Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

# On Occupation Measures for Total-Reward MDPs

Eric V. Denardo,[*] Eugene A. Feinberg[†] and Uriel G. Rothblum[‡]

## Abstract

*This paper is based on our recent contribution [3] that studies Markov Decision Processes (MDPs) with Borel state and action spaces and with the expected total rewards. The initial state distribution is fixed. According to [3], for a given randomized stationary policy, its occupation measure as a convex combination of occupation measures for simpler policies. If this is possible for a given policy, we say that the policy can be split. In particular, we are interested in splitting a randomized stationary policy into (nonrandomized) stationary policies or into a randomized stationary policies that are nonrandomized on a given subset of states. Though [3] studies Borel-state MDPs with expected total rewards, some of its results are new for finite state and action discounted MDPs. This paper focuses on these results.*

## 1. Introduction

For a Markov Decision Process (MDP), an occupation measure is a measure on the product of the state and action sets such that, for each measurable subset of this product, it is equal to the expected total number of events when state-action pairs belong to the subset. For MDPs with expected total rewards, occupation measures play an important role, because, if occupation measures coincide for two policies, the objective functions are equal for these policies for all reward functions. Furthermore, the set of occupation measures possesses several important properties including convexity. Typically for MDPs with multiple criteria and constraints, the first and the most important step for finding an optimal policy is to compute its occupation measure by solving an optimization problem in the set of all occupation measures.

We have recently studied in [3] representations of occupation measures for randomized stationary policies via convex combinations of occupation measures for nonrandomized (or nonrandomized on a subset of states) stationary policies in discrete-time total-reward MDPs with Borel state and action spaces. Such representations play an important role in the convex-analytical approach to MDPs [2, 10] which is the major method for the analysis of MDPs with multiple criteria and constraints [1, 9]. When such representation is possible, we say that a given policy can be split into the corresponding policies. This terminology is consistent with Altman's [1, p. 109] definition of splitting at a state. This paper focuses on two types of splitting: splitting at a state and finite splitting.

This paper focuses on the results from [3] that are new for discounted MDPs with finite state and action sets. We provide a new formula for a splitting at a state, describe splitting a randomized stationary policies via nonrandomized stationary policies and formulate an algorithm that implements such splitting, and describe a new class of optimal policies for constrained problems. Since occupation measures become finite-dimensional vectors for finite state and action MDPs, this paper deals with occupation vectors.

Splitting at a state is representing the occupation measure of a randomized stationary policy as a convex combination of occupation measures of randomized stationary policies with the following properties: (i) these policies are nonrandomized at this state, and (ii) they coincide with this policy outside of this state. Altman [1, p. 109] provided an explicit formula for splitting at a state for transient countable-state MDPs. [3, Theorem 5.1] gives necessary and sufficient conditions when splitting at a state is possible and unique. As follows from this theorem, for discounted MDPs splitting is always possible and unique. Formula (4.2) below expresses the splitting measure via the initial policy $\sigma$ and the occupation measures for splitting policies. It also expresses the splitting measure via the occupation

[*] Eric V. Denardo is with Center for System Science, Yale University, PO Box 208267, New Haven, CT 06520, USA.

[†] Eugene A. Feinberg is with the Department of Applied Math and Statistics, State University of New York At Stony Brook, Stony Brook, NY 11794 USA. efeinberg@notes.cc.sunysb.edu

[‡] Uriel G. Rothblum is with Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa 32000, Israel.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

measures for the initial and splitting policies.

Finite splitting deals with a randomized stationary policy $\sigma$ that uses a finite number of actions at some finite subset $Y$ of states. For finite state and action MDPs, finite splitting means that the occupation vector for each policy can be presented as a convex combination of a relatively small number of occupation measures for stationary policies and and consequent policies in this combination differ only at one state. In particular, this result implies that for a multiple-criterion problem with $K$ constraints, an optimal policy can be presented as a mixture of no more than $K+1$ nonrandomized stationary policies and consequent policies in the mix differ only at one state. Without the property that consequent policies in the mix differ only at one state, this result was described in Feinberg and Shwartz [7, Corollary 5.3] for discounted countable state MDPs.

## 2. Definition of the model

We consider a Markov decision process (MDP) $\{X, A, A(\cdot), p, r\}$, where: (i) $X$ is a finite state set, (ii) $A$ is a finite action set, (iii) $A(x) \subseteq A$ are sets of actions available at states $x \in X$; (iv) $p(y|x,a)$ is a transition probability, i.e., the probability that the next state is $y \in X$, if the current state is $x \in X$ and the action $a \in A(x)$ is selected, $\sum_{y\in X} p(y|x,a) = 1$ for all $x \in X$ and $a \in A(x)$, (v) $r(x,a)$ is a one-step reward if an action $a \in A(x)$ is selected at a state $a \in A(x)$. For constrained problems, $r(x,a)$ is a vector with the elements $r_k(x,a)$, $k = 0, 1, \dots, K$, where $K$ is the number of constraints.

As usual, a *policy* $\pi$ is a sequence of transition probabilities $\pi_t(a_t|h_t)$ concentrated on the sets $A(x_t)$, where $h_t = x_0, a_0, \dots, a_{t-1}, x_t$ is the observed history. If transition probabilities $\pi_t$ depend only on the current state and time, i.e. $\pi_t(\cdot|h_t) = \pi_t(\cdot|x_t)$ for all $t = 0, 1, \dots$, then the policy $\pi$ is called *randomized Markov*. If for a randomized Markov policy $\pi$ decisions do not depend on the time parameter, i.e. $\pi_t(\cdot|x) = \pi_s(\cdot|x)$, $x \in X$, the policy $\pi$ is called *randomized stationary*. For a randomized Markov policy $\pi$, we write $\pi(\cdot|x)$ instead of $\pi_t(\cdot|x)$. If each measure $\pi_t(\cdot|h_t)$ is concentrated at one point, the policy is called *nonrandomized*. A nonrandomized Markov and nonrandomized stationary policies are called *Markov* and *stationary*, respectively. A stationary policy is defined by a mapping $\phi$ from $X$ to $A$ such that $\phi(x) \in A(x)$ for all $x \in X$.

Let $R\Pi$ be the set of all policies, $\Pi$ be the set of nonrandomized policies, $RM$ be the set of randomized Markov policies, $M$ be the set of Markov Policies, $RS$ be the set of all randomized stationary policies, and $S$ be the set of stationary policies.

According to the Ionescu Tulcea theorem [8, p. 178], an initial distribution $\mu$ on $X$ and a policy $\pi$ define a unique probability measure $P^\pi_\mu$ on the space of trajectories $H_\infty = (X \times A)^\infty$ which is called *a strategic measure*. We denote by $E^\pi_\mu$ expectations with respect to $P^\pi_\mu$. We consider a $\sigma$-field on $H_\infty$ defined as a product of Borel $\sigma$-fields on $X$ and $A$. Throughout this paper, we fix the initial distribution $\mu$.

For a constant $\beta \in [0,1)$ called the discount factor and $\pi \in R\Pi$, define the e*xpected discounted rewards*

$$V^\pi(\mu) := E^\pi_\mu \sum_{n=0}^{\infty} \beta^n r(x_n, a_n).$$

For any policy $\pi$ denote by $Q^\pi_\mu$ the occupation vector defined for $x \in X$ and $a \in A(x)$ as

$$Q^\pi_\mu(x,a) := \sum_{n=0}^{\infty} \beta^n P^\pi_\mu\{x_n = x, a_n = a\}. \tag{2.1}$$

Then

$$V^\pi(\mu) = \sum_{x\in X} \sum_{a\in A(x)} r(x,a) Q^\pi_\mu(x,a). \tag{2.2}$$

Therefore, if $Q^\pi_\mu = Q^\sigma_\mu$ then $V^\pi(\mu) = V^\sigma(\mu)$ for any reward function $r$. Of course, if $P^\pi_\mu = P^\sigma_\mu$ then, according to (2.1), $Q^\pi_\mu = Q^\sigma_\mu$. In other words, if strategic measures are equal then occupation vectors are equal too.

For $x \in X$, set

$$q^\pi_\mu(x) := \sum_{a\in A(x)} Q^\pi_\mu(x,a).$$

## 3. Properties of strategic measures

For a set of policies $\Delta$, define $L^\Delta_\mu := \{P^\pi_\mu | \pi \in \Delta\}$ the set of strategic measures for the policies from $\Delta$. Obviously, $L^\Delta_\mu \subseteq L^{\Delta'}_\mu$ when $\Delta \subseteq \Delta'$. According to [6, Theorem 3.2], the set $L^\Delta_\mu$ is a measurable subset of $(\mathscr{P}(H_\infty), \mathscr{M}(H_\infty))$ when $\Delta = R\Pi, \Pi, RM, M, RS$, or $S$.

According to [4, Sections 3.5 and 5.5], the set $L^{R\Pi}_\mu$ is convex in the following strong case. For a probability measure $\nu$ on $L^{R\Pi}_\mu$, define the probability measure $P^\nu$ on $H_\infty$ by

$$P^\nu(E) := \int_{L^{R\Pi}_\mu} P(E)\nu(dP), \tag{3.1}$$

where $E$ are measurable subsets of $H_\infty$. Then $P^\nu \in L^{R\Pi}_\mu$; see [4, Sections 3.5 and 5.5]. In other words, there exists a policy $\pi$ such that $P^\pi_\mu = P^\nu$. A policy $\pi$ is called *mixed* if there exists a probability measure $\nu$ on $L^\Pi_\mu$ such that $\nu(L^\Pi_\mu) = 1$ and $P^\pi_\mu = P^\nu$ with $P^\nu$ defined by (3.1). A policy $\pi$ is called *mixed Markov* if there exists a probability measure $\nu$ on $L^{R\Pi}_\mu$ such that $\nu(L^M_\mu) = 1$ and $P^\pi_\mu = P^\nu$ with $P^\nu$ defined by (3.1). In other words, a policy $\pi$ is called mixed Markov if for some probability measure $\nu$ on $L_\mu^M$

$$P_\mu^\pi(E) = \int_{L_\mu^M} P(E)\nu(dP), \tag{3.2}$$

for all measurable subsets $E$ of $H_\infty$. We notice the sets $M$ and $\Pi$ are continuum even if the sets $X$ and $A$ are finite. However, for finite-horizon problems, the sets $\Pi$ and $M$ are finite and we can rewrite (3.2) as

$$P_\mu^\pi(h_n) = \sum_{\phi\in M} \nu(\phi) P_\mu^\phi(h_n), \tag{3.3}$$

where $n$ is the length of the horizon, for a mixed-Markov policy. The same formula holds for a general mixed policy with the summation in (3.3) changed over $\Pi$ instead of over $M$.

Similarly, a policy $\pi$ is called *mixed stationary* if for some probability distribution $\nu$ on $S$

$$P_\mu^\pi(E) = \sum_{\phi\in S} \nu(\phi) P_\mu^\phi(E), \tag{3.4}$$

for all measurable subsets $E$ of $H_\infty$.

According to [6, Theorem 5.1], any policy is a mixed policy and any randomized Markov policy is a mixed Markov policy. In other words: (i) for any policy $\pi$ there exists a probability measure $\nu$ on $L_\mu^\Pi$ such that (3.2) holds with $L_\mu^M$ substituted with $L_\mu^\Pi$, and (ii) for any randomized Markov policy $\pi$ there exists a probability measure $\nu$ on $L_\mu^M$ such that (3.2) holds. However, the similar statement does not hold for randomized stationary policies, because there may exist a randomized stationary policy $\pi$ such that equality (3.4) does not hold for any probability measure $\nu$ on $L_\mu^S$. In particular, Remark 3.1 in Feinberg [5] provides an example of an MDP with two stationary policies, i. e., $S = \{\phi^1, \phi^2\}$, such that $P_\mu^\pi \neq \alpha P_\mu^{\phi^1} + (1-\alpha)P_\mu^{\phi^2}$ for all $\pi \in RS\setminus S$ and for all $\alpha \in (0,1)$.

The situation is different when we consider a representation of an occupation vector $Q_\mu^\pi$, where $\pi \in R\Pi$, via a convex combination of vectors from $\{Q_\mu^\phi | \phi \in S\}$. Since the set of vectors $\{Q_\mu^\pi | \pi \in R\Pi\}$ is a convex hull of $\{Q_\mu^\phi | \phi \in S\}$, Caratheodory's theorem implies that such representations exist. Theorem 5.1 shows that stationary policies in such convex representations can be chosen in a way that sequential policies differ only in one decision.

## 4. Splitting at a state

For a randomized stationary policy $\sigma$, for a state $y \in X$, and for an action $a \in A(y)$, we denote by $\sigma[y,a]$ the randomized stationary policy that coincides with $\sigma$ at any state $x \neq y$ and always selects the action $a$ at $y$.

According to Altman [1, p. 108], for $y \in X$, *a probability vector* $\gamma^*(a)$, $a \in A(y)$, *splits a randomized stationary policy* $\sigma$ *at the state* $y$ if for any $x \in X$ and any $b \in A(x)$

$$Q_\mu^\sigma(x,b) = \sum_{a\in A(y)} \gamma^*(a) Q_\mu^{\sigma[y,a]}(x,b). \tag{4.1}$$

Altman [1, p. 109] provided an explicit formula for $\gamma^*$ that splits a randomized stationary policy at a state. [3, Theorem 5.1] provides necessary and sufficient conditions when a policy can be split and provides the splitting formula in a simpler form. The following theorem follows from [3, Theorem 5.1].

**Theorem 4.1.** *Consider a randomized stationary policy* $\sigma$ *and a state* $y \in X$.

*(i) If* $q_\mu^\sigma(y) = 0$ *then any probability vector on* $A(y)$ *splits* $\sigma$ *at* $y$.

*(ii) If* $q_\mu^\sigma(y) > 0$ *then*

$$\gamma^*(a) := \frac{\frac{\sigma(a|y)}{q_\mu^{\sigma[y,a]}(y)}}{\sum_{b\in A(y)} \frac{\sigma(b|y)}{q_\mu^{\sigma[y,b]}(y)}} = \frac{Q_\mu^\sigma(y,a)}{q_\mu^{\sigma[y,a]}(y)}, \quad a \in A(y), \tag{4.2}$$

*is the unique probability vector that splits* $\sigma$ *at* $y$.

## 5. Finite splitting at multiple states

For a set of policies $\Delta \subseteq R\Pi$, let $\mathscr{O}_\mu^\Delta := \{Q_\mu^\pi | \pi \in \Delta\}$ be the set of occupation vectors generated by the policies $\pi$ from $\Delta$. Obviously, $\mathscr{O}_\mu^\Delta \subseteq \mathscr{O}_\mu^{\Delta'}$ when $\Delta \subseteq \Delta'$. It is well-known, see e.g., [1], that $\mathscr{O}_\mu^{R\Pi} = \mathscr{O}_\mu^{RS}$. In addition: (i) $\mathscr{O}_\mu^{R\Pi}$ is a convex polytope consisting of all vectors $Q$ such that for all $x \in X$

$$\sum_{a\in A(x)} Q(x,a) = \mu(x) + \beta \sum_{y\in X}\sum_{a\in A(y)} p(x|y,a) Q(y,a)$$

and $Q(x,a) \geq 0$, $a \in A(x)$, and (ii) $\mathscr{O}_\mu^S$ is the set of extreme points of $\mathscr{O}_\mu^{R\Pi}$.

For a policy $\sigma^{R\Pi}$ and for a state $x \in X$, we define $A^\sigma(x) = \{a \in A(x) | Q_\mu^\sigma(x,a) > 0\}$. We denote by $S^\sigma$ the set of stationary policies $\phi \in S$ such that $\phi(x) \in A^\sigma(x)$ if $A^\sigma(x) \neq \emptyset$. We also set $X^\sigma = \{x \in X | A^\sigma(x) \neq \emptyset\}$.

The following theorem describes splitting at multiple states.

**Theorem 5.1.** *Let* $M = \sum_{y\in X} |A^\sigma(y)|$ *and* $m = M - |X^\sigma|$ *for an arbitrary policy* $\sigma$. *Then there exist* $(m+1)$ *stationary policies* $\phi^1, \dots, \phi^{m+1}$ *from* $S^\sigma$, *and there exist*

*$(m+1)$ nonnegative numbers $\alpha_1,\ldots,\alpha_{m+1}$ such that $\sum_{j=1}^{m+1}\alpha_j = 1$ and*

$$Q_\mu^\sigma = \sum_{j=1}^{m+1} \alpha_j Q_\mu^{\phi^j}. \tag{5.1}$$

*The stationary policies $\phi^1,\ldots,\phi^{m+1}$ can be selected in such a way that $\phi^i \neq \phi^j$ when $i \neq j$, and for each $i = 1,\ldots,m$ there is exactly one state $y^i \in Y$ such that $\phi^i(y^i) \neq \phi^{i+1}(y^i)$. In addition, any policy from $S^\sigma$ can be selected as $\phi^1$.*

The following example demonstrates that it may be impossible for all $\alpha_i$ to be positive in (5.1). This example also demonstrates that it may be impossible to select the stationary policies $\phi^1,\ldots,\phi^{m+1}$ in (5.1) in such a way that, in addition to $\phi^i$ and $\phi^{i+1}$, $i = 1,\ldots,m$, the stationary policies $\phi^{m+1}$ and $\phi^1$ also differ only at one state.

**Example 5.2.** Let $X = \{1,2\}$, $A = \{a^1,a^2\}$, $A(1) = A(2) = A$, $\mu(1) = \mu(2) = 0.5$, $p(x|x,a) = 1$ for all $(x,a) \in X \times A$, and there is a discount factor $\beta = 0.5$. Let $\pi$ be a randomized stationary policy with $\pi(a^i|x) = \pi(a^i|y) = 0.5$, $i = 1,2$. Then straightforward computations imply that $Q_\mu^\pi(x,a) = 0.5$ for all $(x,a) \in X \times A$. For a stationary policy $\phi$ we have that $Q_\mu^\phi(x,\phi(x)) = 1$ for all $x \in X$. It is easy to verify that the ordered sets $\{\alpha_1,\alpha_2,\alpha_3\}$ and $\{\phi^1,\phi^2,\phi^3\}$ of constants and stationary policies satisfy (5.1) if and only if $\alpha_1 = \alpha_3 = 0.5$, $\alpha_2 = 0$, $\phi^1(x) \neq \phi^3(x)$ for all $x \in X$, and either $\phi^2 = \{\phi^1(1),\phi^3(2)\}$ or $\phi^2 = \{\phi^3(1),\phi^1(2)\}$. ■

Next, we provide an algorithm that, for a policy $\sigma \in R\Pi$, constructs an equivalent mixture of randomized stationary policies in the form described in Theorem 5.1.

For a policy $\pi \in R\Pi$, the algorithm operates with the finite set

$$Z(\pi) := \{x \in X(\pi) : |A^\pi(x)| > 1\}.$$

**Algorithm 5.3.** *Input:* a policy $\sigma \in R\Pi$. *Outputs:* a natural number $m = \sum_{y\in X}|A^\sigma(y)| - |X^\sigma|$, nonnegative numbers $\alpha_1,\ldots,\alpha_{m+1}$ satisfying $\sum_{j=1}^{m+1}\alpha_j = 1$, and stationary policies $\phi^1,\ldots,\phi^{m+1}$ from $S^\sigma$ such that (5.1) holds.

1. Set $j := 1$, $\pi := \sigma$, compute $Z(\pi)$, set $Q^\pi(x,a) = Q_\mu^\pi(x,a)$ for $x \in Z(\pi)$, $a \in A^\pi(x)$, and select any stationary policy $\phi^1 \in S^\pi$.

2. While $|Z(\pi)| > 1$ do steps 2a – 2h:

2a compute $q_\mu^{\phi^j}(x)$ for $x \in Z(\pi)$ and set

$$\alpha := \min\{\frac{Q^\pi(x,\phi^j(x))}{q_\mu^{\phi^j}(x)} | x \in Z(\pi)\}$$

and

$$G(\pi) := \left\{x \in Z(\pi) | \frac{Q^\pi(x,\phi^j(x))}{q_\mu^{\phi^j}(x)} = \alpha\right\};$$

2b for $x \in G(\pi)$ set $A^\pi(x) := A^\pi(x) \setminus \{\phi^j(x)\}$;

2c set $\alpha_j := \alpha$;

2d set $k := |G(\pi)|$ and select $\phi^{j+k} : Y \to A$ such that $\phi^{j+k} \in A^\pi(x)$ when $x \in G(\pi)$ and $\phi^{j+k}(x) = \phi^j(x)\}$ when $x \in X \setminus G(\pi)$;

2e if $k > 1$ then order the elements of $G(\pi)$ in any way, $G(\pi) = \{x^1,\ldots,x^k\}$, and for $i = 1,\ldots,k-1$ set $\alpha_{j+i} := 0$ and set

$$\phi^{j+i}(x) := \begin{cases} \phi^{j+k}(x) & \text{if } x = x^\ell \text{ for } \ell = 1,\ldots,i, \\ \phi^j(x) & \text{if } x \in X \setminus G(\pi) \text{ or} \\ & x = x^\ell \text{ for } \ell = i+1,\ldots,k-1; \end{cases}$$

2f set $Z(\pi) := Z(\pi) \setminus \{x \in G(\pi) : |A^\pi(x)| = 1\}$;

2g for $x \in Z(\pi)$ set

$$Q^\pi(x,\phi^j(x)) := Q^\pi(x,\phi^j(x)) - \alpha q_\mu^{\phi^j}(x); \tag{5.2}$$

2h set $j := j+k$.

3. If $Z(\pi) = \emptyset$ then set $m := j-1$, $\alpha_{m+1} := 1 - \sum_{i=1}^m \alpha_i$, and stop, because formula (5.1) holds.

4. Let $Z(\pi) = \{x^*\}$ and $A^\pi(x^*) = \{a^0,\ldots,a^\ell\}$, where $a^0 = \phi^j(x^*)$. For $i = 1,\ldots,\ell$ define $\phi^{j+i}(x^*) = a^i$ and $\phi^{j+i}(x) = \phi^j(x)$ when $x \neq x^*$.

5. For $i = 0,\ldots,\ell-1$ compute $q_\mu^{\phi^{j+i}}(x^*)$ and set

$$\alpha_{j+i} := \frac{Q^\pi(x^*,a^i)}{q_\mu^{\phi^{j+i}}(x^*)}.$$

6. Set $m := j+\ell-1$, $\alpha_{m+1} := 1 - \sum_{i=1}^m \alpha_i$ and stop, because formula (5.1) holds. ■

We notice that the values of $Q^\pi(x,a)$, and the sets $A^\pi(x)$ and $Z(\pi)$ are computed only one time by using their definitions. This happens at step 1 for $\pi = \sigma$. Then they are iteratively modified at steps 2b, 2g, and 2f. We remark that it is possible to simplify the algorithm either by slightly increasing the number of operations it performs or by not providing the values of $\phi^j$ when $\alpha_j = 0$. In particular, it is possible to change the condition $|Z(\pi)| > 1$ to $|Z(\pi)| > 0$ in step 2 and exclude steps 4 – 6. Steps 4 – 6 are introduced only to utilize computational advantages of formula (5.1) for splitting at a state compared to $\ell$ executions of step 3 of the algorithm. Step 6 can be excluded by setting $i = 0,\ldots,\ell$ in step 5. Step 2e computes the stationary policies $\phi^j$

when $\alpha_j = 0$. These steps can be excluded and the modified algorithm would compute the coefficients and the mappings for equation (5.1) rewritten in the form

$$Q_\mu^\sigma = \sum_{j=1}^{\hat{m}+1} \hat{\alpha}_j Q_\mu^{\hat{\phi}^j},$$

with $\hat{m} \le m$ equal to the number of positive instances of $\alpha_j$ in (5.1), $\hat{\alpha}_j = \alpha_{N(j)}$ and $\hat{\phi}^j = \phi^{N(j)}$ for $j = 1,\ldots,\hat{m}$, where $N(0) = 0$ and $N(k+1) = \min\{i \ge N(k)+1 \,|\, \alpha_i > 0\}$ for $k = 0,\ldots,\hat{m}-1$.

## 6. Constrained optimization

In this section, we consider a problem with reward functions $r_k$, $k = 0,\ldots,K$, where $K < \infty$. The objection criteria are the expected total discounted rewards

$$V_k(\mu,\pi) = E_x^\pi \sum_{n=0}^{\infty} \beta^n r_k(x_n,a_n), \qquad k = 1\ldots,K,$$

where $0 \le \beta < 1$.

For a given initial distribution $\mu$ and for numbers $c_1,\ldots,c_K$, we consider the problem

$$\text{maximize} \quad V_0(\mu,\pi) \tag{6.1}$$

$$\text{subject to} \quad V_k(\mu,\pi) \ge c_k, \qquad k = 1,\ldots,K. \tag{6.2}$$

Problem (6.1)–(6.2) was studied by Feinberg and Shwartz [7] for a countable state problem. First, we recall some definitions from [7] adapted to the case of finite state and action sets.

Let $M = 0,1,\ldots$ . A randomized stationary policy $\pi$ is called *M-randomized stationary* if

$$\sum_{x\in X}\sum_{a\in A(x)} \mathbf{I}\{\pi(a|x) > 0\} \le |X| + M.$$

For a finite nonnegative integer $N$, a randomized stationary policy $\pi$ is called *a strong* $(M,N)$*-policy* if: (a) it is stationary from time $N$ onwards, i.e. $\pi_n(\phi(x)|x) = 1$ for some stationary policy $\phi$ for all $x \in X$ and for all $n \ge N$, and (b) for all states it uses no more than $M$ additional actions than a stationary policy would use, and (c) for all time-state pairs at epochs $n = 0,\ldots,N-1$, it uses no more than $M$ additional actions than a (nonrandomized) Markov policy would use, i.e., the following two properties hold:

$$\sum_{x\in X}\sum_{a\in A(x)} \mathbf{I}\{\sum_{n=0}^{N-1} \pi_n(a|x) > 0\} \le |X| + M$$

and

$$\sum_{n=0}^{N-1}\sum_{x\in X}\sum_{a\in A(x)} \mathbf{I}\{\pi_n(a|x) > 0\} \le N\cdot|X| + M.$$

A policy $\pi$ is called $M$-mixed stationary if there exist $(M+1)$ stationary policies $\phi^1,\ldots,\phi^{M+1}$ and $(M+1)$ nonnegative numbers $\alpha_1,\ldots,\alpha_{M+1}$ such that $\sum_{i=1}^{M} \alpha_i = 1$ and

$$P_\mu^\pi = \sum_{i=1}^{M+1} \alpha_i P_\mu^{\phi^i}.$$

Feinberg and Shwartz [7] proved the following results for a countable state space $X$:

**Theorem 6.1.** *(Feinberg and Shwartz [7, Theorem 2.1]) If problem (6.1)–(6.2) is feasible then*

*(i) there exists an optimal K-randomized stationary policy;*

*(ii) for some finite N there exists an optimal strong $(K,N)$-policy.*

**Theorem 6.2.** *(Feinberg and Shwartz [7, Theorem 5.1]) For any M-randomized stationary policy $\pi$ there exists an M-mixed stationary policy $\sigma$ with $Q_\mu^\sigma = Q_\mu^\pi$.*

**Corollary 6.3.** *(Feinberg and Shwartz [7, Corollary 5.3]) If problem (6.1)–(6.2) is feasible then there exists an optimal K-mixed stationary policy.*

The following statement follows from Theorem 5.1.

**Corollary 6.4.** *For any M-randomized stationary policy $\sigma$, there exists an m-mixed stationary policy $\pi$ with $Q_\mu^\pi = Q_\mu^\sigma$, where $m = \sum_{x\in X^\sigma} |A^\sigma(x)| - |X^\sigma| \le M$. In addition, the strategic measure $P_\mu^\pi$ can be presented in the form*

$$P_\mu^\pi = \sum_{i=1}^{m+1} \alpha_i P_\mu^{\phi^i}, \tag{6.3}$$

*where $\alpha_1,\ldots,\alpha_{m+1}$ are nonnegative numbers such that $\sum_{i=1}^{m+1} \alpha_i = 1$, and $\phi^1,\ldots,\phi^{m+1}$ are stationary policies such that $\phi^i \ne \phi^j$ when $i \ne j$ and $\phi^i$ and $\phi^{i+1}$ differ exactly at one point $x^i$, $i = 1,\ldots,m$.*

We notice that Corollary 6.4 is a stronger result than Theorem 6.2 because Corollary 6.4 specifies that the policies $\phi^i$ and $\phi^{i+1}$ in the mix differ only at one state. This may be interpreted in the way that the structure of the policy $\phi^i$ changes very little when $i$ increases by 1. The following theorem strengthens Corollary 6.3.

**Theorem 6.5.** *If problem (6.1)–(6.2) is feasible then for some $m = 0,\ldots,K$ there exists an optimal m-mixed stationary policy $\pi$ whose strategic measure $P_\mu^\pi$ can be presented in the form (6.3), where $\alpha_1,\ldots,\alpha_{m+1}$ are nonnegative numbers such that $\sum_{i=1}^{m+1} \alpha_i = 1$, and $\phi^1,\ldots,\phi^{m+1}$ are stationary policies such that $\phi^i \ne \phi^j$ when $i \ne j$ and $\phi^i$ and $\phi^{i+1}$ differ exactly at one point $x^i$, $i = 1,\ldots,m$.*

**Acknowledgments.** Research of the second author was partially supported by NSF Grant DMI-0600538 and by New York Foundation for Science, Technology, and Innovation (NYSTAR).

## References

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

[2] Borkar, V.S. 2002. Convex analytic methods in Markov decision processes. E.A. Feinberg and A. Shwartz, eds. *Handbook on Markov Decision Processes.* Kluwer Academic Publishers, Boston, 347–375.

[3] Denardo, E.V., Feinberg, E.A., Rothblum, U.G. 2008. Splitting randomized stationary policies in total-reward Markov Decision Processes. Preprint. Department of Applied Mathematics and Statistics, Stony Brook University, http://www.ams.sunysb.edu/ feinberg/public/DFR.pdf

[4] Dynkin, E.B., Yushkevich, A.A. 1979. *Controlled Markov Processes and their Applications.* Springer-Verlag, New York, USA.

[5] Feinberg, E.A. 1987. Sufficient classes of strategies in discrete dynamic programming I: decomposition of randomized startefies and embedded models. *SIAM Theory Prob. Appl.* **31** 658-668.

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

[7] Feinberg, E.A., Shwartz, A. 1996. Constrained discounted dynamic programming, *Mathematics of Operations Research* **21** 922-945.

[8] Hernández-Lerma, O., Lasserre, J.B. 1996. *Discrete-Time Markov Control Processes. Basic Optimality Criteria.* Springer, New York, USA.

[9] Piunovskiy, 1997. *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[10] Piunovskiy, A.B. 1998. Controlled random sequences: methods of convex analysis and problems with functional constraints. *Russian Math. Surveys* **53** 1233-1293.