Journal of Mathematical Analysis and Applications **247,** 45–66 (2000)
doi:10.1006/jmaa.2000.6819, available online at http://www.idealibrary.com on IDEAL®

Multiple Objective Nonatomic Markov Decision Processes with Total Reward Criteria[1]

Eugene A. Feinberg

E-mail: Eugene.Feinberg@sunysb.edu

and

Aleksey B. Piunovskiy

Division of Statistics and Operations Research, Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 3BX, United Kingdom
E-mail: piunov@liverpool.ac.uk

Submitted by William F. Ames

Received September 20, 1999

We consider a Markov decision process with an uncountable state space and multiple rewards. For each policy, its performance is evaluated by a vector of total expected rewards. Under the standard continuity assumptions and the additional assumption that all initial and transition probabilities are nonatomic, we prove that the set of performance vectors for all policies is equal to the set of performance vectors for (nonrandomized) Markov policies. This result implies the existence of optimal (nonrandomized) Markov policies for nonatomic constrained Markov decision processes with total rewards. We provide two examples of applications of our results to constrained multiple objective problems in inventory control and finance.

1. INTRODUCTION

This paper studies a nonhomogeneous Markov decision process with an uncountable state space and with a finite number of reward functions. For each reward function we consider expected total rewards. For each policy, these expected total rewards form a performance vector.

The main result of the paper, Theorem 1, shows that if the initial state distribution and transition probabilities do not have atoms then, under

[1] This research was partially supported by NSF Grants DMI-9500746 and DMI-9908258.

0022-247X/00 $35.00

standard continuity, compactness, and convergent conditions, for each policy there exists a (nonrandomized) Markov policy with the same performance vector. This result implies the existence of optimal Markov policies (Theorem 2) for problems with multiple total reward criteria and constraints.

At first glance, this result contradicts the well-known phenomenon that optimal policies in problems with constraints are typically randomized. Example 1 in Frid [11] demonstrates that randomized policies can be significantly better than nonrandomized strategies in problems with constraints; see also examples in books on Markov decision processes with constraints by Kallenberg [12], Piunovskiy [17], and Altman [2]. In these examples the state and action sets are finite, but they can be easily modified to countable state problems with arbitrary action sets.

Since the use of randomization procedures can improve the system performance, the natural question is how to obtain the best performance and to limit the number of randomization procedures. When the state space is countable, the number of randomization procedures can be limited by the number of constraints. This fact was established by Borkar [6] for problems with average rewards per unit time and Feinberg and Shwartz [10] for problems with total discounted rewards. More particular results were proved earlier by Ross [19], Sennott [21], and Altman [1].

Though comprehensive results for coutable state models were described in Borkar [6] for average rewards per unit time and in Feinberg and Shwartz [10] for the total discounted rewards, significantly less is known for models with uncountable state spaces; see Tanaka [24] and Piunovskiy [17]. In particular, the results, that the number of randomization procedures is limited by the number of constraints, are not available for models with uncountable state spaces. If such results were available they would imply the existence of optimal nonrandomized policies for models with nonatomic initial distributions and transition probabilities. Indeed, if these probabilities are nonatomic, the probability to visit any finite state is 0. Therefore, randomized actions could be replaced with a nonrandomized one on the finite state where they may be required by a general theorem, which, as we mentioned, is not currently available for models with uncountable state spaces.

This paper describes an important phenomenon that there are nonrandomized optimal policies for problems with multiple total reward criteria when the state space is uncountable and transition and initial probabilities are nonatomic. To illustrate possible applications, we consider inventory and financial examples for which we establish the existence of optimal nonrandomized Markov policies for problems with multiple expected total cost criteria and constraints; see Liu and Esogbue [13] for various criteria for inventory systems.

We introduce the model, major assumptions, and formulate the major results in Section 2. These results are that for nonatomic Markov decision processes with multiple criteria the following two statements hold: (i) the set of performance vectors for all policies coincides with the set of performance vectors for nonrandomized Markov policies (Theorem 1) and (ii) nonrandomized Markov policies are optimal for constrained problems (Theorem 2). In Section 3 we prove our main result, Theorem 1, for a one-step model. In Section 4 we construct an equivalent one-step model for a multiple-step model. In Section 5 we prove our main theorem by using the results of Sections 3 and 4. Section 6 deals with applications.

In the completion of this section we recall several measure-theoretical definitions. Two measurable spaces $(\Omega, \mathscr{F})$ and $(\Omega_1, \mathscr{F}_1)$ are called *isomorphic* if there is a one-to-one measurable mapping f of Ω onto Ω_1 such that f^{-1} is also measurable. A *Polish* space is a complete separable metric space. For a topological space Ω, we denote by $\mathscr{B}(\Omega)$ its Borel σ-field. A measurable space is called Borel if it is isomorphic to a Borel subset of a Polish space endowed with the Borel σ-field. Any Borel space is either finite or countable or isomorphic to $[0, 1]$; see Dynkin and Yushkevich [8, Appendix 1] or Bertsekas and Shreve [5, Corollary 7.16.1].

If Ω is finite or countable, we consider a discrete topology on Ω. In this case, $\mathscr{B}(\Omega)$ is the set of all subsets of Ω.

For a Polish space $(\Omega, \mathscr{B}(\Omega))$, we denote by $\mathscr{P}(\Omega)$ the set of probability measures on $(\Omega, \mathscr{B}(\Omega))$. Let $\{P_n,\ n = 1, 2, \dots\}$ and P be probability measures on a Polish space $(\Omega, \mathscr{B}(\Omega))$. The measures P_n converge *weakly* to P if $\int_\Omega f(\omega) P_n(d\omega) \to \int_\Omega f(\omega) P(d\omega)$ for any bounded continuous function f on Ω. We always consider a *weak topology* on $\mathscr{P}(\Omega)$ which is the weakest topology generated by the weak convergence. The weak topology is metrizable and the space $\mathscr{P}(\Omega)$ is Polish in this topology; see Bertsekas and Shreve [5, Proposition 7.20]. If Ω is compact then $\mathscr{P}(\Omega)$ is compact in the weak topology; see Bertsekas and Shreve [5, Proposition 7.22].

Let q be a nonnegative measure on a Borel space $(\Omega, \mathscr{F})$. A point $\omega \in \Omega$ is called an atom if $q(\{\omega\}) > 0$. A measure is called nonatomic if it does not have atoms. Obviously, if q is a nonatomic measure on a Borel space $(\Omega, \mathscr{F})$ and $q(\Omega) > 0$ then this space is uncountable and therefore it is isomorphic to $[0, 1]$.

2. DESCRIPTION OF THE MODEL AND THE MAIN RESULT

We consider a Markov decision process (MDP) $\{X, A, A.(\cdot), p., r.\}$, where

(i) X is a Polish state space;

(ii) A is a Polish 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 $A_t(x)$ are nonempty measurable subsets of A;

(iv) $p_t(dy|x, a)$ are measurable transition probabilities from $X \times A$ to X at steps $t = 0, 1, \ldots$;

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

As usual, a policy π is a sequence of measurable transition probabilities $\pi_t(da|h_t)$ concentrated on the sets $A_t(x_t)$, where $h_t = x_0, a_0, \ldots, a_t, x_t$ is the observed history. If transition probabilities π 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 Markov and it is denoted by φ. Let Δ be the set of all policies and Δ^M be the set of all Markov policies. A policy π is called randomized Markov if $\pi_t(da_t|h_t) = \pi_t(da_t|x_t)$ for all $h_t = x_0 a_0 \cdots x_t \in (X \times A)^t \times X$.

Condition 1. A is compact; the graph $\mathrm{Gr}\ A_t \triangleq \{(x, a): x \in X, a \in A_t(x)\}$ is closed in $X \times A$ for all $t = 0, 1, \ldots$. (In particular, all the sets $A_t(x)$ are compact.)

For a fixed t, a measurable mapping $\varphi_t(\cdot): X \to A$ is called a selector if $\varphi_t(x) \in A_t(x)$ for all $x \in X$. The existence of at least one policy is equivalent to the existence of at least one selector for each $t = 0, 1, \ldots$ [8, Sect. 3.1]. According to [3, Corollary I.1.1], Condition 1 implies the upper semicontinuity of the multifunction $x \to A_t(x)$ for each $t = 0, 1, \ldots$. Hence, in view of [25, Theorem 4.1], there exists at least one selector.

According to the Ionescu Tulcea theorem [8, Sect. 5.4; 5, Proposition 7.45], a policy π and an initial distribution μ on X define a unique probability measure P_μ^π on the space of trajectories $H_\infty = (X \times A)^\infty$. We denote by E_μ^π expectations with respect to P_μ^π. If $\pi^1 = \pi^2$ (P^{π^1}-a.s. or P^{π^2}-a.s.) then $P^{\pi^1} = P^{\pi^2}$ and the strategies π^1 and π^2 are equivalent (indistinguishable). If the initial distribution is concentrated at the point x, we use notations P_x^π and E_x^π.

For a number c we define $c^+ = \max\{c, 0\}$ and $c^- = \min\{c, 0\}$. For a policy π, an initial distribution μ, and $n = 1, \ldots, N$, we define

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

$$R^n_-(\mu,\pi) = R(\mu,\pi) = E^\pi_\mu\left[\sum_{t=0}^{\infty}\left(r^n_t(x_t,a_t)\right)^-\right],$$

and $R^n(\mu,\pi) = R^n_+(\mu,\pi) + R^n_-(\mu,\pi)$. Everywhere in this paper, $\infty - \infty$ is assumed equal $-\infty$.

The performance of a policy π is evaluated by a vector

$$\mathbf{R}\left(P^\pi_\mu\right) = R(\mu,\pi) = \left(R^1(\mu,\pi), R^2(\mu,\pi), \ldots, R^N(\mu,\pi)\right). \quad (1)$$

Condition 2. For any initial state $x \in X$ and for any policy π

$$\left|E^\pi_x\left[r^n_t(x_t,a_t)\right]\right| \le c_t, \qquad n = 1,\ldots,N,$$

where c_t is a summable sequence, $\sum_{t=0}^{\infty} c_t < \infty$.

Condition 2 guarantees the correctness of formula (1). For instance, Condition 2 is satisfied for discounted model where $r^n_t(x,a) = \beta^t r^n(x,a), |r^n(x,a)| \le c < \infty$, $\beta \in (0,1)$.

Unless it is specified explicitly, we assume that the initial distribution $\mu(dx)$ of x_0 has been fixed. In order to simplify the notations, we usually write P^π, E^π, and $R(\pi)$ instead of P^π_μ, E^π_μ, and $R(\mu,\pi)$, respectively. We define the sets of performance vectors for all policies

$$\mathscr{V} \triangleq \{R(\pi) | \pi \in \Delta\} \quad (2)$$

and for Markov policies

$$\mathscr{V}^M \triangleq \{R(\varphi) | \varphi \in \Delta^M\}.$$

We also define the set of all strategic measures

$$\mathscr{D} \triangleq \{P^\pi | \pi \in \Delta\}$$

and its subset generated by Markov policies

$$\mathscr{D}^M \triangleq \{P^\varphi | \varphi \in \Delta^M\}.$$

If the initial distribution μ is not fixed, we write $\mathscr{V}(\mu)$, $\mathscr{V}^M(\mu)$, $\mathscr{D}(\mu)$, and $\mathscr{D}^M(\mu)$. We replace μ with x if measure μ is concentrated at x.

The set of strategic measures $\mathscr{D}$ is convex and a measurable subset of $\mathscr{P}(H_\infty)$; see Dynkin and Yushkevich [8, Sect. 5.5]. According to a well-known result by Derman and Strauch [7] and Strauch [22], for any policy π there exists a randomized Markov policy σ with the same marginal distributions of couples (x_t, a_t); see, e.g., Dynkin and Yushkevich [8, Sects. 4.8 and 5.2] or Puterman [18, Theorem 5.5.1]. Therefore $R(\sigma) = R(\pi)$ and one may replace the set of all policies with the set of all randomized Markov policies in (2).

Condition 3. The transition probabilities $p_t(dy|x, a)$ are weakly continuous in $(x, a) \in \mathrm{Gr}\, A_t$, $t = 0, 1, \ldots$.

Condition 4. The function $r_t: X \times A \to \mathbb{R}^N$ is continuous and bounded on $\mathrm{Gr}\, A_t$ for every $t = 0, 1, \ldots$.

LEMMA 1. (i) *The set $\mathscr{D}$ is convex.*

(ii) *If Condition* 2 *holds, then the performance set $\mathscr{V}$ is convex.*

(iii) *If Conditions* 1 *and* 3 *hold then the set $\mathscr{D}$ is compact.*

(iv) *If Conditions* 1–4 *hold then $\mathscr{V}$ is compact.*

Proof. The convexity of $\mathscr{D}$ was pointed out earlier. The compactness of $\mathscr{D}$ under Conditions 1 and 3 was proved in Balder [4] and Schäl [20]. Under Condition 2 the mapping $\mathbf{R}: \mathscr{D} \to \mathbb{R}^N$ is affine. Conditions 2 and 4 imply that this mapping is continuous. Hence, $\mathscr{V}$ is also convex and compact under Conditions 1–4. ▌

The main result of this paper is Theorem 1 that states the sufficiency of the class of Markov policies. It is worth noting that Conditions 1–5 do not imply that the set $\mathscr{D}^M$ is either convex or compact. In fact, this set may be neither convex nor compact. For example, if $X = [0, 1]$, $A = A_t(x) = \{0, 1\}$, and μ is the Lebesgue measure then $\mathscr{D}^M$ is neither convex nor compact in a one-step model; see Piunovskiy [17, Sect. 3.4.2]. Nevertheless, $\mathscr{V}^M = \mathscr{V}$ is convex and compact under the following additional condition.

Condition 5. The measure $\mu(\cdot)$ is nonatomic and for every $t = 0, 1, \ldots$, for every $x \in X$, and for every $a \in A_t(x)$ the measure $p_t(\cdot|x, a)$ is nonatomic.

THEOREM 1. *Let Conditions* 1–5 *be satisfied. Then $\mathscr{V}^M = \mathscr{V}$.*

We consider an important application of Theorem 1. For given numbers $C^2, \ldots, C^N$, we consider a constrained optimization problem

$$\text{maximize } R^1(\pi) \tag{3}$$

subject to

$$R^n(\pi) \geq C^n, \qquad n = 2, \ldots, N. \tag{4}$$

The following result follows from Theorem 1 and Lemma 1.

THEOREM 2. *Let Conditions* 1–5 *be satisfied. If problem* (3), (4) *is feasible then there exists an optimal Markov policy.*

3. ONE-STEP MODEL

Suppose that the life time of the controlled process equals one step. In this case, the symbol $+\infty$ is replaced by 0 in all the constructions. In other words, in this section our performance vector is $E^{\pi}[r_0(x_0, a_0)]$. We omit the index $t = 0$ in this section.

In a one-step model, the set of Markov policies coincides with the set of all selectors $\varphi(x) \in A(x)$. Since there is no history, the set of all policies coincides with the set of randomized Markov policies. Strategic measures are defined on $X \times A$.

In this section, we prove Theorem 1 for a one-step model. The starting point of our approach is based on the following observation.

LEMMA 2. *If the probability measure $\mu(dx)$ is nonatomic then there is a collection of measurable sets $\{X_\alpha \subseteq X,\ \alpha \in [0,1]\}$ such that $\mu(X_0) = 0$, $\mu(X_1) = 1$, and $\mu(X_\alpha)$ is a continuous function of α.*

Proof. Since μ is a nonatomic measure, the Borel space X is isomorphic to the interval $[0,1]$. Let $\Psi\colon X \to [0,1]$ be an isomorphism. Then Ψ can be interpreted as a random variable with the distribution function

$$F(b) \triangleq \mu(\{x : \Psi(x) \le b\}), \qquad b \in [-\infty, \infty].$$

Since μ has no atoms, the function F is continuous. We define quantiles $q(\alpha) = \inf\{b \ge 0 | F(b) = \alpha\}$. Since F is continuous, $F(q(\alpha)) = \alpha$ for any $\alpha \in [0,1]$.

We set

$$X_\alpha \triangleq \{x : \Psi(x) < q(\alpha)\}. \tag{5}$$

Obviously, $\mu(X_\alpha) = \alpha$. ∎

A subset B of a metric space is called connected if for any two points $c, d \in B$ there is a continuous mapping f of the interval $[0,1]$ into B such that $f(0) = c$ and $f(1) = d$.

LEMMA 3. *If the measure $\mu(\cdot)$ is nonatomic then the set $\mathscr{D}^M$ is connected.*

Proof. If Gr A does not contain a selector, $\mathscr{D}^M = \varnothing$ and the lemma is obvious. Let us fix two arbitrary selectors $\varphi^0(x)$ and $\varphi^1(x)$. Then P^{φ^0} and P^{φ^1} are the corresponding strategic measures on $X \times A$.

Let

$$\varphi^{\alpha}(x) = \begin{cases} \varphi^1(x), & \text{if } x \in X_\alpha, \\ \varphi^0(x), & \text{if } x \notin X_\alpha, \end{cases}$$

where X_α are the sets given by formula (5). Consider the mapping

$$\gamma(\alpha) \triangleq P^{\varphi^\alpha}$$

from $[0,1]$ into $\mathscr{D}^M$. Clearly,

$$\gamma(0) = P^{\varphi^0}, \qquad \gamma(1) = P^{\varphi^1}.$$

In the first case, $X_0 = \varnothing$. In the second case, $\mu(X_1) = 1$.

It is sufficient to show that the mapping $\gamma: [0,1] \to \mathscr{D}^M$ is continuous. Let c be a bounded measurable function on $X \times A$, $|c(x,a)| \le C < \infty$. We observe that $\varphi^\alpha(x) = \varphi^\beta(x)$ when $x \notin X_\alpha \triangle X_\beta$ and $\mu(X_\alpha \triangle X_\beta) = |\alpha - \beta|$. Thus

$$\begin{aligned} &\left|E^{\varphi^\alpha}[c(x,a)] - E^{\varphi^\beta}[c(x,a)]\right| \\ &\quad = \left|E^{\varphi^\alpha}\left[\mathbf{I}\{x \in X_\alpha \triangle X_\beta\}c(x,a)\right] - E^{\varphi^\beta}\left[\mathbf{I}\{x \in X_\alpha \triangle X_\beta\}c(x,a)\right]\right| \\ &\quad \le 2|\alpha - \beta|C, \end{aligned} \tag{6}$$

where $\mathbf{I}$ is the indicator function. Therefore, the mapping $\gamma: [0,1] \to \mathscr{D}^M$ is continuous. ∎

We consider the following relaxation of Conditions 1–5.

Condition 6. The following assumptions hold:

(i) Gr A is measurable, contains at least one selector (measurable function), and all sets $A(x)$ are compact, $x \in X$;

(ii) the functions $r^k(x,a)$, $k = 1, \ldots, N$, are bounded, measurable, and continuous in a;

(iii) the fixed initial measure μ is nonatomic;

(iv) the $\mathscr{V}(\nu)$ are convex and compact for all initial measures ν such that $\nu \ll \mu$.

LEMMA 4. *Conditions* 1, 4, *and* 5 *imply Condition* 6.

Proof. Conditions 1 and 5 imply respectively (i) and (iii). For one-step models, (ii) follows from Condition 4. Lemma 1 implies (iv). ∎

If the sets of feasible actions $A(x)$ are substituted with subsets $B(x) \subseteq A(x)$, $x \in X$, then the corresponding MDP is called a submodel. We denote by $\mathscr{V}_B(\nu)$ and $\mathscr{V}_B^M(\nu)$ the performance sets for all policies and for all selectors in the submodel where ν is an initial probability distribution. Obviously, $\mathscr{V}_B(\nu) \subseteq \mathscr{V}(\nu)$ and $\mathscr{V}_B^M(\nu) \subseteq \mathscr{V}^M(\nu)$. As usual, we may omit the fixed initial distribution in our notations. So, we set $\mathscr{V}_B = \mathscr{V}_B(\mu)$ and $\mathscr{V}_B^M = \mathscr{V}_B^M(\mu)$.

Let $\partial\mathscr{V}$ be the boundary of $\mathscr{V}$. By Lemma 1, $\mathscr{V}$ is closed and convex. Therefore $\mathscr{V}$ has a supporting hyperplane at each boundary point. We recall that a subset $E \subset \partial E'$ of a closed convex set E' in an Euclidean space is called exposed if E is an intersection of E' and its supporting hyperplane. Note that $\dim E < \dim E'$.

LEMMA 5. *Let Condition* 6 *hold. For any exposed subset* E *of* $\mathscr{V}$ *there exists a submodel with sets of available actions* $B(x)$ *such that*: (i) *this submodel satisfies Condition* 6 *and* (ii) $\mathscr{V}_B = E$.

Proof. Consider a supporting hyperplane that defines E. Let $E = \{z \in \mathbb{R}^n : \sum_{n=1}^N b_n z_n = d\} \cap \mathscr{V}$. Then $\sum_{n=1}^N b_n R^n(\sigma) = d$ for any policy σ such that $R(\sigma) \in E$. In addition, let the signs of b_n be selected in a way that $\sum_{n=1}^N b_n R^n(\pi) \le d$ for any $\pi \in \Delta$. Therefore, if we consider the objective function $r^*(x,a) = \sum_{n=1}^N b_n r^n(x,a)$ then a policy π is optimal for the maximization problem with this objective function and with the initial distribution μ if and only if $R(\pi) \in E$.

Let $\tilde{r}(x) = \max_{a \in A(x)} r^*(x,a)$, $x \in X$. Then the sets $B^*(x) = \{a \in A(x) | r^*(x,a) = \tilde{r}(x)\}$ are compact. In view of [5, Propositions 7.47 and 7.50], $\tilde{r}$ is a universally measurable function and there exists a universally measurable mapping $\varphi: X \to A$ such that $\varphi(x) \in A(x)$ for all $x \in X$ and $r^*(x, \varphi(x)) = \tilde{r}(x)$. Let ψ be an arbitrary selector. We consider a Borel mapping $\varphi': X \to A$ such that $\mu(\{x : \varphi(x) \ne \varphi'(x)\}) = 0$. Let Y be a Borel subset of X such that $\mu(Y) = 1$ and $\varphi(x) = \varphi'(x)$ when $x \in Y$. We define compact sets

$$B(x) = \begin{cases} B^*(x), & \text{if } x \in Y, \\ \{\psi(x)\}, & \text{if } \in X \setminus Y. \end{cases}$$

We notice that $\operatorname{Gr} B = \Gamma_1 \cup \Gamma_2$ where

$$\Gamma_1 = \{(x,a) | r^*(x, \varphi'(x)) = r^*(x,a),\ x \in Y\}$$

and

$$\Gamma_2 = \{(x,a) | a = \psi(x),\ x \in X \setminus Y\}.$$

The set Γ_1 is measurable because it can be presented as a set on which a Borel function is equal to 0. The set Γ_2 is measurable because it is a graph of a Borel function. Therefore, $\operatorname{Gr} B$ is measurable. In addition, $\operatorname{Gr} B$ contains the selector

$$\tilde{\varphi}(x) = \begin{cases} \varphi'(x), & \text{if } x \in Y; \\ \psi(x), & \text{if } x \in X \setminus Y. \end{cases}$$

For an arbitrary policy π, $r^*(x,\pi) \le \tilde{r}(x)$ (μ-a.s.), where $r^*(x,\pi) = \int_X r^*(x,a)\pi(da|x)$. We also have that $r^*(x,\varphi'(x)) = \tilde{r}(x)$ (μ-a.s.). Therefore, $r^*(x,\pi) = \tilde{r}(x)$ (μ-a.s.) if and only if π is an optimal policy for the reward function r^* and the initial distribution μ. Therefore, if π is a policy in the submodel with action sets $B(\cdot)$ then $R(\pi) \in E$. Thus $\mathscr{V}_B \subseteq E$.

Let $v \in E$. Then $v = E^{\pi}[r(x,a)]$ for a policy π which is optimal for the MDP with the reward function r^* and initial distribution μ. Therefore, $r^*(x,\pi) = \tilde{r}(x)$ (μ-a.s.). Let Z be a Borel subset of X such that $\mu(Z) = 1$ and $r^*(x,\pi) = \tilde{r}(x)$ for $x \in Z$. Then $\pi(B^*(x)|x) = 1$ for every $x \in Z$. We define a policy σ in the submodel with action sets $B(\cdot)$,

$$\sigma(C|x) = \begin{cases} \pi(C|x) & \text{if } x \in Z \cap Y, \\ \gamma(C|x) & \text{otherwise,} \end{cases}$$

where γ is an arbitrary policy in this submodel. Then σ is a policy in this submodel and $R(\sigma) = R(\pi) = v$. Thus $\mathscr{V}_B \supseteq E$. So, $\mathscr{V}_B = E$.

The constructed submodel meets Conditions 6(1)–(iii). We verify Conditions 6(iv). Lemma 1(ii) implies that $\mathscr{V}_B(\nu)$ is convex. For an arbitrary initial distribution ν and for an arbitrary policy π, we define $r^*(\nu,\pi) = \int_X \int_A r^*(x,a)\pi(da|x)\nu(dx)$. Let us fix an arbitrary initial measure $\nu \ll \mu$. Let $\tilde{d} = \max_{\pi} r^*(\nu,\pi)$. If $\tilde{\pi}$ is a policy in the submodel with action sets $B(\cdot)$ then $r^*(\nu,\tilde{\pi}) = \tilde{d}$. So $R(\nu,\tilde{\pi}) \in E'$ where $E' = \mathscr{V}(\nu) \cap \{z \in \mathbb{R}^N : \sum_{n=1}^N b_n z_n = \tilde{d}\}$ is the exposed subset of the convex compact set $\mathscr{V}(\nu)$. Hence, $\mathscr{V}_B(\nu) \subseteq E'$. Let $v \in E'$. Then $v = R(\nu,\pi)$ for a policy π which is optimal for the MDP with the reward function r^* and the initial distribution ν. Similarly to the previous paragraph we obtain $R(\nu,\sigma) = R(\nu,\pi) = v$ for a policy σ in the submodel. Thus $E' \subseteq \mathscr{V}_B(\nu)$. So, $\mathscr{V}_B(\nu) = E'$ is compact. ∎

For a selector φ and a measurable set $Y \subseteq X$, we consider a submodel with the sets of available actions

$$A[\varphi,Y](x) = \begin{cases} \{\varphi(x)\}, & \text{if } x \in Y; \\ A(x), & \text{if } x \in X \setminus Y. \end{cases}$$

We notice that $A[\varphi,X](x) = \{\varphi(x)\}$ and $A[\varphi,\varnothing](x) = A(x)$, $x \in X$.

LEMMA 6. *If Condition* 6 *holds then it holds for each submodel with available sets of actions* $A[\varphi,Y](x)$ *where* $x \in X$ *and* $Y \in \mathscr{B}(X)$.

Proof. Conditions 6(i)–(iii) are obvious. We verify Condition 6(iv).

Let ν be the initial measure and $\nu \ll \mu$. For any measurable $E \subseteq X$ we consider the following version of condition probabilities

$$\nu(Z|E) = \begin{cases} \nu(Z \cap E)/\nu(E), & \text{if } \nu(E) > 0; \\ \nu(E), & \text{if } \nu(E) = 0. \end{cases}$$

We define two measures $\nu'(Z) = \nu(Z|X \setminus Y)$ and $\nu''(Z) = \nu(Z|Y)$.

If π is a policy in the submodel then

$$R(\pi, \nu) = \nu(X \setminus Y) R(\pi, \nu') + \nu(Y) R(\varphi, \nu'').$$

Therefore

$$\mathscr{V}_{A[\varphi, Y]}(\nu) = \nu(X \setminus Y) \mathscr{V}(\nu') + \nu(Y) R(\varphi, \nu''). \tag{7}$$

Since $\nu' \ll \nu \ll \mu$, $\mathscr{V}(\nu')$ is convex and compact. Formula (7) implies that $\mathscr{V}_{A[\varphi, Y]}(\nu)$ is convex and compact too. ∎

LEMMA 7. *If Condition* 6 *holds then for any selector* φ *and for any* $v \in \mathscr{V}$ *there exists a measurable subset* Y *of* X *such that* $v \in \partial \mathscr{V}_{A[\varphi, Y]}$.

Proof. We consider a collection of sets X_α whose existence is stated in Lemma 2. We denote $\mathscr{V}_{A[\varphi, X_\alpha]}$ by $\mathscr{V}^\alpha$. Then $\mathscr{V}^0 = \mathscr{V}$ and $\mathscr{V}^1 = \{R(\varphi)\}$. In addition, $\mathscr{V}^\alpha \supseteq \mathscr{V}^\beta$ if $\alpha \le \beta$. By Lemma 6, each set $\mathscr{V}^\alpha$ is convex and compact, $\alpha \in [0, 1]$.

We have that $v \in \mathscr{V} = \mathscr{V}^0$. Let

$$\alpha^* = \begin{cases} \inf\{\alpha \ge 0 : v \notin \mathscr{V}^\alpha\}, & \text{if } v \ne R(\varphi); \\ 1, & \text{if } v = R(\varphi). \end{cases}$$

It is clear that $\alpha^* \in [0, 1]$. We shall show that

$$v \in \partial \mathscr{V}^{\alpha^*}. \tag{8}$$

For two points $e = (e^1, \ldots, e^N)$ and $g = (g^1, \ldots, g^N)$ in $\mathbb{R}^N$, we define the distance

$$d(e, g) = \max\{|e^i - g^i| : i = 1, \ldots, N\}.$$

Besides, for $E \subset \mathbb{R}^N$, $d(e, E) = \min_{g \in E}\{d(e, g)\}$.

For an arbitrary policy π and for an arbitrary $\alpha \in [0, 1]$, we consider policy π^α defined at epoch 0 by

$$\pi_0^\alpha(E|x) = \begin{cases} \pi_0(E|x), & \text{if } x \in X \setminus X_\alpha; \\ \mathbf{I}\{\varphi(x) \in E\}, & \text{if } x \in X_\alpha. \end{cases}$$

Then for $1 \geq \alpha \geq \beta \geq 0$ we have

$$\begin{aligned} &\left| R^n(\pi^\alpha) - R^n(\pi^\beta) \right| \\ &\quad = \left| \int_{X_\alpha \setminus X_\beta} \left(r^n(x, \varphi(x)) - \int_A r^n(x,a)\, \pi_0(da|x) \right) \mu(dx) \right| \\ &\quad \leq \mu\left(X_\alpha \setminus X_\beta\right) 2 \sup_{x,a} |r^n(x,a)| \leq 2C(\alpha - \beta), \end{aligned} \tag{9}$$

where $|r^n(x,a)| \leq C$, $n = 1, \ldots, N$, and the existence of a finite constant C follows from Condition 6(ii). Inequality (9) implies that for any policy $\pi \in \Delta$ and for any $\alpha, \beta \in [0,1]$

$$d\left(R(\pi^\alpha), R(\pi^\beta)\right) \leq 2C|\alpha - \beta|. \tag{10}$$

If (8) does not hold then either $v \notin \mathscr{V}^{\alpha^*}$ or $v \in \mathscr{V}^{\alpha^*} \setminus \partial \mathscr{V}^{\alpha^*}$.

Let $v \notin \mathscr{V}^{\alpha^*}$. Then $\alpha^* > 0$. We set $d = d(v, \mathscr{V}^{\alpha^*})$. Since $v \notin \mathscr{V}^{\alpha^*}$, $d > 0$. Let $\alpha = \max\{0, \alpha^* - \frac{d}{4(C+1)}\} < \alpha^*$. Inequality (10) implies that $\mathscr{V}^\alpha$ is within the $\frac{d}{2}$-neighborhood of $\mathscr{V}^{\alpha^*}$. Thus $d(v, \mathscr{V}^\alpha) > 0$ and therefore $v \notin \mathscr{V}^\alpha$. Since $\alpha^* > \alpha \geq 0$, we get a contradiction.

Now let $v \in \mathscr{V}^{\alpha^*} \setminus \partial \mathscr{V}^{\alpha^*}$. Then $\alpha^* < 1$. We set $D = d(v, \partial \mathscr{V}^{\alpha^*})$. Since $\mathscr{V}^{\alpha^*}$ is compact, $\partial \mathscr{V}^{\alpha^*}$ is compact. Since $v \notin \partial \mathscr{V}^{\alpha^*}$, $D > 0$. We set $\alpha = \min\{1, \alpha^* + \frac{D}{4(C+1)}\} > \alpha^*$. Then inequality (10) implies that $\mathscr{V}^{\alpha^*}$ belongs to the $\frac{D}{2}$-neighborhood of $\mathscr{V}^\alpha$ and $d(v, \partial \mathscr{V}^\alpha) > 0$. This implies $v \in \mathscr{V}^\alpha$. Since $\mathscr{V}^\alpha \subseteq \mathscr{V}^\beta \subseteq \mathscr{V}^{\alpha^*}$ for any $\beta \in [\alpha^*, \alpha]$ and $v \notin \mathscr{V}^\beta$ for some $\beta \in [\alpha^*, \alpha]$, we have $v \notin \mathscr{V}^\alpha$ which contradicts $v \in \mathscr{V}^\alpha$.

Formula (8) is proved. ∎

THEOREM 3. *If a one-step model satisfies Condition* 6 *then* $\mathscr{V} = \mathscr{V}^M$.

Proof. Let $n = \dim \mathscr{V}$. We observe that $n \leq N$. We prove this lemma by induction in n.

Let $n = 1$. In this case either $N = 1$ or for some $j = 1, \ldots, N$ there are constants c_i and d_i such that $R_i(\pi) = c_i + d_i R_j(\pi)$ for all $\pi \in \Delta$, $i = 1, \ldots, N$. Thus, if Theorem 3 holds for $N = 1$ then it holds for $n = 1$.

We prove the lemma for $N = 1$. In this case, $R(\pi) = R^1(\pi)$ and $\mathscr{V}$ is a closed interval in $\mathbb{R}^1$. Let $\mathscr{V} = [c, d]$.

We consider a policy π such that $R(\pi) = d$. For any policy π there exists a Markov policy φ^0 such that $R(\varphi^0) \geq R(\pi)$; see Feinberg [9]. Therefore, $d \in \mathscr{V}^M$. Similarly, $c \in \mathscr{V}^M$. Lemma 3 implies that $[c, d] \in \mathscr{V}^M$. Since $\mathscr{V}^M \subseteq \mathscr{V}$, we have $\mathscr{V}^M = \mathscr{V}$.

Let Theorem 3 hold when $\dim \mathscr{V} \leq n$ for some $n = 1, 2, \ldots$. Let $\dim \mathscr{V} = n + 1$ and $v \in \mathscr{V}$. By Lemma 7, $v \in \partial \mathscr{V}_{A[\varphi, Y]}$ for some selector φ and

for some Borel $Y \subseteq X$. By Lemma 6, Condition 6 holds for this submodel. Consider a hyperplane that supports $\mathscr{V}_{A[\varphi, Y]}$ at v. Lemma 5 implies that there is a submodel with the sets of available actions $B(\cdot) \subseteq A[\varphi, Y](\cdot)$ such that the following properties hold for this submodel: (i) Condition 6 holds, (ii) $v \in \mathscr{V}_B$, and (iii) $\dim \mathscr{V}_B \leq n$. The inductive assumption implies that $v = R(\psi)$ for some $\psi \in \Delta^M$. ▮

COROLLARY 1. *If Conditions* 1, 4, *and* 5 *hold then* $\mathscr{V} = \mathscr{V}^M$.

Proof. The corollary follows from Lemma 4 and Theorem 3. ▮

The following example demonstrates that if all assumptions of Corollary 1 hold except the assumption that the initial measure μ is nonatomic then $\mathscr{V}$ may not be equal to $\mathscr{V}^M$.

Let $X = [0, 1]$, $A = A(x) = \{0, 1\}$ for all $x \in X$. Let $\hat{x} \in X$ and $\mu(\{\hat{x}\}) = \beta > 0$. We consider some $d > 0$ such that

$$\mu\{(X \cap [\hat{x} - d, \hat{x} + d]) \setminus \{\hat{x}\}\} \leq \beta/3.$$

We set $r(x, 0) \equiv 0$;

$$r(x, 1) \triangleq \begin{cases} 1 - |\hat{x} - x|/d, & \text{if } |\hat{x} - x| \leq d; \\ 0, & \text{if } |\hat{x} - x| > d. \end{cases}$$

Let $\varphi^0(x) \equiv 0$ and $\varphi^1(x) \equiv 1$. Obviously, $R(\varphi^0) = 0$ and $R(\varphi^1) \geq \beta$. Since $\mathscr{V}$ is convex $\mathscr{V} \supseteq [0, \beta]$. Let us show that there does not exist a selector with the property $R(\varphi) = \beta/2$. Indeed, if $\varphi(\hat{x}) = 1$ then $R(\varphi) \geq \beta$ and if $\varphi(\hat{x}) = 0$ then $R(\varphi) \leq \beta/3$. So the set $\mathscr{V}^M$ is not convex. In view of Lemma 1, $\mathscr{V} \neq \mathscr{V}^M$.

4. REDUCTION OF A MULTIPLE-STEP MODEL TO A ONE-STEP MODEL

We define the set of all strategic measures

$$\mathbf{D} = \{P_\nu^\pi | \pi \in \Delta;\ \nu \in \mathscr{P}(X)\}$$

and its subsets

$$\mathbf{U} = \{P_\nu^\pi | \pi \in \Delta,\ \nu \in \mathscr{P}(X), \text{ and } \pi \text{ is nonrandomized at the epoch } t = 0\},$$

$$\mathscr{U}(x) = \{P_x^\pi | \pi \in \Delta \text{ and } \pi \text{ is nonrandomized at the epoch } t = 0\}, \quad x \in X.$$

According to Dynkin and Yushkevich [8, Sects. 3.5 and 5.5], $\mathbf{D}$ is a measurable subset of $\mathscr{P}(H_\infty)$.

LEMMA 8. **U** *and* $\mathscr{U}(x)$ *are measurable subsets of* **D**.

Proof. Since $\mathscr{D}(x)$ are measurable (Dynkin and Yushkevich [8, Sects. 3.5 and 5.5]) and $\mathscr{U}(x) = \mathbf{U} \cap \mathscr{D}(x)$, we need only to prove that **U** is measurable. Let Γ_1 be the subset of all probability measures $P \in \mathscr{P}(H_\infty)$ such that $P(da_0|x_0)$ is degenerated ($P(dx_0)$-a.s.). In other words, if Θ is a set of measures ρ on A such that $\rho(\Gamma) \in \{0, 1\}$ for any $\Gamma \in \mathscr{B}(A)$ then

$$\Gamma_1 = \{P \in \mathscr{P}(H_\infty) | P(da_0|dx_0) \in \Theta \ (P(dx_0)\text{-a.s.})\}.$$

By Lemma 6.1 in Parthasarathy [16, Chap. 2], Θ is a measurable subset of $\mathscr{P}(A)$.

Let Y_1 and Y_2 be two Polish spaces. We consider a natural projection f of $\mathscr{P}(Y_1 \times Y_2)$ on $\mathscr{P}(Y_1)$ defined by $f(P)(dy_1) = P(dy_1 \times Y_2)$ for all $P \in \mathscr{P}(Y_1 \times Y_2)$. Then mapping f is measurable. Indeed, the σ-field on $\mathscr{P}(Y_1)$ is defined as a sigma-field generated by sets $\Gamma(C, c) = \{P \in \mathscr{P}(Y_1) | P(C) \geq c\}$ for all $C \in \mathscr{B}(Y_1)$ and all real c. We have that $f^{-1}(\Gamma(C, c)) = \Gamma(C, c) \times \mathscr{P}(Y_2)$. Since the product of two Borel sets is a Borel set in a product topology, $f^{-1}(\Gamma(C, c))$ is measurable. So, the measurability of f is established.

We have that $\mathbf{U} = \mathbf{D} \cap \Gamma_1$. So, the measurability of Γ_1 implies the lemma. We define $\Gamma_2 \in \mathscr{P}(X \times A)$ by

$$\Gamma_2 = \{P \in \mathscr{P}(X \times A) | P(da_0|dx_0) \in \Theta \ (P(dx_0)\text{-a.s.})\}.$$

First, we observe that Γ_2 is a measurable subset of $\mathscr{P}(X \times A)$ in view of Sudderth [23, Lemma 2]. Second, for any $P \in \mathscr{P}(H_\infty)$ we consider its projection F on $\mathscr{P}(X \times A)$,

$$F(P)(dx_0\, da_0) = P\left(dx_0\, da_0 \times (X \times A)^\infty\right).$$

In view of the general fact explained in the previous paragraph, F is a measurable mapping of $\mathscr{P}(H_\infty)$ on $\mathscr{P}(X \times A)$. Since $\Gamma_1 = F^{-1}(\Gamma_2)$, Γ_1 is measurable. ∎

We consider the one-step model with the state space X, the state of available actions $\mathscr{U}(x) \subseteq \mathbf{D}$, and the reward vector-function $R(x, P) = \mathbf{R}(P)$, $P \in \mathbf{D}$. We remark that if π is a nonrandomized policy in the original model then the Ionescu Tulcea theorem implies that $x \to P_x^\pi$ is a measurable selector in the new model. We denote by $\tilde{\mathscr{V}}(\nu)$ the set of performance vectors for the new model when an initial distribution $\nu \in \mathscr{P}(X)$ is fixed.

LEMMA 9. $\mathscr{V}(\nu) = \tilde{\mathscr{V}}(\nu)$ *for any* $\nu \in \mathscr{P}(X)$.

Proof. Each of the sets $\mathscr{V}(\nu)$ and $\tilde{\mathscr{V}}(\nu)$ is empty if and only if for some $t = 0, 1, \ldots$ the set of measurable selectors in the original model is

empty. Therefore, $\mathscr{V}(\nu) = \varnothing$ if and only if $\tilde{\mathscr{V}}(\nu) = \varnothing$. According to the result by Derman and Strauch [7] mentioned before Condition 3, for a fixed initial distribution ν and for any policy σ there exists a randomized Markov policy π such that $R(\nu, \pi) = R(\nu, \sigma)$.

Let $v \in \mathscr{V}(\nu)$. We consider a randomized Markov policy π for which $R(\nu, \pi) = v$. For every pair $x \in X$, $a \in A_0(x)$, we denote by $u^{a,\pi}(x) = P_x^\sigma$ the strategic measure corresponding to the policy σ which is concentrated at the point $a \in A_0(x)$ for $t = 0$ and coincides with π for $t > 0$. The mapping

$$g(x, a) = u^{a,\pi}(x) : X \times A \to \mathbf{D} \tag{11}$$

is measurable in accordance with the Ionescu Tulcea theorem; see Neveu [15, Sect. V.1].

We wish to construct a measurable stochastic kernel γ^π from X to $\mathbf{D}$ such that $\gamma^\pi(\mathscr{U}(x)|x) = 1$ for all $x \in X$ and

$$\int_X \nu(dx) \int_{\mathbf{D}} \gamma^\pi(du|x) R(x, u) = R(\nu, \pi). \tag{12}$$

Let us introduce the stochastic kernel

$$\hat{\pi}_0(\Gamma|x) \triangleq \pi_0(\Gamma_x|x),$$

where $\Gamma \in \mathscr{B}(X \times A)$ is an arbitrary set, and $\Gamma_x = \{a \in A_0(x) | (x, a) \in \Gamma\}$ is the section of the set Γ at $x \in X$. For each nonnegative measurable function l on $X \times A$

$$\int_{X \times A} l(y, a) \hat{\pi}_0(d(y, a)|x)$$

$$= \int_{X \times A} l(y, a) \mathbf{I}\{x = y\} \hat{\pi}_0(d(y, a)|x) = \int_A l(x, a) \pi_0(da|x). \tag{13}$$

In view of Bertsekas and Shreve [5, Proposition 7.29], the last integral in (13) is a measurable function of $x \in X$. Therefore $\hat{\pi}_0$ is the measurable kernel. Using the mapping $g: X \times A \to \mathbf{D}$ defined in (11), we introduce the desired stochastic kernel γ^π by

$$\gamma^\pi(\Gamma|x) = \hat{\pi}_0(g^{-1}(\Gamma)|x),$$

where $\Gamma \in \mathscr{B}(\mathbf{D})$. The measurability of γ^π follows from the measurability of $\hat{\pi}_0$ and g. Since $\gamma^\pi(\Gamma|x) = \gamma^\pi(\Gamma \cap \mathscr{U}(x)|x)$ for any $\Gamma \in \mathscr{B}(\mathbf{D})$, we get

$$\gamma^\pi(\mathbf{D}|x) = \gamma^\pi(\mathscr{U}(x)|x) = \pi_0(A_0(x)|x) = 1.$$

Let us check equality (12). Clearly, for any $x \in X$

$$P_x^\pi = \int_A u^{a,\pi}(x)\,\pi_0(da|x) = \int_A g(x,a)\,\pi_0(da|x). \tag{14}$$

According to properties of images of probability measures (Meyer [14, Theorem T12, Chap. 2]), we have

$$\int_{\mathbf{D}} u\gamma^\pi(du|x) = \int_{X\times A} g(y,a)\,\hat{\pi}_0(d(y,a)|x) = \int_A g(x,a)\,\pi_0(da|x), \tag{15}$$

where the second equality is (13). We have from (14), (15) that

$$P_x^\pi = \int_{\mathscr{U}(x)} u\gamma^\pi(du|x).$$

Hence we have

$$R(x,P_x^\pi) = \int_{\mathscr{U}(x)} R(x,u)\,\gamma^\pi(du|x);$$

$$R(\nu,\pi) = \int_X R(x,P_x^\pi)\,\nu(dx) = \int_X \nu(dx)\int_{\mathscr{U}(x)} \gamma^\pi(du|x)\,R(x,u).$$

Formula (12) is proved. So, $\mathscr{V}(\nu) \subseteq \tilde{\mathscr{V}}(\nu)$.

Let $\gamma(du|x)$ be a stochastic kernel concentrated on $\mathscr{U}(x)$. Then the measure

$$P = \int_X \nu(dx)\int_{\mathbf{D}} u\gamma(du|x)$$

belongs to $\mathbf{D}$; see Dynkin and Yushkevich [8, Sects. 3.5 and 5.5]. From Fubini theorem, Meyer [14, Theorem T16, Chap. 2],

$$\int_X \nu(dx)\int_{\mathbf{D}} \gamma(du|x)\,R(x,u) = R(\nu,\pi)$$

for some policy π in the initial model. Hence, $\tilde{\mathscr{V}}(\nu) \subseteq \mathscr{V}(\nu)$. ∎

LEMMA 10. *If the original model satisfies Conditions* 1–5 *then the new one-step model satisfies Condition* 6.

Proof. We check each of the four Conditions 6.

(i) First of all, we show that the graph $\mathrm{Gr}\,\mathscr{U}$ of the multifunction $x \to \mathscr{U}(x)$ is closed in $X \times \mathbf{D}$. Let

$$\mathrm{Gr}\,\mathscr{U} \ni (x^n, P^n) \underset{n\to\infty}{\to} (x,P) \in X \times \mathbf{D}, \qquad n = 1,2,\ldots. \tag{16}$$

If $Q \in \mathbf{D}$ then $Q \in \mathscr{U}(y)$ if and only if $Q(\Gamma^X \times \Gamma^A \times (X \times A)^\infty) = \mathbf{I}\{y \in \Gamma^X, a \in \Gamma^A\}$ for some point $a \in A_0(y)$ and for all $\Gamma^X \in \mathscr{B}(X)$ and all $\Gamma^A \in \mathscr{B}(A)$. We consider points $a_n \in A$ such that $\mathbf{I}\{x^n \in \Gamma^X, a^n \in \Gamma^A\} = P^n(\Gamma^X \times \Gamma^A \times (X \times A)^\infty)$ for all $\Gamma^X \in \mathscr{B}(X)$ and all $\Gamma^A \in \mathscr{B}(A)$. Then (16) implies the existence of the limit $a = \lim_{n \to \infty} a^n$. According to Condition 1, $a \in A_0(x)$. Therefore, $P(\Gamma^X \times \Gamma^A \times (X \times A)^\infty) = \mathbf{I}\{x \in \Gamma^X, a \in \Gamma^A\}$. So, $a \in A_0(x)$ and $P \in \mathscr{U}(x)$. Thus, Gr $\mathscr{U}$ is closed.

Since Gr $\mathscr{U}$ is closed, the sets $\mathscr{U}(x)$ are closed for all $x \in X$. According to Lemma 1, $\mathscr{D}(x)$ is compact. Therefore, its closed subset $\mathscr{U}(x)$ is compact too. The measurability of the set $\mathbf{D}$ was explained before Lemma 8. Since Condition 1 implies the existence of a Markov policy φ (Dynkin and Yushkevich [8, Sect. 3.1]) and the mapping $x \to P_x^\varphi$ is measurable according to the Ionescu Tulcea theorem (Neveu [15, Sect. V.1]), the Gr $\mathscr{U}$ contains a measurable selector.

Conditions 2 and 4 imply that $R(x, P)$ is continuous in P for each $x \in X$. So, Condition (ii) holds. Condition (iii) is obvious, and Lemmas 1 and 9 imply that Condition (iv) holds for any initial measure ν. ▌

Let $\tilde{\Delta}^M$ be the set of nonrandomized Markov policies for the one-step model introduced in this section. In fact, $\tilde{\Delta}^M$ is the set of nonrandomized policies in the one-step model. Let $\tilde{\mathscr{V}}^M$ be the set of performance vectors for the policies from this set when the original initial distribution μ is fixed. The following result follows from Lemmas 9 and 10 and from Theorem 3.

THEOREM 4. *Let the original model satisfy Conditions* 1–5. *Then* $\tilde{\mathscr{V}}^M = \tilde{\mathscr{V}} = \mathscr{V}$.

5. PROOF OF THEOREM 1

Unless it is specified, we consider the original infinite-step model in this section.

LEMMA 11. *Let Conditions* 1–5 *be satisfied. For any policy* π *there exist a Markov policy* φ *and randomized Markov policies* γ^m, $m = 0, 1, \ldots$, *such that*: (i) $R(\gamma^m) = R(\pi)$ *and* (ii) *each policy* γ^m *is nonrandomized at steps* $t = 0, \ldots, m$ *and it coincides with* φ *at these steps, i.e.,* $\gamma_t^m(\Gamma|x) = \mathbf{I}\{\varphi_t(x) \in \Gamma\}$ *for* $t = 0, 1, \ldots, m$ *and* $\Gamma \in \mathscr{B}(A)$.

Proof. We fix a policy π. We construct policies γ^m and φ. Let $m = 0$. According to Theorem 4, there exists a measurable mapping ϕ from X to $\mathbf{U}$ such that $\hat{R}(\phi) = R(\pi)$. Here $\hat{R}$ is the performance vector in the one-step model introduced before Lemma 9. According to Dynkin and Yushkevich [8, Sect. 3.6], there exists a policy $\tilde{\sigma}$ in the original model such

that $P_x^{\tilde{\sigma}} = \phi(x)$ (μ-a.s.). Therefore, $R(\tilde{\sigma}) = R(\pi)$. We show that this policy can be selected nonrandomized at step 0. It means that $\tilde{\sigma}_0(\Gamma|x_0) = \mathbf{I}\{\varphi_0(x_0) \in \Gamma\}$ for some Markov policy φ and for all $\Gamma \in \mathscr{B}(A)$.

We consider the set Θ of all probability measures on A concentrated at one point. This set is a measurable subset of $\mathscr{P}(A)$; see Parthasarathy [16, Lemma 6.1, Chap. 2]. Therefore, the set $Y = \{x \in X | \tilde{\sigma}_0(\cdot|x) \in \Theta\}$ is a measurable subset of X. Since $P_x^{\tilde{\sigma}} \in \mathbf{U}$ (μ-a.s.), $\mu(x_0 \in Y) = 1$. Let ψ be an arbitrary selector from X to A_0. We can redefine $\tilde{\sigma}_t(\cdot|x)$ being equal to $\mathbf{I}\{\psi(x) \in \Gamma\}$ when $t = 0$, $x \in X \setminus Y$, and remaining unchanged on Y at epoch 0 and in all states at epochs $1, 2, \ldots$. Since $\mu(Y) = 1$, the measure $P_\mu^{\tilde{\sigma}}$ remains unchanged after this modification and policy $\tilde{\sigma}$ is nonrandomized at step 0. We set $\varphi_0(x)$ to be equal to the point where the measure $\tilde{\sigma}_0(\cdot|x)$ is concentrated.

By using the well-known procedure described by Derman and Strauch [7] (see also Strauch [22, Theorem 4.1]) we consider a randomized Markov policy γ^0 such that $R(\gamma^0) = R(\tilde{\sigma})$. According to this procedure, we can select $\gamma_0^0(\cdot) = \tilde{\sigma}_0(\cdot)$. This implies that $\gamma_0^0(\Gamma|x) = \mathbf{I}\{\varphi_0(x) \in \Gamma\}$ for all $\Gamma \in \mathscr{B}(A)$. The step $m = 0$ is completed.

Let the Markov policy φ be defined at the steps $0, 1, \ldots, m$. We consider the probability measure $\tilde{\mu}(dx_0) = P_\mu^{\varphi}(dx_{m+1})$ on X. Let $\tilde{A}_t = A_{m+1+t}$, $\tilde{r}_t = r_{m+1+t}$, and $\tilde{p}_t = p_{m+1+t}$, $t = 0, 1, \ldots$. We consider a new model with the action sets $\tilde{A}_t$, reward functions $\tilde{r}_t$, and transition probabilities $\tilde{p}_t$ at steps $t = 0, 1, \ldots$. The initial state distribution is $\tilde{\mu}$. Condition 5 implies that this measure is nonatomic.

We consider a Markov policy $\tilde{\pi}$, $\tilde{\pi}_t(da_t|x_t) = \gamma_{m+1+t}^m(da_{m+1+t}|x_{m+1+t})$, where $t = 0, 1, \ldots$. Let $\tilde{R}(\tilde{\pi})$ be the performance vector in the new model. We observe that

$$R(\pi) = R(\gamma^m) = E_\mu^{\gamma^m}\left\{\sum_{t=0}^{m} r_t(x_t, a_t) + \sum_{t=m+1}^{\infty} r_t(x_t, a_t)\right\}$$

$$= E_\mu^{\varphi}\left\{\sum_{t=0}^{m} r_t(x_t, a_t)\right\} + \tilde{R}(\tilde{\pi}).$$

We apply our result for $m = 0$ to the new model with the initial distribution $\tilde{\mu}$. We have that there is a Markov policy $\tilde{\gamma}$ such that this policy is not randomized at step 0 and $\tilde{R}(\tilde{\gamma}) = \tilde{R}(\tilde{\pi})$. We define $\varphi_{m+1}(x)$ to be equal to the point where the measure $\tilde{\gamma}_0(\cdot|x)$ is concentrated; for any $\Gamma \in \mathscr{B}(A)$

$$\gamma_t^{m+1}(\Gamma|x) = \begin{cases} \mathbf{I}\{\varphi_t(x) \in \Gamma\}, & \text{when } t = 0, \ldots, m+1; \\ \tilde{\gamma}_{t-m-1}(\Gamma|x), & \text{when } t > m+1. \end{cases}$$

Then

$$R(\gamma^{m+1}) = E_\mu^\varphi\left\{\sum_{t=0}^{m} r_t(x_t, a_t)\right\} + \tilde{R}(\tilde{\gamma})$$

$$= E_\mu^\varphi\left\{\sum_{t=0}^{m} r_t(x_t, a_t)\right\} + \tilde{R}(\tilde{\pi}) = R(\pi).$$

So, we have constructed the Markov policy φ and randomized Markov policies γ^m satisfying conditions of the lemma. ∎

Proof of Theorem 1. We fix an arbitrary policy π and consider the policy φ and policies γ^m, $m = 0, 1, \ldots$, from Lemma 11. We have that $R(\gamma^m) = R(\pi)$ and Condition 2 implies that for all $n = 1, \ldots, N$ and for all $m = 0, 1, \ldots$.

$$|R^n(\gamma^m) - R^n(\varphi)| \le 2\sum_{t=m+1}^{\infty} c_t.$$

Therefore, $R(\varphi) = R(\pi)$. ∎

6. APPLICATIONS

EXAMPLE 1. We consider a discrete-time single-product inventory system with finite capacity. The amount of inventory in the system is limited above by some number M. The demand at epoch $t = 0, 1, \ldots$ is ξ_t and $\xi_0, \xi_1, \ldots$ is a sequence of independent and identically distributed random variables. We assume that the distribution of ξ_t has no atoms and the ξ_t are bounded above with probability 1. The latter assumption and the nonnegativity assumption mean that $P\{0 \le \xi_t \le C\} = 1$ for some $0 < C < \infty$.

In general, back orders are allowed. However, when the inventory level declines below some level B, at least D units of inventory have to be ordered, where $D \ge C$. We assume that $D \le M - B$. Orders are placed after the demand is known and it is possible to order up to the full capacity M of the system.

Let $h(x)$ be the holding cost of the amount of x during one period of time. If $x < 0$ then $h(x)$ is the cost of back orders during one unit of time. We assume that the function $h(x)$ is continuous, $x \in]-\infty, \infty[$, and $h(0) = 0$. Ordering costs of u units are $K(u)$ when $u > 0$. We assume that $K(u)$ is a continuous function of $u \in [0, \infty[$. In the literature, typical examples of functions h and K are

$$h(x) = \begin{cases} bx, & \text{if } x \ge 0; \\ -b'x, & \text{if } x \le 0; \end{cases}$$

where $b' > b > 0$ and $K(u) = k + du$ for some $k, d > 0$ when $u > 0$.

Let the initial inventory be y, $B \le y \le M$. Then the initial state of the system $x_0 = y - \xi_0$. The random variable x_0 has the nonatomic distribution $\mu(x_0 \le c) = P\{\xi_0 \ge y - c\}$. The dynamics of the system is defined by the equation $x_{t+1} = x_t + a_t - \xi_{t+1}$, where $t = 0, 1, \ldots$ and a_t is the amount of inventory ordered which is the decision parameter, $0 \le a_t \le M - x_t$.

First, we describe a Markov decision process for this situation. Then we shall introduce the objective functions. Let $X = [B - C, M]$ be the state space and $A = [-1, M - B + C]$ be the action set. The sets of available actions for $t = 0, 1, \ldots$ are

$$A_t(x) = A(x) = \begin{cases} \{-1\} \cup [0, M - x], & \text{if } x \ge B; \\ [D, M - x], & \text{if } x < B. \end{cases}$$

We explain why we consider action $a = -1$. According to Conditions 1 and 4, the sets $A(x)$ are compact and the cost/reward functions are continuous. However, ordering costs $K(u)$ are not continuous at point $u = 0$ in many applications where $K(u) = k + du$ for small positive u and $K(0) = 0$. So, in our model $a = -1$ means that there is no order and $a = 0$ means that the order of size 0 has been placed. We define $K(-1) = 0$. The introduction of action $a = -1$ is possible because we have convexity assumptions neither on A nor on $A(x)$. We also define transition probabilities $p(x_{t+1} \le c | x_t, a) = P\{\xi_{t+1} \ge x_t + a - c\}$ $t = 0, 1, \ldots$.

We consider reward functions $r_t^1(x, a) = r^1(x, a) = -h(x)\mathbf{I}\{x \ge 0\}$, $r_t^2(x, a) = r^2(x, a) = -h(x)\mathbf{I}\{x \le 0\}$, and $r_t^3(x, a) = r^3(x, a) = -K(a)$, $t = 0, 1, \ldots$. Since $h(0) = 0$, the reward functions r^1 and r^2 are well-defined. In addition, all three functions are continuous in (x, a). For a discount factor $\beta \in [0, 1[$, we define for $n = 1, 2, 3$ the expected total discounted criteria

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

We observe that Conditions 1–5 hold for this model. We can consider various criteria which are functions of R^1, R^2, and R^3. For example, we can define $R^4 = R^1 + R^3$ as a criterion that characterizes operational costs which are the sums of ordering and holding costs. The criterion R^2 is related to backorders and it characterizes the quality of service. As an example, we can consider optimization of R^4 subject to constraints on R^2. Theorems 1 and 2 imply that (nonrandomized) Markov policies for this problem are as good as general policies.

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

nonatomic, weakly continuous, and for each $z \in]-\infty, \infty[$ there is $f(z)$ such that (i) $q_t(\{z_{t+1} \le f(z_t)\}|z_t) = 1$ and (ii) $f(\cdot)$ is bounded from above on every bounded interval $[z_1, z_2]$.

At each epoch $t = 1, \ldots, T$, the investor has two options: to sell the whole portfolio or to keep it. If the portfolio is sold at epoch t, the gain is z_t. The goal is to maximize the expected gain under the constraint that with at least probability $P > 0$ the gain is greater or equal than a given level C. We remark that if the value of the portfolio is negative, the investor should not sell it. In this case the policy to hold the portfolio forever yields better results.

We construct a Markov decision process for this problem. Let $D_1 = f(z_0)$ and $D_{t+1} = f(D_t)$. We set $D = \max\{D_t | t = 1, \ldots, T\}$. We consider the state space $X = \{0, 1\} \times]-\infty, D]$ and the action set $A = \{0, 1, 2\}$. Action 0 means to hold the portfolio, action 1 means to sell the portfolio at the price less than or equal to C, and action 2 means to sell the portfolio at the price greater than or equal to C. It is obvious that this problem is not feasible if $D \le C$. So, we consider the nontrivial case $D > C$.

The state of the system is $x_t = (0, z_{t+1})$ if the portfolio has not been sold and $x_t = (1, z_{t+1})$ otherwise, $t = 0, 1, \ldots, T-1$. In particular, $x_0 = (0, z_1)$ has a nonatomic distribution.

For $t = 0, \ldots, T-1$ we set $A_t(0, z) = \{0\}$ if $z < 0$, $A_t(0, z) = \{0, 1\}$ if $0 \le z < C$, $A(0, C) = \{0, 1, 2\}$, and $A_t(0, z) = \{0, 2\}$ if $C < z \le D$. If $x = (1, z)$ or $t > T - 1$, the control sets $A_t(x)$ are not important. We set $A_t(x) = \{0\}$ in these cases.

If at epoch $t = 0, \ldots, T-2$, the system is in state $x_t = (0, z)$ and action 0 is selected then the next state is $(0, y)$, where y has the distribution $q_{t+1}(dy|z)$. In all other situations, the system moves from state x_t to the state $(1, u)$ where u has a uniform distribution on $[D-1, D]$. The selection of q_{t+1} in the latter cases satisfies the nonatomic and continuity conditions.

For $t = 0, \ldots, T-1$, we define $r_t^1((0, z), i) = z$ for $i > 0$ and $r_t^2((0, z), 2) = 1$. We also set $r_t^n(x) = 0$ in all other situations, $n = 1, 2$. The problem of maximization of the expected gain subject to the constraint that the gain exceeds C with at least probability P is equivalent to the maximization of $R^1(\pi)$ subject to the constraint that $R^2(\pi) \ge P$. This model satisfies the conditions of Theorem 2. Therefore, if this problem is feasible, there exists an optimal Markov policy.

REFERENCES

1. E. Altman, Denumerable constrained Markov decision processes and finite approximations, *Math. Oper. Res.* **19** (1994), 169–191.

2. E. Altman, "Constrained Markov Decision Processes," Chapman & Hall, London, 1999.
3. J. P. Aubin and A. Cellina, "Differential Inclusions," Springer-Verlag, Berlin, 1984.
4. E. J. Balder, On compactness of the space of policies in stochastic dynamic programming, *Stochastic Process. Appl.* **32** (1989), 141–150.
5. D. P. Bertsekas and S. E. Shreve, "Stochastic Optimal Control," Academic Press, New York, 1978.
6. V. S. Borkar, Ergodic control of Markov chains with constraints—The general case, *SIAM J. Control Optim.* **32** (1994), 176–186.
7. C. Derman and R. E. Strauch, A note on memoryless rules for controlling sequential processes, *Ann. Math. Statist.* **37** (1966), 276–278.
8. E. B. Dynkin and A. A. Yushkevich, "Controlled Markov Processes and Their Applications," Springer-Verlag, New York, 1979.
9. E. A. Feinberg, Non-randomized Markov and semi-Markov strategies in dynamic programming, *SIAM Theory Probab. Appl.* **27** (1982), 116–126.
10. E. A. Feinberg and A. Shwartz, Constrained discounted dynamic programming, *Math. Oper. Res.* **21** (1996), 922–945.
11. E. B. Frid, On optimal strategies in control problems with constraints, *SIAM Theory Probab. Appl.* **17** (1972), 188–192.
12. L. C. M. Kallenberg, "Linear Programming and Finite Markovian Control Problems," Mathematical Centre Tracts, Vol. 148, Mathematisch Centrum, Amsterdam, 1983.
13. B. Liu and A. O. Esogbue, "Decision Criteria and Optimal Inventory Processes," Kluwer Academic, Boston, 1999.
14. P.-A. Meyer, "Probability and Potentials," Blaisdell, Waltham, 1966.
15. J. Neveu, "Mathematical Foundations of the Calculus of Probability," Holden–Day, San Francisco, 1965.
16. K. R. Parthasarathy, "Probability Measures on Metric Spaces," Academic Press, New York, 1967.
17. A. B. Piunovskiy, "Optimal Control of Random Sequences in Problems with Constraints," Kluwer Academic, Boston, 1997.
18. M. L. Puterman, "Markov Decision Processes," Wiley, New York, 1994.
19. K. W. Ross, Randomized and past dependent policies for Markov decision processes with finite action set, *Oper. Res.* **37** (1989), 474–477.
20. M. Schäl, On dynamic programming: Compactness of the space of policies, *Stochastic Process. Appl.* **3** (1975), 345–364.
21. L. I. Sennott, Constrained discounted Markov decision chains, *Probab. Engrg. Inform. Sci.* **5** (1991), 463–475.
22. R. E. Strauch, Negative dynamic programming, *Ann. Math. Statist.* **37** (1966), 871–890.
23. W. D. Sudderth, On the existence of good stationary strategies, *Trans. Amer. Math. Soc.* **135** (1969), 339–414.
24. K. Tanaka, On discounted dynamic programming with constraints, *J. Math. Anal. Appl.* **155** (1991), 264–277.
25. D. H. Wagner, Survey of measurable selection theorems, *SIAM J. Control Optim.* **15** (1977), 859–903.