MATHEMATICS OF OPERATIONS RESEARCH
Vol. 37, No. 1, February 2012, pp. 129–153
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.1110.0525

Splitting Randomized Stationary Policies in Total-Reward Markov Decision Processes

Eugene A. Feinberg
Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794,
eugene.feinberg@sunysb.edu

Uriel G. Rothblum
Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel,
rothblum@ie.technion.ac.il

This paper studies a discrete-time total-reward Markov decision process (MDP) with a given initial state distribution. A (randomized) stationary policy can be *split* on a given set of states if the occupancy measure of this policy can be expressed as a convex combination of the occupancy measures of stationary policies, each selecting deterministic actions on the given set and coinciding with the original stationary policy outside of this set. For a stationary policy, necessary and sufficient conditions are provided for splitting it at a single state as well as sufficient conditions for splitting it on the whole state space. These results are applied to constrained MDPs. The results are refined for absorbing (including discounted) MDPs with finite state and actions spaces. In particular, this paper provides an efficient algorithm that presents the occupancy measure of a given policy as a convex combination of the occupancy measures of finitely many (stationary) deterministic policies. This algorithm generates the splitting policies in a way that each pair of consecutive policies differs at exactly one state. The results are applied to constrained problems to efficiently compute an optimal policy by computing and splitting a stationary optimal policy.

Key words: Markov decision processes; occupancy measures; splitting occupancy measures; constrained Markov decision processes
MSC2000 subject classification: Primary: 90C40; secondary: 97K60, 60J20, 60J22
OR/MS subject classification: Primary: dynamic programming/optimal control; secondary: deterministic Markov, finite state, infinite state
History: Received April 25, 2008; revised December 11, 2010, and October 16, 2011. Published online in *Articles in Advance* January 9, 2012.

1. Introduction. This paper is concerned with a discrete-time Markov decision process (MDP) with a given distribution of an initial state and with total-reward criteria. It investigates whether and how a stationary policy can be replaced by another policy that is defined as a random selection among policies that are deterministic on a prescribed set of states and coincide with the original stationary policy outside of that set. Contributions are presented for MDPs with finite state and actions sets, for MDPs with countable state sets, and for MDPs with Borel state and action spaces.

An MDP is said to be *absorbing* if its expected lifetime is finite under every policy. In particular, a discounted MDP can be presented as an absorbing MDP. For an absorbing MDP with a fixed initial state distribution, the *occupancy measure* of a policy specifies the expectation of the number of visits to each measurable set of state-action pairs. The expected total reward for the policy can be expressed as the integral of the one-step reward function with respect to the policy's occupancy measure. Thus, optimizing the expected total rewards is reduced to optimizing a linear function over the set of occupancy measures. This is the basic idea of the *convex-analytic approach* which provides useful methods for solving MDPs with multiple criteria and constraints. The convex-analytical approach was introduced by Derman; see Derman [12] and references therein. Most of the later developments are recapped in monographs by Kallenberg [35], Borkar [6], Piunovskiy [42], Altman [1], and Hernández-Lerma and Lasserre [32], and in surveys by Piunovskiy [43] and Borkar [7]. It has various applications including to the Hamiltonian cycle problem; see Feinberg [21], Filar [28, §§3.3, 3.4], and Ejov et al. [15]. The convex-analytic approach is also applicable to average rewards per unit time with expected state-action frequencies playing the role of occupancy measures (Derman [12], Kallenberg [35], Borkar [6], Piunovskiy [42], Altman [1]).

Two policies that have the same occupancy measure have the same expected total reward for any one-step reward function. The gist of splitting over a set of states S is to match the occupancy measure of a given stationary policy σ to the occupancy measure of a policy that randomizes (or, in other words, mixes) over policies that make deterministic decisions on S and match σ on the remaining states. When a policy is split, its expected total reward under each one-step reward function can be expressed as a convex combination of the expected total rewards of the splitting policies.

The principal results of this paper are as follows:

(i) For absorbing Borel MDPs that satisfy certain compactness and continuity conditions, Theorem 4.1 states that a stationary policy can be split on the entire state space into deterministic policies.

(ii) Theorem 5.1 presents necessary and sufficient conditions for a stationary policy to be split at a single state of a Borel MDP.

(iii) For absorbing MDPs with finite state and action sets, Theorem 6.1 establishes the existence of finite splitting of a stationary policy on the entire state space via deterministic policies that can be ordered so that successive policies differ only at a single state; Algorithm 1 efficiently executes such splitting. Furthermore, Theorem 7.1 presents conditions under which the existence result extends to MDPs with Borel state and action sets.

(iv) For constrained absorbing MDP with finite state and action sets, Algorithm 2 efficiently computes an optimal policy in the form of a mixture of deterministic policies that can be ordered in a way described in (iii). For constrained countable state discounted MDPs, Theorem 10.2 presents sufficient conditions for the existence of optimal policies with the above structure.

(v) For constrained absorbing Borel MDPs with compact action sets, Theorem 9.2(ii) establishes the existence of a stationary optimal policy and an optimal policy being a mixture of deterministic policies.

The conclusions of (i)—splitting on the entire space—has been established by Borkar [6, p. 37] for countable state discounted MDPs with compact action sets and by Piunovskiy [43, Theorem 6] for Borel state discounted MDPs that satisfy certain compactness and weak-continuity conditions. Related results for nonstationary policies appear in Feinberg [16, 20]. We do not know whether Theorem 4.1 holds without the continuity and compactness conditions. The conditions in (ii) are necessary and sufficient for splitting at a single state for total-reward MDPs. Splitting at a state was introduced by Hordijk and Spieksma [34, p. 415] and by Spieksma [47, p. 168] for average-reward MDPs satisfying certain conditions. The term "splitting" was introduced by Spieksma [47, p. 168]. For total rewards, splitting at a state was established by Altman [1, p. 109] for transient countable state MDPs. A countable state MDP is *transient* if under each policy the expected number of visits to each state is finite. A countable absorbing MDP is transient, whereas the reverse implication holds for finite state MDPs, but not for MDPs with infinitely countable state spaces. The results in (iii) are new even for discounted MDPs. Finite splitting of state-action frequencies were extended in Feinberg [23] to unichain finite state and action average-reward MDPs. Algorithm 2, mentioned in (iv), is the first algorithm of polynomial complexity for finding optimal mixtures of deterministic policies for constrained MDPs. Previous approaches (see Feinberg [19] and Altman [1, Chapter 9]) for computing such policies had exponential bounds on the running time. The result in (v) closes a gap between the known facts on countable state constrained absorbing MDPs (Altman [1]) and on Borel-state constrained discounting MDPs (Piunovskiy [42, 43], Hernández-Lerma and González-Hernández [30]).

The concept of embedded MDPs, introduced in Feinberg [17] and discussed at the end of §4, is useful for a reduction of the state space of an MDP. In particular, it is used in the current paper to reduce splitting at a subset of the state space to splitting on the entire state space for an MDP whose state space is this subset.

The strategic measure of a policy is the probability measure it induces on the set of state-action sequences. If strategic measures coincide for two policies, then their occupancy measures coincide, and, in addition, any probabilistic criterion defined on the space of state-action sequences coincide for these two policies (not just expected total rewards, as takes place for equal occupancy measures). The occupancy measure of a policy can be viewed as a projection of its strategic measure and it contains limited information about the underlying stochastic sequence, while the strategic measures provide complete probabilistic description of this sequence. Variants of the convex-analytic method were applied to strategic measures; see §3 or Feinberg [20] for representations of strategic measures for general past-dependent and Markov policies that are similar to splitting of occupancy measures. Those results do not require continuity or compactness assumptions as does Theorem 4.1 which concerns splitting of occupancy measures. The study of splitting of occupancy measures is important primarily for situations when similar results do not take place for strategic measures.

2. The model. Before we introduce an MDP, we summarize some definitions and notation that we use. For a topological space B, we always consider its *Borel σ-field* $\mathcal{B}$ (the smallest σ-field containing all open subsets of B); sets in $\mathcal{B}$ are then called *Borel sets*. A *standard Borel space* is a pair $(B, \mathcal{B})$ with B being a nonempty Borel subset of a Polish (complete, separable, metric) space. To simplify notations, when referring to a standard Borel space $(B, \mathcal{B})$, the reference to $\mathcal{B}$ is usually omitted. Also, everywhere in this paper "measurable" means "Borel measurable."

For a standard Borel space B, denote by $\mathcal{P}(B)$ and $\mathcal{Q}(B)$, respectively, the sets of all probability measures and all finite nonnegative measures on $(B, \mathcal{B})$. The minimal σ-field on $\mathcal{Q}(B)$, containing the sets $\{\nu \in \mathcal{Q}(B) \mid \nu(E) \le c\}$ for all $c \in [-\infty, \infty]$ and for all $E \in \mathcal{B}$, is denoted by $\mathcal{R}(B)$. Also, $\mathcal{M}(B)$ is the σ-field on $\mathcal{P}(B)$

$$\begin{array}{ccccc} S & \Rightarrow & M & \Rightarrow & \Pi \\ \Downarrow & & \Downarrow & & \Downarrow \\ RS & \Rightarrow & RM & \Rightarrow & R\Pi \end{array}$$

FIGURE 1. Relationships between classes of policies.

formed by the intersections of the sets in $\mathscr{R}(B)$ with $\mathscr{P}(B)$. Then $(\mathscr{P}(B), \mathscr{M}(B))$ is a standard Borel space; see Dynkin and Yushkevich [14, Appendix 5]. We also notice that $\mathscr{P}(B)$ is a convex subset of the linear space of all signed finite measures on $(B, \mathscr{B})$. Similarly, $(\mathscr{Q}(B), \mathscr{R}(B))$ is a standard Borel space and $\mathscr{Q}(B)$ is a convex cone in the linear space of all signed finite measures on $(B, \mathscr{B})$. Everywhere in this paper we follow the convention that $0 \cdot \infty = 0$.

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

(i) X is a standard Borel space (the state space);

(ii) A is a standard Borel space (the action space);

(iii) $A(x)$ is a nonempty subset of A for each state $x \in X$ (the set of actions available at x), where the graph $\mathrm{Gr}(A) = \{(x, a)\colon x \in X, a \in A(x)\}$ is a measurable subset of $X \times A$ and where there exists a measurable mapping $\varphi\colon X \to A$ with $\varphi(x) \in A(x)$ for all $x \in X$;

(iv) $p(\cdot \mid x, a)$ is a probability measure on X for each $(x, a) \in X \times A$ (the transition probability), such that $p(B \mid \cdot, \cdot)$ is a measurable function on $X \times A$ for each measurable subset B of X; and

(v) $r(\cdot, \cdot)$ is a real-valued measurable function on $X \times A$ that is bounded from above (the reward function).

A *policy* π is a sequence of measurable transition probabilities $\pi_t(da_t \mid h_t)$ concentrated on the sets $A(x_t)$, where $h_t = x_0, a_0, \ldots, a_{t-1}, x_t$ is the observed history. If transition probabilities π_t depend only on the current state and time, that is, $\pi_t(\cdot \mid h_t) = \pi_t(\cdot \mid x_t)$ for all $t = 0, 1, \ldots$, then the policy π is called *Markov*. If for a Markov policy π, decisions do not depend on the time parameter, that is, $\pi_t(\cdot \mid x) = \pi_s(\cdot \mid x)$ for all $x \in X$, then the policy π is called *stationary*. For a stationary policy π, we write $\pi(\cdot \mid x)$ instead of $\pi_t(\cdot \mid x)$. If each measure $\pi_t(\cdot \mid h_t)$ is concentrated at one point, then the policy is called *nonrandomized*. Nonrandomized stationary policies are also called deterministic. A deterministic policy is defined by a measurable mapping ϕ 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, let Π be the set of nonrandomozed policies, let RM be the set of Markov policies, let M be the set of nonrandomized Markov policies, let RS be the set of all stationary policies, and let S be the set of deterministic policies. For readers' convenience, we observe that classes of nonrandomized policies are denoted by a single letter, and their extensions that allow randomization are denoted by the same letter preceded by R. The relationships among the above classes are summarized in Figure 1, where $\Rightarrow$ stands for $\subseteq$.

According to the Ionescu Tulcea theorem (Hernández-Lerma and Lasserre [31, p. 178]), 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. We consider a σ-field on H_∞ defined as a product of Borel σ-fields on X and A. Throughout this paper, we fix the initial distribution μ.

For a constant $\beta \in [0, 1)$, called the *discount factor*, an initial distribution μ on X, and a policy $\pi \in R\Pi$, define the *expected total discounted reward*

$$\tilde{V}_\beta^\pi(\mu) := E_\mu^\pi \sum_{n=0}^{\infty} \beta^n r(x_n, a_n). \tag{1}$$

Following Hordijk [33] and Altman [1], we consider a slightly more general situation than an MDP with the expected total discounted reward criterion. We set $\beta = 1$ and assume that there is a special (possibly empty) measurable subset $\mathscr{Z}$ of the state space X such that the process stops when it reaches $\mathscr{Z}$. In other words, $p(x \mid x, a) = 1$ and $r(x, a) = 0$ for all $x \in \mathscr{Z}$ and $a \in A(x)$. Consider the stopping time $T = \inf\{n \geq 0 \mid x_n \in \mathscr{Z}\}$, with the minimum over the empty set defined as $+\infty$. If $\mathscr{Z} = \varnothing$, then $T = \infty$. Since $T = \sum_{n=0}^{\infty} \mathbf{I}\{n < T\}$, where $\mathbf{I}$ is the indicator function, for any policy π,

$$E_\mu^\pi T = E_\mu^\pi \sum_{n=0}^{\infty} \mathbf{I}\{n < T\}.$$

For a fixed initial distribution μ, an MDP is called *absorbing* (or absorbing to $\mathscr{Z}$), if $E_\mu^\pi T < \infty$ for any policy π. For an absorbing MDP, the *expected total reward* is

$$V^\pi(\mu) := E_\mu^\pi \sum_{n=0}^{\infty} r(x_n, a_n) = E_\mu^\pi \sum_{n=0}^{T-1} r(x_n, a_n), \tag{2}$$

where we follow everywhere the convention that a sum from i to $j < i$ equals 0. In addition, if $r(x,a) = 1$, when $x \in X \setminus \mathcal{X}$ and $a \in A(x)$, then $V^\pi(\mu) = E^\pi_\mu T$. According to Dynkin and Yushkevich [14, §§4.4 and 5.5], if the reward function r is nonnegative and $V^\pi(\mu) < \infty$ for all policies $\pi \in R\Pi$, then $\sup\{V^\pi(\mu) \mid \pi \in R\Pi\} < \infty$. Thus, if an MDP is absorbing, then $E^\pi_\mu T \le K_\mu$ for some $K_\mu < \infty$. According to Blackwell [4], for positive dynamic programming (that is, the reward function r is nonnegative and $\sup\{V^\pi(\mu) \mid \pi \in R\Pi\} < \infty$), the equality $\sup\{V^\pi(\mu) \mid \pi \in R\Pi\} = \sup\{V^\phi(\mu) \mid \phi \in S\}$ holds. Therefore, an MDP is absorbing if and only if there exists a finite constant $K_\mu < \infty$ such that $E^\phi_\mu T \le K_\mu$ for all deterministic policies ϕ.

We recall (Altman [1, p. 137]) that an absorbing MDP is a more general concept than a discounted MDP, because a discounted MDP can be transformed into an absorbing MDP by adding an additional state x^* to X, by choosing the absorbing set $\mathcal{X} = \{x^*\}$, and by setting the transition probabilities from $x \in X$:

$$p^*(Y \mid x,a) := \begin{cases} \beta p(Y \mid x,a) & \text{if } Y \text{ is a measurable subset of } X,\\ 1-\beta & \text{if } Y = \{x^*\}. \end{cases}$$

Then $V^\pi(\mu) = \tilde{V}^\pi_\beta(\mu)$ for any policy π and for any reward function r, where $V^\pi(\mu)$ is the expected total reward for the absorbing MDP with the transition probabilities p^*, and $\tilde{V}^\pi_\beta(\mu)$ is the expected total discounted reward for the original MDP; see, e.g., Altman [1, p. 137]. We remark that an absorbing MDP can be defined as an MDP with substochastic transition probabilities p (that is, $p(X \mid x,a) \le 1$ for each $x \in X$ and $a \in A(x)$) and without considering explicitly the set $\mathcal{X}$; see, e.g., Feinberg and Sonin [26], and Denardo et al. [11]. However, this setup is not followed in the current paper, and we assume $p(X \mid x,a) = 1$ for each $x \in X$ and $a \in A(x)$.

For any policy π denote by Q^π_μ the occupancy measure on $X \times A$ defined by

$$Q^\pi_\mu(Y \times B) := E^\pi_\mu \sum_{n=0}^{T-1} \mathbf{I}\{x_n \in Y, a_n \in B\} = \sum_{n=0}^{\infty} P^\pi_\mu\{x_n \in Y \setminus \mathcal{X}, a_n \in B\}, \tag{3}$$

where Y and B are measurable subsets of X and A, respectively. In particular, $Q^\pi_\mu(\mathcal{X} \times A) = 0$ and, according to the arguments provided on p. 112 in Altman [1], for a bounded above function r,

$$V^\pi(\mu) = \int_X \int_A r(x,a) Q^\pi_\mu(dx, da). \tag{4}$$

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 (3), $Q^\pi_\mu = Q^\sigma_\mu$. In other words, if strategic measures are equal, then occupancy measures are equal too.

We also define the measure q^π_μ on X by setting

$$q^\pi_\mu(Y) := Q^\pi_\mu(Y \times A)$$

for any measurable subset Y of X. Observe that

$$q^\pi_\mu(Y) = \mu(Y) + E^\pi_\mu \sum_{n=1}^{T-1} \mathbf{I}\{x_n \in Y\}.$$

Therefore, the measure μ is absolutely continuous with respect to q^π_μ.

To simplify notation, we shall write $Q^\pi_\mu(Y,B)$ instead of $Q^\pi_\mu(Y \times B)$, and for $x \in X$ identify the set $\{x\}$ with the point x; e.g., we write $Q^\pi_\mu(x,B)$ and $q^\pi_\mu(x)$ instead of $Q^\pi_\mu(\{x\},B)$ and $q^\pi_\mu(\{x\})$, respectively. Also, for the probability measure δ_x on X that is concentrated at $x \in X$, we replace δ_x with x; e.g., we write P^π_x, Q^π_x, and q^π_x instead of $P^\pi_{\delta_x}$, $Q^\pi_{\delta_x}$, and $q^\pi_{\delta_x}$, respectively.

3. Strategic measures. For a set of policies Δ, let $L^\Delta_\mu := \{P^\pi_\mu \mid \pi \in \Delta\}$ (the set of strategic measures for the policies in Δ). Obviously, $L^\Delta_\mu \subseteq L^{\Delta'}_\mu$ when $\Delta \subseteq \Delta'$. According to Feinberg [20, Theorem 3.2], the set L^Δ_μ is a measurable subset of $(\mathcal{P}(H_\infty), \mathcal{M}(H_\infty))$ when $\Delta \in \{R\Pi, \Pi, RM, M, RS, S\}$.

According to Dynkin and Yushkevich [14, §§3.5 and 5.5], the set $L^{R\Pi}_\mu$ is convex in the following strong sense. For a probability measure ν on $L^{R\Pi}_\mu$, define the probability measure P^ν on H_∞ by

$$P^\nu(E) := \int_{L^{R\Pi}_\mu} P(E)\, \nu(dP), \tag{5}$$

where E is a measurable subset of H_∞. Then $P^\nu \in L^{R\Pi}_\mu$. In other words, there exists a policy π such that $P^\pi_\mu = P^\nu$. This convexity property of the set of strategic measures is relevant to Kuhn's theorem on sufficiency of behavioral policies in stochastic games (Kuhn [39], Aumann [2]).

Let Δ be a set of policies satisfying the condition that L_μ^Δ is a measurable set; that is, $L_\mu^\Delta \in \mathscr{M}(H_\infty)$. A policy π is called *a mixture of policies from* Δ, if there exists a probability measure ν on the set of strategic measures $L_\mu^{R\Pi}$ such that $\nu(L_\mu^\Delta) = 1$ and for all measurable subsets E of H_∞:

$$P_\mu^\pi(E) = \int_{L_\mu^\Delta} P(E)\,\nu\,(dP). \tag{6}$$

A policy π is called *a mixture of deterministic policies* if there exists a probability measure ν on $L_\mu^{R\Pi}$ such that $\nu(L_\mu^S) = 1$ and, for all measurable subsets E of H_∞, (6) holds with $\Delta = S$. These definitions can be interpreted as a random selection, up front, of a policy from Δ (or, in particular, from S) using the probability measure ν.

According to Feinberg [20, Theorem 5.2], any policy is a mixture of nonrandomized policies, and any Markov policy is a mixture of nonrandomized Markov policies. In other words,

(i) for any policy π there exists a probability measure ν on L_μ^Π such that (6) holds with $\Delta = \Pi$; and

(ii) for any Markov policy π there exists a probability measure ν on L_μ^M such that (6) holds with $\Delta = M$.

For a countable state MDP, Krylov [36] establishes a result similar to (i). For Borel state controlled stochastic processes (not necessarily MDPs), Gikhman and Skorohod [29, proof of Theorem 1.2] provide a version of (i) and Feinberg [16] provides a version of (i) and (ii), when integration in formulae similar to (6) is taken over an artificial set introduced by Aumann [2] rather than directly over the set of the corresponding strategic measures. Feinberg [17, 18] formulates general sufficient conditions when the results similar to (i) and (ii) hold for particular classes of policies.

Feinberg [17, Remark 3.1] provides an example of an MDP with two deterministic policies, that is, $S = \{\phi^1, \phi^2\}$, such that $P_\mu^\pi \neq \alpha P_\mu^{\phi^1} + (1-\alpha)P_\mu^{\phi^2}$ for all $\pi \in RS$ and for all $\alpha \in (0,1)$. Thus, even for a simple MDP with one state and two actions, strategic measures of stationary policies may not be represented via strategic measures for deterministic policies in the way similar to (i) and (ii).

According to Theorem 4.1 from the following section, under Schäl's [45] compactness Condition (S) or (W), a representation similar to (i) and (ii) holds for occupancy measures of stationary policies via occupancy measures of deterministic policies. This contrasts to statements (i) and (ii) that hold for general and Markov policies without any compactness or any continuity conditions.

4. Occupancy measures. For a policy $\pi \in R\Pi$, consider the occupancy measures Q_μ^π and q_μ^π, where the former is a measure on $X \times A$ and the latter is its projection on X.

LEMMA 4.1. *Let $q_\mu^\pi(X) < \infty$ for a policy $\pi \in R\Pi$. Then there exists a q_μ^π-a.e. unique stationary policy σ such that*

$$Q_\mu^\pi(Y, B) = \int_Y \sigma(B \mid x)\, q_\mu^\pi\,(dx) \tag{7}$$

for any measurable sets Y and B from X and A, respectively.

PROOF. We recall that a transition probability σ from X to A is a measurable mapping $\sigma[x]$ of X to $\mathscr{P}(A)$. Following our previous notations, we write $\sigma(B \mid x)$, $x \in X$, instead of $\sigma[x](B)$ for any measurable subset B of A. By dividing both sides of (7) by $q_\mu^\pi(X)$, we see that the existence of a transition probability σ satisfying (7) follows from the existence of a conditional distribution for a probability measure on $X \times A$; see Dynkin and Yushkevich [14, Appendix 4]. Since $Q_\mu^\pi((X \times A)\backslash \mathrm{Gr}(A)) = 0$, the transition probability σ can be defined in such a way that $\sigma(A(x) \mid x) = 1$, $x \in X$. In other words, σ is a stationary policy satisfying (7). The q_μ^π-a.e. uniqueness of σ means that if $\sigma = \sigma_1$ and $\sigma = \sigma_2$ satisfy (7), then $q_\mu^\pi(\{x \in X \mid \sigma_1(\cdot \mid x) \neq \sigma_2(\cdot \mid x)\}) = 0$, where $\sigma_1(\cdot \mid x) = \sigma_2(\cdot \mid x)$ if and only if $\sigma_1(B \mid x) = \sigma_2(B \mid x)$ for any measurable subset B of A. The q_μ^π-a.e. uniqueness of σ follows from the following two observations: (i) if $\sigma = \sigma_1$ and $\sigma = \sigma_2$ satisfy (7) for some measurable subset B of A, then $\sigma_1(B \mid x) = \sigma_2(B \mid x)$ q_μ^π-a.e.; and (ii) there exists a countable set $\{B_1, B_2, \dots\}$ of measurable subsets of A such that any two measures on A are equal if and only if they are equal on B_i for all $i = 1, 2, \dots$. We observe that (i) follows directly from (7), and (ii) holds because A is a standard Borel space. If A is a real line $(-\infty, \infty)$, then we can select $B_i = (-\infty, z_i]$, $i = 1, 2 \dots$, where $\{z_1, z_2, \dots\}$ is the set of rational numbers. Any standard Borel space is either countable or isomorphic to the real line; see, e.g., Dynkin and Yushkevich [14, Appendix 1]. Thus, if A is uncountable, then B_i, $i = 1, 2 \dots$, can be selected as isometric images of $(-\infty, z_i]$. For a countable $A = \{a^1, a^2, \dots\}$, we can set $B_i = \{a^i\}$, $i = 1, 2, \dots$. □

For an absorbing MDP, $q_\mu^\pi(X) < \infty$ for any policy π, and the corresponding condition in Lemma 4.1 holds. Lemma 4.1 also holds when the measure q_μ^π is locally finite, but we do not need this fact in this paper.

The following Lemma 4.2 states that if an MDP is absorbing, then for the fixed initial measure μ, the occupancy measures for π and σ, where σ is defined by (7), coincide. This result was originally established

by Borkar [5] for countable state discounted MDPs and by Krylov [37, 38] for controlled discounted diffusion processes. For countable state MDPs, this result is proved in Altman [1, p. 102]. For Borel-state discounted MDPs, this result is proved in Piunovskiy [42, pp. 141, 307]. Example 4.3 in Feinberg and Sonin [27] (see also Altman [1, p. 103]) shows that the absorbing assumption is essential. In particular, Lemma 4.2 does not hold for countable state transient MDPs; see §1 for the definition of transience.

LEMMA 4.2. *For an arbitrary policy π in an absorbing MDP, let σ be a q_μ^π-a.e. unique stationary policy satisfying (7). Then $Q_\mu^\sigma = Q_\mu^\pi$.*

PROOF. Altman's [1, p. 102] proof remains correct for Borel-state MDPs after the standard measurability considerations, described in Piunovskiy [42, p. 307], are added to it. □

For a set of policies $\Delta \subseteq R\Pi$, let $\mathcal{O}_\mu^\Delta := \{Q_\mu^\pi \mid \pi \in \Delta\}$ be the set of occupancy measures generated by the policies π from Δ. Obviously, $\mathcal{O}_\mu^\Delta \subseteq \mathcal{O}_\mu^{\Delta'}$ when $\Delta \subseteq \Delta'$.

LEMMA 4.3. *For an absorbing MDP, the set $\mathcal{O}_\mu^{RS}$ coincides with the set of all finite measures Q on $X \times A$ satisfying $q(\mathcal{X}) = 0$ and*

$$q(Y) = \mu(Y) + \int_X \int_A p(Y \mid x, a)\, Q(dx, da) \tag{8}$$

for any measurable subset Y of X, where $q(Y) = Q(Y, A)$.

PROOF. For countable state MDPs, this lemma is proved in Altman [1, Lemma 7.1]. The proof remains correct for the uncountable state case; see Piunovskiy [42, Lemma 25, p. 141], where this lemma is proved for Borel-state discounted MDPs. □

COROLLARY 4.1. *For an absorbing MDP, $\mathcal{O}_\mu^{RS} = \mathcal{O}_\mu^{R\Pi}$ and this set is convex.*

PROOF. Lemma 4.2 states the equality $\mathcal{O}_\mu^{RS} = \mathcal{O}_\mu^{R\Pi}$, and Lemma 4.3 implies the convexity of $\mathcal{O}_\mu^{RS}$. □

Consider a mapping g of L_μ^{RS} on $\mathcal{O}_\mu^{RS}$ defined as $g(P_\mu^\pi) := Q_\mu^\pi$ for all $\pi \in R\Pi$.

LEMMA 4.4. *For an absorbing MDP, the mapping g: $L_\mu^{RS} \to \mathcal{O}_\mu^{RS}$ is a one-to-one correspondence; that is, $P_\mu^\pi = P_\mu^\sigma$ if and only if $Q_\mu^\pi = Q_\mu^\sigma$ for any $\pi, \sigma \in RS$.*

PROOF. It is obvious that $P_\mu^\pi = P_\mu^\sigma$ implies $Q_\mu^\pi = Q_\mu^\sigma$ for any policies π and σ, because $Q_\mu^\delta(C) = \sum_{n=0}^\infty P_\mu^\delta\{(x_n, a_n) \in C \backslash (\mathcal{X} \times A)\}$ for any policy δ, where C is any measurable subset of $X \times A$. We shall prove that $Q_\mu^\pi = Q_\mu^\sigma$ implies $P_\mu^\pi = P_\mu^\sigma$ for any $\pi, \sigma \in RS$. For any policy δ we define marginal distributions $P_\mu^{\delta,n}(B) = P_\mu^\delta(x_n \in B \backslash \mathcal{X})$, where B is an arbitrary measurable subset of X. Then $q_\mu^\delta = \sum_{n=0}^\infty P_\mu^{\delta,n}$ and measures $P_\mu^{\delta,n}$ are absolutely continuous with respect to q_μ^δ, $n = 0, 1, \ldots$.

Let $Q_\mu^\pi = Q_\mu^\sigma$ for some $\pi, \sigma \in RS$. Then $q_\mu^\pi = q_\mu^\sigma$. Denote $q_\mu = q_\mu^\pi$. Lemma 4.2 implies that $\pi(\cdot \mid x) = \sigma(\cdot \mid x)$ q_μ-a.e. This equality implies that $\pi(\cdot \mid x) = \sigma(\cdot \mid x)$ $P_\mu^{\pi,n}$-a.e. and $P_\mu^{\sigma,n}$-a.e. for all $n = 0, 1, \ldots$. In addition, $P_\mu^{\pi,0} = P_\mu^{\sigma,0} = \mu$. By induction, we have that $P_\mu^\pi(C_n) = P_\mu^\sigma(C_n)$ for any measurable subset C_n of $X \times (A \times X)^{n-1}$. Kolmogorov's theorem on finite-dimensional distributions implies that $P_\mu^\pi = P_\mu^\sigma$. □

As defined in the beginning of §2, $\mathcal{Q}(X \times A)$ is the set of finite measures on $X \times A$, and $\mathcal{R}(X \times A)$ is the Borel σ-field on $\mathcal{Q}(X \times A)$. The following lemma claims that the sets of occupancy measures generated by stationary and by deterministic policies are measurable subsets of $\mathcal{Q}(X \times A)$.

LEMMA 4.5. *For an absorbing MDP, $\mathcal{O}_\mu^{RS} \in \mathcal{R}(X \times A)$ and $\mathcal{O}_\mu^S \in \mathcal{R}(X \times A)$.*

PROOF. According to Feinberg [20, Theorem 3.2], L_μ^{RS} and L_μ^S are measurable subsets of the standard Borel space $(\mathcal{P}(H_\infty), \mathcal{M}(H_\infty))$. Consider the one-to-one mapping g defined in Lemma 4.4. The formula

$$Q_\mu^\pi(Y, B) = \sum_{n=0}^\infty P_\mu^\pi\{x_n \in Y, a_n \in B\}, \tag{9}$$

where Y is a measurable subset of $X \backslash \mathcal{X}$ and B is a measurable subset of A, implies that the mapping g is measurable. The Lusin-Purves theorem (Cantón et al. [8, p. 279]) implies that the sets $\mathcal{O}_\mu^{RS} = g(L_\mu^{RS})$ and $\mathcal{O}_\mu^S = g(L_\mu^S)$ belong to $\mathcal{R}(X \times A)$. □

LEMMA 4.6. *For an absorbing MDP, $\mathcal{O}_\mu^S$ is the set of extreme points of the set $\mathcal{O}_\mu^{RS}$.*

PROOF. By Piunovskiy [42, Theorem 19], $\mathcal{O}_\mu^S$ is the set of the extreme points of the set $\mathcal{O}_\mu^{RS} = \mathcal{O}_\mu^{R\Pi}$ for discounted MDPs. The proof of this theorem in Piunovskiy [42] relies on the properties of occupancy measures and the statement of Lemma 4.2 for discounted MDPs. Since all of these facts hold for absorbing MDPs, the statement of the lemma holds for absorbing MDPs. □

Next we shall consider conditions under which the set of occupancy measures $\mathcal{O}_\mu^{RS}$ is compact in some natural topology. Since the set of all occupancy measures $\mathcal{O}_\mu^{R\Pi}$ is the projection of the set of all strategic measures $L_\mu^{R\Pi}$, it is natural to look at conditions when $L_\mu^{R\Pi}$ is compact. Such conditions, namely conditions (W) and (S) below, were introduced by Schäl [45].

Condition (W)

(i) The sets $A(x)$ are compact for all $x \in X$.

(ii) The set-valued mapping $A(\cdot)$ is upper semicontinuous; that is, for any open subset G of A, the set $\{x \in X \mid A(x) \subseteq G\}$ is open in X.

(iii) The transition probability $p(\cdot \mid x, a)$ is weakly continuous on $(X \times A)$; that is, if $(x_i, a_i) \to (x, a)$, then $\int_X f(y)p\,(dy \mid x_i, a_i) \to \int_X f(y)p\,(dy \mid x, a)$ for any bounded continuous function f on X.

Condition (S)

(i) The same as (i) in Condition (W).

(ii) The transition probability $p(\cdot \mid x, a)$ is setwise continuous in $a \in A(x)$; that is, $p(B \mid x, a_i) \to p(B \mid x, a)$ as $a_i \to a$ for any measurable subset B of X.

We consider the w-topology on $\mathcal{P}(X \times A)$. The w-topology is the coarsest topology on $\mathcal{P}(X \times A)$ in which all mappings $\mu \to \int_{X\times A} f(x, a)\mu\,(dx\,da)$ are continuous for bounded continuous functions f.

Lemma 4.7. *Consider an absorbing MDP. If either Condition* (W) *or Condition* (S) *holds, then* $\mathcal{O}_\mu^{RS}$ *is compact in the* w*-topology on* $\mathcal{P}(X \times A)$.

Proof. Consider the w-topology on $\mathcal{P}(H_\infty)$. This is the coarsest topology on H_∞ in which all mappings $p \to E_p f(h_n)$ are continuous for all bounded continuous functions f on $H_n = (X \times A)^n$ for all $n = 0, 1, \ldots$, where $p \in \mathcal{P}(H_\infty)$ and E_p is the expectation operator with respect to the measure p. Conditions (W) and (S) imply that the set of strategic measures $L_\mu^{R\Pi}$ is compact in the w-topology (Balder [3], Nowak [40]).

We shall show that $P_\mu^\pi \to Q_\mu^\pi$ is a continuous mapping. Let $\pi_1, \pi_2, \ldots$, be a sequence of policies such that $P_\mu^{\pi_k}$ w-converges to a strategic measure P_μ^π as $k \to \infty$. This means that for any $N = 0, 1, \ldots$, and for any bounded continuous function $f(x_0, a_0, \ldots, x_N, a_N)$,

$$E_\mu^{\pi_k} f(x_0, a_0, \ldots, x_N, a_N) \to E_\mu^\pi f(x_0, a_0, \ldots, x_N, a_N). \tag{10}$$

We need to show that $\int_X \int_A g(x, a)\, Q_\mu^{\pi_k}\,(dx, da) \to \int_X \int_A g(x, a)\, Q_\mu^\pi\,(dx, da)$ for any bounded continuous function g on $X \times A$, which, in view of (4) and (2), is equivalent to

$$E_\mu^{\pi_k} \sum_{n=0}^{\infty} g(x_n, a_n) \to E_\mu^\pi \sum_{n=0}^{\infty} g(x_n, a_n). \tag{11}$$

Fix a bounded continuous function g on $X \times A$. Let $|g(x, a)| \le K < \infty$ for all $x \in X$ and all $a \in A$.

For any finite $N = 1, 2, \ldots$, formula (10) with $f(x_0, a_0, \ldots, x_N, a_N) = \sum_{n=0}^N g(x_n, a_n)$ yields $E_\mu^{\pi_k} \sum_{n=0}^N g(x_n, a_n) \to E_\mu^\pi \sum_{n=0}^N g(x_n, a_n)$. Since the MDP is absorbing, $E_\mu^\sigma T \le K_\mu < \infty$ for some constant K_μ. Therefore,

$$\left| E_\mu^\sigma \sum_{n=N+1}^{\infty} g(x_n, a_n) \right| \le E_\mu^\sigma \sum_{n=N+1}^{\infty} |g(x_n, a_n)| \le K P_\mu^\sigma\{T \ge N+1\} \le \frac{K E_\mu^\sigma T}{N+1} \le \frac{K \cdot K_\mu}{N+1},$$

where the third inequality follows from the Markov inequality. Thus,

$$\left| E_\mu^{\pi_k} \sum_{n=0}^{\infty} g(x_n, a_n) - E_\mu^\pi \sum_{n=0}^{\infty} g(x_n, a_n) \right| \le \left| E_\mu^{\pi_k} \sum_{n=0}^{N} g(x_n, a_n) - E_\mu^\pi \sum_{n=0}^{N} g(x_n, a_n) \right| + \frac{2K \cdot K_\mu}{N+1},$$

and this inequality implies (11). So $P_\mu^\pi \to Q_\mu^\pi$ is a continuous mapping, and therefore $\mathcal{O}_\mu^{R\Pi}$ is compact as a continuous image of a compact set $L_\mu^{R\Pi}$. □

When $A(x) = A$ for all $x \in X$, Condition (W)(ii) obviously holds. In this case, for Condition (W), Lemma 4.7 is presented in Piunovskiy [43, Theorem 6] for discounted MDPs.

Theorem 4.1. *Consider an absorbing MDP. If either Condition* (W) *or Condition* (S) *holds, then for any policy* π *there exists a probability measure* ρ *on* $\mathcal{O}_\mu^S$ *such that for any measurable subset* C *of* $X \times A$,

$$Q_\mu^\pi(C) = \int_{\mathcal{O}_\mu^S} Q(C)\,\rho\,(dQ). \tag{12}$$

PROOF. According to Corollary 4.1 and Lemma 4.7, $\mathcal{O}_\mu^{R\Pi} = \mathcal{O}_\mu^{RS}$ is a convex compact subset of the set of signed finite measures under the w-topology. By Rudin [44, Theorem 3.10], the set of signed finite measures on the Borel σ-field of $X \times A$ is locally convex under the w-topology. Lemmas 4.5 and 4.6 imply that $\mathcal{O}_\mu^S$ is the set of extreme points of $\mathcal{O}_\mu^{RS}$, and $\mathcal{O}_\mu^S$ is measurable. Thus, (12) follows from the Choquet–Bishop–de Leeuw theorem (Phelps [41, p. 17]). □

In some sense, Theorem 4.1, which deals with occupancy measures and stationary policies, is similar to Feinberg [20, Theorem 5.1] described in statements (i) and (ii) in §3. Currently, it is not clear whether Theorem 4.1 holds without continuity and compactness assumptions (W) or (S). Corollary 5.2 below presents a particular case of Theorem 4.1 that requires neither continuity nor compactness assumptions. If the MDP is not absorbing, the statement of Theorem 4.1 does not hold even when the state and action sets are finite; see Kallenberg [35, Theorem 3.3.4, Example 3.3.3] and Example 5.1 below.

Let Y be a measurable subset of $X\backslash\mathfrak{X}$. We denote by τ_Y the time of the first visit to Y, namely $\tau_Y := \min\{n \ge 0 \mid x_n \in Y\}$. For a policy π, any measurable set $Z \subseteq X\backslash\mathfrak{X}$, and for any measurable set $B \subseteq A$, we define the expected total time spent in $Z \times B$ until the first visit to Y as

$$g_x^{Y,\pi}(Z,B) = E_x^\pi \sum_{n=0}^{\tau_Y - 1} \mathbf{I}\{x_n \in Z, a_n \in B\}, \quad x \in X\backslash\mathfrak{X},$$

and $g_\mu^{Y,\pi}(Z,B) = \int_X g_x^{Y,\pi}(Z,B)\mu(dx)$. Then, standard calculations imply that for any stationary policy π and for any measurable sets $Z \subseteq X\backslash(Y \cup \mathfrak{X})$ and $B \subseteq A$,

$$Q_\mu^\pi(Z,B) = g_\mu^{Y,\pi}(Z,B) + \int_Y \int_A \left[\int_{X\backslash(Y\cup\mathfrak{X})} g_x^{Y,\pi}(Z,B) p(dx \mid y,a) \right] Q_\mu^\pi(dy,da). \tag{13}$$

Let us fix a stationary policy σ and a measurable subset Y of $X\backslash\mathfrak{X}$. Denote by $RS(Y,\sigma)$ the set of stationary policies such that the policy σ is always used at states $x \in X\backslash(Y \cup \mathfrak{X})$. Let $S(Y,\sigma)$ be the set of stationary policies that select deterministic actions on Y and act like σ outside of $Y \cup \mathfrak{X}$. Set $\mathcal{O}_\mu^S(Y,\sigma) = \{Q_\mu^\pi \mid \pi \in S(Y,\sigma)\}$ and $\mathcal{O}_\mu^{RS}(Y,\sigma) = \{Q_\mu^\pi \mid \pi \in RS(Y,\sigma)\}$.

Observe that $\mathcal{O}_\mu^S(Y,\sigma)$ and $\mathcal{O}_\mu^{RS}(Y,\sigma)$ are measurable subsets of the sets of all finite measures on $X \times A$. Indeed, for a finite nonnegative measure Q on $X \times A$, define its marginal measure q on X: $q(Z) = Q(Z \times A)$ for any measurable subset Z of X. We also define the measure $j_{Y,\sigma}(Q)$ by $j_{Y,\sigma}(Z \times B) = Q((Z \cap Y), B) + \int_{Z\backslash(Y\cup\mathfrak{X})} \sigma(B \mid y)\, q(dy)$. From Dubins and Freedman [13, 2.1], we have that $Q \to j_{Y,\sigma}(Q)$ is a measurable mapping of $(\mathcal{Q}(B), \mathcal{R}(B))$ to itself. In addition, $Q_\mu^\Delta(Y,\sigma) = \{Q \in \mathcal{O}_\mu^\Delta \mid Q = j_{Y,\sigma}(Q)\}$, where $\Delta = S$ or $\Delta = RS$. Since $\mathcal{O}_\mu^\Delta \in \mathcal{R}(X \times A)$, we have $Q_\mu^\Delta(Y,\sigma) \in \mathcal{R}(X \times A)$, where $\Delta \in \{S, RS\}$.

In view of Theorem 4.1, the natural question is whether for any stationary policy $\pi \in RS(Y,\sigma)$ there exists a probability measure ρ^* on $\mathcal{O}_\mu^S(Y,\sigma)$ such that for any measurable subset C of $X \times A$,

$$Q_\mu^\pi(C) = \int_{\mathcal{O}_\mu^S(Y,\sigma)} Q(C)\, \rho^*(dQ). \tag{14}$$

Equality (13) implies that (14) holds for all measurable subsets of $X \times A$ if and only if it holds for all measurable subsets of $Y \times A$. We observe that (14) is a straightforward generalization of (12), because (14) becomes (12) when $Y = \varnothing$. The following three sections of this paper deal with two particular situations when (14) holds without explicit assumptions that either Condition (W) or Condition (S) holds: (i) $Y = \{y\}$ is a singleton, and (ii) the set Y is finite and for each $x \in Y$ the set $A(x)$ is finite.

We conclude this section with the construction of an embedded MDP introduced in Feinberg [17] for countable state MDPs and general policies. Here we need this construction only when a stationary policy σ is fixed on $X\backslash Y$. Again, we fix a policy $\sigma \in \Pi^{RS}$ and a measurable subset Y of X. Let $\tau_Y^1 = \inf\{n \ge 1 \mid x_n \in Y\}$. We consider an MDP with the state space $Y \cup \mathfrak{X}$, action sets $A(x)$, $x \in Y \cup \mathfrak{X}$, and transition probabilities p_Y^σ such that $p_Y^\sigma(x \mid x,a) = 1$ if $x \in \mathfrak{X}$, and for any measurable subset Y' of $Y \cup \mathfrak{X}$,

$$p_Y^\sigma(Y' \mid x,a) = \begin{cases} p(Y' \mid x,a) + \displaystyle\int_{X\backslash Y} P_z^\sigma\{\tau_Y < \infty, x_{\tau_Y} \in Y'\} p(dz \mid x,a) & \text{if } x \in Y \text{ and } Y' \subseteq Y, \\ p(\mathfrak{X} \mid x,a) + \displaystyle\int_{X\backslash Y} P_z^\sigma\{\tau_Y = \infty\} p(dz \mid x,a) & \text{if } x \in Y \text{ and } Y' = \mathfrak{X}. \end{cases}$$

For a stationary policy $\tilde\pi$ in the embedded MDP, we denote by $\sigma[Y,\tilde\pi]$ the stationary policy for the original MDP that coincides with $\tilde\pi$ on Y and coincides with σ on $X\backslash Y$. Let $\tilde Q_\nu^{\tilde\pi}$ be the occupancy measure in the

embedded model for the policy $\tilde{\pi}$ and for the initial (possibly, subprobability) measure ν on Y, where $\nu(Y') = P_\mu^\sigma\{\tau_Y < T, x_{\tau_Y} \in Y'\}$ for any measurable subset Y' of Y. We denote by $\tilde{\mathscr{O}}_\nu^{RS}(Y,\sigma)$ and $\tilde{\mathscr{O}}_\nu^{S}(Y,\sigma)$, respectively, the sets of all occupancy measures for stationary and deterministic policies in the embedded MDP.

It is easy to conclude from (3) that

$$Q_\mu^{\sigma[Y,\tilde{\pi}]}(Y' \times B) = \tilde{Q}_\nu^{\tilde{\pi}}(Y' \times B) \tag{15}$$

for any measurable subsets $Y' \subseteq Y$ and $B \subseteq A$. Thus, $Q_\mu^{\sigma[Y,\tilde{\pi}]}(C) = \tilde{Q}_\nu^{\tilde{\pi}}(C)$ for any measurable set $C \subseteq Y \times A$. This equality and (13) imply that l: $Q_\mu^{\sigma[Y,\tilde{\pi}]} \to \tilde{Q}_\nu^{\tilde{\pi}}$ is a one-to-one measurable mapping of the set $\mathscr{O}_\mu^{RS}(Y,\sigma)$ on the set $\tilde{\mathscr{O}}_\nu^{RS}(Y,\sigma)$. Being applied to the subset $\mathscr{O}_\mu^{S}(Y,\sigma)$ of $\mathscr{O}_\mu^{RS}(Y,\sigma)$, this correspondence is a one-to-one measurable mapping of $\mathscr{O}_\mu^{S}(Y,\sigma)$ onto $\tilde{\mathscr{O}}_\nu^{S}(Y,\sigma)$. In particular, for a deterministic policy $\tilde{\pi}$ in the embedded model, $l^{-1}(\tilde{Q}_\nu^{\tilde{\pi}}) = Q_\mu^{\sigma[Y,\tilde{\pi}]}$.

We observe that $\pi = \sigma[Y,\tilde{\pi}]$ for any stationary policy $\pi \in RS(Y,\sigma)$, where $\tilde{\pi}(\cdot \mid x) = \pi(\cdot \mid x)$ for any $x \in Y$. We rewrite (12) for the embedded MDP: for any measurable subset C of $Y \times A$,

$$\tilde{Q}_\nu^{\tilde{\pi}}(C) = \int_{\tilde{\mathscr{O}}_\nu^{S}(Y,\sigma)} Q(C)\,\tilde{\rho}\,(dQ), \tag{16}$$

where $\tilde{\rho}$ is a probability measure on the set $\tilde{\mathscr{O}}_\nu^{S}(Y,\sigma)$ of occupancy measures in the embedded models corresponding to deterministic policies. The following lemma shows that formulae (14) and (16) are equivalent. In particular, formula (14) holds if the corresponding embedded MDP satisfies the conditions of Theorem 4.1.

LEMMA 4.8. *Let Y be a measurable subset of $X\backslash\mathfrak{X}$ and let σ be a stationary policy. Formula (14) holds for any measurable set $C \subseteq X \times A$ if and only if formula (16) holds for any measurable set $C \subseteq Y \times A$. In addition, ρ^* is induced by ρ and the correspondence l; that is, $\rho^*(U) = \tilde{\rho}(l(U))$ for any measurable subset U of $\mathscr{O}_\mu^{S}(Y,\pi)$ or, equivalently, $\tilde{\rho}(\tilde{U}) = \rho^*(l^{-1}(\tilde{U}))$ for any measurable subset $\tilde{U}$ of $\tilde{\mathscr{Q}}_\nu^{S}(Y,\sigma)$.*

PROOF. Let $\pi \in RS(Y,\sigma)$. Then $\pi = \sigma[Y,\tilde{\pi}]$, where $\tilde{\pi}(\cdot \mid x) = \pi(\cdot \mid x)$ for all $x \in Y$. Let C be a measurable subset of $Y \times A$. Since $Q_\mu^{\sigma[Y,\tilde{\pi}]}(C) = \tilde{Q}_\nu^{\tilde{\pi}}(C)$, the change of variables in Lebesque integrals implies that (14) holds if and only if (16) holds. To complete the proof, we need to show that if (14) holds for any measurable $C \subseteq Y \times A$, then it holds for any measurable $C \subseteq X \times A$ or equivalently for any measurable $C \subseteq (X\backslash(Y \cup \mathfrak{X})) \times A$. The latter is equivalent to the validity of (14) for any $C = Z \times B$, where Z and B are measurable subsets of $X\backslash(Y \cup \mathfrak{X})$ and A, respectively. This is true because (13) implies

$$\begin{aligned} Q_\mu^\pi(Z,B) &= g_\mu^{Y,\sigma}(Z,B) + \int_Y \int_A \left[\int_{X\backslash(Y\cup\mathfrak{X})} g_x^{Y,\sigma}(Z,B) p\,(dx \mid y,a) \right] Q_\mu^\pi\,(dy,da) \\ &= \int_{\mathscr{O}_\mu^S(Y,\sigma)} \left\{ g_\mu^{Y,\sigma}(Z,B) + \int_Y \int_A \left(\int_{X\backslash(Y\cup\mathfrak{X})} g_x^{Y,\sigma}(Z,B) p\,(dx \mid y,a) \right) Q\,(dy,da) \right\} \rho^*(dQ) \\ &= \int_{\mathscr{O}_\mu^S(Y,\sigma)} Q\,(Z,B)\,\rho^*(dQ), \end{aligned}$$

where the first equality holds because $g_\mu^{Y,\pi}(Z,B)$ does not depend on decisions at states $x \in Y$ and therefore, $g_\mu^{Y,\pi}(Z,B) = g_\mu^{Y,\sigma}(Z,B)$, the second equality follows from the validity of (14) for all measurable $C \subseteq Y \times A$ and from the change of the order of integration, and the last equality follows from the first equality applied to the situation when all the decisions $\pi(\cdot \mid x)$ are deterministic when $x \in Y$. □

5. Splitting at a state. We recall that an MDP with a countable state space X is transient (Altman [1, p. 75]) if $E_\mu^\pi \sum_{n=0}^\infty \mathbf{I}\{x_n = x\} < \infty$ for each $x \in X\backslash\mathfrak{X}$. In particular, if X is countable, then an absorbing MDP is transient but not vice versa. A finite-state MDP is transient if and only if it is absorbing.

For a stationary policy σ, for a state $y \in X\backslash\mathfrak{X}$, and for an action $a \in A(y)$, we denote by $\sigma[y,a]$ the stationary policy that coincides with σ at any state $x \neq y$ and always selects the action a at y. The Ionescu Tulcea theorem (Hernández-Lerma and Lasserre [31, p. 178]) implies that the mapping $a \to P_\mu^{\sigma[y,a]}$ is measurable. Formula (3) implies that the mapping $P_\mu^\pi \to Q_\mu^\pi$ is measurable. Thus, the mapping $a \to Q_\mu^{\sigma[y,a]}$ is measurable; that is, $a \to Q_\mu^{\sigma[y,a]}(C)$ is a measurable function on $A(y)$ for any measurable subset C of $X \times A$.

Henceforth in this section, let state $y \in X\backslash\mathfrak{X}$ be fixed. According to Altman [1, p. 108], *a probability measure γ^* on $A(y)$ splits a stationary policy σ at the state y* if for any measurable subset C of $X \times A$,

$$Q_\mu^\sigma(C) = \int_{A(y)} Q_\mu^{\sigma[y,a]}(C)\gamma^*\,(da). \tag{17}$$

For transient countable state MDPs, Altman [1, p. 109] provided an explicit formula for γ^* satisfying (17); see (27) below. We observe that, since $a \to Q_\mu^{\sigma[y,a]}$ is a measurable mapping of $A(y)$ on $\mathcal{O}_\mu^S(\{y\}, \sigma)$, it is possible to change the measure and to write (17) in the form of (14) with $Y = \{y\}$. Thus, (17) can be viewed as a particular case of (14).

Let τ be the first epoch when the process hits y; i.e., $\tau = \min\{n \geq 0 \mid x_n = y\}$. In view of the notation used in the previous section, $\tau = \tau_Y$ with $Y = \{y\}$. We observe that $P_\mu^\pi\{\tau < \infty\} = P_\mu^\sigma\{\tau < \infty\}$ for any stationary policy π that coincides with σ outside of y. We also observe that, for any policy π, the following two properties hold: (i) $q_\mu^\pi(y) \geq P_\mu^\pi\{\tau < \infty\}$, and (ii) $q_\mu^\pi(y) > 0$ if and only if $P_\mu^\pi\{\tau < \infty\} > 0$. By setting $\pi = \sigma[y, a]$, $a \in A(y)$, we have: (a) $q_\mu^{\sigma[y,a]}(y) > 0$ if and only if $P_\mu^\sigma\{\tau < \infty\} > 0$, and (b)

$$q_\mu^{\sigma[y,a]}(y) \geq P_\mu^{\sigma[y,a]}\{\tau < \infty\} = P_\mu^\sigma\{\tau < \infty\}, \quad a \in A(y). \tag{18}$$

If $P_\mu^\sigma\{\tau < \infty\} = 0$, then $P_\mu^{\sigma[y,a]}\{\tau < \infty\} = P_\mu^\sigma\{\tau < \infty\} = 0$ for all $a \in A(y)$, because the policies σ and $\sigma[y, a]$) coincide at any time $t < \tau$. Thus, $P_\mu^\sigma\{\tau < \infty\} = 0$ implies $P_\mu^\sigma = P_\mu^{\sigma[y,a]}$, and therefore $Q_\mu^\sigma = Q_\mu^{\sigma[y,a]}$ for all $a \in A(y)$. So, when $P_\mu^\sigma\{\tau < \infty\} = 0$, we have that $Q_\mu^{\sigma[y,a]}(C) = Q_\mu^\sigma(C)$ in formula (17), and therefore (17) holds for any probability distribution γ^* on $A(y)$. So, we shall concentrate on the case $P_\mu^\sigma\{\tau < \infty\} > 0$.

If $P_\mu^\sigma\{\tau < \infty\} > 0$, formula (18) implies that

$$\int_{A(y)} \frac{\sigma(da \mid y)}{q_\mu^{\sigma[y,a]}(y)} \leq \int_{A(y)} \frac{\sigma(da \mid y)}{P_\mu^\sigma\{\tau < \infty\}} = \frac{1}{P_\mu^\sigma\{\tau < \infty\}} < \infty.$$

On the Borel set $A(y)$, consider the finite measure γ defined as

$$\gamma(B) := \int_B \frac{\sigma(da \mid y)}{q_\mu^{\sigma[y,a]}(y)} \tag{19}$$

for measurable subsets B of $A(y)$. We recall that two probability measures P and Q, defined on the same measurable space, are called equivalent if $P(B) > 0$ if and only if $Q(B) > 0$.

THEOREM 5.1. *Consider a stationary policy σ and a state y. Let $D := \{a \in A(y) \mid q_\mu^{\sigma[y,a]}(y) = \infty\}$.*

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

(ii) *If $q_\mu^\sigma(y) = \infty$, then $\sigma(D \mid y) = 1$ and a probability measure γ^* splits σ at y if and only if γ^* is equivalent to $\sigma(\cdot \mid y)$.*

(iii) *If $0 < q_\mu^\sigma(y) < \infty$ and $\sigma(D \mid y) = 0$, then γ^*, with*

$$\gamma^*(B) := \frac{\gamma(B)}{\gamma(A(y))} = \frac{\int_B \sigma(da \mid y)/q_\mu^{\sigma[y,a]}(y)}{\int_{A(y)} \sigma(da \mid y)/q_\mu^{\sigma[y,a]}(y)} \tag{20}$$

for any measurable subset B of $A(y)$, is the unique probability measure on $A(y)$ that splits σ at y.

(iv) *If $0 < q_\mu^\sigma(y) < \infty$ and $\sigma(D \mid y) > 0$, then σ cannot be split at y.*

PROOF. (i) Condition $q_\mu^\sigma(y) = 0$ is equivalent to $P_\mu^\sigma\{\tau < \infty\} = 0$. This implies that $P_\mu^{\sigma[y,a]} = P_\mu^\sigma$, and therefore $Q_\mu^{\sigma[y,a]} = Q_\mu^\sigma$ for all $a \in A(y)$.

For the remaining cases $q_\mu^\sigma(y) > 0$, which is equivalent to $P_\mu^\sigma\{\tau < \infty\} > 0$. Lemma 4.8 implies that a given probability measure γ^* splits σ at y if and only if for any measurable subset B of $A(y)$,

$$Q_\nu^\sigma(y, B) = \int_{A(y)} Q_\nu^{\sigma[y,a]}(y, B)\, \gamma^*(da), \tag{21}$$

where $\nu(Y) = \mathbf{I}\{y \in Y\} P_\mu^\sigma\{\tau < \infty\}$ for any measurable subset Y of X. Since $Q_\nu^\pi(y, B) = P_\mu^\sigma\{\tau < \infty\} Q_y^\pi(y, B)$ for $\pi = \sigma$ and for $\pi = \sigma[y, a]$, $a \in A(y)$, Equation (21) is equivalent to the same equation with the initial measure ν being replaced with the initial state y.

Since $Q_y^\sigma(y, B) = q_y^\sigma(y)\sigma(B \mid y)$ and $Q_y^{\sigma[y,a]}(y, B) = q_y^{\sigma[y,a]}(y)\mathbf{I}\{a \in B\}$, (21) is equivalent to

$$q_y^\sigma(y)\sigma(B \mid y) = \int_B q_y^{\sigma[y,a]}(y)\, \gamma^*(da). \tag{22}$$

The rest of the proof is based on the validity of (22) for cases (ii)–(iv). Before we consider these cases, we derive some useful formulae.

Let $\tau^1 := \tau_y^1 = \min\{n \geq 1 \mid x_n = y\}$ be the first time, except time 0, when the process x_n hits y. Then $P_y^\pi\{\tau^1 < \infty\}$ is the probability to return to state y. For a stationary policy π, the total number of visits to y, starting from y, has a geometric distribution with the parameter $P_y^\pi\{\tau^1 = \infty\}$. Since the expectation of this number is $q_y^\pi(y)$, we have $q_y^\pi(y) = 1/(1 - P_y^\pi\{\tau^1 < \infty\})$. From the total probability formula,

$$P_y^\sigma\{\tau^1 < \infty\} = \int_{A(y)} P_y^{\sigma[y,a]}\{\tau^1 < \infty\}\sigma(da \mid y). \tag{23}$$

Thus,

$$q_y^\sigma(y) = \frac{1}{1 - P_y^\sigma\{\tau^1 < \infty\}} = \frac{1}{\int_{A(y)}(1 - P_y^{\sigma[y,a]}\{\tau^1 < \infty\})\sigma(da \mid y)} = \frac{1}{\int_{A(y)} \sigma(da \mid y)/q_y^{\sigma[y,a]}(y)}. \tag{24}$$

We observe that for any stationary policy π,

$$q_\mu^\pi(y) = P_\mu^\pi\{\tau < \infty\}q_y^\pi(y). \tag{25}$$

Formulae (24), (25), and $P_\mu^\sigma\{\tau < \infty\} = P_\mu^{\sigma[y,a]}\{\tau < \infty\}$ imply

$$q_y^\sigma(y) = \frac{1}{P_\mu^\sigma\{\tau < \infty\}\int_{A(y)} \sigma(da \mid y)/q_\mu^{\sigma[y,a]}(y)} = \frac{1}{P_\mu^\sigma\{\tau < \infty\}\gamma(A(y))}. \tag{26}$$

(ii) Equality (24) implies $\sigma(D \mid y) = 1$, because otherwise $q_y^\sigma(y) < \infty$, and, in view of (25), $q_\mu^\sigma(y) < \infty$. For $B = A(y)\backslash D$, the left-hand side of (22) equals 0, because $\sigma(D \mid y) = 1$ (recall that we follow the standard convention that $0 \times \infty = 0$). Since $q_y^\pi(y) \geq 1 > 0$ for any policy π, equality (22) holds for this particular B if and only if $\gamma^*(D) = 1$. This implies that (22) holds for any measurable subset B of $A(y)$ if and only if γ^* and $\sigma(\cdot \mid y)$ are equivalent.

(iii) Let γ^* be defined by (20). Equations (19), (26), and $P_\mu^\sigma\{\tau < \infty\} = P_\mu^{\sigma[y,a]}\{\tau < \infty\}$ imply that for any measurable subset B of $A(y)$,

$$\int_B q_y^{\sigma[y,a]}(y)\gamma^*(da) = \frac{\int_B (q_y^{\sigma[y,a]}(y)/q_\mu^{\sigma[y,a]}(y))\sigma(da \mid y)}{P_\mu^\sigma\{\tau < \infty\}\gamma(A(y))} = \frac{\int_B (q_\mu^{\sigma[y,a]}(y)/q_\mu^{\sigma[y,a]}(y))\sigma(da \mid y)}{\gamma(A(y))} = \frac{\sigma(B \mid y)}{\gamma(A(y))},$$

where $q_\mu^{\sigma[y,a]}(y)$ cancel each other in the middle formula because $\sigma(D \mid y) = 0$. Thus, (22) is proved. Therefore, (21) holds and γ^* splits σ at y. If some measure γ^* splits σ at y, then the inequality $q_y^\pi(y) \geq 1 > 0$, where π is an arbitrary policy, and equality (22) imply that the measure γ^* is absolutely continuous with respect to the measure $\sigma(\cdot \mid y)$. The Radon-Nikodym theorem implies that there exists a $\sigma(\cdot \mid y)$-measurable function f on $A(y)$ with $\gamma^*(B) = \int_B f(a)\sigma(da \mid y)$ and f is $\sigma(\cdot \mid y)$-a.s. unique. Formula (22) implies that f satisfies ($\sigma(\cdot \mid y)$-a.s.) the equality $f(a) = q_y^\sigma(y)/q_y^{\sigma[y,a]}(y)$. This and $q_y^\sigma(y) = \gamma(A(y))$, that in view of $P_y^\sigma\{\tau < \infty\} = 1$ follows from (26), imply that γ^* satisfies (19). Thus, γ^* is unique.

(iv) Suppose σ is split at y. Formula (25) implies $q_y^\sigma(y) < \infty$. Recall that $q_y^\sigma(y) \geq 1$. Thus, $0 < q_y^\sigma(y)\sigma(D \mid y) < \infty$. Since $q_y^{\sigma[y,a]}(y) = \infty$ when $a \in D$, the right-hand side of (22) with $B = D$ is either 0 or ∞. This contradiction completes the proof. □

The following example illustrates statement (iv) of Theorem 5.1.

Example 5.1. Let $X = \{1, 2\}$ and $\mathcal{X} = \{1\}$; that is, state 1 is absorbing. Let $A(2) = \{a, b\}$ and $p(1 \mid 2, a) = p(2 \mid 2, b) = 1$. In this example there are only two deterministic policies: ϕ^a with $\phi^a(2) = a$ and ϕ^b with $\phi^b(2) = b$. From state 2 the policy ϕ^a moves the process to the absorbing state, while policy ϕ^b moves the process back to state 2. Thus, $Q_2^{\phi^a}(2) = 1$ and $Q_2^{\phi^b}(2) = \infty$. Consider the stationary policy σ that selects actions a and b at state 2 with probabilities 0.5. This policy defines a Markov chain with the recurrent state 1 and transient state 2. In particular, $Q_2^\sigma(2) = 2$. Since 2 is not a convex combination of 1 and ∞, σ cannot be split.

Remark 5.1. As mentioned above, Altman [1, p. 109] established splitting of a stationary policy at a state for transient countable state MDPs. The splitting measure is presented in Altman [1] in the form

$$\gamma^*(B) = \frac{\int_B (1 - P_y^{\sigma[y,a]}\{\tau^1 < T\})\sigma(da \mid y)}{1 - P_y^\sigma\{\tau^1 < T\}}. \tag{27}$$

Formulas (24), (25), and $q_y^{\sigma[y,a]}(y) = 1/(1 - P_y^{\sigma[y,a]}\{\tau^1 < \infty\})$ imply that representations (20) and (27) are equivalent. However, the advantage of (20) is that splitting of Q_μ^σ into $Q_\mu^{\sigma[y,a]}$, $a \in A(y)$, is represented in (20)

only via σ and $Q_\mu^{\sigma[y,a]}(y,A)=q_\mu^{\sigma[y,a]}(y)$, $a\in A(y)$. Statements (i) and (iii) of Theorem 5.1 cover transient countable state MDPs. We also observe that, if $q_\mu^\sigma(y)>0$, then $\sigma(da\mid y)=Q_\mu^\sigma(y,da)/q_\mu^\sigma(y)$, and (20) can be rewritten as

$$\gamma^*(B)=\frac{\int_B Q_\mu^\sigma(y,da)/q_\mu^{\sigma[y,a]}(y)}{\int_{A(y)} Q_\mu^\sigma(y,da)/q_\mu^{\sigma[y,a]}(y)}. \tag{28}$$

If $A(y)$ is countable, then (17) becomes $Q_\mu^\sigma=\sum_{a\in A(y)}\gamma^*(a)Q_\mu^{\sigma[y,a]}$, where

$$\gamma^*(a)=\frac{\sigma(a\mid y)/q_\mu^{\sigma[y,a]}(y)}{\sum_{a\in A(y)}\sigma(a\mid y)/q_\mu^{\sigma[y,a]}(y)},\quad a\in A(y). \tag{29}$$

We notice that (22) is equivalent to

$$\gamma^*(da)=\frac{q_\mu^\sigma(y)}{q_\mu^{\sigma[y,a]}(y)}\sigma(da\mid y)=\frac{Q_\mu^\sigma(y,da)}{q_\mu^{\sigma[y,a]}(y)}, \tag{30}$$

where the second equality follows from $Q_\mu^\sigma(y,da)=q_\mu^\sigma(y)\sigma(da\mid y)$. For a countable set $A(y)$, (30) becomes

$$\gamma^*(a)=\frac{q_\mu^\sigma(y)}{q_\mu^{\sigma[y,a]}(y)}\sigma(a\mid y)=\frac{Q_\mu^\sigma(y,a)}{q_\mu^{\sigma[y,a]}(y)},\quad a\in A(y). \tag{31}$$

Formula (30) is simpler than (28).

Remark 5.2. As follows from the proof of Theorem 5.1, the initial measure μ in the definition (20) of γ^* and in (30) can be replaced with any probability measure ν such that $P_\nu^\sigma\{\tau<\infty\}>0$ or equivalently, $q_\nu^\sigma(y)>0$. In particular, after this replacement, the statements of Theorem 5.1 and Corollary 5.1 remain valid.

For an absorbing MDP, $q_\mu^\pi(X)<\infty$ for any policy π. Therefore, if the MDP is absorbing, then the situations described in Theorem 5.1(ii) and (iv) are impossible. So, we have the following statement.

Corollary 5.1. *For an absorbing MDP, consider a stationary policy σ and a state y. If $q_\mu^\sigma(y)>0$, then the probability measure γ^*, defined by (20), is the unique probability measure on A that splits σ at y. If $q_\mu^\sigma(y)=0$, then any probability measure on $A(y)$ splits σ at y.*

Consider a particular case when the policy σ is deterministic at all states except x. Then Theorem 5.1 implies the following statement that assumes neither Condition (W) nor Condition (S).

Corollary 5.2. *Consider an absorbing MDP such that there is a state y such that the sets $A(x)$ are singletons when $x\neq y$. Then for any policy π, there exists a probability measure ρ on $\mathscr{O}_\mu^S$ such that (12) holds.*

Proof. In view of Lemma 4.2, for any policy π there exists $\sigma\in RS$ such that $Q_\mu^\sigma=Q_\mu^\pi$. If $q_\mu^\sigma(y)=0$, then (12) holds for any probability measure ρ because $Q_\mu^\gamma=Q_\mu^\sigma$ for any policy γ. If $q_\mu^\sigma(y)\neq 0$, then there is a one-to-one correspondence between probability distributions γ on $A(x)$ and ρ on $\mathscr{O}_\mu^S$. This correspondence is defined by $\rho(\{Q_\mu^\phi\mid\phi(y)\in B,\phi\in S\})=\gamma(B)$ for any measurable subset B of $A(y)$. Let ρ^* be the probability measure on $\mathscr{O}_\mu^S$ corresponding to γ^* defined in (20). Then $Q_\mu^\sigma=\int_{A(y)}Q_\mu^{\sigma[y,a]}(C)\gamma^*(da)=\int_{\mathscr{O}_\mu^S}Q(C)\rho^*(dQ)$ for all measurable subsets C of $X\times A$, where the first equality follows from Theorem 5.1(iii) and the second inequality follows from the definition of ρ^*. □

6. Finite splitting at multiple states when state and action sets are finite. We recall that finite state and action MDPs are absorbing if and only if they are transient. For such MDPs, transience and absorbing are routinely used in the literature on unconstrained MDPs when the conditions hold, respectively, for all initial distributions (and not just for a specified one) or, equivalently, for an initial distribution that assigns positive probabilities to all states. These definitions are more restrictive than the definitions used in this paper. Of course, the aforementioned equivalence also holds for the restrictive definitions; also, the restrictive definitions are known to be equivalent to the assumption that absolute values of all eigenvalues of the transition matrices associated with the deterministic policies are less than 1.

In this section we focus on absorbing MDPs with finite state and action sets and show how the occupancy measure for an arbitrary stationary policy can be presented as a convex combination of occupancy measures for deterministic policies. In other words, we show how to split a stationary policy into deterministic policies. Our results are constructive and yield accompanying computational methods.

When a finite state and action MDP is absorbing, the occupancy measure of each policy σ is represented by $Q_\mu^\sigma = \{Q_\mu^\sigma(x,a)\colon x \in X, a \in A(x)\}$; the coordinates $Q_\mu^\sigma(\cdot,\cdot)$ are then referred to as *occupancy values*. Also, $q_\mu^\sigma(x) := \sum_{a\in A(x)} Q_\mu^\sigma(x,a)$ for each $x \in X$. In particular, $Q_\mu^\sigma(x,a) = q_\mu^\sigma(x)\sigma(a \mid x)$ for each $a \in A(x)$, when σ is a stationary policy.

Properties of occupancy measures used in this section are next recorded. First, the occupancy values of all policies are finite. Second, $Q = \{Q(x,a)\colon x \in X, a \in A(x)\}$ represents the occupancy values of a policy if and only if $Q(x,a) = 0$ when $x \in \mathscr{X}$ and, with $X^* := X\backslash\mathscr{X}$,

$$\sum_{a\in A^\sigma(x)} Q(x,a) - \sum_{y\in X^*}\sum_{a\in A(y)} p(x \mid y,a)Q(y,a) = \mu(x), \quad x \in X^*, \tag{32}$$

$$Q(x,a) \ge 0, \quad x \in X^*,\ a \in A(x); \tag{33}$$

see Altman [1, Lemma 7.1] or Lemma 4.3 above. Third, a feasible solution Q of (32)–(33), represents the occupancy values of any stationary policy π with

$$\pi(a \mid x) = \begin{cases} \dfrac{Q(x,a)}{\sum_{b\in A(x)} Q(x,b)} & \text{if } \displaystyle\sum_{b\in A(x)} Q(x,b) > 0, \\ \text{arbitrary} & \text{otherwise.} \end{cases} \tag{34}$$

Here, arbitrary means any selection that satisfies $\sum_{a\in A(x)} \pi(a \mid x) = 1$; see Altman [1, Theorem 8.1(ii)] or Lemma 4.2 above. Finally, the solution set of (32)–(33) is bounded and the extreme points of this polytope are represented by the occupancy values of deterministic policies; see Altman [1, Lemma 8.3(ii)] and Kallenberg [35, Theorem 3.3.3].

The following definitions refer to finite state and action MDPs but will later be extended to general MDPs. For a stationary policy σ and a state $x \in X$, let $A^\sigma(x) := \{a \in A(x) \mid \sigma(a \mid x) > 0\}$. We say that a deterministic policy ϕ is a *restriction of* a stationary policy σ if $\phi(x) \in A^\sigma(x)$ for all $x \in X$.

For a finite set E, we denote by $|E|$ the number of its elements. For $m = 0, 1, \ldots,$ a stationary policy π is called *m-randomized stationary* if

$$\sum_{x\in X}(|A^\pi(x)| - 1) \le m; \tag{35}$$

and it is called *exactly m-randomized stationary* if the inequality in (35) holds as an equality. The notions of deterministic, 0-randomized stationary, and exactly 0-randomized stationary policies coincide.

We shall use the standard convention that $\{\phi^i, \ldots, \phi^j\} = \varnothing$ for $j < i$.

THEOREM 6.1. *Consider an absorbing MDP with finite state and action sets. Let σ be an exactly m-randomized stationary policy with $m \ge 0$ and let ϕ^1 be a restriction of σ. Then there exist restrictions of σ, $\phi^2, \ldots, \phi^{m+1}$, and nonnegative numbers $\alpha_1, \ldots, \alpha_{m+1}$ such that $\phi^1, \ldots, \phi^{m+1}$ are distinct, $\sum_{j=1}^{m+1} \alpha_j = 1$,*

$$Q_\mu^\sigma = \sum_{j=1}^{m+1} \alpha_j Q_\mu^{\phi^j} \tag{36}$$

and for $i = 1, \ldots, m$ there is exactly one state $x^i \in X$ with $\phi^i(x^i) \ne \phi^{i+1}(x^i)$.

The final property of the deterministic policies $\phi^1, \ldots, \phi^{m+1}$ in Theorem 6.1 implies that they are "similar" in the sense that they can be distinct in at most m states. Without this requirement, splitting formula (36) can be viewed as an instance of (14).

The proof of Theorem 6.1 will be preceded by three lemmas. The first lemma records elementary properties of occupancy values.

LEMMA 6.1. *Consider an absorbing MDP with finite state and action sets. Let σ and π be two stationary policies.*

(a) *If $A^\sigma(x) \subseteq A^\pi(x)$ for all $x \in X$ with $q_\mu^\sigma(x) > 0$, then $Q_\mu^\sigma(z,a) > 0$ implies $Q_\mu^\pi(z,a) > 0$ for $z \in X$ and $a \in A(z)$.*

(b) *If $|A^\sigma(x)| = 1$ and $A^\pi(x) = A^\sigma(x)$ for every state x with $q_\mu^\sigma(x) > 0$, then $Q_\mu^\sigma = Q_\mu^\pi$.*

(c) *If ϕ is a restriction of σ, then $q_\mu^\phi(x) > 0$ implies $q_\mu^\sigma(x) > 0$ each $x \in X$.*

PROOF. Recall that $Q^{\gamma}_{\mu}(z,a)=0$ for $z\in\mathfrak{X}$ and any policy γ.

(a) Assume that $A^{\sigma}(x)\subseteq A^{\pi}(x)$ for all $x\in X$ with $q^{\sigma}_{\mu}(x)>0$, and consider $z\in X\backslash\mathfrak{X}$ with $Q^{\sigma}_{\mu}(z,a)>0$. Then under the Markov chain defined by σ on X, z is accessible from some $y\in X\backslash\mathfrak{X}$ with $\mu(y)>0$, i.e., there is a finite path of states from y to z that has positive probability. Every state u on such a path, including $u=z$, has $q^{\sigma}_{\mu}(u)>0$ and $q^{\pi}_{\mu}(u)>0$ (as $A^{\sigma}(x)\subseteq A^{\pi}(x)$ for x with $q^{\sigma}_{\mu}(x)>0$). It now follows that $Q^{\pi}_{\mu}(z,a)=q^{\pi}_{\mu}(z)\pi(a\,|\,z)>0$ when $a\in A^{\pi}(z)$.

(b) If $q^{\sigma}_{\mu}(u)=0$ for some $u\in X$, then under the Markov chain defined by σ on X, either $u\in\mathfrak{X}$ or u is not accessible from every state $y\in X\backslash Z$ with $\mu(y)>0$. In either case, if σ is changed arbitrarily for states u with $q^{\sigma}_{\mu}(u)=0$, the occupancy values do not change. Thus, $Q^{\sigma}_{\mu}=Q^{\pi}_{\mu}$ for every policy π that differs from σ only at states u with $q^{\sigma}_{\mu}(u)=0$.

(c) Follows from (a). □

We shall use the standard convention that $a/0=\infty$ for $a>0$.

LEMMA 6.2. *Consider an absorbing MDP with finite state and action sets. Let σ be a stationary policy for which $|A^{\sigma}(x)|>1$ for at least one state x with $q^{\sigma}_{\mu}(x)>0$ and let ϕ be a restriction of σ. Then*

$$\alpha:=\min_{x\in X,\,q^{\sigma}_{\mu}(x)>0}\frac{Q^{\sigma}_{\mu}(x,\phi(x))}{q^{\phi}_{\mu}(x)} \tag{37}$$

is well defined (that is, no ratio $0/0$ appears in (37)) and $0<\alpha<1$.

PROOF. As Q^{ϕ}_{μ} solves (32), it is nonzero and therefore $q^{\phi}_{\mu}(x)>0$ for some x; in view of Lemma 6.1(c), such x satisfies $q^{\sigma}_{\mu}(x)>0$. Also, as $Q^{\sigma}_{\mu}(x,\phi(x))=q^{\sigma}_{\mu}(x)\sigma(\phi(x)\,|\,x)>0$ for each x, the numerator of each ratio in (37) is positive. These conclusions assure that α is well defined, positive and finite. Next assume that $\alpha\geq 1$ and we will establish a contradiction. As $Q^{\phi}_{\mu}(x,a)=0$ when either $q^{\phi}_{\mu}(x)=0$ or $a\neq\phi(x)$, the assumptions $\alpha\geq 1$ and Lemma 6.1(c) imply that

$$Q^{\sigma}_{\mu}(x,a)-Q^{\phi}_{\mu}(x,a)\begin{cases}\geq Q^{\sigma}_{\mu}(x,a)-\alpha Q^{\phi}_{\mu}(x,a)\geq 0 & \text{if } q^{\phi}_{\mu}(x)>0 \text{ and } a=\phi(x),\\ =Q^{\sigma}_{\mu}(x,a)\geq 0 & \text{otherwise.}\end{cases}$$

Next, for the state x with $|A^{\sigma}(x)|>1$ and $q^{\sigma}_{\mu}(x)>0$, there is an action $b\in A^{\sigma}(x)\backslash\{\phi(x)\}$; this action satisfies $Q^{\sigma}_{\mu}(x,b)=q^{\sigma}_{\mu}(x)\sigma(b\,|\,x)>0$, $Q^{\phi}_{\mu}(x,b)=q^{\phi}_{\mu}(x)\phi(b\,|\,x)=0$, and hence, $Q^{\sigma}_{\mu}(x,b)-Q^{\phi}_{\mu}(x,b)>0$. Define $Q'(x,a)=Q^{\sigma}_{\mu}(x,a)-Q^{\phi}_{\mu}(x,a)$ for $x\in X$ and $a\in A^{\sigma}(x)$. Then all elements of Q' are nonnegative and at least one of them is positive. Furthermore, as Q^{σ}_{μ} and Q^{ϕ}_{μ} satisfy (32), Q' satisfies the corresponding homogenous equation (where μ is replaced by 0). It follows that Q' is a nonzero direction of recession of the polyhedron defined by (32)–(33), contradicting its boundedness. This contradiction proves that $\alpha<1$. □

LEMMA 6.3. *Consider an absorbing MDP with finite state and action sets. Let σ be a stationary policy for which $|A^{\sigma}(x)|>1$ for at least one state x with $q^{\sigma}_{\mu}(x)>0$, let ϕ be a restriction of σ, and let α be defined by (37). Then,*

(a) *the following defines a stationary policy*:

$$\pi(a\,|\,x)=\begin{cases}\dfrac{Q^{\sigma}_{\mu}(x,a)-\alpha Q^{\phi}_{\mu}(x,a)}{q^{\sigma}_{\mu}(x)-\alpha q^{\phi}_{\mu}(x)} & \text{if } q^{\sigma}_{\mu}(x)>0 \text{ and } |A^{\sigma}(x)|>1,\\ 1 & \text{if } a=\phi(x) \text{ and either } q^{\sigma}_{\mu}(x)=0 \text{ or } |A^{\sigma}(x)|=1,\\ 0 & \text{otherwise,}\end{cases} \tag{38}$$

which satisfies

$$Q^{\pi}_{\mu}=\frac{Q^{\sigma}_{\mu}-\alpha Q^{\phi}_{\mu}}{1-\alpha}; \tag{39}$$

(b) *there exists a state x with $|A^{\sigma}(x)|>1$, $q^{\sigma}_{\mu}(x)>0$, and $\alpha=Q^{\sigma}_{\mu}(x,\phi(x))/q^{\phi}_{\mu}(x)$; and*

(c) *if σ is m-randomized for $m>0$, then π that satisfies (38) is $(m-1)$-randomized.*

PROOF. (a) Let $Q':=(Q^{\sigma}_{\mu}-\alpha Q^{\phi}_{\mu})/(1-\alpha)$. The definition of α in (37) and Lemmas 6.2 and 6.1(c) assure that Q' is well defined and its elements are nonnegative. Furthermore, as Q^{σ}_{μ} and Q^{ϕ}_{μ} satisfy (32), so does Q'. Thus, Q' satisfies (32)–(33), implying that it represents the occupancy values of any stationary policy π defined by (34) with Q' replacing Q. We next show that the right-hand side of (38) is an instance of (34).

Define $q'(x)=\sum_{a\in A(x)}Q'(x,a)$, $x\in X$. Then $q'(x)=\sum_{a\in A(x)}(Q^{\sigma}_{\mu}(x,a)-\alpha Q^{\phi}_{\mu}(x,a))/(1-\alpha)=(q^{\sigma}_{\mu}(x)-\alpha q^{\phi}_{\mu}(x))/(1-\alpha)$ for each $x\in X$; in particular, $q'(x)\geq 0$. If $q'(x)=0$, then trivially (38) is an instance of (34). Consider the alternative case where $q'(x)>0$. Then $0<q'(x)=(q^{\sigma}_{\mu}(x)-\alpha q^{\phi}_{\mu}(x))/(1-\alpha)\leq q^{\sigma}_{\mu}(x)/(1-\alpha)$, assuring $q^{\sigma}_{\mu}(x)>0$. We consider two subcases: (i) and (ii).

(i) $|A^\sigma(x)| > 1$. Then (38) gives $\pi(a \mid x) = Q'(x,a)/q'(x)$ which is the unique choice under (34).

(ii) $|A^\sigma(x)| = 1$. Observe that $[Q'(x,a) > 0] \Rightarrow [Q^\sigma_\mu(x,a) > 0] \Rightarrow [a \in A^\sigma(x)] \Rightarrow [a = \phi(x)]$. Thus, $\pi(a \mid x)$ in (34) equals 1 for $a = \phi(x)$ and it equals 0 for all other a's. Hence $\pi(a \mid x)$ in (34) and $\pi(a \mid x)$ in (38) are equal.

(b) The proof of (b) is by contradiction. Assume that the conclusion of part (b) is false; that is: (1) $Q^\sigma_\mu(x,\phi(x)) > \alpha Q^\phi_\mu(x,\phi(x))$ for all x such that $q^\sigma_\mu(x) > 0$ and $|A^\sigma(x)| > 1$, and (2) $Q^\sigma_\mu(x,\phi(x)) = \alpha Q^\phi_\mu(x,\phi(x))$ for some x with $|A^\sigma(x)| = 1$ and $q^\sigma_\mu(x) > 0$. Of course, $|A^\sigma(x)| = 1$ means that $A^\sigma(x) = \{\phi(x)\}$. Let π be a stationary policy defined in (38). In view of (39), (1) implies that $\pi(a \mid x) q^\pi_\mu(x) = Q^\pi_\mu(x,a) > 0$ whenever $q^\sigma_\mu(x) > 0$, $|A^\sigma(x)| > 1$, and $a \in A^\sigma(x)$. Consequently,

$$[|A^\sigma(x)| > 1,\ q^\sigma_\mu(x) > 0 \text{ and } a \in A^\sigma(x)] \ \Rightarrow\ [a \in A^\pi(x)].$$

Observing that if $|A^\sigma(x)| = 1$, then $A^\sigma(x) = \{\phi(x)\} = A^\pi(x)$ (the second equality following from (38)), we conclude that

$$[q^\sigma_\mu(x) > 0 \text{ and } a \in A^\sigma(x)] \ \Rightarrow\ [a \in A^\pi(x)].$$

By Lemma 6.1(a), this implication assures that $Q^\pi_\mu(x,a) > 0$ whenever $Q^\sigma_\mu(x,a) > 0$. But, in view of (39), (2) implies that $Q^\pi_\mu(x,\phi(x)) = 0$ for some x with $A^\sigma(x) = \{\phi(x)\}$ and $Q^\sigma_\mu(x,\phi(x)) = q^\sigma_\mu(x) > 0$, a contradiction which completes the proof of (b).

(c) The definition of π in (38) implies that $A^\pi(x) \subseteq A^\sigma(x)$ for each $x \in X$ and (38) together with part (b) show that the inclusion is strict for some $x \in X$. Consequently, $\sum_{x\in X} |A^\pi(x)| < \sum_{x\in X} |A^\sigma(x)|$ which implies that if σ is m-randomized, then π is $(m-1)$-randomized. □

Proof of Theorem 6.1 The proof is based on the induction in m. For $m = 0$, the statement of the theorem is trivial. Suppose that for some $n = 1, 2, \ldots$, the conclusion of the theorem holds for $m = 0, 1, \ldots, n-1$. We shall prove that it also holds for $m = n$. So, let σ be an exactly n-randomized policy and let ϕ^1 be a restriction of σ.

We first consider the case where some state z has $|A^\sigma(z)| > 1$ and $q^\sigma_\mu(z) = 0$. For such z, let $a^1 = \phi^1(z)$ and choose any $a^2 \in A^\sigma(z) \backslash \{a^1\}$. Define the stationary policy π by $\pi(a^2 \mid z) = \sigma(a^1 \mid z) + \sigma(a^2 \mid z)$, $\pi(a^1 \mid z) = 0$ and $\pi(a \mid x) = \sigma(a \mid x)$ when $x \neq z$ or $x = z$ and $a \notin \{a^1, a^2\}$. Also, define the stationary policy ϕ^2 by $\phi^2(z) = a^2$ and $\phi^2(x) = \phi^1(x)$ for $x \neq z$. Then $A^\pi(z) = A^\sigma(z) \backslash \{a^1\}$, $A^\pi(x) = A^\sigma(x)$ for every $x \neq z$, π is exactly $(m-1)$-randomized, ϕ^2 is a restriction of π, and ϕ^2 differs from ϕ^1 in exactly one state. Next, by the induction assumption, there exist restrictions of π, $\phi^3, \ldots, \phi^{n+1}$, and nonnegative numbers $\alpha_2, \ldots, \alpha_{n+1}$, such that $\phi^2, \ldots, \phi^{n+1}$ are distinct, $\sum_{j=2}^{n+1} \alpha_j = 1$,

$$Q^\pi_\mu = \sum_{j=2}^{n+1} \alpha_j Q^{\phi^j}_\mu,$$

and for $i = 2, \ldots, n$, there is exactly one state where ϕ^i and ϕ^{i+1} differ. Now, as $A^\pi(x) \subseteq A^\sigma(x)$ for each x, we have that $\phi^2, \ldots, \phi^{n+1}$ are also restrictions of σ. As $\phi^1(z) = a^1$ and $a^1 \notin A^\pi(z) \supseteq A^{\phi^i}(z)$ for $i = 2, \ldots, n+1$, we have that $\phi^1 \notin \{\phi^2, \ldots, \phi^{n+1}\}$, implying that $\phi^1, \ldots, \phi^{n+1}$ are distinct. Finally, with $\alpha_1 = 0$, we have that $\sum_{j=1}^{n+1} \alpha_j = \sum_{j=2}^{n+1} \alpha_j = 1$, and using Lemma 6.1(b), $Q^\sigma_\mu = Q^\pi_\mu = \sum_{j=2}^{n+1} \alpha_j Q^{\phi^j}_\mu = \sum_{j=1}^{n+1} \alpha_j Q^{\phi^j}_\mu$. Thus, the induction hypothesis has been established for $m = n$.

We next consider the alternative case where no state z has $|A^\sigma(z)| > 1$ and $q^\sigma_\mu(z) = 0$. As $m > 0$ assures that $|A^\sigma(z)| > 1$ for some state z, it follows that such a state has $q^\sigma_\mu(z) > 0$; consequently, σ and $\phi = \phi^1$ satisfy the assumptions of Lemmas 6.2–6.3. Let α and π be defined, respectively, by the conclusions of Lemmas 6.2 and 6.3(a) applied to σ and $\phi = \phi^1$. Let $G = \{x \in X \mid Q^\sigma_\mu(x,\phi(x)) = \alpha q^\phi_\mu(x),\ q^\sigma_\mu(x) > 0, |A^\sigma(x)| > 1\}$. Lemma 6.3(b) assures that $k := |G| \geq 1$. Also, (38) together with the assumption that no state z has $|A^\sigma(z)| > 1$ and $q^\sigma_\mu(z) = 0$ assure that $A^\pi(x) = A^\sigma(x) \backslash \{\phi(x)\}$ for $x \in G$ and $A^\pi(x) = A^\sigma(x)$ for $x \in X \backslash G$. In particular, π is exactly $(n-k)$-randomized.

Enumerate the elements of G, say, $G = \{x^1, \ldots, x^k\}$ and define deterministic policies $\phi^2, \ldots, \phi^{k+1}$ sequentially for $i = 2, \ldots, k+1$ by

$$\phi^i(x) := \begin{cases} \phi^{i-1}(x) & \text{if } x \neq x^{i-1}, \\ \text{any element of } A^\sigma(x) \backslash \{\phi^{i-1}(x)\} & \text{if } x = x^{i-1}. \end{cases} \tag{40}$$

For $1 \leq i < j \leq k+1$, $\phi^i(x^i) = \phi(x^i) \neq \phi^{i+1}(x^i) = \phi^j(x^i)$. This implies $\phi^1, \ldots, \phi^{k+1}$ are distinct. Also, the construction assures that $\phi^2, \ldots, \phi^{k+1}$ are all restrictions of σ and for $i = 1, \ldots, k$, ϕ^i and ϕ^{i+1} differ only in

state x^i. Finally, the explicit expressions for the $A^\pi(x)$'s assure that ϕ^{k+1} is a restriction of π. Next, by the induction assumption, there exist deterministic policies $\phi^{k+2}, \ldots, \phi^{n+1}$ that are restrictions of π, and nonnegative numbers $\beta_2, \ldots, \beta_{n+1}$, such that $\phi^{k+1}, \ldots, \phi^{n+1}$ are distinct, $\sum_{j=k+1}^{n+1} \beta_j = 1$,

$$Q_\mu^\pi = \sum_{j=k+1}^{n+1} \beta_j Q_\mu^{\phi^j},$$

and for $i = 1, \ldots, n$, there is exactly one state where ϕ^i and ϕ^{i+1} differ. As $A^\pi(s) \subseteq A^\sigma(x)$ for each x, we have that $\phi^{k+1}, \ldots, \phi^{n+1}$ are restrictions of σ. Also, for $1 \le i \le k < j \le n+1$, $\phi^i(x^i) = \phi(x^i) \notin A^\pi(x) \supseteq A^{\phi^j}(x)$. This implies $\{\phi^1, \ldots, \phi^k\} \cap \{\phi^{k+1}, \ldots, \phi^{n+1}\} = \varnothing$, and therefore $\phi^1, \ldots, \phi^{n+1}$ are distinct. Next, for $i = 1, \ldots, n$, there is exactly one state where ϕ^i and ϕ^{i+1} differ. Finally, set $\alpha_1 = \alpha$, $\alpha_i = 0$ for $i = 2, \ldots, k$, and $\alpha_i = (1-\alpha)\beta_i$ for $i = k+1, \ldots, n+1$. We then have that $\sum_{j=1}^{n+1} \alpha_j = \alpha + \sum_{j=k+1}^{n+1} \alpha_j = \alpha + \sum_{j=k+1}^{n+1} (1-\alpha)\beta_j = \alpha + (1-\alpha) = 1$, and (39) implies that $Q_\mu^\pi = (Q_\mu^\sigma - \alpha Q_\mu^\phi)/(1-\alpha)$, and therefore

$$Q_\mu^\sigma = \alpha Q_\mu^\phi + (1-\alpha) Q_\mu^\pi = \sum_{j=1}^{n+1} \alpha_j Q_\mu^{\phi^j};$$

thus, the induction hypothesis has been established for $m = n$. □

Consider the inductive step of Theorem 6.1 when $\{z \in X\colon |A^\sigma(z)| > 1 \text{ and } q_\mu^\sigma(z) = 0\} = \varnothing$. With π as the constructed policy, if $|A^\pi(z)| > 1$ for $z \in X$, then $q_\mu^\pi(z) > 0$; consequently, $\{z \in X\colon |A^\pi(z)| > 1 \text{ and } q_\mu^\pi(z) = 0\} = \varnothing$.

The following example demonstrates that it may be impossible to require for all α_i's to be positive in (36) (and in (42)). This example also demonstrates that it may be impossible to select the deterministic policies $\phi^1, \ldots, \phi^{m+1}$ in Theorem 6.1 in such a way that, in addition to ϕ^i and ϕ^{i+1}, $i = 1, \ldots, m$, the mappings ϕ^{m+1} and ϕ^1 also differ only at one state.

Example 6.1. Let $X = \{1, 2\}$, $A = \{a^1, a^2\}$, $A(1) = A(2) = A$, $\mu(1) = \mu(2) = 0.5$, $p(x \mid x, a) = 1$ for all $(x, a) \in X \times A$, and there is a discount factor $\beta = 0.5$. Of course, instead of the discount factor, we can consider an absorbing state $\mathfrak{T}$ to which the model moves with probability 0.5 each time before it reaches $\mathfrak{T}$. Let π be a stationary policy with $\pi(a^i \mid 1) = \pi(a^i \mid 2) = 0.5$, $i = 1, 2$. Then straightforward computations imply that $Q_\mu^\pi(x, a) = 0.5$ for all $(x, a) \in X \times A$. For a deterministic policy ϕ we have that $Q_\mu^\phi(x, \phi(x)) = 1$ for all $x \in X$. It is easy to verify that $m = 2$ and the ordered sets $\{\alpha_1, \alpha_2, \alpha_3\}$ and $\{\phi^1, \phi^2, \phi^3\}$ satisfy the properties stated in Theorem 6.1 if and only if $\alpha_1 = \alpha_3 = 0.5$, $\alpha_2 = 0$, $\phi^1(x) \ne \phi^3(x)$ for all $x \in X$, and either $\{\phi^2(1), \phi^2(2)\} = \{\phi^1(1), \phi^3(2)\}$ or $\{\phi^2(1), \phi^2(2)\} = \{\phi^3(1), \phi^1(2)\}$.

The proof of Theorem 6.1 is constructive and motivates the following algorithm that splits a given stationary policy into deterministic policies satisfying the properties described in Theorem 6.1.

Algorithm 1

Input: A stationary policy σ and a restriction ϕ^1 of σ.

Output: A nonnegative integer m, restrictions $\phi^2, \ldots, \phi^{m+1}$ of σ and nonnegative numbers $\alpha_1, \ldots, \alpha_{m+1}$ that satisfy the conclusions of Theorem 6.1.

Initiation:

1. Compute $\{q_\mu^\sigma(x)\colon x \in X \backslash \mathfrak{T}\}$ and $\{Q_\mu^\sigma(x, a)\colon x \in X \backslash \mathfrak{T} \text{ and } a \in A(x)\}$.
2. Set $q(x) \leftarrow q^\sigma(x)$, $A^*(x) \leftarrow A^\sigma(x)$ for $x \in X \backslash \mathfrak{T}$, $\phi \leftarrow \phi^1$, $j \leftarrow 1$, $U \leftarrow \{x \in X \backslash \mathfrak{T}\colon |A^*(x)| > 1, q(x) = 0\}$, $V \leftarrow \{x \in X \backslash \mathfrak{T}\colon |A^*(x)| > 1, q(x) > 0\}$, and $Q(x, a) \leftarrow Q_\mu^\sigma(x, a)$ for $x \in V$ and $a \in A^*(x)$.

Preliminary step:

3. While $U \ne \varnothing$ do Steps 3(a)–(d):
   (a) Select any $z \in U$ and any $a \in A^*(z) \backslash \{\phi(z)\}$; set $\alpha_j \leftarrow 0$.
   (b) Define deterministic policy ϕ^{j+1} by: $\phi^{j+1}(z) = a$ and $\phi^{j+1}(x) = \phi^j(x)$ for $x \ne z$.
   (c) Set $A^*(z) \leftarrow A^*(z) \backslash \{\phi(z)\}$, $\phi(z) \leftarrow \phi^{j+1}(z)$ (do not change the other $A^*(x)$'s and $\phi(x)$'s), and $j \leftarrow j + 1$ (in particular, do not change the $Q(x, a)$'s or the $q(x)$'s).
   (d) If $|A^*(z)| = 1$, then set $U \leftarrow U \backslash \{z\}$.

Recursive step:

4. While $V \ne \varnothing$, do Steps 4(a)–(d):
   (a) Set

$$\alpha_j \leftarrow \min_{x \in V} \frac{Q(x, \phi(x))}{q_\mu^\phi(x)}, \qquad G \leftarrow \left\{x \in V\colon \frac{Q(x, \phi(x))}{q_\mu^\phi(x)} = \alpha_j\right\}, \quad k \leftarrow |G|,$$

and enumerate the elements of G, say $G = \{x^1, \ldots, x^k\}$.

(b) For $i = 1, \ldots, k$, do:

$$\phi^{j+i}(x) \leftarrow \begin{cases} \phi^{j+i-1}(x) & \text{if } x \neq x^i, \\ \text{arbitrary } a \in A^\sigma(x) \backslash \{\phi^{j+i-1}(x)\} & \text{if } x = x^i. \end{cases}$$

(c) For $i = 1, \ldots, k-1$, do $\alpha_{j+i} = 0$.

(d) Set $A^*(x) \leftarrow A^*(x) \backslash \{\phi(x)\}$ and $\phi(x) \leftarrow \phi^{j+k}(x)$ for $x \in G$ (do not change the other $A^*(x)$'s and $\phi(x)$'s), $V \leftarrow V \backslash \{x \in G\colon |A^*(x)| = 1\}$, $j \leftarrow j+k$, and $Q(x, \phi(x)) \leftarrow Q(x, \phi(x)) - \alpha_j Q_\mu^\phi(x, \phi(x)$ for $x \in V$ (do not change the other $Q(x, a)$'s).

Final step:

5. Set $m \leftarrow j-1$ and $\alpha_{m+1} \leftarrow 1 - \sum_{i=1}^m \alpha_i$. Output $m, \phi^2, \ldots, \phi^{m+1}, \alpha_1, \ldots, \alpha_{m+1}$.

Before presenting a formal result about Algorithm 1, we make two observations.

Observation 1. There are two alternatives for representing the input policy σ: through the $\sigma(a \mid x)$'s or through the $Q_\mu^\sigma(x, a)$'s; for an example of the latter, see §8 which discusses the solution of constrained MDPs. When σ is presented by the $\sigma(a \mid x)$'s, the $A^\sigma(x)$'s are easily available; however, the $q_\mu^\sigma(x)$'s and $Q_\mu^\sigma(x, a)$'s require computation that is explained in the second observation. When σ is represented by the $Q_\mu^\sigma(x, a)$'s, one has $q_\mu^\sigma(x) = \sum_{a \in A(x)} Q_\mu^\sigma(x, a)$ for $x \in X \backslash \mathcal{X}$ and $A^\sigma(x) = \{a \in A(x) \mid Q_\mu^\sigma(x, a) > 0\}$ for $x \in X \backslash \mathcal{X}$ with $q_\mu^\sigma(x) > 0$; in addition, one can have $A^\sigma(x)$ as an arbitrary element from $A(x)$ for $x \in X \backslash \mathcal{X}$ with $q_\mu^\sigma(x) = 0$ (these selections are consistent with (34)). It is noted that there is no need to determine actual values of the $\sigma(a \mid x)$'s as they are not used in the preliminary or recursive steps of Algorithm 1 (though they could have been determined from (34)).

Observation 2. The algorithm requires the computation of Q_μ^ϕ for each deterministic policy that is generated, and possibly the computation of Q_μ^σ for the input policy σ (see the first observation). Let γ be a stationary policy (including $\gamma = \phi$ and $\gamma = \sigma$). For an absorbing MDP, $Q_\mu^\gamma(x, a) = 0$ for $x \in \mathcal{X}$ and $Q_\mu^\gamma(x, a) = q_\mu^\gamma(x)\gamma(a \mid x)$ for $x \in X \backslash \mathcal{X}$, where the values $\{q_\mu^\gamma(x)\colon x \in X \backslash \mathcal{X}\}$ are the solution of the system of linear equations

$$q_\mu^\gamma(x) = \mu(x) + \sum_{y \in X \backslash \mathcal{X}} \left(\sum_{a \in A(y)} p(x \mid y, a)\gamma(a \mid y) \right) q_\mu^\gamma(y), \quad x \in X \backslash \mathcal{X}; \tag{41}$$

see Altman [1, Lemma 7.1]. In particular, when γ is a deterministic policy, that is, $\gamma \in S$, the coefficients multiplying $q_\mu^\phi(y)$ simplify to $p(x \mid y, \phi(y))$. Whether γ is deterministic or not, the computation of q_μ^γ is available by solving a nonsingular $|X \backslash \mathcal{X}| \times |X \backslash \mathcal{X}|$ system of linear equations (41).

Let $N := |X \backslash \mathcal{X}|$ and $M = \sum_{x \in X \backslash \mathcal{X}} |A(x)|$. Evidently, a stationary policy can be exactly m-randomized stationary only for $m \leq M - N$.

Theorem 6.2. *With input (σ, ϕ^1), Algorithm 1 finishes in finite time with output $(m, \phi^2, \ldots, \phi^{m+1}, \alpha_1, \ldots, \alpha_{m+1})$ that satisfies the conclusions of Theorem 6.1; furthermore; m is the unique integer for which σ is exactly m-randomized. Executing the algorithm requires at most $O(m \cdot N^{2.376})$ arithmetic operations and $m \leq M - N$.*

Proof. Assume that σ is an exactly m'-randomized stationary policy (with $0 \leq m' \leq M - N$). Following the initiation step, $\sum_{x \in X \backslash \mathcal{X}} |A^*(x)| = \sum_{x \in X \backslash \mathcal{X}} |A^\sigma(x)| = m' + N$ and the execution of each preliminary and recursive step reduces $\sum_{x \in X \backslash \mathcal{X}} |A^*(x)|$ by a positive number that equals the increase of j. It follows that the algorithm must terminate in finite time; at termination, $|A^*(x)| = 1$ for each x ($U = V = \varnothing$), which implies that $j = 1 + [(m' + N) - N] = m' + 1$. The number of generated deterministic policies is equal to the total increase in j, that is, $m' = j - 1 = m$, and the number of generated nonnegative numbers is $m + 1$.

The induction used in the proof of Theorem 6.1 yields a recursion that generates the splitting asserted in the statement of that theorem. This recursion is implemented in Algorithm 1 with some modifications. We next review the differences between the "algorithm" and the "recursion." In particular, the comparison demonstrates that the algorithm and the induction generate the same sequences of deterministic policies and nonnegative numbers; consequently, the proof of Theorem 6.1 assures that the output of the algorithm satisfies the asserted conclusions.

Modification 1. Given a stationary policy σ, the inductive step generates a policy π which is then treated by the induction hypothesis. The algorithm does not compute the policy π explicitly and, instead of this, it directly computes the occupancy values for π in Step 4(d). It is straightforward to compute the policy π by using (34), but such computations are not needed.

Modification 2. The inductive step considers two possible cases: (i) $U \equiv \{z \in X\backslash\mathcal{X}\colon\ A^\sigma(z)| > 1$ and $q_\mu^\sigma(z) = 0\} \neq \varnothing$, and (ii) $U = \varnothing$. Implementing the recursion seems to require consideration of situations with $U \neq \varnothing$ following $U = \varnothing$, but the paragraph following the proof of Theorem 6.1 explains that this will never happen. The algorithm explicitly excludes such situations, as $U \neq \varnothing$ is restricted to the preliminary step.

Modification 3. Instead of the recursion (39), the algorithm uses its simpler normalized version

$$Q \leftarrow (1-\alpha_j)Q^\pi \leftarrow Q - \alpha_j Q_\mu^\phi,$$

with the initial values $Q = Q_\mu^\sigma$. This allows us to simplify the formulae in the algorithm and simplify the calculations of the coefficients α_j. The algorithm computes the coefficients α_j described in Theorem 6.1 directly in Step 4(a) by using the ratios $Q(x,a)/q_\mu^\pi(x)$ instead of the ratios in (37) and the following calculations of the coefficient α_j as some products.

The most computation-demanding parts in executing Algorithm 1 is the initial computation of Q_μ^σ, if required, and the computation of Q_μ^ϕ at the recursive step. Each such computation requires the solution of an $N \times N$ system of linear equations (in fact, the algorithm may be specified in the way that the size of the system is $|V| \times |V|$, where $|V| \le N$, but we do not discuss this here); this can be done by Gaussian elimination in $O(N^3)$ arithmetic operations, by the Strassen algorithm (Strassen [48]) in $O(N^{2.807})$ arithmetic operations and by the Coppersmith-Winograd algorithm (Coppersmith and Winograd [10]) in $O(N^{2.376})$ arithmetic operations. All other computations are dominated by the above. As the number of iterations is bounded by m and the computation of at most $(m+1)$ occupancy measures $Q_\mu^\sigma, Q_\mu^{\phi^1}, \ldots, Q_\mu^{\phi^m}$ is required, the asserted complexity bound follows. □

The normalization of the occupancy rates in the algorithm, explained in Modification 3, can be interpreted through normalization of the initial distribution. Specifically, $(1-\alpha_1)Q_\mu^\pi = Q_\nu^\pi$ for every policy π and initial distribution μ, where ν is the initial distribution on X such that $\nu(x) = (1-\alpha_1)\mu(x)$ for all $x \in X\backslash\mathcal{X}$. Thus, the algorithm generates a sequence of occupancy measures for the policies σ or π defined in the inductive proof of Theorem 6.1, but with respect to altering initial distributions.

Consider the question of how to represent the occupancy values of an m-randomized stationary policy σ in the form of (36), without the specification of ϕ^1 and without the extra requirement about the "similarity" of consecutive restrictions. The equivalent question is how to represent the occupancy values of σ by (36) with m replaced by some $m' \le m$. In fact, the existence of such a splitting follows from Carathéodory's theorem about the representation of a point in a polytope as a convex combination of vertices of that polytope (recall that the occupancy arrays of deterministic policies are the vertices of the set of occupancy arrays of all stationary policies which equals the feasible set of (32)–(33)). In those cases, where one is only interested in the variant of (36), Algorithm 1 can be simplified by excluding restrictions that get weight 0; in particular, whenever $q(x) = 0$, one can replace $A^*(x)$ with any singleton in $A(x)$ and Step 3(b) can be simplified by determining only ϕ^{t+1}. While the simplified algorithm will void some of the operations of Algorithm 1, it does not have a better (worse case) complexity bound. The simplified algorithm, as well as Algorithm 1 with its full detail, construct the decomposition asserted by Carathéodory's theorem.

A geometric interpretation of the "adjacency condition" of consecutive restrictions in the generated sequence of Theorem 6.1 is that the occupancy values of consecutive restrictions are vertices that share a common edge. It is noted that the Carathéodory theorem cannot be extended, in general, to include this requirement—specifically, in general, a point in a d-dimensional polytope need not be representable by $d+1$ vertices that can be enumerated so that each consecutive pair share an edge.

If the set V consists of one point, it is possible to find the remaining values of α_i by splitting at the single state to which V is equal. In this case the algorithm will perform fewer operations, though its computational complexity bounds will remain the same. To implement this, the condition $V \neq \varnothing$ in Step 4 can be changed to $|V| > 1$ and the following operations can be added between Steps 4 and 5:

If $|V| = 1$ consider the state x^* such that $V = \{x^*\}$, consider its action set $A^*(x^*) = \{a^0, \ldots, a^l\}$, where $a^0 = \phi^t(x^*)$, and select the restrictions $\phi^{t+1}, \ldots, \phi^{t+l}$ of σ as $\phi^{t+j}(x^*) = a^j$, $j = 1, \ldots, l$, and $\phi^{t+j}(x^*) = \phi^t(x)$ for $x \neq x^*$ (observe that $t + l - 1 = m + 1$), compute

$$\alpha_{t+i} := \frac{Q(x^*, a^i)}{q^\phi}, \quad i = 1, \ldots l-1,$$

and set $t \leftarrow t + l - 1$.

It is possible to use embedded MDPs and solve linear equations (41) for $x \in V \subseteq X\backslash\mathcal{X}$ instead of $x \in X\backslash\mathcal{X}$. This requires recomputing the transition probabilities when $A^*(x)$ becomes a singleton and state x should be eliminated. The details are straightforward.

7. Finite splitting for MDPs with general state and action sets. This section extends the results of the previous section to MDPs with Borel state and action sets. Let σ be a stationary policy. For a stationary policy σ and a state $x \in X$, let $A^\sigma(x) := \{a \in A(x) \mid \sigma(a \mid x) > 0\}$. We say that a stationary policy π is *discrete* at x if $\sigma(A^\sigma(x) \mid x) = 1$. We say that a stationary policy *uses a finite number of actions at* x if it is discrete at x and $A^\sigma(x)$ is finite. Suppose that Y is a finite subset of $X\backslash\mathscr{T}$ such that the policy σ uses a finite number of actions at any state $y \in Y$. We denote by F_Y^σ the set of functions $f\colon Y \to A$ such that $f(y) \in A^\sigma(y)$ for all $y \in Y$. For $f \in F_Y^\sigma$, consider the stationary policy $\sigma[Y, f]$ that coincides with σ outside of Y and selects the actions $f(y)$ at $y \in Y$. In other words, for a measurable subset B of A,

$$\sigma[Y, f](B \mid x) = \begin{cases} \mathbf{I}\{f(x) \in B\} & \text{if } x \in Y, \\ \sigma(B \mid x) & \text{otherwise.} \end{cases}$$

For $m = 0, 1, \ldots,$ a stationary policy π is called (*exactly*) *m-randomized stationary* if there exists a finite set of states Y such that (i) π uses only one action at each state $x \in X\backslash Y$; (ii) π uses a finite number of actions at each state $x \in Y$; and (iii) with Y replacing X, (35) holds (with equality).

The following theorem generalizes Theorem 6.1.

THEOREM 7.1. *Assume that σ is a stationary policy and Y is a finite subset of $X\backslash\mathscr{T}$ such that σ uses a finite number of actions at each state $y \in Y$ and, in addition, $q_\mu^{\sigma[Y,f]}(Y) < \infty$ for all $f \in F_Y^\sigma$. Let $m := \sum_{y\in Y}(|A^\sigma(y)| - 1)$ and $f^1 \in F_Y^\sigma$. Then there exist mappings $f^2, \ldots, f^{m+1}$ in F_Y^σ and nonnegative numbers $\alpha_1, \ldots, \alpha_{m+1}$ such that $f^1, \ldots, f^{m+1}$ are distinct, $\sum_{j=1}^{m+1} \alpha_j = 1$,*

$$Q_\mu^\sigma = \sum_{j=1}^{m+1} \alpha_j Q_\mu^{\sigma[Y,f^j]}, \tag{42}$$

and for $i = 1, \ldots, m$ there is exactly one state $y^i \in Y$ such that $f^i(y^i) \neq f^{i+1}(y^i)$.

PROOF. Consider two cases: (i) $q_\mu^\sigma(Y) = 0$, and (ii) $q_\mu^\sigma(Y) \neq 0$. If case (i) takes place, then $Q_\mu^{\sigma[Y,f]} = Q_\mu^\sigma$ for any $f \in F_Z^\sigma$, and (42) holds for any nonnegative $\alpha_1, \ldots, \alpha_{m+1}$ whose sum is 1 and for any $f^1, \ldots, f^{m+1}$ in F_Z^σ. In particular, one can set $\alpha_1 = 1$, $\alpha_j = 0$, $j = 2, \ldots, m+1$, and choose f^j, $j = 1, \ldots, m+1$, by applying the construction of Step 3 of Algorithm 1 with $U = Y$, $A^*(x) = A^\sigma(x)$, $x \in Y$, and $\phi^1 = f^1$, and setting $f^j = \phi^j$ for $j > 1$. Next assume that case (ii) takes place. For $x \in Y$, reduce the action sets $A(x)$ to $A^\sigma(x)$. After this reduction, consider the embedded MDP with the state set Y corresponding to σ by using the construction of the paragraphs following (14) (a construction that applies to any stationary policy). Observe that $\nu(Y) > 0$, where ν is the initial state distribution in the embedded MDP. The assumption $q_\mu^{\sigma[Y,f]}(Y) < \infty$ for all $f \in F_Y^\sigma$ implies $\tilde{q}_\nu^\phi(Y) := \tilde{Q}_\nu^\phi(Y, A) < \infty$ for any deterministic policy ϕ in the embedded MDP. As explained in §3 with the reference to Blackwell [4], this implies that the embedded MDP is absorbing. Thus Theorem 6.1 can be applied to the embedded MDP. Formulas (42) and (36) are particular cases of (14) and (16), respectively, when ρ^* and $\tilde{\rho}$ are discrete measures concentrated at $m+1$ points. Lemma 4.8 then implies that (42) and the conclusions of this theorem hold for the original MDP with $f^2, \ldots, f^{m+1}$ corresponding to the constructed deterministic policies in the embedded MDP. □

Splitting formula (42) without the final requirement of Theorem 7.1 is an instance of (14).

The following corollary demonstrates that the assumptions that the state and action sets are finite are not needed in Theorem 6.1.

COROLLARY 7.1. *Consider an absorbing MDP. Let σ be an exactly m-randomized stationary policy and let ϕ^1 be a restriction of σ. If $m > 0$, then there exist deterministic policies $\phi^2, \ldots, \phi^{m+1}$ and nonnegative numbers $\alpha_1, \ldots, \alpha_{m+1}$ satisfying the properties described in Theorem 6.1.*

PROOF. Consider the finite set $Y = \{x \in X\backslash\mathscr{T}\colon |A^\sigma(x)| > 1\}$ and apply Theorem 7.1. □

Example 5.1 demonstrates that the assumption $q_\mu^\sigma(Y) < \infty$ is not sufficient for the validity of Theorem 7.1. Also, for absorbing (and, in particular, discounted) MDPs, the assumption $q_\mu^{\sigma[Y,f]}(Y) < \infty$ for all $f \in F_Y^\sigma$ holds with $Y = X\backslash\mathscr{T}$, because for such MDPs, $q_\mu^\pi(Y) \le q_\mu^\pi(X) < \infty$ for any policy π. In addition, this assumption holds for transient countable state MDPs, because $q_\mu^\pi(Y) = \sum_{y\in Y} q_\mu^\pi(\{y\}) < \infty$ for any policy π, when Y is finite.

Algorithm 1 can be applied to situations described in Theorem 7.1, including infinite state and action MDPs. In those cases, stationary policies $\sigma[Y, f^j]$ play the role of deterministic policies ϕ^j in finite state-action versions of the algorithms. In view of Lemma 4.8, it is sufficient to consider only $x \in Y$. Computations of occupancy values $Q_\mu^{\sigma[Y,f^j]}(x, a)$ and $Q_\mu^\sigma(x, a)$ for $x \in Y$ can be done either by simulation, or by using embedding, or by problem-specific methods.

8. Constrained absorbing MDPs with finite state and action sets. Consider an absorbing MDP with finite state and action sets and $(K+1)$ reward functions $r_0(\cdot,\cdot),\dots,r_K(\cdot,\cdot)$ defined on $X\times A$, where K is a positive integer. For an absorbing MDP we assume that $r_k(x,a)=0$ when $z\in\mathcal{Z}$, $k=0,1,\dots,K$. For $k=0,\dots,K$, let $V_k^\pi(\mu)$ be defined by (2) with r_k replacing r. For numbers $c_1,\dots,c_K$, consider the problem of finding an optimal policy for the problem

$$\text{maximize}\{V_0^\pi(\mu)\mid V_k^\pi(\mu)\ge c_k,\ k=1,\dots,K\}. \tag{43}$$

It is well known that if problem (43) is feasible, then there exists an optimal policy that is a mixture of $K+1$ deterministic policies, and such a policy can be found from a solution of the following LP (Feinberg [19], Altman [1, p. 133]):

$$\text{maximize}\left\{\sum_{\phi\in S}V_0^\phi u_\phi \;\middle|\; \begin{array}{ll}\displaystyle\sum_{\phi\in S}V_k^\phi u_\phi\ge c_k, & k=1,\dots,K,\\ \displaystyle\sum_{\phi\in S}u_\phi=1, & \\ u_\phi\ge 0, & u_\phi\in S.\end{array}\right\}. \tag{44}$$

However, the set of deterministic policies S grows exponentially with the growth of the problem size and the above LP is typically computationally intractable (Altman [1, p. 117]).

As in §6, let $X^*=X\setminus\mathcal{Z}$. It is also known that, if problem (43) is feasible, then there is an optimal policy that is K-randomized stationary. A standard method for determining such a policy consists of two steps. The first step is to find an optimal basic feasible solution to the following LP:

$$\text{maximize}\sum_{x\in X^*}\sum_{a\in A(x)}r_0(x,a)Q(x,a) \tag{45}$$

subject to

$$\sum_{x\in X^*}\sum_{a\in A(x)}r_k(x,a)Q(x,a)\ge c_k,\quad k=1,\dots,K, \tag{46}$$

$$\sum_{a\in A(x)}Q(x,a)-\sum_{y\in X^*}\sum_{a\in A(y)}p(x\mid y,a)Q(y,a)=\mu(x),\quad x\in X^*, \tag{47}$$

$$Q(x,a)\ge 0,\quad x\in X^*,\ a\in A(x). \tag{48}$$

The second step is then to determine a K-randomized stationary policy whose occupancy values coincide with the found optimal solution of the LP, e.g., by using (34) with deterministic selection of any action when $\sum_{b\in A(x)}Q(x,b)=0$ (see Kallenberg [35, Algorithm VII, Chapter 3]).

The following efficient algorithm finds an optimal policy to (43) which is a mixture of deterministic policies and has further structure.

Algorithm 2
Input: Problem (43).

Output: Either a finite list of deterministic policies and corresponding nonnegative numbers whose sum is 1 or the response "infeasible."

1. Find an optimal basic solution Q of LP (45)–(48) or conclude that it is not feasible; in the latter case output "infeasible."

2. Split Q by applying Algorithm 1 using ϕ^1 as any deterministic policy having $Q(x,\phi^1(x))>0$ for each $x\in X^*$ such that $Q(x,a)>0$ for some $a\in A(x)$.

Let m be a nonnegative integer. A policy π is called m-deterministic if there exist $(m+1)$ deterministic policies $\phi^1,\dots,\phi^{m+1}$ and $(m+1)$ nonnegative numbers $\alpha_1,\dots,\alpha_{m+1}$ such that $\sum_{i=1}^{m+1}\alpha_i=1$ and

$$P_\mu^\pi=\sum_{i=1}^{m+1}\alpha_i P_\mu^{\phi^i}; \tag{49}$$

this definition can be interpreted as randomly selecting, up front, one of the given $(m+1)$ deterministic policies with α_i as the probability of selecting ϕ^i. A policy is deterministic if and only if it is 0-deterministic. Notice

that (49) specializes (6) with $\Delta = S$ when ν has a finite support, and this definition is also applicable to MDPs with infinite state and action sets. In particular, (49) refers to the strategic measures, rather than the occupancy measures used in (36). Also, the above definition does not require the "adjacency property" of Theorem 6.1.

We recall that an algorithm is *weakly* (*strongly*) *polynomial*, if the number of arithmetic operations that it performs is bounded above by a polynomial in the number of binary bits (real numbers) in a representation of the input.

THEOREM 8.1. *For an absorbing MDP with finite state and action sets, the following statements hold*:

(i) *Algorithm 2 terminates in Step 1 if and only if (43) is infeasible. In this case, the output is "infeasible."*

(ii) *If problem (43) is feasible, then the output* $(m, \phi^1, \dots, \phi^{m+1}, \alpha_1, \dots, \alpha_{m+1})$ *of Algorithm 2 defines an m-deterministic optimal policy that randomly selects, up front, the deterministic policy* ϕ^i *with probability* α_i, $i = 1, 2, \dots, m+1$, *where* $m = 1, \dots, \min\{K, \sum_{x \in X^*} |A(x)| - |X^*|\}$, *and any two sequential deterministic policies* ϕ^i *and* ϕ^{i+1}, $i = 1, 2, \dots, m$, *differ only at one state.*

(iii) *Algorithm 2 is weakly polynomial.*

PROOF. By the paragraphs preceding Algorithm 2, if problem (43) is feasible, then for some nonnegative integer $m \le \min\{K, \sum_{x \in X^*} |A(x)| - |X^*|\}$ the output of Step 1 will have at most $|X^*| + m$ nonzero variables, and it corresponds to an exactly m-randomized stationary optimal policy. The output of Algorithm 2 has the desired properties guaranteed by Algorithm 1, as described in Theorem 6.2. Finally, the complexity bound of Algorithm 2 follows from Theorem 6.2 (by which Algorithm 1 is strongly polynomial) and the existence of weakly polynomial algorithms for solving LPs. □

9. Constrained MDPs with Borel state and compact action sets. Consider an MDP with Borel state and action sets and $(K+1)$ reward functions $r_0, \dots, r_K$ defined on $X \times A$, where K is a positive integer and each of these functions is measurable on $X \times A$ and bounded from above. For $k = 0, \dots, K$, let $V_k^\pi(\mu)$ be defined by (2) with r_k replacing r. The following condition implies that $V_k^\pi(\mu)$ are well defined, relaxing the assumption that the MDP is absorbing; see Schäl [45] and Feinberg [22].

General Convergence Condition. The inequality $E_\mu^\pi \sum_{t=0}^\infty r_k^+(x_t, a_t) < \infty$ holds for all $\pi \in RM$ and for all $k = 0, \dots, K$, where $z^+ = \max\{z, 0\}$ for a real number z.

For real numbers $c_1, \dots, c_K$, we consider problem (43). The main result of this section, Theorem 9.2, describes sufficient conditions for the existence of a stationary optimal policy and an optimal policy, which is a mixture of deterministic policies, for this problem, when the MDP is absorbing. Theorem 9.2 is, in fact, an application of the results of §4.

We say that an MDP satisfies *Complete Condition (W)* if it satisfies Condition (W) of §4 and, in addition, each function $r_k(\cdot, \cdot)$ is upper-semicontinuous on $X \times A$. We say that an MDP satisfies *Complete Condition* (S), if it satisfies Condition (S) of §4 and, in addition, for each $x \in X$, the functions $r_k(x, \cdot)$ is upper-semicontinuous on $A(x)$. (Our Complete Conditions (W) and (S) correspond, respectively, to Conditions (W) and (S) in Balder [3] and Schäl [45, 46].) We shall also consider the following condition.

Condition (C). The convergence $\sup_{N \ge n} \sup_{\phi \in M} E_\mu^\phi \sum_{t=n+1}^N r_k(x_t, a_t) \to 0$ as $n \to \infty$ takes place for all $k = 0, \dots, K$.

The General Convergence Condition and Condition (C) are equivalent to the similar conditions introduced in slightly different forms in Schäl [45], where the supremum in Condition (C) is taken over all policies $\phi \in R\Pi$ and the General Convergence Condition is assumed for all $\pi \in R\Pi$. However, it is possible to consider the smaller sets M and RM respectively, because of the sufficiency of Markov and nonrandomized Markov policies; see Feinberg [16].

The next result provides sufficient conditions for the existence of an optimal policy for problem (43); for related results, see Piunovskiy [42, 43], Hernández-Lerma and González-Hernández [30], and Feinberg and Piunovskiy [24].

THEOREM 9.1. *Consider an MDP satisfying Condition* (C), *the General Convergence Condition, and either Complete Condition* (S) *or Complete Condition* (W). *If problem* (43) *is feasible, then there exists an optimal policy.*

PROOF. By Schäl [45] and Balder [3], the set of all strategic measures $L_\mu^{R\Pi}$ is compact in the corresponding topologies and the mappings $P_\mu^\pi \to V_k^\pi(\mu)$ are upper-semicontinuous on $L_\mu^{R\Pi}$. Thus, the sets

$B_k = \{P_\mu^\pi\colon V_k^\pi(\mu) \ge c_k\}$, $k = 1, \ldots, K$, and $\bigcap_{k=1}^K B_k$ are compact and the objective function $V_0(P_\mu^\pi) := V_0^\pi(\mu)$ achieves a maximum on this set. □

THEOREM 9.2. *Consider an absorbing MDP satisfying Condition* (C) *and either Complete Condition* (S) *or Complete Condition* (W).

(i) *For each policy π there exist a stationary policy σ and a $(K+1)$-deterministic policy γ such that $V_k^\pi(\mu) = V_k^\sigma(\mu) = V_k^\gamma(\mu)$ for all $k = 0, \ldots, K$.*

(ii) *If problem* (43) *is feasible, then there exists a stationary optimal policy and there exists a K-deterministic optimal policy.*

PROOF. (i) For an arbitrary policy π, consider a stationary policy σ defined in Lemma 4.1. Lemma 4.2 implies that $Q_\mu^\sigma = Q_\mu^\pi$, and therefore,

$$V_k^\sigma(\mu) = \int_X \int_A r_k(x,a)\, Q_\mu^\sigma(dx, da) = \int_X \int_A r_k(x,a)\, Q_\mu^\pi(dx, da) = V_k^\pi(\mu).$$

Next, by Corollary 4.1 and Lemma 4.7, the set $\mathscr{V}(\mu) := \{V^\pi(\mu))\colon \pi \in R\Pi)\}$, where $V^\pi(\mu) = (V_0^\pi(\mu), \ldots, V_K^\pi(\mu))$ is a $(K+1)$-dimensional projection of the convex compact $\mathscr{O}_\mu^{R\Pi}$. According to Theorem 4.1, any occupancy measure Q_μ^π is a barycenter of the set $\mathscr{O}_\mu^S$. Therefore, $V^\pi(\mu)$ is a barycenter of the set $\mathscr{V}^S(\mu) = \{V^\phi(\mu)\colon \phi \in S\}$. Carathéodory's theorem implies that this vector is a convex combination of at most $(K+2)$ points from $\mathscr{V}^S(\mu)$.

(ii) By Theorem 9.1, there exists an optimal policy, say π. Any policy σ with $V_k^\pi(\mu) = V_k^\sigma(\mu)$ for $k = 1, \ldots, K$ is then optimal, and part (i) assures the existence of a stationary optimal policy satisfying this condition. To verify the second part of (ii) observe that $V^\pi(\mu)$ belongs to the boundary of the convex set $\mathscr{V}(\mu)$, because, if a neighborhood of this point (in the Euclidian metric) belongs to $\mathscr{V}(\mu)$, then there exists a feasible policy σ such that $V_0^\sigma(\mu) > V_0^\pi(\mu)$. Statement (i) assures the existence of a $(K+1)$-deterministic policy σ with $V^\sigma(\mu) = V^\pi(\mu)$. If σ is not exactly $(K+1)$-deterministic, it is K-deterministic optimal, and the proof is complete. Let σ be exactly $(K+1)$-deterministic. Consider a supporting hyperplane P to the convex set $\mathscr{V}(\mu)$ such that $V^\sigma(\mu) \in P$. All $(K+2)$ points from $\mathscr{V}^S(\mu)$, whose convex combination is $V^\sigma(\mu)$, belong to P. Since the dimension of P is K, according to Carathéodory's theorem, it is possible to select the coefficients α_i, $i = 1, \ldots, K+2$, in such a way that at most $(K+1)$ of them are positive. □

Theorem 9.2 does not imply the existence of a K-deterministic optimal policy whose composing stationary policies satisfy the "adjacency property" of the conclusions of Theorems 6.1 and 8.1.

By Schäl [45], Condition (C) is implied by the following condition:

$$\sup_{\phi \in M} E_\mu^\phi \sum_{t=n}^\infty r_k^+(x_t, a_t) \to 0 \qquad \text{as } n \to \infty, \quad k = 0, \ldots, K.$$

In particular, Condition (C) is satisfied in the following two cases: (i) the functions r_k are nonpositive, and (ii) the MDP is discounted. Also, since the functions r_k are bounded from above, the following condition,

$$\sup_{\phi \in M} \sum_{t=n}^\infty P_\mu^\phi\{x_t \in X \setminus \mathscr{Z}\} \to 0 \qquad \text{as } n \to \infty, \quad k = 0, \ldots, K.$$

implies that the MDP is absorbing and that Condition (C) holds. Finally, Feinberg [22, Example 6.8] demonstrates that optimal policies may not exist for an absorbing MDP with a countable state space and finite action sets without Condition (C).

10. Constrained discounted MDPs with countable state spaces. In this section, we consider countable state discounted MDPs with $(K+1)$ reward functions $r_0, \ldots, r_K$ defined on $X \times A$, where K is a positive integer. The discount factor β will be assumed fixed, and we shall suppress indexing by β. For each policy π and $k = 0, \ldots, K$, we shall use the notation $\tilde{V}_k^\pi(\mu)$ for the expected total discounted reward defined by the right-hand side of (1), with r_k replacing r. For numbers $c_1, \ldots, c_K$, we consider the variant of (43) given by

$$\text{maximize}\{\tilde{V}_0^\pi(\mu) \mid \tilde{V}_k^\pi(\mu) \ge c_k,\ k = 1, \ldots, K\}. \tag{50}$$

For a countable state space X, Conditions (W) and (S) are equivalent, as explained in detail in the appendix of Chen and Feinberg [9]. Therefore, for bounded above one-step rewards, Complete Conditions (W) and (S) can be written in the following form: (1) X is countable; (2) A is a metric space; (3) for each $x \in X$, $A(x)$ is a compact subset of A; (4) for all $x, y \in X$, $p(y \mid x, \cdot)$ is continuous on $A(x)$; (5) for all $x \in X$ and $a \in A(x)$, $p(X \mid x, a) = 1$; (6) for each $x \in X$, the functions $r_0(x, \cdot), r_1(x, \cdot), r_2(x, \cdot), \ldots$, are bounded from above and upper semi-continuous in $a \in A(x)$. We assume in this section that conditions (1)–(6) hold.

Problem (50) was studied by Feinberg and Shwartz [25]. First, we recall some definitions from Feinberg and Shwartz [25].

For finite nonnegative integers m and N, a stationary policy π is called *a strong* (m, N)*-policy* if: (a) it is deterministic from time N onward; that is, $\pi_n(\phi(x) \mid x) = 1$ for some deterministic policy ϕ for all $x \in X$ and for all $n \geq N$, and (b) for all states it uses no more than m additional actions than a deterministic policy would use, and 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; that is, for each $x \in X$ there exists a finite subset $B(x)$ of A(x) such that $\pi_n(B(x) \mid x) = 1$ for all $n = 0, 1, \ldots$, and for all $x \in X$, and the following two properties hold:

$$\sum_{x \in X} (|B(x)| - 1) \leq m,$$

and

$$\sum_{n=0}^{N-1} \sum_{x \in X} \left[\left(\sum_{a \in B(x)} \mathbf{I}\{\pi_n(a \mid x) > 0\} \right) - 1 \right] \leq m.$$

Feinberg and Shwartz [25, Theorem 2.1] proved that if problem (50) is feasible, then

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

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

Similar results were established by Borkar for (the more difficult) average reward criterion (see Borkar [7, Lemma 11.24, Theorem 11.6], references to Borkar's papers in Borkar [7], and the remarks on parallel treatments of average rewards per unit time and discounted rewards in Borkar [7, p. 368]).

Feinberg and Shwartz [25, Theorem 5.1]) also showed that for any m-randomized stationary policy π, where $m = 0, 1, \ldots$, there exists an m-deterministic policy σ with $Q_\mu^\sigma = Q_\mu^\pi$. Consequently, they concluded (Feinberg and Shwartz [25, Corollary 5.3]) that if problem (50) is feasible, then there exists a K-deterministic optimal policy. This conclusion follows from the second claim in Theorem 9.2(ii), which deals with absorbing MDPs with Borel state spaces. The next result follows immediately from Theorem 7.1.

THEOREM 10.1. *For any m-randomized stationary policy σ, where $m = 0, 1, \ldots$, there exists an m-deterministic policy π with $Q_\mu^\pi = Q_\mu^\sigma$ such that the deterministic policies $\phi^1, \ldots, \phi^{m+1}$ in (49) are distinct and for each $i = 1, \ldots, m$, there exists only one state x^i such that $\phi^i(x^i) \neq \phi^{i+1}(x^i)$.*

Theorem 10.1 is a stronger result than Feinberg and Shwartz [25, Theorem 5.1] because it specifies that the policies ϕ^i and ϕ^{i+1} differ only at one state. The following result strengthens Corollary 5.3 from Feinberg and Shwartz [25] described above.

THEOREM 10.2. *If problem (50) is feasible, then for some $m = 0, \ldots, K$, there exists an m-deterministic optimal policy π such that the deterministic policies $\phi^1, \ldots, \phi^{m+1}$ in (49) are distinct, and for each $i = 1, \ldots, m$, there exists only one state x^i such that $\phi^i(x^i) \neq \phi^{i+1}(x^i)$.*

PROOF. In view of Feinberg and Shwartz [25, Theorem 2.1(i)], consider an exactly m-randomized stationary optimal policy σ, where $m \leq K$. For the policy σ, consider an m-deterministic policy π whose existence is stated in Theorem 10.1. The policy π is also optimal. □

To the best of our knowledge, the validity of the conclusions of Feinberg and Shwartz [25, Theorem 2.1] and the validity of Theorem 10.2 for uncountable state constrained discounted MDPs satisfying either Complete Condition (W) or Complete Condition (S) are open questions.

Acknowledgments. The authors are grateful to Eric Denardo for his initial participation in this project and his insightful inputs. The authors express their appreciation to two anonymous referees for their insightful reviews of the original submission and to Pelin Gulsah Canbolat for a thorough reading of parts of this paper and for making many valuable suggestions. The research of Eugene Feinberg was partially supported by the National Science Foundation [Grants CMMI-0600538, CMMI-0900206, CMMI-0928490].

References

[1] Altman, E. 1999. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Boca Raton, FL.
[2] Aumann, R. 1964. Mixed and behavior strategies in infinite extensive games. *Ann. Math. Stud.* **52** 627–650.
[3] Balder, E. J. 1989. On compactness of the space of policies in stochastic dynamic programming. *Stochastic Processes Their Appl.* **32** 141–150.
[4] Blackwell, D. 1967. Positive dynamic programming. *Proc. 5th Berkeley Sympos. Math. Statistics Probab.*, Vol. 1. University of California Press, Berkeley, 415–418.
[5] Borkar, V. S. 1988. A convex analytic approach to Markov decision processes. *Probab. Theory Related Fields* **78** 583–602.
[6] Borkar, V. S. 1990. *Topics in Controlled Markov Chains.* Longman Scientific and Technical, Harlow, UK.
[7] Borkar, V. S. 2002. Convex analytic methods in Markov decision processes. E. A. Feinberg, A. Shwartz, eds. *Handbook on Markov Decision Processes*. Kluwer Academic Publishers, Boston, 347–375.
[8] Cantón, A., A. Granados, C. Pommerenke. 2004. Borel images and analytic functions. *Michigan Math. J.* **52** 279–287.
[9] Chen, R. C., E. A. Feinberg. 2010. Compactness of the space of nonrandomized policies in countable-state sequential decision processes. *Math. Methods Oper. Res.* **71** 307–323.
[10] Coppersmith, D., S. Winograd. 1990. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.* **9** 251–280.
[11] Denardo, E., H. Park, U. G. Rothblum. 2007. Risk-sensitive and risk-neutral multiarmed bandits. *Math. Oper. Res.* **32** 374–394.
[12] Derman, C. 1970. *Finite State Markovian Decision Processes*. Academic Press, New York.
[13] Dubins, L., D. Freedman. 1964. Measurable sets of measures. *Pacific J. Math.* **14** 1211–1222.
[14] Dynkin, E. B., A. A. Yushkevich. 1979. *Controlled Markov Processes and Their Applications*. Springer-Verlag, New York.
[15] Ejov, V., J. A. Filar, M. Haythorpe, G. T. Nguyen. 2009. Refined MDP-based branch-and-fixed algorithm for the Hamiltonian cycle problem. *Math. Oper. Res.* **34** 758–768.
[16] Feinberg, E. A. 1982. Nonrandomized Markov and semi-Markov strategies in dynamic programming. *SIAM Theory Probab. Appl.* **27** 116–126.
[17] Feinberg, E. A. 1986. Sufficient classes of strategies in discrete dynamic programming I: Decomposition of randomized strategies and embedded models. *SIAM Theory Probab. Appl.* **31** 658–668.
[18] Feinberg, E. A. 1991. Nonrandomized strategies in stochastic decision processes. *Ann. Oper. Res.* **29** 315–332.
[19] Feinberg, E. A. 1994. Constrained semi-Markov decision processes with average rewards. *Math. Methods Oper. Res.* **39** 257–288.
[20] Feinberg, E. A. 1996. On measurability and representation of strategic measures in Markov decision processes. T. S. Ferguson, L. S. Shapley, J. B. MacQueen, eds. *Statistics, Probability and Game Theory Papers in Honor of David Blackwell*. IMS Lecture Notes–Monograph Series, Vol. 30. Institute of Mathematical Statistics, Hayward, CA, 29–43.
[21] Feinberg, E. A. 2000. Constrained discounted Markov decision processes and Hamiltonian cycles. *Math. Oper. Res.* **25** 130–140.
[22] Feinberg, E. A. 2002. Total reward criteria. E. A. Feinberg, A. Shwartz, eds. *Handbook of Markov Decision Processes*. Methods and Applications. Kluwer, Boston, 173–207.
[23] Feinberg, E. A. 2009. Adaptive computation of optimal nonrandomized policies in constrained average-reward MDPs. *Proc. 2009 IEEE Sympos. Adapt. Dynam. Programming and Reinforcement Learn. (ADPLR 2009), Nashville, TN*, 96–100.
[24] Feinberg, E. A., A. B. Piunovskiy. 2002. Nonatomic total reward Markov decision processes with multiple criteria. *J. Math. Anal. Appl.* **273** 93–111.
[25] Feinberg, E. A., A. Shwartz. 1996. Constrained discounted dynamic programming. *Math. Oper. Res.* **21** 922–945.
[26] Feinberg, E. A., I. M. Sonin. 1983. Stationary and Markov policies in countable state dynamic programming. J. V. Prokhorov, K. Itô, eds. *Probability Theory and Mathematical Statistics*. Lecture Notes in Mathematics, Vol. 1021. Springer-Verlag, Berlin, 111–129.
[27] Feinberg, E. A., I. M. Sonin. 1996. Notes on equivalent stationary policies in Markov decision processes with total rewards. *Math. Methods Oper. Res.* **44** 205–221.
[28] Filar, J. A. 2006. *Controlled Markov Chains, Graphs & Hamiltonicity*. Foundations and Trends in Stochastic Systems, Vol. 1. Now, Boston.
[29] Gikhman, I. I., A. V. Skorohod. 1979. *Controlled Stochastic Processes*. Springer, New York.
[30] Hernández-Lerma, O., J. González-Hernández. 2000. Constrained Markov control processes in Borel spaces: The discounted case. *Math. Methods Oper. Res.* **52** 271–285.
[31] Hernández-Lerma, O., J. B. Lasserre. 1996. *Discrete-Time Markov Control Processes. Basic Optimality Criteria*. Springer, New York.
[32] Hernández-Lerma, O., J. B. Lasserre. 1999. *Further Topics on Discrete-Time Markov Control Processes*. Springer, New York.
[33] Hordijk, A. 1977. *Dynamic Programming and Markov Potential Theory.* Mathematical Centre Tracts, 2nd ed., Vol. 51. Mathematisch Centrum, Amsterdam.
[34] Hordijk, A., F. M. Spieksma. 1990. Constrained admission control to a queueing system. *Adv. Appl. Probab.* **21** 409–431.
[35] Kallenberg, L. C. M. 1983. *Linear Programming and Finite Markov Control Problems.* Mathematical Centre Tracts, Vol. 148. Mathematisch Centrum, Amsterdam.
[36] Krylov, N. V. 1965. The construction of an optimal strategy for a finite controlled chain. *SIAM Theory Probab. Appl.* **10** 45–54.
[37] Krylov, N. V. 1985. Once more about the connection between elliptic operators and Itô's stochastic equations. N. V. Krylov, R. Sh. Liptser, A. A. Novikov, eds. *Statistics and Control of Stochastic Processes*. Optimization Software, New York, 69–101.
[38] Krylov, N. V. 1987. An approach in the theory of controlled diffusion processes. *SIAM Theory Probab. Appl.* **31** 604–626.
[39] Kuhn, H. W. 1953. Extensive games and the problem of information. H. W. Kuhn, W. W. Tucker, eds. *Contributions to the Theory of Games*, Vol. II. Princeton University Press, Princeton, NJ, 193–216.
[40] Nowak, A. S. 1988. On the weak topology in the space of probability measures induced by policies. *Bull. Polish Acad. Sci. Math.* **36** 181–186.
[41] Phelps, R. R. 2001. *Lectures on Choquet's Theorem.* Lecture Notes in Mathematics, 2nd ed., Vol. 1757. Springer, Berlin.
[42] Piunovskiy, A. B. 1997. *Optimal Control of Random Sequences in Problems with Constraints*. Kluwer Academic, Dordrecht, The Netherlands.

[43] Piunovskiy, A. B. 1998. Controlled random sequences: Methods of convex analysis and problems with functional constraints. *Russian Math. Surveys* **53** 1233–1293.
[44] Rudin, W. 1973. *Functional Analysis*. McGraw Hill, New York.
[45] Schäl, M. 1975. On dynamic programming: Compactness of the space of policies. *Stochastic Processes Their Appl.* **3** 345–364.
[46] Schäl, M. 1993. Average optimality in dynamic programming with general state space. *Math. Oper. Res.* **18** 163–172.
[47] Spieksma, F. M. 1990. Geometrically ergodic Markov chains and the optimal control of queues. Ph.D. thesis, Leiden University, Leiden, The Netherlands.
[48] Strassen, V. 1969. Gaussian elimination is not optimal. *Numer. Math.* **13** 354–356.