Nonatomic total rewards Markov decision processes with multiple criteria

E. A. Feinberg [1]
Department of Applied Mathematics and Statistics
State University of New York at Stony Brook
Stony Brook, NY 11794-3600, USA
Eugene.Feinberg@sunysb.edu

A. B. Piunovskiy
Department of Mathematical Sciences
M & O Building, The University of Liverpool
Liverpool, L69 7ZL, UK
piunov@liverpool.ac.uk

Abstract

We consider a Markov decision process with an uncountable state space for which the vector performance functional has the form of expected total rewards. Under the single condition that the fixed initial distribution and transition probabilities are nonatomic, we prove that for any policy there is a nonrandomized Markov policy such that for these two policies the performance vectors are equal.

1 Introduction

This paper deals with a Markov Decision Process (MDP) with Borel state and action spaces and multiple performance criteria. Each criterion has the form of expected total rewards. A standard natural approach to such problems is to optimize one of these criteria under inequality constrains on other criteria. The well-known phenomenon for problems with constrains is that optimal strategies, if they exist, may be randomized and nonrandomized optimal strategies may not exist; see Altman [1], Frid [8], or Piunovskiy [10].

An MDP is called nonatomic if all transition probabilities are nonatomic and a nonatomic initial distribution is fixed. Feinberg and Piunovskiy [7] proved that if a nonatomic MDP satisfies continuity and compactness conditions then there exists an optimal nonrandomized Markov policy for a multiple criterion problem with constraints. This result was established in [7] as a corollary of the following fact: the performance set for nonrandomized Markov policies coincides with the performance set for all policies in a nonatomic MDP satisfying continuity and compactness conditions. Continuity and compactness conditions were essential for the proofs in Feinberg and Piunovskiy [7].

In this paper we prove that the performance set for nonrandomized Markov policies coincides with the performance set for all policies for an arbitrary nonatomic MDP with a vector criterion of expected total rewards. We recall that Borel space $(\Omega, \mathcal{B}(\Omega))$ is a measurable space isomorphic to a Polish space; see Bertsekas and Shreve [3], Dynkin and Yushkevich [5], or Feinberg and Piunovskiy [7] for details. Our proofs are based on the following result.

Lyapunov's Theorem (Barra [2]) Let $\{P_1, P_2, ..., P_N\}$ be a finite collection of nonatomic probability distributions on the Borel space $(\Omega, \mathcal{B}(\Omega))$. Then, for each random variable $\hat{\pi}(x)$ with values in $[0,1]$, there exists a measurable set $\Gamma \in \mathcal{B}(X)$ such that

$$P_n(\Gamma) = \int_X P_n(dx)\hat{\pi}(x), \quad n = 1, 2, ..., N.$$

2 Model description, notations, and main result

We consider an MDP $\{X, A, A(\cdot), p, r\}$, where

(i) X is a Borel state space;

(ii) A is a Borel action space;

(iii) $A_t(x)$ are sets of actions available at states $x \in X$ at epochs $t = 0, 1, \ldots$; it is assumed that for each t

[1] Research of this coauthor was partially supported by NSF Grant DMI-9908258

the graph $\mathrm{Gr}(A_t) = \{(x,a) : \ x \in X, a \in A(x)\}$ is a measurable subset of $A \times X$ and there exists a measurable mapping $\varphi : \ X \to A$ with $\varphi(x) \in A_t(x)$ for all $x \in X$. (In other words, the graphs of A_t are measurable and multifunctions $x \to A_t(x)$ can be uniformized, see [5, 12].);

(iv) $p_t(dy|x,a)$ are measurable transition probabilities from $X \times A$ to X at steps $t = 1, 2, \ldots$;

(v) $r_t(x,a) = (r_t^1, r_t^2, \ldots, r_t^N)$ are N-dimensional vectors of measurable rewards with values in $[-\infty, \infty]$ at steps $t = 0, 1, \ldots$ where N is a positive integer and $(x,a) \in X \times A$.

As usually, a policy π is a sequence of measurable transition probabilities $\pi_t(da|h_t)$ concentrated on the sets $A_t(x_t)$, where $h_t = x_0, a_0, \ldots, a_{t-1}, x_t$ is the observed history. Δ is the set of all policies. If transition probabilities π_t depend only on the current time and the current state, i.e. $\pi_t(\cdot|h_t) = \pi_t(\cdot|x_t)$ for all $t = 0, 1, \ldots$, then the policy π is called randomized Markov. If the measure π_t, for all $t = 0, 1, \ldots$, is concentrated at the point $\varphi_t(x_t) \in A_t(x_t)$, then the policy is called nonrandomized Markov and is denoted by φ. Δ^M is the set of all nonrandomized Markov policies.

According to the Ionescu Tulcea Theorem [5] an initial distribution μ on X and a policy π define a unique probability measure P_μ^π on the space of trajectories $H_\infty = (X \times A)^\infty$ which is called a strategic measure. We denote by E_μ^π expectations with respect to P_μ^π. Since the initial distribution μ is fixed, we usually omit the index μ. Let $\mathcal{D}$ be the set of all strategic measures with the initial measure μ and let $\mathcal{D}^M$ be the set of all strategic measures generated by nonrandomized Markov policies and the initial measure μ.

For a Borel space $(\Omega, \mathcal{B}(\Omega))$, we denote by $\mathcal{P}(\Omega)$ the set of all probability measures on it. The set $\mathcal{P}(\Omega)$ and the minimal σ-field thereon, with respect to which all functions $P(E)$ are measurable for all $E \in \mathcal{B}(\Omega)$), form a Borel space; Dynkin and Yushkevich [5, Appendix 5]. We also notice that $\mathcal{P}(\Omega)$ is a convex subset of the linear space of all signed finite measures on $(\Omega, \mathcal{B}(\Omega))$.

In what follows, $C^+ = \max\{C, 0\}$, $C^- = \min\{C, 0\}$;

$$R_+^n(P^\pi) = E^\pi \left[\sum_{t=0}^{\infty} (r_t^n(x_t, a_t))^+ \right],$$

$$R_-^n(P^\pi) = E^\pi \left[\sum_{t=0}^{\infty} (r_t^n(x_t, a_t))^- \right],$$

$$R^n(P^\pi) = R_+^n(P^\pi) + R_-^n(P^\pi),$$

where throughout this paper $+\infty - \infty = -\infty$. The performance of a policy π is evaluated by a vector

$$\mathbf{R}(P^\pi) = (R^1(P^\pi), R^2(P^\pi), \ldots, R^N(P^\pi)).$$

Let us introduce performance spaces

$$\mathcal{V} = \{\mathbf{R}(P^\pi), \ \pi \in \Delta\}, \quad \mathcal{V}^M = \{\mathbf{R}(P^\varphi), \ \varphi \in \Delta^M\}.$$

Infinite values $R^n(P^\pi) = \pm\infty$ can be obtained for some policies. Let $\mathbb{R}^N$ be the Euclidean space of N-dimensional vectors with finite coordinates. Then subsets $\mathcal{V} \cap \mathbb{R}^N$ and $\mathcal{V}^M \cap \mathbb{R}^N$ consist of vectors with finite elements.

We assume that the following condition is satisfied.

Condition 1
The measure $\mu(\cdot)$ is nonatomic; the measures $p_{t+1}(\cdot|x, a)$ are nonatomic for all $t = 0, 1, 2, \ldots$, $x \in X$, $a \in A_t(x)$.

Theorem 1 *Under Condition 1, $\mathcal{V} = \mathcal{V}^M$.*

Let for $n = 1, \ldots, 2N$

$$\tilde{R}^n(P^\pi) = \begin{cases} R_+^k(P^\pi), & \text{if } n = 2k-1, \\ -R_-^k(P^\pi), & \text{if } n = 2k. \end{cases}$$

If Theorem 1 holds for the vector $\tilde{\mathbf{R}}(P^\pi)$, it holds for the original performance vector $\mathbf{R}(P^\pi)$. Thus, we assume further without loss of generality that for all t the functions $r_t(\cdot)$ are nonnegative.

The set of strategic measures $\mathcal{D}$ is a measurable convex subset of $\mathcal{P}(H_\infty)$; Dynkin and Yushkevich [5]. Since functions r_t^n are nonnegative, $\mathbf{R}$ is an affine mapping. Therefore, $\mathcal{V} = \mathbf{R}(\mathcal{D})$ is convex.

3 One-step model

Suppose that $r_t^n(\cdot) = 0$ for all $t \geq 1$, $n = 1, 2, \ldots, N$. To put it differently, the control process ends after we chose the stochastic kernel π_0. The index $t = 0$ is omitted everywhere in this section. The set of all nonrandomized Markov policies for this model coincides with the set of all nonrandomized policies.

Lemma 1 *In the one-step model with the nonatomic measure μ, the set $\mathcal{V}^M \cap \mathbb{R}^N$ is convex.*

Proof: Let us fix two arbitrary nonrandomized policies $\varphi^1(x)$ and $\varphi^2(x)$ such that the vectors $\mathbf{R}(P^{\varphi^1}) \neq \mathbf{R}(P^{\varphi^2})$ are finite. For an arbitrary $\alpha \in]0,1[$, we consider

$$v = \alpha \mathbf{R}(P^{\varphi^1}) + (1-\alpha)\mathbf{R}(P^{\varphi^2}).$$

To prove Lemma 1, it is sufficient to show that $v = \mathbf{R}(P^\varphi)$ for some (nonrandomized) policy φ. Note that $v = \mathbf{R}(P^\pi)$ for some policy π which uses no more than

two actions $\varphi^1(x)$, $\varphi^2(x)$ in each state $x \in X$. Let $\hat{\pi}(x) \triangleq \pi(\varphi^1(x)|x)$. We are going to construct such a set $\Gamma \in \mathcal{B}(X)$ that the policy

$$\varphi(x) = \begin{cases} \varphi^1(x), & \text{if } x \in \Gamma, \\ \varphi^2(x), & \text{if } x \in X \setminus \Gamma \end{cases}$$

meets the equality $\mathbf{R}(P^{\varphi}) = \mathbf{R}(P^{\pi})$.

We define a finite partition $\{Y_i\}$ of X by

$$\begin{aligned} Y_i \triangleq \{x \in X :\ & r^1(x,\varphi^1(x)) \sim_1 r^1(x,\varphi^2(x)), \\ & r^2(x,\varphi^1(x)) \sim_2 r^2(x,\varphi^2(x)), \dots, \\ & r^N(x,\varphi^1(x)) \sim_N r^N(x,\varphi^2(x))\}, \end{aligned} \tag{1}$$

where $\sim_n \in \{>, <, =\}$, $n = 1, \dots, N$. Each set Y_i is measurable. We shall construct Γ in the form $\Gamma = \cup_i \Gamma_i$ where each Γ_i is a measurable subset of Y_i. To do it, we need to construct all Γ_i. We consider only the most significant case when all inequalities in (1) are strict inequalities and $\mu(Y_i) > 0$. The reasonings in other cases are similar.

Let us introduce the collection of nonatomic probability measures on the Borel space $(Y_i, \mathcal{B}(Y_i))$ by the formula

$$P_n(E) \triangleq \frac{\int\limits_E [r^n(x,\varphi^1(x)) - r^n(x,\varphi^2(x))]\mu(dx)}{\int\limits_{Y_i} [r^n(x,\varphi^1(x)) - r^n(x,\varphi^2(x))]\mu(dx)},$$

$$n = 1, 2, \dots, N.$$

By Lyapunov's theorem, there exists a set $\Gamma_i \in \mathcal{B}(Y_i)$ such that

$$P_n(\Gamma_i) = \int\limits_{Y_i} P_n(dx)\hat{\pi}(x), \quad n = 1, 2, \dots, N. \tag{2}$$

We define

$$\varphi(x) = \begin{cases} \varphi^1(x), & \text{if } x \in \Gamma_i; \\ \varphi^2(x), & \text{if } x \in Y_i \setminus \Gamma_i. \end{cases}$$

Then

$$P_n(\Gamma_i) = \frac{\int\limits_{Y_i} r^n(x,\varphi(x))\mu(dx) - \int\limits_{Y_i} r^n(x,\varphi^2(x))\mu(dx)}{\int\limits_{Y_i} [r^n(x,\varphi^1(x)) - r^n(x,\varphi^2(x))]\mu(dx)},$$

$$n = 1, 2, \dots, N;$$

$$\int\limits_{Y_i} P_n(dx)\hat{\pi}(x) =$$

$$\frac{\int\limits_{Y_i} r^n(x,\varphi^1(x))\hat{\pi}(x)\mu(dx) - \int\limits_{Y_i} r^n(x,\varphi^2(x))\mu(dx)}{\int\limits_{Y_i} [r^n(x,\varphi^1(x)) - r^n(x,\varphi^2(x))]\mu(dx)} +$$

$$+ \frac{\int\limits_{Y_i} r^n(x,\varphi^2(x))[1 - \hat{\pi}(x)]\mu(dx)}{\int\limits_{Y_i} [r^n(x,\varphi^1(x)) - r^n(x,\varphi^2(x))]\mu(dx)}, \ n = 1, 2, \dots, N.$$

Theorem 16.10 from Billingsley [4] has been used in the last formula. Therefore, according to (2) and by the definition of $\hat{\pi}$,

$$\int\limits_{Y_i} r^n(x,\varphi(x))\mu(dx) = \int\limits_{Y_i} \int\limits_{\bar{A}(x)} r^n(x,a)\pi(da|x)\mu(dx), \quad n = 1, 2, \dots, N. \tag{3}$$

Since formula (3) holds for any Y_i, such that $\mu(Y_i) > 0$ in the finite partition $\{Y_i\}$,

$$\mathbf{R}(P^{\varphi}) = \int\limits_X r(x,\varphi(x))\mu(dx) =$$

$$\int\limits_X \int\limits_{\bar{A}(x)} r(x,a)\pi(da|x)\mu(dx) = \mathbf{R}(P^{\pi}),$$

as we wished to prove. ■

Before we establish the validity of Theorem 1 for a one-step model, let us prove its simplified version.

Lemma 2 *In the one-step model with the nonatomic measure μ,*

$$\mathcal{V} \cap \mathbb{R}^N = \mathcal{V}^M \cap \mathbb{R}^N.$$

Proof: Since $\mathcal{V} \supseteq \mathcal{V}^M$, it is sufficient to prove that $\mathcal{V} \cap \mathbb{R}^N \subseteq \mathcal{V}^M \cap \mathbb{R}^N$. According to Feinberg [6, Theorem 5.2], for any fixed $\pi \in \Delta$, there exists a probability measure ν on the set $\mathcal{D}^M$ such that for any $E \in \mathcal{B}(E)$

$$P^{\pi}(E) = \int\limits_{\mathcal{D}^M} Q(E)\nu(dQ).$$

Therefore, if $\mathbf{R}(P^{\pi}) \in \mathcal{V} \cap \mathbb{R}^N$ then

$$\mathbf{R}(P^{\pi}) = \int\limits_{\mathcal{V}^M \cap \mathbb{R}^N} r\tilde{\nu}(dr),$$

where $\tilde{\nu}$ is the image of ν under the mapping $\mathbf{R}(\cdot)$. Hence $\mathbf{R}(P^{\pi})$ is the finite expectation of a random

variable in $\mathbb{R}^N$ with respect to a probability concentrated on the convex set $\mathcal{V}^M \cap \mathbb{R}^N$; see Lemma 1. Thus $\mathbf{R}(P^\pi) \in \mathcal{V}^M \cap \mathbb{R}^N$ and $\mathcal{V} \cap \mathbb{R}^N \subseteq \mathcal{V}^M \cap \mathbb{R}^N$. ■

Proof of Theorem 1 for a one-step model. First, we observe that Lemma 2 is valid also for subprobability measures μ. Second, in order to prove Theorem 1 for a one-step MDP, it is sufficient to show that for any policy π there exists a nonrandomized policy φ such that

$$\mathbf{R}(P^\varphi) = \mathbf{R}(P^\pi). \tag{4}$$

Let K^π be the number of coordinates of the vector $\mathbf{R}(P^\pi)$ with infinite values. We prove (4) by induction.

For $K^\pi = 0$, formula (4) follows from Lemma 2. Let (4) be valid for $K^\pi = K \geq 0$. We prove this formula for $K^\pi = K + 1$.

Without loss of generality we assume that the N-th coordinate of the vector $\mathbf{R}(P^\pi)$ is infinite:

$$R^N(P^\pi) = \int_X \mu(dx) \int_{A(x)} r^N(x,a)\pi(da|x) = +\infty.$$

Let $R(x) = \int_{A(x)} r^N(x,a)\pi(da|x)$. Then $\int_X R(x)\mu(dx) = \infty$. One can prove that there exists a sequence $\{Y_1, Y_2, \ldots\}$ of disjoint measurable subsets of X such that $\int_{Y_i} R(x)\mu(dx) > 1$ for all $i = 1, 2, \ldots$.

We will construct the mapping $\varphi : X \to A$ separately on the sets Y_i, $i = 1, 2, \ldots$. We fix an arbitrary i. There is a positive number M for which

$$\int_{Y_i} \mu(dx) \int_{A(x)} \min\{r^N(x,a), M\}\pi(da|x) > 1. \tag{5}$$

Since μ is a finite measure, the expression in the left hand side of (5) is finite. We replace the reward function $r^N(x,a)$ with $\min\{r^N(x,a), M\}$ and apply the induction assumption to the new reward function. We have that there exists a measurable mapping $\varphi : Y_i \to A$ such that $\varphi(x) \in A(x)$ for all $x \in X$ and for all $n = 1, 2, \ldots, N-1$,

$$\int_{Y_i} \mu(dx) r^n(x, \varphi(x)) = \int_{Y_i} \mu(dx) \int_{A(x)} r^n(x,a)\pi(da|x), \tag{6}$$

$$\int_{Y_i} \mu(dx) \min\{r^N(x, \varphi(x)), M\} = \int_{Y_i} \mu(dx) \int_{A(x)} \min\{r^N(x,a), M\}\pi(da|x). \tag{7}$$

From (5) and (7) we have that

$$\int_{Y_i} \mu(dx) r^N(x, \varphi(x)) \geq \int_{Y_i} \mu(dx) \min\{r^N(x, \varphi(x)), M\}$$

$$= \int_{Y_i} \mu(dx) \int_{A(x)} \min\{r^N(x,a), M\}\pi(da|x) > 1. \tag{8}$$

Since Y_i is an arbitrary element of the partition, function φ is defined on X. Then formula (6) implies

$$R^n(P^\varphi) = R^n(P^\pi), \quad n = 1, 2, \ldots, N-1$$

and (8) implies that $R^N(P^\varphi) = R^N(P^\pi) = \infty$. ■

4 Proof of Theorem 1

In this section we extend Theorem 1 from one-step to infinite-step MDPs. For $T = 1, 2, \ldots$, we define T-horizon rewards

$$R^n(P^\pi, T) = E^\pi \sum_{t=0}^{T-1} r_t^n(x_t, a_t) \quad n = 1, \ldots, N$$

and
$\mathbf{R}(P^\pi, T) = (R^1(P^\pi, T), R^2(P^\pi, T), \ldots, R^N(P^\pi, T))$.

Lemma 3 *For any policy π and for any $T = 1, 2, \ldots$ there exists a randomized Markov policy γ such that (i) γ is nonrandomized at steps $0, 1, \ldots, T-1$; (ii) $\mathbf{R}(P^\gamma) = \mathbf{R}(P^\pi)$; (iii) $\mathbf{R}(P^\gamma, T) = \mathbf{R}(P^\pi, T)$.*

Proof: For any policy π, there exists a randomized Markov policy σ such that $P^\sigma(dx_t da_t) = P^\pi(dx_t da_t)$ for all $t = 0, 1, \ldots$ and therefore $\mathbf{R}(P^\sigma) = \mathbf{R}(P^\pi)$ and $\mathbf{R}(P^\sigma, s) = \mathbf{R}(P^\pi, s)$ for all $s = 1, 2, \ldots$; Strauch [11, Theorem 4.1] Therefore, without loss of generality we can assume that π is a Markov policy.

First, we construct a randomized Markov policy which is nonrandomized at epoch 0 and satisfies (ii) and (iii). To do it, we consider a one-step MDP, introduced in Feinberg and Piunovskiy [7, Section 4], with the state space X, set of actions $\mathbf{D}$, sets $\mathcal{U}(x)$ of available actions at states $x \in X$, where $\mathbf{D}$ is the set of all strategic measures in the original MDP and $\mathcal{U}(x)$ is the set of all strategic measures in the original MDP such that: (a) the initial distribution is concentrated at x, (b) the policy is nonrandomized at step 0. By Lemma 8 in Feinberg and Piunovskiy [7], sets $\mathcal{U}(x)$ and the set of all strategic measures $\mathbf{U}$, generated by policies nonrandomized at step 0, are measurable. One can prove also that the graph of $\mathcal{U}$ is measurable.

For each $x \in X$ and for each $P \in \mathcal{U}(x)$ we consider one-step rewards $\tilde{r}^n(x,P) = R^n(P)$ and $\tilde{r}^{N+n}(x,P) = R^n(P,T)$, $n = 1,\dots,N$. We have a one-step model with $2N$ criteria $\tilde{R}^n(\tilde{P}) = \tilde{E}\tilde{r}^n(x_0,u_0)$, where $\tilde{P}$ is a strategic measure in the new one-step model, $\tilde{E}$ is the expectation in the new model, and u_0 is an action selected in the new model.

We assume that the initial measure μ, which was fixed for the original MDP, is also fixed for the new MDP. Then the new MDP satisfies Condition 1.

Lemma 9 in Feinberg and Piunovskiy [7] implies that there is a policy γ in the new MDP such that $\tilde{R}^n(\tilde{P}^\gamma) = R^n(P^\pi)$ and $\tilde{R}^{N+n}(\tilde{P}^\gamma) = R^n(P^\pi,T)$, $n = 1,\dots,N$. By Theorem 1 applied to the new one-step MDP we have that in this MDP there exists a nonrandomized policy ϕ such that $\tilde{R}^n(\tilde{P}^\phi) = \tilde{R}^n(\tilde{P}^\gamma)$, $n = 1,\dots,2N$. As explained on p. 62 in Feinberg and Piunovskiy [7], for the nonrandomized policy ϕ in the new MDP there exists a randomized Markov policy γ^0 such that $R^n(P^{\gamma^0}) = \tilde{R}^n(\tilde{P}^\phi)$ and $R^n(P^{\gamma^0},T) = \tilde{R}^{N+n}(\tilde{P}^\phi)$, $n = 1,\dots,N$ and this policy is nonrandomized at step 0. Therefore, $\mathbf{R}(P^{\gamma^0}) = \mathbf{R}(P^\pi)$ and $\mathbf{R}(P^{\gamma^0},T) = \mathbf{R}(P^\pi,T)$.

Then we can consider the nonatomic measure $\mu_1(Y) = P_\mu^{\gamma^0}(x_1 \in Y)$. We repeat the previous arguments applied to policy γ^0 on the horizon $1,2,\dots$ and construct a Markov policy γ^1 such that it is nonrandomized at the first step and

$$E_{\mu_1}^{\gamma^1}\sum_{t=1}^{T-1} r_t^n(x_t,a_t) = E_{\mu_1}^{\gamma^0}\sum_{t=1}^{T-1} r_t^n(x_t,a_t),$$

$$E_{\mu_1}^{\gamma^1}\sum_{t=1}^{\infty} r_t^n(x_t,a_t) = E_{\mu_1}^{\gamma^0}\sum_{t=1}^{\infty} r_t^n(x_t,a_t)$$

for all $n = 1,\dots,N$. We define γ^1 at step 0 being equal to γ^0 at that step. Then

$$\mathbf{R}(P^{\gamma^1}) = \mathbf{R}(P^{\gamma^0}) = \mathbf{R}(P^\pi),$$

$$\mathbf{R}(P^{\gamma^1},T) = \mathbf{R}(P^{\gamma^0},T) = \mathbf{R}(P^\pi,T),$$

and randomized Markov policy γ^1 is nonrandomized at steps 0 and 1.

By repeating this construction $T-2$ times more, we obtain the policy $\gamma = \gamma^{T-1}$ satisfying conditions (i –iii) of the Lemma. ■

Proof of Theorem 1. Fix an arbitrary policy π. To prove the theorem, it is sufficient to show that $\mathbf{R}(P^\sigma) = \mathbf{R}(P^\pi)$ for some Markov policy σ.

Consider a sequence $\epsilon_k \searrow 0$. We define $T_1 > 0$ such that for all $n = 1,\dots,N$

$$R^n(P^\pi,T_1) \ge \begin{cases} R^n(P^\pi) - \epsilon_1, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_1}, & \text{otherwise.} \end{cases}$$

Let γ be a policy which existence was stated in Lemma 3 for $T = T_1$. We set $\gamma^1 = \gamma$.

Suppose that for some natural k and for some $T_k \ge k$, we have a randomized Markov policy γ^k such that

(a) γ^k is nonrandomized at steps $0,\dots,T_k-1$;

(b) for all $n = 1,\dots,N$

$$R^n(P^{\gamma^k},T_k) \ge \begin{cases} R^n(P^\pi) - \epsilon_k, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_k}, & \text{otherwise;} \end{cases}$$

(c) $\mathbf{R}(P^{\gamma^k}) = \mathbf{R}(P^\pi)$.

We select $T_{k+1} > T_k$ such that for all $n = 1,\dots,N$

$$R^n(P^{\gamma^k},T_{k+1}) \ge \begin{cases} R^n(P^\pi) - \epsilon_{k+1}, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_{k+1}}, & \text{otherwise.} \end{cases}$$

By applying Lemma 3 to the MDP with the horizon $T_k, T_k+1,\dots$ and with the initial distribution $\tilde{\mu}(Y) = P_\mu^{\gamma^k}(x_{T_k} \in Y)$, we have that there exists a randomized Markov policy γ^{k+1} which satisfies (a – c) with k increased by 1. At steps $0,1,\dots,T_k-1$ this policy is defined being equal to γ^k and on steps $T_k, T_k+1,\dots$ this policy is constructed by using lemma 3.

We define a nonrandomized Markov policy γ which coincides with γ^k on steps $0,1,\dots,T_k-1$ for all $k = 1,2,\dots$. Since $T_k < T_{k+1}$ and $\gamma_t^k = \gamma_t^{k+1}$ for $t < T_k$, $k = 1,2,\dots$, this definition is correct. Inequality (b) implies that

$$R^n(P^\gamma,T_k) \ge \begin{cases} R^n(P^\pi) - \epsilon_k, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_k}, & \text{otherwise,} \end{cases} \tag{9}$$

and equality (c) implies that

$$R^n(P^\gamma,T_k) \le R^n(P^\pi) \tag{10}$$

for all $n = 1,\dots,N$. Since $\epsilon_k \searrow 0$ and all one-step rewards are nonnegative, (9) and (10) imply $\mathbf{R}(P^\gamma) = \lim_{k\to\infty} \mathbf{R}(P^\gamma,T_k) = \mathbf{R}(P^\pi)$. ■

5 Application

A popular way to study a multiciterion problem is to replace it by a constrained one. Namely, after choosing some constants $d_2,\dots,d_N$, it is natural to consider the following optimization problem

$$R^1(P^\pi) \to \sup_\pi, \quad R^n(P^\pi) \ge d_n, \ n = 2,\dots,N. \tag{11}$$

In the rest of this paper, we consider expected total discounted criteria

$$R^n(P^\pi) = E^\pi \left[\sum_{t=0}^{\infty} \beta^t r^n(x_t, a_t) \right],$$

where $\beta \in (0,1)$ is a discount factor. According to Hernandez-Lerma and Gonzalez-Hernandez [9], problem (11) has a solution if the following condition is satisfied for a homogeneous model in which $A_t(\cdot)$, p_t, and r_t^n do not depend on the time parameter t.

Condition 2

1. The graph $\mathrm{Gr}(A)$ *is closed for each* $t = 0, 1, \ldots$.
2. All the functions $r^n(\cdot)$ *are bounded above,* $n = 1, 2, \ldots, N$.
3. For each constant r *all the sets*

$$\{(x,a) \in \mathrm{Gr}(A) | r^1(x,a) \geq r\}$$

are compact.
4. The functions $r^n(\cdot)$, $n = 2, \ldots, N$, *are upper semicontinuous.*
5. Transition probabilities $p(dy|x,a)$ *are weakly continuous.*
6. There exists such a policy π *that* $R^n(P^\pi) \geq d_n$, $n = 2, \ldots, N$.

Hence, if conditions 1,2 are satisfied then there exists a nonrandomized Markov policy which is optimal for problem (11).

Let us consider an inventory system with finite capacity M. The demand at epoch $t = 0, 1, \ldots$ is ξ_t (positive i.i.d. random variables). We assume that the distribution of ξ_t has no atoms and $P\{\xi_t < \infty\} = 1$. Orders are placed after the demand is known and it is possible to order up to the full capacity of the system.

Let $h(x)$ be the holding cost of the amount x during one period of time, and $K(a)$ be the ordering cost of a units. We assume that the function h is continuous, bounded below, and $h(x) \to \infty$ as $|x| \to \infty$, and $K(a)$ is bounded below and lower semicontinuous on $[0, \infty)$. We remark that our assumptions cover the following particular functions

$$h(x) = \begin{cases} h_1 x, & \text{if } x \geq 0, \\ -h_2 x, & \text{otherwise,} \end{cases}$$

and

$$K(a) = \begin{cases} k_0 + k_1 a, & \text{if } a > 0, \\ 0, & \text{if } a = 0, \end{cases}$$

where $h_0, h_1 > 0$ and $k_0, k_1 \geq 0$ are some coefficients.

Let the initial inventory be y. Then the initial state of the system is $x_0 = y - \xi_0$. The dynamics of the system is defined by the equation $x_{t+1} = x_t + a_t - \xi_{t+1}$.

One can consider different reward functions associated with this inventory system. For instance, we may set $r^1(x,a) = -h(x)$ and $r^2(x,a) = -K(a)$. Then R^1 is the criterion characterizing holding and backordering costs, and R^2 characterizes operational costs.

Theorem 1 implies that (nonrandomized) Markov policies for this multicriterion problem are as good as general policies. If we fix appropriate constant d_2 then Conditions 2 are satisfied for the constrained problem (11) with $N = 2$. Thus, there exists an optimal nonrandomized Markov policy for this problem.

References

[1] Altman E. Constrained Markov Decision Processes, Chapman & Hall/CRC, Boca Raton, 1999.

[2] Barra J.R. Mathematical Basis of Statistics, Academic Press, New York, 1981.

[3] Bertsekas D.P., Shreve S.E. Stochastic Optimal Control. – Academic Press: N.Y. etc., 1978.

[4] Billingsley P. Probability and Measure. Wiley, New York, 1986.

[5] Dynkin E.B., Yushkevich A.A. Controlled Markov Processes and their Applications. – Springer-Verlag: N.Y.-Berlin, 1979.

[6] Feinberg E.A. On measurability and representation of strategic measures in Markov decision processes. Statistics, Probability and Game Theory Papers in Honor of David Blackwell (ed. T.Ferguson), IMS Notes - Monograph Series, 1996, V.30, p.29-43.

[7] Feinberg E.A., Piunovskiy A.B. Multiple objective nonatomic Markov decision processes with total reward criteria. *J.Math.Anal.Appl.* **247** (2000), 45-66.

[8] Frid E.B. On optimal strategies in controlled problems with constraints. SIAM Theory Probab. Appl., 1972, **17**, 188-192.

[9] Hernandez-Lerma O., Gonzalez-Hernandez J. Constrained Markov control processes in Borel spaces: the discounted case, *Math. Meth. Oper. Res.* **52** (2000), to appear.

[10] Piunovskiy A.B. Optimal Control of Random Sequences in Problems with Constraints. Kluwer, Dordrecht, 1997.

[11] Strauch R.E. Negative dynamic programming, *Ann. Math. Statist.* **37** (1966), 871-890.

[12] Wagner D.H., Survey of measurable selection theorems, *SIAM J. Control and Optim.* **15** (1977), 859-903.