U(x)$ are compact for all $x \in \mathbf{X}$.*

Proof. We fix some $x \in \mathbf{X}$. The action sets, transition probabilities, and reward functions satisfy condition (S) in Schäl (1975). By Theorem 6.6 in Schäl (1975), the set $\mathcal{P}_x = \{\mathbb{P}_x^\pi : \pi \in \Pi\}$ is compact and the mappings $\mathbb{P}_x^\pi \to \mathbb{E}_x^\pi r_{mk}(x_n, a_n)$ are continuous in the ws^∞-topology for any $m = 1, \ldots, M$, $k = 1, \ldots, K$, and $n = 0, 1, \ldots$. Therefore, the mappings $\mathbb{P}_x^\pi \to D_{mk}(x, \pi)$ are continuous, since if a sequence of continuous functions converges uniformly to some function on a compact set, then the limit is a continuous function. This implies that $\mathbb{P}_x^\pi \to V_m(x, \pi)$ are continuous mappings, $m = 1, \ldots, M$. Hence $\mathbb{P}_x^\pi \to V(x, \pi)$ is a continuous mapping of a compact into $\mathbb{R}^{M+1}$. Therefore, $U(x)$ is compact for each $x \in \mathbf{X}$. ■

4. Finite horizon models.

Since for a given x the set $U(x)$ is compact, if problem (2.1)–(2.2) has a feasible solution, it has a solution. Since this set is convex, an optimal policy is either Pareto optimal in the set of feasible policies, or is dominated by such a Pareto optimal policy. Theorem 6.7 states that, for any Pareto optimal policy, there exists a policy π which is equivalent at x, such that for some $N < \infty$ and for some stationary policy ϕ one has $\pi_n = \phi$ for all $n \geq N$. If N and ϕ are known, this result reduces the constrained infinite horizon problem with weighted discounted rewards to a constrained finite horizon problem with expected total rewards.

Constrained finite horizon problems were considered by Derman and Klein (1965) and Kallenberg (1981). It was shown that, for a given initial distribution, there exists an optimal randomized Markov policy which can be constructed from the solution of an LP program. Derman and Klein (1965) and Kallenberg (1981) formulated two different LPs for the solution of this problem.

In this section, we consider this problem by a different method than Derman and Klein (1965) or Kallenberg (1981). For the analysis of this problem, Derman and Klein (1965) used the reduction to an infinite horizon model with average rewards per unit time. Kallenberg (1981) used the direct analysis of occupation probabilities. We introduce a method based on the reduction of finite horizon problems to discounted infinite horizon problems.

Let R_m, $m = 0, \ldots, M$, be arbitrary rewards. Let $\mathbf{1}\{y = x\} = 1$, if $y = x$, and $\mathbf{1}\{y = x\} = 0$, if $y \neq x$. Consider the following LP:

$$\text{maximize} \quad \sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}(y)} \sum_{n=0}^{N-1} R_0(y, n, a) z_{y,n,a} \tag{4.1}$$

$$\text{subject to} \quad \sum_{a \in \mathbf{A}(y)} z_{y,0,a} = \mathbf{1}\{y = x\}, \quad y \in \mathbf{X}, \tag{4.2}$$

$$\sum_{a \in \mathbf{A}(y)} z_{y,n,a} - \sum_{u \in \mathbf{X}} \sum_{a \in \mathbf{A}(u)} p(y|u, a) z_{u,n-1,a} = 0, \quad y \in \mathbf{X},\ n = 1, \ldots, N-1, \tag{4.3}$$

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathbf{A}(y)} \sum_{n=0}^{N-1} R_m(y, n, a) z_{y,n,a} \geq c_m, \quad m = 1, \ldots, M, \tag{4.4}$$

$$z_{y,n,a} \geq 0, \quad y \in \mathbf{X},\ n = 0, \ldots, N-1,\ a \in \mathbf{A}(y). \tag{4.5}$$

Theorem 4.1. *Consider problem (2.1)–(2.2) with expected total rewards V_m defined by (2.7). This problem is feasible if and only if LP (4.1)–(4.5) is feasible. If z is an optimal basic solution of*

LP (4.1)–(4.5) then the formula

$$\phi_n(a|y) = \begin{cases} \dfrac{z_{y,n,a}}{\sum_{a'\in\mathbf{A}(y)} z_{y,n,a'}}, & \textit{if } \sum_{a'\in\mathbf{A}(y)} z_{y,n,a'} > 0, \\ \mathbf{1}\{a = a(y)\}, & \textit{otherwise,} \end{cases} \tag{4.6}$$

where $a(y) \in \mathbf{A}(y)$ are arbitrary, $n = 0, \ldots, N-1$, and $y \in \mathbf{X}$, defines an optimal randomized Markov policy ϕ of order M.

In order to prove Theorem 4.1, we consider a constrained problem (2.1)–(2.2) for a new finite model, whose details are given below, with the expected discounted rewards

$$\bar{V}_m(x, \pi) = \mathbb{E}_x^\pi \sum_{n=0}^{\infty} \beta^n r_m(x_n, a_n) \tag{4.7}$$

for some nonnegative $\beta < 1$. Consider the following LP:

$$\text{maximize} \quad \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}(y)} r_0(y,a) z_{y,a} \tag{4.8}$$

$$\text{subject to} \quad \sum_{a\in\mathbf{A}(y)} z_{y,a} - \beta \sum_{u\in\mathbf{X}} \sum_{a\in\mathbf{A}(u)} p(y|u,a) z_{u,a} = \mathbf{1}\{y = x\}, \quad y \in \mathbf{X}, \tag{4.9}$$

$$\sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}(y)} r_m(y,a) z_{y,a} \ge c_m, \quad m = 1, \ldots, M, \tag{4.10}$$

$$z_{y,a} \ge 0, \quad y \in \mathbf{X}, a \in \mathbf{A}(y). \tag{4.11}$$

Theorem 4.2. *(Kallenberg (1983), Heyman and Sobel (1984)). Consider problem (2.1)–(2.2) with the expected total discounted rewards defined by (4.7) for some nonnegative $\beta < 1$. This problem is feasible if and only if the LP (4.8)–(4.11) is feasible. If z is an optimal basic solution of LP (4.8–4.11) then the formula*

$$\phi(a|y) = \begin{cases} \dfrac{z_{y,a}}{\sum_{a'\in\mathbf{A}(y)} z_{y,a'}}, & \textit{if } \sum_{a'\in\mathbf{A}(y)} z_{y,a'} > 0; \\ \mathbf{1}\{a = a(y)\}, & \textit{otherwise;} \end{cases} \tag{4.12}$$

where $a(y) \in \mathbf{A}(y)$ are arbitrary and $y \in \mathbf{X}$, defines an optimal M–randomized stationary policy ϕ.

We note that Kallenberg (1983) and Heyman and Sobel (1984) do not formulate the property that the randomized stationary policy ϕ defined by (4.12) is M-randomized stationary. This follows from the fact that the number of constraints is $|X| + M$ and each equality (4.9) defines at least one basic solution, cf. Ross (1989) for similar arguments.

Proof of Theorem 4.1. We consider an MDP with state space $\bar{\mathbf{X}}$, action sets $\bar{\mathbf{A}}(\cdot)$, transition probabilities $\bar{p}(\cdot|\cdot,\cdot)$, and reward functions $\bar{r}_m$, $m = 1, \ldots, M$, where

(i) $\bar{\mathbf{X}} = (\mathbf{X} \times \{0, \ldots, N-1\}) \cup \{0\}$;

(ii) $\bar{\mathbf{A}}(x, n) = \mathbf{A}(x)$ for $x \in \mathbf{X}$, $n = 0, \ldots, N-1$, and $\bar{\mathbf{A}}(0) = \{\bar{a}\}$ for some fixed arbitrary $\bar{a} \in \mathbf{A}$;

(iii) $\bar{p}((u, n+1)|(y, n), a) = \bar{p}(u|y, a)$ for $n = 0, \ldots, N-2$ and $\bar{p}(0|(y, N-1), a) = \bar{p}(0|0, \bar{a}) = 1$, where $u, y \in \mathbf{X}$, $a \in \mathbf{A}(y)$, and all other transition probabilities equal 0;

(iv) $\bar{r}_m(0, \bar{a}) = 0$ and $\bar{r}_m((y, n), a) = \beta^{-n} R_m(y, n, a)$ for $m = 1, \ldots, M$, $y \in \mathbf{X}$, $n = 0, \ldots, N-1$, and $a \in \mathbf{A}(y)$.

There is a natural one-to-one correspondence

$$\phi_n(\cdot|y) = \phi(\cdot|y, n) \quad n = 0, \ldots, N-1, y \in \mathbf{X}$$

between randomized Markov policies in the original finite horizon model and randomized stationary policies in the new infinite horizon discounted model. For every $m = 0, 1, \ldots,$ this mapping is also a one-to-one correspondence between randomized Markov policies of order m in the original finite horizon model and m-randomized stationary policies in the new infinite horizon discounted model. This correspondence preserves the values of all criteria. By Theorem 4.2 applied to the new model, since the state and action sets are finite and V_m, $m = 0, \ldots, M$, are the total expected discounted rewards with the same discount factor β, there exists an optimal randomized stationary policy for problem (2.1)–(2.2), if this problem has a feasible policy. Therefore, Theorem 4.2 implies Theorem 4.1. We note that, in order to get LP (4.1)–(4.5) directly from LP (4.7)–(4.11), one has to consider variables $z_{y,n,a} = \beta^n z_{u,a}$, where $y \in \mathbf{X}$, $n = 0, \ldots, N-1$, $u = (y, n)$, $a \in \mathbf{A}(y)$, and a variable $z_0 = z_{0,\bar{a}}$. Then LP (4.7)–(4.11) transforms to LP (4.1)–(4.5) with the additional constraint

$$\sum_{y \in \mathbf{X}} \sum_{u \in \mathbf{X}} \sum_{a \in \mathbf{A}(u)} p(y|u, a) z_{y,N-1,a} = z_0.$$

Constraints (4.2)–(4.3) imply that the left hand side of this equality equals 1. This constraint becomes $z_0 = 1$. Since the variable z_0 is absent in (4.1)–(4.5), the variable and the constraint may be omitted. ■

Algorithm 4.3. *(Computation of an optimal randomized Markov policy of order M for a finite horizon model).*

(i) Solve LP (4.1)–(4.5).

(ii) If this LP is not feasible, there is no optimal policy. If this LP is feasible, compute an optimal randomized Markov policy of order M by (4.6).

We remark that if one is interested in the solution of a finite horizon problem with respect to a given initial distribution δ_y, $y \in \mathbf{X}$, one should consider problem (4.1)–(4.5) when the right hand side of (4.2) is replaced by δ_y.

5. Unconstrained problems with weighted discounted rewards.

For unconstrained problems, we have $M = 0$ and $V(x,\pi) = V_0(x,\pi)$, where $x \in \mathbf{X}$ and $\pi \in \Pi$. For a set $\Delta \subseteq \Pi$, we define $V_\Delta(x) = \sup\{V(x,\pi) : \ \pi \in \Delta\}$. A policy π is called optimal if $V(x,\pi) = V_\Pi(x)$ for all $x \in X$. To simplify the notation, throughout in this section, whenever we deal with unconstrained problems, we omit index $m = 0$ in the criteria, and in the reward functions. Assume that the discount factors are ordered so that $\beta_1 > \beta_2 > \cdots > \beta_K$. We can do it without loss of generality because, if $\beta_k = \beta_{k+1}$ for some k, we consider the reward function $r_k + r_{k+1}$ and lower K by 1.

We consider an unconstrained model with weighted discounted rewards. Recall the definition (2.6) of $D_k(x,\pi)$ and define the action sets $\Gamma_k(x)$, $k = 0, 1, \dots, K$, recursively as follows. Set $\Gamma_0(x) = \mathbf{A}(x)$ for $x \in \mathbf{X}$. Given $\Gamma_k \geq 0$, let Π_k be the set of policies whose actions are in the sets $\Gamma_k(x)$, $x \in \mathbf{X}$. For $x \in \mathbf{X}$ we define

$$D_{k+1}(x) = \sup_{\pi \in \Pi_k} D_{k+1}(x,\pi)$$

and

$$\Gamma_{k+1}(x) = \left\{ a \in \Gamma_k(x) : D_{k+1}(x) = r_{k+1}(x,a) + \beta_{k+1} \sum_{z \in \mathbf{X}} p(z|x,a) D_{k+1}(z) \right\}, \quad x \in \mathbf{X}.$$

We set $\Gamma(x) = \Gamma_K(x)$, $x \in \mathbf{X}$.

Theorem 5.1. *(Feinberg and Shwartz (1991), Theorem 3.8). Consider an unconstrained MDP with an infinite horizon and weighted discounted reward V defined by (2.5)–(2.6) with $M = 0$. For each initial state x there exists an optimal (N,∞)-stationary policy ϕ. The stationary policy ϕ^N which ϕ uses when the time parameter is greater than or equal to N may be chosen as an arbitrary policy satisfying the condition $\phi^N(x) \in \Gamma(x)$ for all $x \in \mathbf{X}$.*

Theorem 5.2. *(Feinberg and Shwartz (1991), Theorem 3.13). Consider an unconstrained problem with weighted discounted rewards. Given initial state $x \in \mathbf{X}$, there exist $N < \infty$ and action sets $\mathbf{A}_t(z) \subseteq \mathbf{A}(z)$, $t = 0, \dots, N-1$ and $z \in \mathbf{X}$, such that $V(x,\pi) = V(x)$ if and only if*

$$a_t \in \mathbf{A}_t(x_t) \quad (\mathbb{P}_x^\pi\text{-a.s.}), \quad t = 0, \dots, N-1, \tag{5.1}$$

and

$$a_t \in \Gamma(x_t) \quad (\mathbb{P}_x^{\pi}-\text{a.s.}), \quad t = N, N+1, \ldots \,. \tag{5.2}$$

Corollary 5.3. *If the policies π^i, $i=1,2$, satisfy (5.1) and (5.2) with $\pi=\pi^i$ and if $\pi_t^1(a|z) = \pi_t^2(a|z)$ for all $t=0,\ldots,N-1$, for all $z \in \mathbf{X}$, and for all $a \in \mathbf{A}$, then $D_k(x,\pi^1) = D_k(x,\pi^2)$ for all $k=1,\ldots,K$.*

Proof. We observe that $\mathbb{P}_x^{\pi^1}(h_N) = \mathbb{P}_x^{\pi^2}(h_N)$ for any $h_N \in H_N$. By Lemma 3.5 in Feinberg and Shwartz (1991) a policy σ is lexicographically-optimal at $z \in \mathbf{X}$ for criteria $D_1, D_2, \ldots, D_K$ if and only if $a_t \in \Gamma(x_t)$ ($\mathbb{P}_z^{\sigma}$)-a.s. for all $t=0,1,\ldots$. This implies that if $\mathbb{P}_x^{\pi^1}(h_N) > 0$ then

$$\mathbb{E}_x^{\pi^1}\left(\sum_{t=N}^{\infty} \beta_k^t r_k(x_t,a_t) \middle| \, h_N\right) = \mathbb{E}_x^{\pi^2}\left(\sum_{t=N}^{\infty} \beta_k^t r_k(x_t,a_t) \middle| \, h_N\right),$$

because both "shifted" policies π^1 and π^2 are lexicographically optimal at x_N. Since $\mathbb{P}_x^{\pi^1}(h_N) = \mathbb{P}_x^{\pi^2}(h_N)$ for all $h_N \in H_N$, this implies the corollary. ∎

Definition 5.4. A set of policies $\Theta \subseteq \Pi$ is called a funnel if there is a number $N < \infty$ and sets $\{\mathbf{A}_n(z) \subseteq \mathbf{A}(z) : \; n=0,\ldots,N, z \in \mathbf{X}\}$ such that $\pi \in \Theta$ if and only if the following conditions hold:

(i) $\pi_n(\mathbf{A}_n(z)|z) = 1$ for all $z \in \mathbf{X}$ and for all $n=0,\ldots,N-1$,

(ii) $\pi_n(\mathbf{A}_N(z)|z) = 1$ for all $z \in \mathbf{X}$ and for all $n \geq N$.

For $\Delta \subseteq \Pi$ we define the sets $D_{mk}(x,\Delta) = \{D_{mk}(x,\pi) : \; \pi \in \Delta\}$, $V_m(x,\Delta) = \{V_m(x,\pi) : \; \pi \in \Delta\}$, and $V(x,\Delta) = \{V(x,\pi) : \; \pi \in \Delta\}$, where $m=0,\ldots,M$ and $k=1,\ldots,K$.

Lemma 5.5. *Consider an unconstrained problem with weighted discounted rewards. Let $\Theta \subseteq \Pi$ be a non-empty funnel and let an initial state x be fixed. There exists a nonempty funnel $\Theta' \subseteq \Theta$ such that*

(i) $V(x,\pi) = V_{\Theta}(x)$ for any $\pi \in \Theta'$,

(ii) $(D_1(x,\Theta'),\ldots,D_K(x,\Theta')) = (D_1(x,\Pi_{\Theta}),\ldots,D_K(x,\Pi_{\Theta}))$, where $\Pi_{\Theta} = \{\pi \in \Theta : \; V(x,\pi) = V_{\Theta}(x)\}$.

Proof. Define an MDP with the state space $\tilde{X} = (\mathbf{X} \times \{0,\ldots,N-1\}) \cup \mathbf{X}$, action set $\mathbf{A}$, feasible action sets

$$\tilde{\mathbf{A}}(z) = \begin{cases} \mathbf{A}_t(y), & \text{if } z=(y,t),\, y \in \mathbf{X},\, t=0,\ldots,N-1, \\ \mathbf{A}_N(z), & \text{if } z \in \mathbf{X}, \end{cases}$$

transition probabilities

$$\tilde{p}(z'|z,a) = \begin{cases} p(y'|y,a), & \text{if } z' = (y', i+1),\ z = (y,i),\ y', y \in \mathbf{X},\ i = 0, \ldots, N-2, \\ p(z'|y,a), & \text{if } z' \in \mathbf{X},\ z = (y, N-1),\ y \in \mathbf{X}, \\ p(z'|z,a), & \text{if } z', z \in \mathbf{X}, \\ 0, & \text{otherwise,} \end{cases}$$

rewards

$$\tilde{r}_k(z,a) = \begin{cases} r_k(y,a), & \text{if } z = (y,i),\ y \in \mathbf{X},\ i = 0, \ldots, N-1, \\ r_k(z,a), & \text{if } z \in \mathbf{X}, \end{cases} \tag{5.3}$$

and discount factors β_k, $k = 1, \ldots, K$.

The set of policies for this model coincides with Θ. Therefore, the value of this model with initial state $(x, 0)$ equals $V_\Theta(x)$. By Theorem 5.2 applied to the new model, there exist $N' \geq N$ and sets $\mathbf{A}'_t(z)$, $z \in \mathbf{X}$ and $t = 0, \ldots, N'$, such that

(a) $\mathbf{A}'_t(z) \subseteq \mathbf{A}_t(z)$ for $t = 0, \ldots, N-1$, $z \in \mathbf{X}$,

(b) $\mathbf{A}'_t(z) \subseteq \mathbf{A}_N(z)$ for $t = N, \ldots, N'$, $z \in \mathbf{X}$,

(c) $\pi \in \Pi_\Theta$ if and only if $a_t \in \mathbf{A}'_t(x_t)$ $(\mathbb{P}^\pi_x)$-a.s. for $t = 0, \ldots, N'-1$ and $a_t \in \mathbf{A}'_{N'}(x_t)$ $(\mathbb{P}^\pi_x)$-a.s. for $t = N', N'+1, \ldots$.

The number N' and the sets $\mathbf{A}'_n(\cdot)$, $t = 0, \ldots, N'$, define a funnel Θ' and, by (a)–(b), $\Theta' \subset \Theta$. From (c) we have that $\Theta' \subseteq \Pi_\Theta$. Therefore, $(D_1(x, \Theta'), \ldots, D_K(x, \Theta')) \subseteq (D_1(x, \Pi_\Theta), \ldots, D_K(x, \Pi_\Theta))$. Let $\pi \in \Pi_\Theta$. By (c), the policy π satisfies the condition $a_t \in \mathbf{A}'_t(x_t)$ $(\mathbb{P}^\pi_x)$-a.s. for $t = 0, \ldots, N'-1$ and $a_t \in \mathbf{A}'_{N'}(x_t)$ $(\mathbb{P}^\pi_x)$-a.s. for $t = N', N'+1, \ldots$. Let ϕ be a policy such that $\phi_t(\mathbf{A}'_t|z) = 1$ for $t = 0, \ldots, N'-1$ and $\phi_t(\mathbf{A}'_{N'}|z) = 1$ for $t = N', N'+1, \ldots$, $z \in \mathbf{X}$. Let $\mathbf{A}'_t(\cdot) = \mathbf{A}'_{N'}(\cdot)$ for $t \geq N'$. For a policy π such that $V(x,\pi) = V(x, \Theta)$, define a policy σ by

$$\sigma_t(a|z) = \begin{cases} \pi_t(a|z), & \text{if } \pi_t(\mathbf{A}'(z)|z) = 1, \\ \phi_t(a|z), & \text{if } \pi_t(\mathbf{A}'(z)|z) \neq 1, \end{cases}$$

$t \in \mathbb{N}_0$, $z \in \mathbf{X}$. We have $\sigma \in \Theta'$ and $D_k(x,\sigma) = D_k(x,\pi)$ for $k = 1, \ldots, K$. Therefore, $(D_1(x, \Theta'), \ldots, D_K(x, \Theta')) \supseteq (D_1(x, \Pi_\Theta), \ldots, D_K(x, \Pi_\Theta))$. ■

The following lemma deals with the constrained problem, so that $V(x,\pi)$ is now a vector in $\mathbb{R}^{M+1}$.

Lemma 5.6. *For any funnel Θ, the set $V(x,\Theta)$ is convex and compact.*

Proof. For any funnel Θ, there exists an MDP with the finite state and action sets such that there is one-to-one correspondence between Θ and the set of policies in this new model. This model is similar to the model defined in the proof of Lemma 5.5 with the only difference that the reward functions r and $\tilde{r}$ in (5.3) depend of two indices $m = 0, \dots, M$ and $k = 1, \dots, K$. By Corollary 3.4 and Lemma 3.5, $V(x, \Theta)$ is convex and compact. ■

6. The existence of optimal (M, N)-policies.

The goal of this section is to show that, if problem (2.1)–(2.2) has a feasible solution for discounted weighted criteria, then for some $N < \infty$ there exists an optimal (M, N)-policy for this problem (Theorem 6.8). The proof is based on a combination of results from Sections 3–5 and on convex analysis.

We remind the reader some notation and definitions from convex analysis; see Stoer and Witzgall (1970). A convex subset W 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 $u_3 \in W$ is only possible for $u_1, u_2 \in W$. A subset W of E is called exposed if there is a supporting plane H of E such that $W = H \cap E$. 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.

Lemma 6.1. *Let Θ be a funnel and W be an exposed subset of $V(x, \Theta)$. There exists a funnel Θ' such that $W = V(x, \Theta')$.*

Proof. Let $\sum\limits_{m=0}^{M} b_m u_m = b$ be a supporting plane of the convex, compact set $V(x, \Theta)$ which contains W and let $\sum\limits_{m=0}^{M} b_m u_m \le b$ for every $u = (u_0, u_1, \dots, u_M) \in V(x, \Theta)$. Then

$$W = \left\{ u \in V(x, \Theta) : \sum_{m=0}^{M} b_m u_m = \max \left\{ \sum_{m=0}^{M} b_m u_m : \ u \in V(x, \Theta) \right\} \right\}$$
$$= \left\{ u \in V(x, \Theta) : \ \sum_{m=0}^{M} b_m u_m = \max \left\{ \sum_{m=0}^{M} b_m V_m(x, \pi) : \pi \in \Theta \right\} \right\}.$$

Therefore, $u \in W$ if and only if $u = \sum\limits_{m=0}^{M} b_m V_m(x, \pi)$, where π is an optimal policy for a weighted discounted criterion $\sum\limits_{m=0}^{M} b_m V_m$ with initial state x. By Lemma 5.5, $W = V(x, \Theta')$ for some funnel $\Theta' \subseteq \Theta$. ■

Corollary 6.2. *Let W be an exposed subset of $U(x)$. There exists a funnel Θ such that $W = V(x, \Theta)$.*

Proof. The set of all policies Π is a funnel defined by $N = 0$ and $\mathbf{A}_0(\cdot) = \mathbf{A}(\cdot)$. ■

Lemma 6.3. *Let E be a proper extreme subset of $U(x)$. There exists a funnel Θ such that $E = V(x, \Theta)$.*

Proof. The proof is based on Lemma 6.1 and on the fact that, if E is a proper extreme subset of a compact convex set W_0, then there is a finite sequence of sets $W_0, W_1, \ldots, W_j$ such that W_{i+1} is an exposed subset of W_i, $i = 0, \ldots, j-1$, and $W_j = E$. This fact follows from Stoer and Witzgall (1970), Propositions (3.6.5) and (3.6.3).

The set $\Theta_0 = \Pi$ is clearly a funnel, defined by $N = 0$ and $\mathbf{A}_0(\cdot) = \mathbf{A}(\cdot)$. By definition, $U(x) = V(x, \Theta_0)$ and we denote $W_0 = U(x)$.

Assume that, for some $i \in \mathbb{N}_0$, we have a funnel Θ_i such that E is a proper extreme subset of $W_i = V(x, \Theta_i)$. By Lemma 5.6, the set W_i is convex and compact. Let W_{i+1} be a proper exposed subset of the convex set W_i such that $W_{i+1} \supseteq E$. By Stoer and Witzgall (1970), Propositions (3.6.5) and (3.6.3), the set W_{i+1} exists and

$$\dim E \leq \dim W_{i+1} < \dim W_i. \tag{6.1}$$

By Lemma 6.1, there exists a funnel Θ_{i+1} such that $W_{i+1} = V(x, \Theta_{i+1})$.

If $E \neq W_{i+1}$, we increase i by 1 and repeat the construction. If $E = W_{i+1}$ for some $i \in \mathbb{N}_0$, the lemma is proved and $\Theta = \Theta_{i+1}$. Otherwise, we get an infinite sequence $\{W_i,\ i \in \mathbb{N}_0\}$. This contradicts (6.1) since $\dim W_0 \leq M + 1$. ■

We remark that, since any exposed subset of a convex set is extreme, the only situation, when an exposed subset E of a convex set U in $\mathbb{R}^{M+1}$ is not proper extreme, is $E = U$ and $\dim U < M+1$.

Corollary 6.4. *If u is an extreme point of $U(x)$ then for some $N < \infty$ there exist an (N, ∞)-stationary policy π such that $V(x, \pi) = u$.*

Proof. If $U(x) = \{u\}$, we have that $V(x, \phi) = u$ for any stationary policy. If $U(x) \neq \{u\}$, we have that $\{u\}$ is a proper extreme subset of $U(x)$. By Lemma 6.3, $\{u\} = V(x, \Theta)$ for some funnel Θ. Let the funnel Θ be generated by the sets $\mathbf{A}_n(\cdot)$, $n = 0, \ldots, N$ for some $N \in \mathbb{N}_0$. Then $V(x, \phi) = u$ for any (N, ∞)-stationary policy $\phi \in \Theta$. ■

For two points $u = (u_1, \ldots, u_M)$ and $v = (v_1, \ldots, v_M)$ in R^M, define the distance $d(u, v) = \sum_{i=1}^{M} |u_i - v_i|$.

Lemma 6.5. *Let E be either an exposed subset or a proper extreme subset of $U(x)$. There exists a stationary policy ϕ with the following property: for any $\epsilon > 0$ there exists $N \in \mathbb{N}_0$ such that for any $u \in E$ there exists a point $v \in E$ satisfying the following conditions:*

(i) v belongs to the ϵ-neighborhood of u,

(ii) $v = V(x, \pi)$ for some policy π satisfying the condition $\pi_t(\phi(z)|z) = 1$ for all $t \geq N$ and all $z \in \mathbf{X}$.

Proof. By Lemmas 6.2 and 6.3, $E = V(x, \Theta)$ for some funnel Θ. Let Θ be generated by the sets $\mathbf{A}_n(\cdot)$, $n = 1, \ldots, N'$, where $N' \in \mathbb{N}_0$. Let ϕ be a stationary policy such that $\phi(z) \in \mathbf{A}_{N'}(z)$ for all $z \in \mathbf{X}$. Let

$$\begin{aligned} \beta &= \max\{\beta_k : \ k = 1, \ldots, K\}. \\ r &= \max\{|r_{mk}(z, a)| : \ m = 0, \ldots, M, \ k = 1, \ldots, K, \ z \in \mathbf{X}, \ a \in \mathbf{A}(z)\}. \end{aligned} \tag{6.2}$$

Note that $\beta \in [0, 1)$ and that if $\pi_i(\cdot) = \sigma_i(\cdot)$ for all $i = 0, \ldots, n$, then $|V_m(x, \pi) - V_m(x, \sigma)| \leq 2Kr\beta^n/(1 - \beta)$. Given $\epsilon > 0$, choose $N \geq N'$ such that $2(M + 1)Kr\beta^N/(1 - \beta) \leq \epsilon$. Then, for any policies π and σ coinciding at steps $0, \ldots, N$, we have that the distance between $V(x, \sigma)$ and $V(x, \pi)$ is not greater than the given ϵ.

Let $u \in E$. Consider a policy $\sigma \in \Theta$ such that $u = V(x, \sigma)$. Define a policy π by $\pi_n = \sigma_n$ for $n = 0, \ldots, N - 1$, and $\pi_n(\phi(z)|z) = 1$ for $n \geq N$. Then $v = V(x, \pi)$ belongs to the ϵ-neighborhood of $V(x, \sigma)$. Since $\sigma \in \Theta$, we have $\pi \in \Theta$ and $V(x, \pi) \in E$. ■

Theorem 6.6. *Let E be either an exposed subset or a proper extreme subset of $U(x)$. For any $u \in E$ there exist a policy π such that:*

(i) $V(x, \pi) = u$;

(ii) there are a stationary policy ϕ and integer $N < \infty$ such that $\pi_t(\phi(z)|z) = 1$ for all $t \geq N$ and all $z \in \mathbf{X}$.

Proof. Since any intersection of extreme sets is an extreme set and any intersection of closed sets is a closed set, there exists a minimal closed extreme subset W of $U(x)$ containing u, $W \subseteq E$. This set is an intersection of all closed extreme subsets of $U(x)$ containing u. If E is an exposed set, it is extreme, but it is possible that $E = U(x)$; Stoer and Witzgall (1970), p. 43.

Let $\dim W = m$, where $m \leq M$. By Caratheodory's theorem, u is a convex combination of $m+1$ extreme points $u_1, \ldots, u_{m+1}$ of W. The minimality of W implies that the convex hull of $\{u_1, \ldots, u_{m+1}\}$ is a simplex and u is a (relatively) inner point of this simplex. We choose $\epsilon > 0$ small enough so that if $\{v_1, \ldots, v_{m+1}\} \subset W$ and each v_i belongs to the ϵ-neighborhood of u_i, $i = 1, \ldots, m+1$, then the following property holds: the convex hull of $v_1, \ldots, v_{m+1}$ is a simplex and u belongs to this simplex.

Either W is a proper extreme subset of $U(x)$ or $W = E = U(x)$ and W is an exposed subset. By Lemma 6.5, we consider an integer $N < \infty$, stationary policy ϕ, and policies π^i, $i = 1, \ldots, m+1$, such that:

(i) $\pi^i_t(\phi(z)|z) = 1$ for all $z \in \mathbf{X}$ and all $t \geq N$,

(ii) $V(x, \pi^i) = v_i$, $i = 1, \ldots, m+1$.

We have that $u = \sum_{i=1}^{m+1} \alpha_i V(x, \pi^i)$ for some nonnegative α_i, $i = 1, \ldots, m+1$, with $\sum_{i=1}^{m+1} \alpha_i = 1$. Lemma 3.1 and Corollary 3.3 imply that there exists a policy π such that $V(x, \pi) = u$ and $\pi_t(\phi(z)|z) = 1$ for all $z \in \mathbf{X}$ and all $t \geq N$. ■

Theorem 6.7. *Let u be a Pareto optimal point of $U(x)$. Then there exist a policy π such that:*

(i) $V(x, \pi) = u$;

(ii) there are a stationary policy ϕ and integer $N < \infty$ such that $\pi_t(\phi(z)|z) = 1$ for all $t \geq N$ and all $z \in \mathbf{X}$.

Proof. We consider two situations: $\dim U(x) \leq M$ and $\dim U(x) = M+1$. If $\dim U(x) \leq M$, then $U(x)$ is an exposed set. If $\dim U(x) = M+1$, a Pareto optimal point u belongs to the (relative) boundary of $U(x)$. In this case, u belongs to some proper extremal subset of $U(x)$. In both cases, Theorem 6.7 follows from Theorem 6.6. ■

Theorem 6.8. *If problem (2.1)–(2.2) is feasible, for some $N < \infty$ there exists an optimal (M, N)-policy for this problem.*

Proof. Assume the problem is feasible. By Lemma 3.5, there exists an optimal solution, say σ. Since $U(x)$ is a convex compact, there exists a Pareto optimal point $u \in U(x)$ such that either $u = V(x, \sigma)$ or u dominates $V(x, \sigma)$. Any policy π, such that $V(x, \pi) = u$, is optimal. By Theorem 6.7, there exists a policy π such that $V(x, \pi) = u$ and $\pi_t(\phi(z)|z) = 1$ for all $z \in \mathbf{X}$ and all $t \geq N$ for some stationary policy ϕ and some finite integer N.

In order to find an optimal policy at epochs $t = 0, \ldots, N-1$, one has to solve a finite horizon problem with the reward functions $R_m(x, n, a)$ defined by (2.4) for $n = 0, \ldots, N-2$ and

$$R_m(x, N-1, a) = \sum_{k=1}^{K} \left\{ \beta_k^{N-1} r_{mk}(x, a) + \beta_k^N D_{mk}(x, \phi) \right\}. \tag{6.3}$$

Let π be a randomized Markov policy of order M which is optimal for the finite horizon problem; see Theorem 4.1. This policy is defined for $n = 0, \ldots, N-1$. We set $\pi_n(\phi(z)|z) = 1$ for all $n \geq N$ and for all $z \in \mathbf{X}$. We have that π is an optimal (M, N)-policy. ∎

7. Multi-criteria problems.

In this section we prove that for weighted discounted problems with M criteria, given any point on the boundary of the performance set $U(x)$, for some $N < \infty$ there exists an (M, N)-policy with this performance (Theorem 7.2). This result implies that for any Pareto optimal policy, for some $N < \infty$ there exists an equivalent (M, N)-policy (Corollary 7.3). We also show that, given an initial point x, for any policy there exists an equivalent $(M+1, N)$-policy for some $N < \infty$ (Theorem 7.5). The proofs follow from Theorem 6.8 and from the following lemma.

Lemma 7.1. *Let $U \subseteq \mathbb{R}^{M+1}$ be convex and compact. Let u^* belong to the boundary of U (if $\dim U \leq M+1$ then U coincides with its boundary). There exist constants d_{mi}, $m, i = 0, \ldots, M$, and constants c_i, $i = 1, \ldots, M$, such that u^* is a unique solution of the problem*

$$\text{maximize} \quad \sum_{i=0}^{M} d_{0i} u_i \tag{7.1}$$

$$\text{subject to} \quad \sum_{i=0}^{M} d_{mi} u_i \geq c_m, \qquad m = 1, \ldots, M, \tag{7.2}$$

$$(u_0, \ldots, u_M) \in U. \tag{7.3}$$

Proof. Let $\sum_{i=0}^{M} d_{0i} u_i = c_0$ be a supporting plane which contains u^* and let $\sum_{i=0}^{M} d_{0i} u_i \leq c_0$ for any $u = (u_0, \ldots, u_M) \in U(x)$. We consider planes $\sum_{i=0}^{M} d_{mi} u_i = c_m$, $i = 1, \ldots, M$, such that $\bigcap_{i=0}^{M} \{u : \sum_{i=0}^{M} d_{mi} u_i = c_m\} = \{u^*\}$. Then u^* is a unique solution of problem (7.1)–(7.3). ∎

Theorem 7.2. *Consider weighted discounted criteria V_m, $m = 0, ..., M$, defined by (2.5). If a vector u belongs to a boundary of $U(x)$ for some $x \in \mathbf{X}$ then for some $N < \infty$ there exists an (M, N)-policy π with $V(x, \pi) = u$.*

Proof. We set $U = U(x)$ and $\tilde{V}_k(x, \pi) = \sum_{i=0}^{M} d_{mi} V_i(x, \pi)$. Then Theorem 6.8 and Lemma 7.1 imply the theorem. ■

Corollary 7.3. *Consider weighted discounted criteria V_m, $m = 0, ..., M$, defined by (2.5). If σ is a Pareto optimal policy at $x \in \mathbf{X}$ then for some $N < \infty$ there exists an (M, N)-optimal policy π with $V(x, \pi) = V(x, \sigma)$.*

Proof. Any Pareto optimal point of a compact convex set belongs to its boundary. ■

Lemma 7.4. *Let $U \subseteq \mathbb{R}^{M+1}$ be convex and compact. For any $u^* \in U$ there exist constants d_{mi}, $m = 0, \ldots, M+1$, $i = 0, \ldots, M$, and constants c_i, $i = 1, \ldots, M+1$, such that u^* is a unique solution of the problem*

$$\begin{aligned} &\text{maximize} && \sum_{i=0}^{M} d_{0i} u_i \\ &\text{subject to} && \sum_{i=0}^{M} d_{mi} u_i \geq c_m, \qquad m = 1, \ldots, M+1, \\ & && (u_0, \ldots, u_M) \in U. \end{aligned}$$

Proof. We consider a plane $\sum_{i=0}^{M} d_{M+1i} u_i = c_{M+1}$ such that u^* belongs to this plane. We set $U^* = U \cap \{u : \sum_{i=0}^{M} d_{M+1i} u_i \geq c_{M+1}\}$. Then u^* belongs to the boundary of U^*. Lemma 7.4 follows from Lemma 7.1 applied to the set U^* and point u^*. ■

Theorem 7.5. *Consider weighted discounted criteria V_m, $m = 0, ..., M$, defined by (2.5). For any policy σ for some $N < \infty$ there exists an $(M + 1, N)$-policy π with $V(x, \pi) = V(x, \sigma)$.*

Proof. The proof is similar to the proof of Theorem 7.2 but we apply Lemma 7.4 instead of Lemma 7.1. ■

The following example illustrates that $M + 1$ cannot be replaced with M in Theorem 7.5.

Example 7.6. Let $\mathbf{X} = \{1\}$, $\mathbf{A}(1) = \{0, 1\}$, $M = 0$, $p(1|1, 0) = p(1|1, 1) = 1$, $r_0(1, 0) = 0$, and $r_0(1, 1) = 1$. Then $U(1)$ is the interval $[0, 2]$. If π is a $(0, N)$-policy for some $N < \infty$ then

$V_0(1, \pi)$ is a rational number. Therefore, if $V_0(1, \sigma)$ is an irrational number for a policy σ then $V_0(1, \pi) \neq V_0(1, \sigma)$ for a policy π which is a $(0, N)$-policy for some $N < \infty$.

We remark the that sets $U(x)$ are convex and compact in the following cases: (i) finite horizon problems (this follows from Corollary 3.4, Lemma 3.5, and the construction in the proof of Theorem 4.1); (ii) infinite horizon problems with the standard total discounted rewards (Corollary 3.4 and Lemma 3.5); (ii) infinite horizon problems with the lower limits of average rewards per unit time (Hordijk and Kallenberg 1984).

For a finite horizon problem, Lemmas 7.1, 7.4, and Theorem 4.1 imply results similar to Theorems 7.2, 7.4, and Corollary 7.3 on the existence of randomized Markov policies π of order M for boundary and Pareto optimal points and of order $M + 1$ for arbitrary points. For a standard discounted infinite horizon problem, Lemmas 7.1, 7.4, and Theorem 4.2 imply results similar to Theorems 7.2, 7.4, and Corollary 7.3 on the existence of M-randomized stationary policies π for boundary and Pareto optimal points and $(M + 1)$-randomized stationary policies π for arbitrary points. Similar results are correct for criteria of lower limits of average rewards per unit time, if all Markov chains on $\mathbf{X}$, defined by stationary policies, have the same number of ergodic classes. This follows from Theorems 7.2, 7.4, Corollary 7.3, and Hordijk and Kallenberg (1984).

8. Computation of optimal constrained policies.

In this section we formulate an algorithm for the approximate solution of problem (2.1)–(2.2). We say that, given $\epsilon \geq 0$, a policy π is ϵ-optimal for problem (2.1)–(2.2) if this policy is feasible and $V_0(x, \pi) \geq V_0(x) - \epsilon$. A policy π is called approximately ϵ-optimal if $V_m(x, \pi) \geq V_m(x) - \epsilon$ for all $m = 0, \ldots, M$. We remark that an approximately ϵ-optimal policy may be infeasible. However, in many applications the constraints have an economical or reliability interpretation. Therefore, from a practical point of view, it is sufficient to find an approximate ϵ-optimal policy for some small positive ϵ.

We consider the following algorithm for the approximate solution of problem (2.1)–(2.2).

Algorithm 8.1. *(Computation of ϵ- and approximately ϵ-optimal (M, N)-optimal policies.)* Let $\epsilon > 0$ be given.

1. Choose an arbitrary stationary policy ϕ.
2. Choose $N \geq 0$ such that $KL\beta^N/(1 - \beta) \leq \epsilon$, where

$$L = r - \min\{r_{mk}(z, \phi(z)) :\ m = 0, \ldots, M,\ k = 1, \ldots, K,\ z \in \mathbf{X}\}$$

where r and β are defined in (6.2).

3. Apply algorithm 4.3 for a finite horizon problem (2.1)–(2.2) with criteria (2.7), where the rewards $R_m(z, n, a)$ are defined by (2.4) for $n = 0, \ldots, N-2$ and $R_m(z, N-1, a)$ are defined by (6.3), where $m = 0, \ldots, M$, $z \in \mathbf{X}$, and $a \in \mathbf{A}$.
4. If the finite horizon problem is feasible, let $\phi_n(\cdot|z)$ be a solution of Algorithm 4.3, where $z \in \mathbf{X}$, $n = 0, \ldots, N-1$. Consider the (M, N)-policy π which coincides with $\phi_n(\cdot|\cdot)$ for $n < N$ and coincides with ϕ for $n \geq N$. This policy is ϵ-optimal.
5. If the finite horizon problem is not feasible, consider a similar finite horizon problem with the constraints c_m in the right hand side of (2.2) replaced by $c_m - \epsilon$, $m = 0, \ldots, M$.
6. If the new problem is feasible, the (M, N)-policy π constructed from its solution similarly to step (iv) is approximately ϵ-optimal.
7. If the new problem is not feasible, the original problem is not feasible.

We note that weighted discounted problems are equivalent to standard discounted problems with an extended state space; Feinberg and Shwartz (1991). Altman (1993, 1991) proved that, under some conditions, optimal and nearly optimal policies for finite horizon approximations of infinite horizon models converge to optimal policies for infinite horizon problems. Under some additional conditions, Altman's results imply the convergence of the ϵ_i-optimal policies for finite horizon weighted discounted problems to the optimal policy for the infinite horizon problem when $\epsilon_i \to 0$ as $i \to \infty$. For example, Theorems 4.1 and 3.1 in Altman (1991) provide the procedure for the construction of an optimal policy, if $V_m(x, \pi[i]) > c_m$ for all $m = 1, \ldots, M$ and for all i large enough, where $\pi[i]$ is a policy obtained from the Algorithm 8.1 for $\epsilon = \epsilon_i \to 0$ as $i \to \infty$, and if the sequence $\pi[i]$ satisfies some convergence conditions.

Acknowledgments

A part of this research was done when the first author visited the Technion. The research of the second author was supported in part by the fund for promotion of research at the Technion. The authors thank Joe Mitchell for useful discussion on the approximation of internal points of convex polytopes.

References

Altman, E. (1993). Asymptotic Properties of Constrained Markov Decision Processes. *ZOR — Methods and Models of Oper. Res.* **37**, 151–170.

______ (1991). Denumerable Constrained Markov Decision Problems and Finite Approximations. Preprint, to appear in *Math. Oper. Res.*

______ and Shwartz, A. (1991). Adaptive Control of Constrained Markov Chains: Criteria and Policies. *Annals of Oper. Res.* **28**, 101–134.

______ and _______ (1991a). Sensitivity of Constrained Markov Decision Processes. *Annals of Operations Research* **32**, 1–22.

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.

Feinberg, E.A. (1982). Controlled Markov Processes with Arbitrary Numerical Criteria. *Theory Probability and its Applications* **27** 486–503.

_______ and Shwartz, A. (1991). Markov Decision Models with weighted Discounted Criteria. Preprint, to appear in *Math. Oper. Res.*

Fernández-Gaucherand, E., Ghosh, M.K., and Marcus, S.I. (1990). Controlled Markov Processes on the Infinite Planning Horizon with a Weighted Cost Criterion. Preprint, to appear in *ZOR — Methods and Models of Oper. Res.*

Filar, J. and Vrieze, O.J. (1992). Weighted Reward Criteria in Competitive Markov Decision Processes. *ZOR — Methods and Models of Oper. Res.* **36** 343–458.

Frid, E.B. (1972). On Optimal Strategies in Control Problems with Constraints. *Theory Probability and its Applications* **17** 188–192.

Ghosh, M.K. and Marcus, S.I. (1991). Infinite Horizon Controlled Diffusion Problems with Some Nonstandard Criteria, *J. Mathematical Systems, Estimation and Control* **1** 45–69.

Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tracts No. 51, Mathematisch Centrum, Amsterdam.

__________ and Kallenberg, L.C.M. (1984). Constrained undiscounted Stochastic Dynamic Programming, *Math. Oper. Res.* **9** 276–289.

Kallenberg, L.C.M. (1981). Unconstrained and Constrained Dynamic Programming over a Finite Horizon. Report No. 81-46, Institute of Applied Math. and Computer Sci., University of Leiden,

Leiden, The Netherlands.

__________ (1981). *Linear Programming and Finite Markovian Control Problems.* Mathematical Centre Tracts No 148, Mathematisch Centrum, Amsterdam.

Krass, D. (1989). *Contributions to the Theory and Applications of Markov Decision Processes,* Ph.D. Thesis, Mathematical Sciences, The Johns Hopkins University.

Krass, D., Filar, J. A., and Sinha, S. S. (1992). A Weighted Markov Decision Process. *Oper. Res.* **40**, 1180–1187.

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.

Ross, K.W. (1989). Randomized and Past Dependent Policies for Markov Decision Processes with Finite Action Set. *Oper. Res.* **37** 474–477.

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.

Sobel, M. J. (1991). Discounting and Risk Neutrality. Preprint.

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

Tanaka, K. (1991). On Discounted Dynamic Programming with Constraints. *J. Math. Anal. Appl.* **155** 264–277.

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