Optimality of Deterministic Policies for Certain Stochastic Control Problems with Multiple Criteria and Constraints

Eugene A. Feinberg¹

State University of New York at Stony Brook, Stony Brook, NY 11794-3600
Eugene.Feinberg@sunysb.edu

For single-criterion stochastic control and sequential decision problems, optimal policies, if they exist, are typically nonrandomized. For problems with multiple criteria and constraints, optimal nonrandomized policies may not exist and, if optimal policies exist, they are typically randomized. In this paper we discuss certain conditions that lead to optimality of nonrandomized policies. In the most interesting situations, these conditions do not impose convexity assumptions on the action sets and reward functions.

1 Introduction

In many applications, the system performance is measured by multiple criteria. For example, in finance, such criteria measure returns and risks, in manufacturing such criteria may be production volumes, quality of outputs, and costs, in service operations performance criteria include service levels and operating costs.

For problems with multiple criteria, the natural approach is to optimize one of the criteria subject to the inequality constraints on the other criteria. In other words, for a problem with $K + 1$ criteria $W_0(\pi), W_1(\pi), \dots, W_K(\pi)$, where π is a policy, the natural approach is to find a policy π that is a solution to the following problem

$$\text{maximize } W_0(\pi) \tag{1}$$

subject to

$$W_k(\pi) \geq C_k, \quad k = 1, \dots, K, \tag{2}$$

where $C_1, \dots, C_K$ are given numbers. For example, since it is possible to consider $W_{k+1}(\pi) = -W_k(\pi)$, this approach can be used to find policies satisfying interval constraints $a \leq W_k(\pi) \leq b$.

Optimal solutions of problem (1, 2), if they exist, are typically randomized with the number of randomization procedures limited by the number of

constraints K ; see [1, 16]. If there are no constraints, i.e. $K = 0$, optimal policies are nonrandomized. The following simple example illustrates that it is possible that any optimal policy for a constrained problem is randomized.

Consider a one-step problem when a decision-maker chooses among two decisions a and b . There are two reward functions r_0 and r_1 defined as $r_0(a) = r_1(b) = 0$ and $r_1(a) = r_0(b) = 1$. The decision-maker selects action a with probability $\pi(a)$ and action b with probability $\pi(b)$, where $\pi(a) + \pi(b) = 1$. The criteria are $W_k(\pi) = \pi(a)r_k(a) + \pi(b)r_k(b)$, $k = 0, 1$. Then the problem

$$\text{maximize } W_0(\pi) \tag{3}$$

subject to

$$W_1(\pi) \geq 1/2 \tag{4}$$

is equivalent to the following linear program (LP)

$$\text{maximize } \pi(b)$$

subject to

$$\begin{aligned} \pi(a) &\geq 1/2, \\ \pi(a) + \pi(b) &= 1, \\ \pi(a) \geq 0, \pi(b) &\geq 0. \end{aligned}$$

This LP has the unique optimal solution $\pi(a) = \pi(b) = 1/2$. Therefore, the optimal policy is randomized.

In many applications, implementation of randomized policies is not natural. In many cases, it is more natural to apply nonrandomized policies when they are optimal. In addition, it appears that the use of randomization procedures increases the variance of the performance criteria. Also, from the computational point of view, finding the best randomized policy in many cases is easy, because this can be done by using linear programming. Finding the best nonrandomized policy may be a computationally intractable problem. For example, finding the best nonrandomized stationary policy for constrained dynamic programming is an NP-hard problem [5]. Typically, when nonrandomized policies are optimal, an optimal nonrandomized policy can be computed by a simple transformation of an optimal randomized policy. Thus, computing optimal nonrandomized policies becomes a computationally tractable problem when nonrandomized policies are optimal.

In this article, we discuss the situations when nonrandomized policies are optimal for problems with multiple criteria and constraints. Of course, nonrandomized policies are at least as good as randomized policies when actions sets are convex subsets of a linear space and reward functions are concave. This situation is trivial and we concentrate on the models when neither the convexity of the action sets nor the concavity of reward functions is assumed. In particular, we consider the following three cases: (i) unichain Markov Decision Processes (MDPs) with average rewards per unit time, (ii) continuous-time

MDPs, and (iii) nonatomic dynamic programming. In case (i), it is possible to achieve optimal state-action frequencies by selecting different actions at different visits to the same state, in case (ii) it is possible to change actions between jumps instead of choosing actions randomly, and in case (iii) the non-atomicity of initial and transition probabilities implies that for any randomized policy there exists a nonrandomized policy with the same performance vector.

2 Discrete-time MDPs with average rewards per unit time

Consider a discrete-time MDP with finite state and action sets and with average rewards per unit time [1, 16, 17]. This model is defined by the objects $\{X, A, A(\cdot), p, K, r, \cdot\}$, where

- (i) X is a finite *state space*;
- (ii) A is a finite *action set*;
- (iii) $A(x)$ are the *sets of available actions* at state $x \in X$, where $A(x) \subseteq A$;
- (iv) $p(y|x, a)$ is the one-step *transition probability*, i.e. $p(\cdot|x, a) \geq 0$ is a probability distribution on X for each $x \in X$ and $a \in A(x)$;
- (v) a finite nonnegative integer K is the *number of constraints*;
- (vi) $r_k(x, a)$, $k = 0, \dots, K$, is the one-step *reward* according to the k^{th} criterion if an action $a \in A(x)$ is selected at a state $x \in X$.

Let $H_n = X \times (A \times X)^n$ be the set of *trajectories* up to time $n = 0, 1, \dots$. A *policy* π is a sequence $(\pi_0, \pi_1, \dots)$ of transition probabilities from H_n to A such that $\pi_n(A(x_n)|x_0, a_0, \dots, x_n) = 1$. A policy π is called *nonrandomized* if $\pi_n(a|h_n) \in \{0, 1\}$ for all $n = 0, 1, \dots$, $h_n \in H_n$, and $a \in A$. Equivalently, nonrandomized policy σ is defined by a sequence of measurable mappings $\sigma_n : H_n \rightarrow A$ such that $\sigma_n(x_0, a_0, \dots, x_n) \in A(x_n)$, $n = 0, 1, \dots$. A policy π is called *randomized stationary* if $\pi_n(a_n|x_0, a_0, \dots, x_n)$ depends only on the value of x_n , i.e. there exists a transition probability π from X to A such that $\pi_n(a_n|x_0, a_0, \dots, x_n) = \pi(a_n|x_n)$. A nonrandomized stationary policy is called *stationary*.

Any initial distribution μ and any policy π define a probability measure P_μ^π on the set of infinite trajectories H_∞ endowed with its Borel σ -field. We denote by E_μ^π expectations with respect to this measure. We shall write P_x^π and E_x^π instead of P_μ^π and E_μ^π if x is the initial state, i.e. $\mu(x) = 1$.

For an initial distribution μ , and policy π , the *average rewards per unit time* are defined as

$$W_k(\mu, \pi) = \liminf_{N \rightarrow \infty} \frac{1}{N} E_\mu^\pi \sum_{n=0}^{N-1} r_k(x_n, a_n), \quad k = 0, 1, \dots, K.$$

For a fixed initial distribution μ , consider problem (1, 2) when $W_k(\pi) = W_k(\mu, \pi)$, $k = 0, \dots, K$. So, the goal is to maximize the expected total rewards

$W_0(\mu, \pi)$ subject to the constraints on the expected total rewards $W_k(\mu, \pi)$, $k = 1, \dots, K$. Though an optimal policy exists for this problem when the state and action spaces are finite, it may have a complicated form. In particular, it may be nonstationary [12, 14]. However, optimal randomized stationary policies exist under the following condition.

Unichain Condition. *Any nonrandomized stationary policy defines a Markov chain on X with one recurrent class (and possible transient states).*

According to [12, 18], if the Unichain Condition holds then there exists a randomized stationary policy that is optimal for any initial distribution. To find such a policy, we need to solve the following LP with variables $u_{x,a}$, where $x \in X$, $a \in A(x)$:

$$\text{maximize } \sum_{x \in X} \sum_{a \in A(x)} r_0(x, a) u_{x,a}$$

subject to

$$\begin{aligned} \sum_{x \in X} \sum_{a \in A(x)} r_k(x, a) u_{x,a} &\geq C_k, & k = 1, \dots, K, \\ \sum_{a \in A(y)} u_{y,a} - \sum_{x \in X} \sum_{a \in A(x)} p(y, a, x) u_{x,a} &= 0, & y \in X, \\ \sum_{x \in X} \sum_{a \in A(x)} u_{x,a} &= 1, \\ u_{x,a} &\geq 0, & x \in X, a \in A(x). \end{aligned}$$

Let u be an optimal solution of the above LP. Let

$$u_x = \sum_{a \in A(x)} u_{x,a}, \quad x \in X.$$

Then, if the Unichain Condition holds, the following formula defines an optimal randomized stationary policy π

$$\pi(a|x) = \begin{cases} \frac{u_{x,a}}{u_x} & \text{if } u_x > 0, \\ \text{arbitrary} & \text{if } u_x = 0. \end{cases}$$

The policy π may not be nonrandomized. If $u_x = 0$, it is natural to select π being nonrandomized at x . If u is a basic optimal solution of the above LP, there are at most $|X| + K$ nonzero state-action pairs (x, a) such that $\pi(a|x) > 0$ [18], where $|E|$ is the number of elements of a finite set E . For a nonrandomized stationary policy, the number of such pairs is $|X|$.

Ross [18] and Altman and Schwartz [2, 3] investigated approaches to construct a nonrandomized stationary policy σ such that $W_k(\mu, \sigma) = W_k(\mu, \pi)$ for any reward function r_k . If such a policy is constructed, then it is optimal. Ross [18] studied the case $K = 1$ when the Unichain Condition holds. Altman and Schwartz [2, 3] also assumed the Unichain Condition and considered the

case of an arbitrary finite K . In addition, they studied MDPs with a countable state space X and finite state MDPs with finite action sets. Altman and Schwartz [2, 3] introduced the concept of time-sharing policies. These policies combine several nonrandomized policies.

For an initial state distribution μ and a policy σ , a state-action frequency is defined as

$$f_\mu^\sigma(x, a) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} P_\mu^\sigma \{x_n = x, a_n = a\}, \quad x \in X, a \in A(x), \quad (5)$$

if this limit exists for all $x \in X$ and for all $a \in A(x)$. In particular, state-action frequencies exist for randomized stationary policies for countable-state MDPs with countable action sets. If for two policies σ_1 and σ_2 the limits in the definition of the vectors of state-action frequencies (5) exist for the initial distribution μ and these vectors are equal for σ_1 and σ_2 , then $W(\mu, \sigma_1) = W(\mu, \sigma_2)$ for any any bounded or nonnegative reward function r . In [18, 2, 3], the policy σ is constructed in the way that the state-action frequencies for σ exist and equal to state-action frequencies for an optimal randomized stationary policy π , and these frequencies do not depend on the initial distribution μ .

For a finite history $h_n = x_0, a_0, \dots, x_n$, define the empirical frequencies

$$\begin{aligned} N(h_n; x, a) &= \sum_{t=0}^{n-1} I\{x_t = x, a_t = a\}, & x \in X, a \in A(x), \\ N(h_n, x) &= \sum_{t=0}^{n-1} I\{x_t = x\}, & x \in X. \end{aligned}$$

For the optimal solution u of the LP, define the nonrandomized policy σ by

$$\sigma_n(h_n) = \operatorname{argmax}_{a \in A(i)} \left\{ u_{i,a} - \frac{N(h_n; x, a) + 1}{N(h_n; x) + 1} \right\},$$

where ties are broken arbitrarily. Then for finite X and A , if the Unichain Condition holds, then σ has the same state-action frequencies as π and therefore σ is an optimal nonrandomized policy. Feinberg and Curry [8] used the above form of the optimal nonrandomized policy σ in a heuristic algorithm for a so-called Generalized Pinwheel problem, which is an NP-hard scheduling problem.

3 Continuous-time MDPs

Consider a continuous-time MDP with a finite or countable state space [6, 7, 13]. This model is defined by the objects $\{X, A, A(\cdot), p, K, r, R\}$, where

- (i) X is a finite or countable *state space*;

(ii) A is a Borel *action set* (a measurable subset of a complete separable metric space);

(iii) $A(x)$ are the *sets of available actions* at state $x \in X$, where $A(x)$ are measurable subsets of A ;

(iv) $q(y|x, a)$ are *jump intensities*, i.e. $q(y|x, a) \geq 0$, $q(x|x, a) = 0$, and $q(x, a) := \sum_{y \in X} q(y|x, a) \leq C < \infty$ for all $x \in X$ and $a \in A(x)$, and, in addition, the functions $q(y|x, a)$ are measurable on $A(x)$ for all $x, y \in X$;

(v) a finite nonnegative integer K is the *number of constraints*;

(vi) $r_k(x, a)$, $k = 0, \dots, K$, is the *reward rate* according to the k^{th} criterion if an action $a \in A(x)$ is selected at a state $x \in X$;

(vii) $R_k(x, a)$, $k = 0, \dots, K$, is the *reward* according to the k^{th} criterion if a jump from state $x \in X$ to state $y \in Y$ occurs while an action $a \in A(x)$ is selected at a state $x \in X$.

Let $\mathbb{R}_+ = [0, \infty)$ and $H_n = X \times (\mathbb{R}_+ \times X)^n$, $n = 0, 1, \dots$. Let $H = \cup_{0 \leq n < \infty} H_n$. The sets H_n are endowed with the σ -fields $\mathcal{H}_n$ which are the products of the σ -fields on X and on $\mathbb{R}_+$, where the σ -field on X is the set of the subsets of X and the σ -field on $\mathbb{R}_+$ is the Borel σ -field on $\mathbb{R}_+$. Let $\mathcal{H}$ be the minimal σ -field on H that contains $\mathcal{H}_n$, $n = 0, 1, \dots$. We interpret $x_0, \xi_0, \dots, x_n \in H_n$ as the sequence of the first $n+1$ states $x_0, \dots, x_n$ and the first n sojourn times $\xi_0, \dots, \xi_{n-1}$ of a multivariate point process with the state space X . The jump epochs are $t_m = \sum_{i=0}^{m-1} \xi_i$, $i = 1, \dots, n$, with $t_0 = 0$.

A *randomized strategy* is a regular transition probability from $H \times \mathbb{R}_+$ to A such that $\pi(A(x_n)|x_0, \xi_0, \dots, x_n, s) = 1$ for any $x_0, \xi_0, \dots, x_n \in H$, $n = 0, 1, \dots$ and for any $s \geq 0$. In other words, $\pi(\cdot|x_0, \xi_0, \dots, x_n, s)$ is a probability measure on $A(x_n)$ for each $x_0, \xi_0, \dots, x_n \in H$ and $0 \leq s < \infty$, and $\pi(E|\cdot)$ is a Borel function on $H \times \mathbb{R}_+$ for any measurable subset B of A . We interpret s as the time elapsed since the last jump. Thus, the decision at any time depends on the previous states and jump epochs, and on the time elapsed since the last jump.

Let $H_\infty = X \times \mathbb{R}_+ \times X \times \mathbb{R}_+ \dots$ be the infinite product of X and $\mathbb{R}_+$. Let $\mathcal{H}_n$ be the countable product of the σ -fields on X and $\mathbb{R}_+$. We define $\Omega = \cup_{0 \leq n \leq \infty} H_n$ and $\mathcal{F}$ as the minimal σ -field containing $\{\mathcal{H}_n, 0 \leq n \leq \infty\}$.

We interpret Ω as the sample space of the multivariate point process without accumulation points. In particular, H_n represent the sets of trajectories with n jumps, $n = 0, 1, \dots$, or $n = \infty$. We set $t_m = \infty$ when $m > n$ and $\omega \in H_n$ with $n < \infty$. In particular, the value $X(t)$ of the process at time $t > 0$ is defined as $X(t) = \sum_{0 \leq n < \infty} I\{t_n < t \leq t_{n+1}\}x_n$.

A randomized strategy π and an initial state distribution μ define a unique probability measure P_μ^π on $(\Omega, \mathcal{F})$ such that $P_\mu^\pi(x_0 = x) = \mu(x)$, $x \in X$, and the compensator of the random measure corresponding to P_μ^π is

$$\nu^\pi(\omega; dt, j) = \sum_{n \geq 0} I\{t_n < t \leq t_{n+1}\} \int_A q(j|x_n, a) \pi(da|x_0, \xi_0, \dots, x_n, t - t_n) dt,$$

where $t_n = \sum_{i=0}^{n-1} t_i$; see [11, 13, 15]. The assumption $q(x, a) \leq C < \infty$ implies that the process defined by P_μ^π does not have accumulation points.

A *nonrandomized strategy* ϕ is defined by a measurable mapping from $H \times \mathbb{R}_+$ to A such that $\phi(x_0, \xi_0, \dots, x_n, s) \in A(x_n)$. For a nonrandomized policy, the measure P_μ^ϕ is defined by $P_\mu^\phi(x_0 = x) = \mu(x)$, $x \in X$, and by the compensator of the random measure corresponding to P_μ^π equal

$$\nu^\phi(\omega; dt, j) = \sum_{n \geq 0} I\{t_n < t \leq t_{n+1}\} q(j|x_n, \phi(x_0, \xi_0, \dots, x_n, t - t_n)) dt.$$

A nonrandomized strategy is called *switching stationary* if $\phi(x_0, \xi_0, \dots, x_n, s) = \phi(x_n, s)$ and, in addition, the function $\phi : X \rightarrow \mathbb{R}_+$ is piecewise-constant in s and has a finite number of discontinuity points, where the discrete topology is considered on X . In other words, there is a finite subset Y of X such that the function $\phi(x, s)$ is constant in s for each $x \in X \setminus Y$ and $\phi(x, s)$ is piecewise-constant in s and has a finite number of jumps for each $x \in Y$. So, the switching stationary policy may change actions between jumps only at a finite number of states and for each such state x , where it changes actions, there is a finite number of times $0 < S_1(x) < S_2(x) < \dots < S_j(x)$ such that the policy changes actions when the time $S_i(x)$ elapsed since the last jump.

Let E_μ^π be the expectation with respect to the measure P_μ^π . Let $N(t)$ be the number of jumps up to time t , $N(t) = \max\{n \geq 0 | t_n \leq t\}$. Then the assumption $q(x, a) \leq C < \infty$ implies that $N(t) < \infty$ for all $0 \leq t < \infty$.

Discounted total rewards. Let $\alpha > 0$ be the discount rate. For a randomized strategy π , the expected total rewards are

$$\begin{aligned} W_k(\mu, \pi) &= E_\mu^\pi \sum_{n=0}^{\infty} [e^{-\alpha t_n} \int_{A(x_n)} R_k(x_n, a, x_{n+1}) \pi(da|\omega, t_{n+1}) \\ &\quad + \int_{t_n}^{t_{n+1}} e^{-\alpha t} \int_{A(x_n)} r_k(x_n, a) \pi(da|\omega, t)], \quad k = 0, \dots, K, \end{aligned}$$

where $\pi(da|\omega, t) = \sum_{n \geq 0} I\{t_n < t \leq t_{n+1}\} \pi(da|x_0, \xi_0, \dots, x_n, t - t_n) dt$.

Consider problem $(\bar{1}, 2)$ with $W_k(\pi) = W_k(\mu, \pi)$ when all the action sets $A(x)$ are compact, the functions R_k and r_k , $k = 0, \dots, K$, are bounded above and continuous in a , and the functions $q(y|x, a)$ and $q(x, a)$ are continuous in a . This problem was studied in [7]. Three natural forms of optimal policies were described in [7] when this problem is feasible. In particular, for a feasible problem, there exists an optimal strategy that randomly selects actions at jump epoch and keeps them unchanged between jumps. The strategies of this type are called randomized stationary policies. A randomized stationary policy is defined by the distributions $\sigma(da|x)$ concentrated on $A(x)$, where the action a is always selected with the probability $\pi(da|x_n)$ when the process jumps to a state x_n and this action is used to control the process until the next jump

epoch. Formula (8.6) in [7] shows how a randomized stationary policy can be presented as a randomized strategy π described above. A *nonrandomized stationary policy* ϕ , where ϕ is a function from X to A satisfying $\phi(x) \in A(x)$ for all $x \in X$, always uses action $\phi(x)$ at state $x \in X$.

As was shown in [7], any m -randomized stationary policy σ can be transformed into a (nonrandomized) switching stationary strategy ϕ such that $W(\mu, \phi) = W(\mu, \sigma)$ for any bounded above or bounded below reward functions r and R . For $x \in X$ consider the actions $a^1, \dots, a^j$ from $A(x)$ such that $\sigma(a^i|x) > 0$, $i = 1, \dots, j$, and $\sum_{i=1}^j \sigma(a^i|x) = 1$. Then $\phi(x, s) = a^1$ if $j = 1$ and for the case $j > 1$

$$\phi(x, s) = a_i \quad \text{for } S_{i-1}(x) < s \leq S_i(x), \quad i = 1, \dots, j, \quad (6)$$

where $S_0(x) = 0$ and $S_i(x) = \sum_{\ell=1}^i s_\ell(x)$, $i = 1, \dots, j$, with

$$s_i(x) = -\frac{1}{\alpha + q(x, a^i)} \ln \left(1 - \frac{\sigma(a^i|x)}{\sum_{\ell=1}^j \sigma(a^\ell|x)} \right). \quad (7)$$

The equality $W(\mu, \phi) = W(\mu, \sigma)$ is based on the fact that the occupation measures for continuous-time MDPs introduced in [7] are equal for σ and ϕ under any initial state distribution. The optimality of nonrandomized strategies can be summarized in the following theorem based on the results from [7].

Theorem 1. *Consider a discounted continuous-time MDP. Let σ be an m -randomized stationary policy, $m = 0, 1, \dots$, and ϕ be an m -switching stationary strategy defined in (6, 7). Then $W_k(\mu, \phi) = W_k(\mu, \sigma)$ for all $k = 0, \dots, K$ and for all initial state distributions μ .*

If problem (1, 2) is feasible, all the action sets $A(x)$, $x \in X$, are compact, the reward functions R_k and r_k , $k = 0, \dots, K$ are bounded above and upper semi-continuous in a , and the functions $q(y|x, a)$ and $q(x, a)$ are continuous in a , then there exist optimal k -randomized stationary policies and optimal k -switching stationary strategies.

Similar to the discrete time, LPs can be used to compute optimal k -randomized stationary policies and optimal k -switching stationary strategies; see [7] for details.

Average rewards per unit time. Let the sets X and A be finite. Consider average rewards per unit time

$$\begin{aligned} W_k(\mu, \pi) = \liminf_{T \rightarrow \infty} \frac{1}{T} E_\mu^\pi \Big[& \sum_{n=0}^{N(T)} \int_{A(x_n)} R_k(x_n, a, x_{n+1}) \pi(da|\omega, t_{n+1}) \\ & + \int_0^T \int_{A(x_n)} r_k(X(t), a) \pi(da|\omega, t) \Big], \quad k = 0, \dots, K. \end{aligned}$$

Let the Unichain Condition hold. Then for any randomized stationary policy π and for any switching stationary strategy π , $W_k(\mu, \pi)$ does not depend on μ . So, for such policies, we set $W_k(\pi) = W_k(\mu, \pi)$, $k = 0, \dots, K$.

For an m -randomized stationary policy σ we can consider the m -switching policy ϕ defined by (6, 7) with $\alpha = 0$. Then, according to [6], $W(\phi) = W(\sigma)$ for any rewards r and R . The following theorem summarizes the results on average rewards per unit time from [6].

Theorem 2. *Let X and A be finite sets and let the Unichain Condition hold. Consider average rewards per unit time.*

(i) *For an m -randomized stationary policy σ consider an m -switching stationary strategy ϕ defined in (6, 7) with $\alpha = 0$. Then $W_k(\phi) = W_k(\sigma)$ for all $k = 0, \dots, K$.*

(ii) *If problem (1, 2) is feasible then there exists an optimal K -switching stationary strategy.*

(iii) *If problem (1, 2) is feasible and $q(x, a) > 0$ for all $x \in X$ and for all $a \in A(x)$ then there exists an optimal K -randomized stationary policy.*

If $q(x, a) = 0$ for some x and a then optimal randomized stationary policies may not exist even if the problem is feasible [6, Example 3.1]. Optimal K -randomized stationary policies and optimal K -switching stationary strategies can be found by using LPs described in [6].

4 Nonatomic MDPs

Consider an MDP $\{X, A, A(\cdot), p, K, r, \cdot\}$, introduced in Section 2 for finite X and A . In this section we consider the situation when X and A are Borel spaces (measurable subsets of complete separable metric spaces) and the standard measurability conditions are satisfied. These conditions are:

(a) the graph of the set-valued mapping $x \rightarrow A(x)$ of X is a measurable subset of $X \times A$ allowing a measurable selection, i.e. there exists a measurable function $\varphi: X \rightarrow A$ such that $\varphi(x) \in A(x)$ for all $x \in X$;

(b) p is a regular transition probability from $X \times A$ to X ;

(c) $r_k(x, a)$ are measurable functions on $X \times A$, $k = 0, \dots, K$.

Let $C^+ = \max\{C, 0\}$, $C^- = \min\{C, 0\}$ and let β be a discount factor, $0 \leq \beta \leq 1$. For a policy π and for an initial probability measure μ on X , we define for $k = 0, \dots, K$ the expected total rewards

$$W_{k,+}(\mu, \pi) = E_{\mu}^{\pi} \left[\sum_{n=0}^{\infty} \beta^n (r_k(x_n, a_n))^+ \right], \quad W_{k,-}(\mu, \pi) = E_{\mu}^{\pi} \left[\sum_{n=0}^{\infty} \beta^n (r_k(x_n, a_n))^- \right],$$

$$W_k(\mu, \pi) = W_{k,+}(\mu, \pi) + W_{k,-}(\mu, \pi), \quad (8)$$

where $+\infty - \infty = -\infty$.

We fix the initial probability distribution μ on X . We recall that a probability measure P on a Borel space Y is called *nonatomic* if $P(\{y\}) = 0$ for any $y \in Y$.

Nonatomicity Assumption. *The initial probability measure μ and all the transition probabilities $p(\cdot|x, a)$, $x \in X$ and $a \in A(x)$ are nonatomic.*

A nonrandomized policy ϕ is called *Markov* if $\phi_n(x_0, a_0, \dots, x_n) = \phi_n(x_n)$ for all $n = 0, 1, \dots$.

The following statement is a particular case of [9, Theorem 2.1], where a nonhomogeneous MDP (the transition probabilities and rewards may depend on time) was considered.

Theorem 3. *If the Nonatomicity Assumption holds then for any policy π there exists a Markov policy ϕ such that $W_k(\mu, \phi) = W_k(\mu, \pi)$ for all $k = 0, \dots, K$.*

Unlike the results in Sections 2 and 3, the Nonatomicity Assumption does not imply that the occupation measures are equal and it is essential that K is finite. The proof of Theorem 3 uses Lyapunov's theorem; see [9] for details. Combined with sufficient conditions for the existence of optimal solutions to problem (1, 2) for $W_k(\pi) = W_k(\mu, \pi)$, Theorem 3 implies the existence of optimal (nonrandomized) Markov policies; see [9, Section 5]

Application to Statistical Decision Theory. Let X and A be Borel sets. Similar to MDPs, X is the state spaces and A is the decision space. For each $x \in X$, a Borel subset $A(x)$ of A (the set of decisions available at x) is given. The sets $A(x)$ satisfy the conditions described above in this section.

The initial distribution μ is not known, but it is known that it is equal to one of the measures $\mu_1, \dots, \mu_N$, where $N = 1, 2, \dots$. For each measure μ_n , $n = 1, \dots, N$, there is a gain function $\rho(\mu_n, x, a) = (\rho_1(\mu_n, x, a), \dots, \rho_M(\mu_n, x, a))$, $M = 1, 2, \dots$, which is a Borel mapping of $X \times A$ to $\mathbb{R}^M$.

A *decision rule* π is a regular transition probability from X to A such that $\pi(A(x)|x) = 1$ for all $x \in X$. A decision rule is called *nonrandomized* if for each $x \in X$ the measure $\pi(\cdot|x)$ is concentrated at one point. A nonrandomized decision rule π is defined by a measurable mapping $\varphi : X \rightarrow A$ such that $\varphi(x) \in A(x)$ and $\pi(\varphi(x)|x) = 1$, $x \in X$. We call such a mapping a *decision function* and denote it by φ .

Using the same agreements as in the definitions of integrals as in (8), for any decision rule π define

$$\mathcal{R}(\mu_n, \pi) = \int_X \int_A \rho(\mu_n, x, a) \pi(da|x) \mu_n(dx), \quad n = 1, \dots, N.$$

Theorem 4. [10, Theorem 1] *If all the measures $\mu_1, \dots, \mu_N$ are nonatomic then for each decision rule π there exists a decision function φ such that $\mathcal{R}(\mu_n, \varphi) = \mathcal{R}(\mu_n, \pi)$ for all $n = 1, \dots, N$.*

Dworetzky, Wald, and Wolfowitz [4] proved Theorem 4 when A is a finite set and weaker results were obtained in [4] for an infinite set A .

Financial Application. The following example from [9] is based on the version of Theorem 3 when transition probabilities and rewards may depend on the time parameter.

An investor has an option to sell a portfolio at epoch $t = 1, \dots, T$, where $T = 1, 2, \dots$. The value of the portfolio at epoch $t = 0, 1, 2, \dots$ is $z_t \in \mathbb{R}_+$. Suppose that $z_0 \geq 0$ is given and the value of z_{t+1} is defined by transition probabilities $q_t(dz_{t+1}|z_t)$, $t = 0, 1, \dots$. We assume that $q_t(\cdot|z_t)$ are nonatomic and weakly continuous.

At each epoch $t = 1, 2, \dots, T$, the investor has two options: to sell the whole portfolio or to keep it. We construct a Markov decision process for this problem. Let $X = \{0, 1\} \times \mathbb{R}$ and $A = \{0, 1\}$. Action 0 (1) means to hold (to sell) the portfolio. The state of the system is $x_t = (0, z_{t+1})$ ($x_t = (1, z_{t+1})$) if the portfolio has not been sold (has been sold). In particular, $x_0 = (0, z_1)$ has a nonatomic distribution μ defined by $\mu(0 \times B) = q_0(B|x_0)$ for any measurable subset B of $\mathbb{R}_+$. For $t = 0, 1, \dots$ we set $A_t((0, z)) = \{0, 1\}$; $A_t((1, z)) = \{0\}$. If at epoch $t = 0, 1, \dots$ the system is in state $x_t = (0, z)$ and action $a_t = 0$ is selected then the next state is $x_{t+1} = (0, y)$, where y has the distribution $q_{t+1}(dy|z)$. It does not matter what going on with the system after the portfolio is sold. To satisfy the Nonatomicity Assumption, we set that the system moves from state x_t to state $(1, y)$, where y has the distribution $q_{t+1}(dy|z)$, when $x_t = (1, z)$ or $x_t = (0, z)$ and $a_t = 1$.

Suppose that N -dimensional vectors $r_t(x, a)$ of measurable additive rewards (or losses) with values in $[-\infty, \infty]$ at steps $t = 0, 1, \dots$ are given. Theorem 4 implies that (nonrandomized) Markov policies for this multicriterion problem are as good as general policies.

We consider the problem when the investor's goal is to maximize the expected discounted value of the sold portfolio under the constraint that with at least probability $P > 0$ the discounted value of the sold portfolio is greater (or equal) than the given level C . For $t = 1, \dots, T - 1$, we define $r_t^1((i, z), a) = \beta^t \cdot (1 - i)az$; $r_t^2((i, z), a) = (1 - i)a \cdot I\{z \geq C/\beta^t\}$, $i = 0, 1$; $\beta \in (0, 1]$ is the given discount factor. We set $r_t^n(x, a) = 0$ when $t \geq T$. Then the problem can be formulated as

$$\text{maximize}_{\pi} W_1(\mu, \pi) \quad \text{s.t.} \quad W_2(\mu, \pi) \geq P. \quad (9)$$

Suppose that $q_t(\cdot|z)$ are concentrated on a finite interval $[0, D]$. Then we can set $X = \{0, 1\} \times [0, D]$ and [9, Condition 5.3] holds. Therefore, [9, Corollary 5.2] implies that if this problem is feasible then there exists an optimal (nonrandomized) Markov policy.

Acknowledgement. This research was partially supported by NSF grant DMI-0600538.

References

1. E. Altman: *Constrained Markov Decision Processes* (Chapman & Hall, Boca Raton London New York Washington, D.C. 1999)

2. E. Altman and A. Shwartz: Markov decision problems and state-action frequencies, *SIAM J. Control and Optimization*, **29**, 786-809, (1991)
3. E. Altman and A. Shwartz: Time-sharing policies for controlled Markov chains, *Operations Research*, **41**, 1116-1124, (1993)
4. A. Dvoretzky, A. Wald and J. Wolfowitz: Elimination of randomization in certain statistical decision procedures and zero-sum two-person games, *Ann. Math. Stat.*, **22**, 1-21, (1951)
5. E.A. Feinberg: Constrained discounted Markov decision processes and Hamiltonian cycles, *Mathematics of Operations Research*, **25**, 130-140, (2000)
6. E.A. Feinberg: Constrained finite continuous-time Markov decision processes with average rewards, *Proceedings of IEEE 2002 Conference on Decisions and Control* (Las Vegas, December 10-13, 2002) 3805-3810
7. E.A. Feinberg: Continuous time discounted jump Markov decision processes: a discrete-event approach, *Mathematics of Operations Research*, **29**, 492-524, (2004)
8. E.A. Feinberg and M.T. Curry: Generalized pinwheel problem, *Mathematical Methods of Operations Research*, **62**, 99-122, (2005)
9. E.A. Feinberg and A.B. Piunovskiy: Nonatomic total reward Markov decision processes with multiple criteria, *Journal of Mathematical Analysis and Applications*, **273**, 93-111, (2002)
10. E.A. Feinberg and A.B. Piunovskiy: On Dvoretzky-Wald-Wolfowitz theorem on nonrandomized statistical decisions, *Theory Probability and its Applications*, **50**, 463-466, (2006)
11. J. Jacod: Multivariate point processes: predictable projection, Radon-Nikodym derivatives, representation of martingales, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **31**, 235-253 (1975)
12. L.C.M. Kallenberg: *Linear Programming and Finite Markovian Control Problems* (Mathematical Center Tracts **148**, Amsterdam, The Netherlands 1983)
13. M. Yu. Kitaev and V.V. Rykov: *Controlled Queueing Systems* (CRC Press Boca Raton New York London Tokyo 1995)
14. D. Krass and O. J. Vrieze: Achieving target state-action frequencies in multichain average-reward Markov decision processes, *Mathematics of Operations Research*, **27**, 545-566, (2002)
15. R. S. Liptser and A.N. Shiryaev: *Theory of Martingales* (Kluwer Academic Publishers, Dordrecht Boston London 1989)
16. A.B. Piunovskiy: *Optimal Control of Random Sequences in Problems with Constraints* (Kluwer Academic Publishers, Dordrecht Boston London 1997)
17. M.L. Puterman: *Markov Decision Processes*, (John Wiley & Sons, New York Chichester Brisbane Toronto Singapore 1994)
18. K.W. Ross: Randomized and past-dependent policies for Markov decision processes with multiple constraints, *Operations Research*, **37**, 474-477, (1989)