SIAM J. Control Optim.
Vol. 57, No. 1, pp. 163–191

SUFFICIENCY OF DETERMINISTIC POLICIES FOR ATOMLESS DISCOUNTED AND UNIFORMLY ABSORBING MDPs WITH MULTIPLE CRITERIA[*]

EUGENE A. FEINBERG[†] AND ALEXEY PIUNOVSKIY[‡]

Abstract. This paper studies Markov decision processes (MDPs) with atomless initial state distributions and atomless transition probabilities. Such MDPs are called atomless. The initial state distribution is considered to be fixed. We show that for discounted MDPs with bounded one-step reward vector-functions, for each policy there exists a deterministic (that is, nonrandomized and stationary) policy with the same performance vector. This fact is proved in the paper for a more general class of uniformly absorbing MDPs with expected total rewards, and then it is extended under certain assumptions to MDPs with unbounded rewards. For problems with multiple criteria and constraints, the results of this paper imply that for atomless MDPs studied in this paper it is sufficient to consider only deterministic policies, while without the atomless assumption it is well-known that randomized policies can outperform deterministic ones. We also provide an example of an MDP demonstrating that if a vector measure is defined on a standard Borel space, then Lyapunov's convexity theorem is a special case of the described results.

Key words. atomless, discounted, Markov decision process, deterministic policy, convex, compact

AMS subject classifications. 90C40, 93E20, 93E03

DOI. 10.1137/18M1194924

1. Introduction.

This paper studies Markov decision processes (MDPs) with multiple criteria when each criterion is evaluated by the expected total discounted rewards or costs. The paper also studies more general uniformly absorbing MDPs. The number of criteria is finite, and the initial state distribution is fixed. For each criterion there is a function of one-step rewards, and the performance of each policy is evaluated by the finite-dimensional vector, whose coordinates are expected total rewards for the corresponding reward functions. For each policy this vector is called a performance vector. An MDP is called atomless if the initial state distribution and transition probabilities are atomless. In general, constrained optimization requires the use of randomized decisions. However, for atomless problems nonrandomized policies are optimal under broad conditions.

The first results of this kind were established by Dvoretzky, Wald, and Wolfowitz [8, 9], who proved that for a one-step problem with multiple atomless initial distributions, multiple reward functions, and finite action sets, the expected reward vector achieved by an arbitrary policy can be achieved by a nonrandomized policy. The case of multiple initial distributions can be reduced to a single initial distribution by using the Radon–Nikodym theorem; see [20] or Example 11.2. Thus, the above-mentioned result from Dvoretzky, Wald, and Wolfowitz [8, 9] can be interpreted as a fact for one-step atomless MDPs. As observed by Feinberg and Piunovskiy [20], this result holds

[*]Received by the editors June 18, 2018; accepted for publication (in revised form) October 24, 2018; published electronically January 8, 2019.
http://www.siam.org/journals/sicon/57-1/M119492.html

Funding: The research of the first author was partially supported by National Science Foundation grant CMMI-1636193.

[†]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, 11794-3600 (eugene.feinberg@stonybrook.edu).

[‡]Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 7ZL, UK (piunov@liverpool.ac.uk).

for infinite action sets; see also Jaśkiewicz and Nowak [25] for the generalization to conditional expectations. The proof in Dvoretzky, Wald, and Wolfowitz [8, 9] is based on Lyapunov's convexity theorem, which states that the range of a finite atomless vector-measure is a convex compact subset of the Euclidean space.

Feinberg and Piunovskiy [18, 19] proved that for atomless MDPs with a given initial state distribution and with multiple expected total rewards, for every policy there is a nonrandomized Markov policy with the same performance vector. In [18] this fact was proved for MDPs with weakly continuous transition probabilities and with weakly continuous reward functions. The proof in [18] is based on geometric arguments. In [19] this fact was proved for arbitrary atomless MDPs with expected total rewards, and the proof was based on Lyapunov's convexity theorem.

In this paper we prove that for an atomless discounted MDP with multiple criteria and bounded reward functions, for each policy there exists a deterministic (that is, nonrandomized and stationary) policy with the same performance vector. In fact, we prove this result for uniformly absorbing MDPs with the expected total rewards. This is a more general class of MDPs than discounted ones. The proof for deterministic policies is much more difficult than the proofs for nonrandomized Markov ones provided in [18] and [19]. In addition, the proofs in this paper use and extend geometric methods introduced in [18] instead of applying Lyapunov's convexity theorem. Example 11.2 demonstrates that Lyapunov's convexity theorem can be interpreted as a one-step version of the main result of this paper.

For discounted MDPs with multiple criteria and constraints, under certain conditions there exist (randomized) stationary optimal policies; see Altman [1], Feinberg and Shwartz [22], Hernández-Lerma and González-Hernández [24], and Piunovskiy [29]. The results of this paper imply the existence of optimal deterministic policies for constrained atomless discounted MDPs and for constrained atomless uniformly absorbing MDPs if optimal policies exist.

The main result of this paper, Theorem 3.8, states that the sets of performance vectors for all policies and for deterministic policies coincide. In order to prove the main result, we deal with three types of subsets of linear spaces: the set of strategic measures, the set of occupancy measures, and the set of performance vectors. For a given policy, the strategic measure is the probability distribution of all state-action trajectories, and the occupancy measure is the measure on the product of the state and action spaces, where the value of this measure on each measurable set is the expected total number of times when the corresponding actions are selected at the corresponding states. The set of performance vectors (strategic measures, occupancy measures) consists of performance vectors (strategic measures, occupancy measures) for all policies. The set of performance vectors is a projection of the set of occupancy measures, and the set of occupancy measures is a projection of the set of strategic measures. Projections inherit certain properties of the sets from which they are projected. These properties include convexity and compactness.

The set of all strategic measures is convex; see Dynkin and Yushkevich [10, section 3.5]. Therefore, the set of all occupancy vectors and the set of all performance vectors are convex. Under certain conditions the set of strategic measures is compact. Schäl [31] introduced two such conditions: (S) and (W). Condition (S) assumes setwise continuity of transition probabilities, and condition (W) assumes weak continuity of transition probabilities. In both cases, appropriate continuity properties are assumed for reward functions. In particular, condition (S) holds for MDPs with finite action sets. Under the above-mentioned conditions, compactness properties also hold for the sets of all occupancy measures and all performance vectors.

For discounted and absorbing MDPs, if the initial distribution is fixed, then for each policy there exists a stationary policy with the same occupancy measure; see [1, 6, 21, 23, 29, 30]. Therefore, the sets of all occupancy measures and all performance vectors coincide with the corresponding sets for all stationary policies. The nontrivial step in proving Theorem 3.8 is to show that the sets of performance vectors for all stationary and for all deterministic policies coincide.

The important and nontrivial step is to prove that for an atomless MDP the set of performance vectors for all deterministic policies is convex. This fact is nontrivial even for the case of one criterion. Example 11.2 demonstrates that for multiple criteria this fact is a nontrivial extension of Lyapunov's convexity theorem for a standard Borel space. In order to prove this fact, we show that the set of occupancy measures endowed with the topology of setwise convergence is path-connected. Therefore, being its projection, the set of performance vectors is a connected subset of the Euclidean space. Thus, for the single-criterion case, this set is a connected subset of a line. Therefore, it is convex. The case of multiple criteria is studied by induction using the dimensionality reduction technique introduced in this paper.

Section 2 of this paper introduces the basic definitions for the discounted case and formulates the main result for discounted MDPs. Section 3 describes absorbing and uniformly absorbing MDPs, formulates the main result for uniformly absorbing MDPs, and shows that a discounted MDP is a particular case of a uniformly absorbing MDP. Section 4 studies the properties of occupancy measures. Section 5 describes condition (S), which is sufficient for compactness of the sets of all strategic measures, all occupancy measures, and all performance vectors. In particular, this condition holds for an MDP with finite action sets. Section 6 describes submodels and dimensionality reduction. Section 7 introduces an MDP generated by two deterministic policies and describes continuity properties for such MDPs. Section 8 establishes path-connectedness of the sets of occupancy measures for all deterministic policies for atomless MDPs. This property implies that the set of all performance vectors for deterministic policies is path-connected. Thus, for a single-criterion problem, this set is convex. The proof of the main theorem is provided in section 9. Section 10 presents the results for unbounded reward vector-functions by using the standard weighted norm approach. These results are used in section 11 to show that for standard Borel spaces Lyapunov's convexity theorem is a special case of the results of this paper.

2. Main result for discounted MDPs.

We start with some definitions. Recall that two measurable spaces $(E, \mathcal{E})$ and $(D, \mathcal{D})$ are called isomorphic if there exists a one-to-one measurable correspondence f between them such that the correspondence f^{-1} is measurable. A Polish space is a complete separable metrizable space. A standard Borel space is a measurable space isomorphic to a Borel subset of a Polish space. Properties of standard Borel spaces can be found in Bertsekas and Shreve [3], Dynkin and Yushkevich [10], Kechris [26], and Srivastava [32]. In particular, a standard Borel space is either finite or countable, or it has the cardinality of the continuum. Two standard Borel spaces with the same cardinality are isomorphic. We always consider Borel σ-fields on topological and metric spaces. In particular, a standard Borel space with the cardinality of the continuum is isomorphic to the interval $[0, 1]$. For two measurable spaces $(E, \mathcal{E})$ and $(D, \mathcal{D})$, a transition probability q defines a probability measure $q(\cdot|d)$ on $(E, \mathcal{E})$ for each $d \in D$ such that $q(C|\cdot)$ is a measurable function on $(D, \mathcal{D})$ for each $C \in \mathcal{E}$. We recall that a measure ν on a standard Borel $(D, \mathcal{D})$ space is called atomless if $\nu(d) = 0$ for all $d \in D$; here and below we omit curly brackets in the expressions like $\nu(\{x\})$ and $p(\{y\}|x, a)$.

A discounted MDP is defined by the following objects:

(i) a standard Borel state space $(\mathbb{X}, \mathcal{X})$,

(ii) a standard Borel action space $(\mathbb{A}, \mathcal{A})$,

(iii) nonempty sets of actions $A(x) \in \mathcal{A}$ available at states $x \in \mathbb{X}$, such that $\mathrm{Gr}_{\mathbb{X}}(A) := \{(x,a) \in \mathbb{X} \times \mathbb{A} : x \in \mathbb{X},\ a \in A(x)\}$ is a measurable subset of $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \otimes \mathcal{A})$,

(iv) a transition probability p from $\mathbb{X} \times \mathbb{A}$ to $\mathbb{X}$,

(v) an initial state distribution μ, which is a probability measure on $(\mathbb{X}, \mathcal{X})$,

(vi) a bounded measurable reward vector-function $r : \mathbb{X} \times \mathbb{A} \mapsto \mathbb{R}^N$, where N is a natural number,

(vii) a discount factor $\beta \in [0, 1)$.

Definition 2.1. *An MDP is called atomless if $\mu(x) = 0$ and $p(y|x,a) = 0$ for all $x, y \in \mathbb{X}$ and $a \in A(x)$.*

If an action $a \in A(x)$ is chosen at a state $x \in \mathbb{X}$, then the process moves to the next state according to the probability distribution $p(\cdot|x,a)$ and the reward vector $r(x,a) = (r^{(1)}(x,a), r^{(2)}(x,a), \ldots, r^{(N)}(x,a))$ is collected according to criteria $1, 2, \ldots, N$. To avoid a trivial situation, when a policy cannot be defined, we always assume that there exists a measurable mapping $\phi : \mathbb{X} \mapsto A$ such that $\phi(x) \in A(x)$ for all $x \in \mathbb{X}$. Such a mapping is called a selector.

Consider the sets of possible finite histories $\mathbb{H}_t := \mathbb{X} \times (\mathbb{A} \times \mathbb{X})^t$ up to time $t = 0, 1, \ldots$. A policy π is a sequence of transition probabilities π_t, $t = 0, 1, \ldots$, from $\mathbb{H}_t$ to A such that $\pi(A(x_t)|h_t) = 1$ for each $h_t = (x_0, a_0, x_1, \ldots, x_t) \in \mathbb{H}_t$. A policy is called nonrandomized if each transition probability $\pi_t(\cdot|h_t)$, $t = 0, 1, \ldots$, is concentrated at one point. A policy π is called Markov if for each $t = 1, 2, \ldots$ the values of probabilities $\pi_t(\cdot|x_0, a_0, \ldots, x_t)$ are the functions of x_t. A Markov policy is called stationary if $\pi_t(\cdot|x) = \pi_s(\cdot|x)$ for all $x \in \mathbb{X}$ and for all $s, t = 0, 1, \ldots$. A transition probability π_t for a stationary policy π is also denoted as π. A nonrandomized Markov policy is defined by a sequence of selectors $\{\phi_t\}_{t=0,1,\ldots}$. These selectors are equal for a nonrandomized stationary policy. A nonrandomized stationary policy ϕ is called deterministic, and we identify it with the selector ϕ. We denote by Π, $\mathbb{M}$, $\mathbb{S}$, and $\mathbb{F}$ the sets of all, nonrandomized Markov, stationary, and deterministic policies, respectively. Observe that $\mathbb{F} \subset \mathbb{M} \subset \Pi$ and $\mathbb{F} \subset \mathbb{S} \subset \Pi$.

The existence of the selector means that $\mathbb{F} \neq \emptyset$. This assumption does not limit the generality of the results of this paper. If $\mathbb{F} = \emptyset$, then $\Pi = \emptyset$; see Dynkin and Yushkevich [10, sections 3.1 and 3.2]. Therefore, if $\mathbb{F} = \emptyset$, then the main result of the paper, Theorem 3.8, is equivalent to the trivial identity $\emptyset = \emptyset$.

The two special features of the introduced model are that (i) the rewards are vector-valued, and (ii) the initial distribution μ is fixed. However, we consider additional initial distributions and initial states in auxiliary results in a few places in this paper. Whenever we consider initial distributions other than μ, we specify them in the notation.

According to the Ionescu Tulcea theorem, an initial probability distribution μ on the state space $\mathbb{X}$ and transition probabilities π_t and p define a unique probability measure P^π on the countable product $\mathbb{H}_\infty := \mathbb{X} \times (\mathbb{A} \times \mathbb{X})^\infty$ endowed with the σ-field $\mathcal{X} \otimes (\mathcal{A} \otimes \mathcal{X})^\infty$. Expectations with respect to this probability are denoted by E^π.

Remark 2.2. The corresponding probabilities and expectations are defined for each initial probability distribution ν on $(\mathbb{X}, \mathcal{X})$. In this case, they are denoted as P^π_ν and E^π_ν. That is, $P^\pi := P^\pi_\mu$ and $E^\pi := E^\pi_\mu$. If a probability measure ν is concentrated at a point $x \in \mathbb{X}$, that is, $\nu(x) = 1$, we shall write P^π_x and E^π_x instead of P^π_ν and E^π_ν, respectively.

For an initial state distribution μ and a policy π, the vector of expected total discounted rewards is

$$v_\beta^\pi := E^\pi \sum_{t=0}^{\infty} \beta^t r(x_t, a_t).$$

For a set of policies $\Delta \subset \Pi$, the set of all performance vectors is $\mathcal{V}_\beta^\Delta := \{v_\beta^\pi : \pi \in \Delta\}$.

Denote $\mathcal{V}_\beta := \mathcal{V}_\beta^\Pi$. It is obvious that $\mathcal{V}_\beta^{\mathbb{F}} \subset \mathcal{V}_\beta \subset \mathbb{R}^N$, and, in general, it is possible that $\mathcal{V}_\beta^{\mathbb{F}} \neq \mathcal{V}_\beta$. For example, if $\mathbb{X}$ and $\mathbb{A}$ are finite sets, then the set $\mathcal{V}_\beta^{\mathbb{F}}$ is finite, while the set $\mathcal{V}_\beta$ may have the cardinality of the continuum. In fact, for problems with finite state and action sets, $\mathcal{V}_\beta$ is a convex hull of $\mathcal{V}_\beta^{\mathbb{F}}$; see, e.g., Feinberg and Rothblum [21, Theorem 6.1]. According to the following theorem, which is the main result of this paper for discounted MDPs, the situation is different for atomless MDPs.

THEOREM 2.3. *For an atomless MDP* $\mathcal{V}_\beta^{\mathbb{F}} = \mathcal{V}_\beta$.

In section 3 we formulate a more general result, which is proved later in this paper.

3. Absorbing MDPs and the main result.

We start this section with the definition of the expected total reward under fairly general conditions and for the case of a single criterion, that is, $N = 1$. In this case, r is a bounded real-valued function, but in formula (3.1) and in Definition 3.1 we do not assume that r is bounded. Then we define absorbing and uniformly absorbing MDPs, formulate the main result of this paper, Theorem 3.8, and show that it is more general than Theorem 2.3, which states the sufficiency of deterministic policies for atomless discounted MDPs.

We recall that the initial state distribution μ is fixed. For an arbitrary nonnegative measurable function r, the expected total reward for a policy π is

$$v^\pi := E^\pi \sum_{t=0}^{\infty} r(x_t, a_t) = \lim_{n\to\infty} E^\pi \sum_{t=0}^{n-1} r(x_t, a_t), \tag{3.1}$$

where the second equality follows from the monotone convergence theorem.

For a number c, let us denote $c^+ := \max\{c, 0\}$ and $c^- := -\min\{c, 0\}$. For a policy $\pi \in \Pi$, we consider positive values v_+^π and v_-^π defined by (3.1) with the rewards $r(x, a)$ replaced by the rewards $r^+(x, a)$ and $r^-(x, a)$, respectively.

DEFINITION 3.1. *If* $\min\{v_+^\pi, v_-^\pi\} < +\infty$, *then the expected total reward* v^π *is well-defined and* $v^\pi := v_+^\pi - v_-^\pi$.

If v^π is well-defined, then the equalities in (3.1) hold because they hold for rewards r^+ and r^- and at least one of the numbers v_+^π and v_-^π is finite.

Now let $N > 1$. Then v_+^π and v_-^π are defined as N-dimensional vectors of the expected total rewards whose coordinates are the expected total rewards for positive and negative parts of the corresponding coordinates of the vector-function r. The vector v^π is well-defined if each of its N coordinates is. In this case, as explained above, $v^\pi := v_+^\pi - v_-^\pi$, and the second equality in (3.1) holds.

Remark 3.2. For an initial probability distribution ν on $(\mathbb{X}, \mathcal{X})$, which can be different from μ, we shall use the notation $v(\nu)$, $v_+(\nu)$, and $v_-(\nu)$, respectively. With a small abuse of notation, we shall write $v(x)$, $v_+(x)$, and $v_-(x)$, respectively, if the probability measure ν is concentrated at the point $x \in \mathbb{X}$.

Now we introduce an absorbing MDP. Let the standard Borel state space of this MDP be denoted by $\bar{\mathbb{X}}$. We use the same notation and assumptions for the standard

Borel action space $\mathbb{A}$, sets of available actions $A(\cdot)$, transition probability p, initial state distribution μ, and reward vector r as in the previous section.

Let T^x denote the first time a stochastic sequence $h = x_0, x_1, \ldots$ with values in $\bar{\mathbb{X}}$ reaches the state $x \in \bar{\mathbb{X}}$, that is, $T^x(h) := \inf\{t = 0, 1, \ldots : x_t = x\}$.

DEFINITION 3.3. *For the initial probability distribution μ, an MDP is called absorbing if there exists a state $\bar{x} \in \bar{\mathbb{X}}$ with the following properties:*

(i) $\mu(\bar{x}) = 0$;

(ii) $A(\bar{x}) = \{\bar{a}\}$ *for some* $\bar{a} \in \mathbb{A}$, $p(\bar{x}|\bar{x}, \bar{a}) = 1$, *and* $r^{(i)}(\bar{x}, \bar{a}) = 0$ *for all* $i = 1, \ldots, N$;

(iii) *there exists a finite constant L such that, for all policies $\pi \in \Pi$,*

$$E^\pi T^{\bar{x}} \leq L. \tag{3.2}$$

Remark 3.4. The state $\bar{x}$ is fictitious in the sense that under every policy this state is absorbing, there is no choice of decisions at $\bar{x}$, and all the rewards are equal to 0 at this state. After the system hits state $\bar{x}$, it is impossible to control it. Therefore, the set $\bar{\mathbb{X}} \setminus \{\bar{x}\}$ plays the same role for absorbing MDPs as the state space $\mathbb{X}$ for discounted MDPs; see the notation in formula (3.3).

Remark 3.5. We make assumption (i) in Definition 3.3 for convenience only. All the results in this paper hold without this assumption. In principle, it is possible to consider initial distributions other than μ. If an MDP is absorbing for an initial distribution ν, which may differ from μ, then this is stated explicitly in this paper. Of course, the value of the upper bound L may depend on the initial distribution. In some publications, including [1, 21], absorbing measurable sets are considered instead of absorbing states. These formulations are equivalent because the states in an absorbing set can be merged into a single state.

Observe that $T^{\bar{x}} = \sum_{t=0}^{\infty} I\{t < T^{\bar{x}}\}$, where I is the indicator function. We recall that assumption (iii) in Definition 3.3 is equivalent to the validity of (3.2) for all deterministic policies $\phi \in \mathbb{F}$ instead of arbitrary policies $\pi \in \Pi$; see Feinberg and Rothblum [21, p. 132]. If we interpret $T^{\bar{x}}$ as the time when the process stops, then (3.2) means that the average lifetime of the process is uniformly bounded for all policies given the initial state distribution μ. For an absorbing MDP, we fix an arbitrary state $\bar{x}$ described in Definition 3.3 and set

$$\mathbb{X} := \bar{\mathbb{X}} \setminus \{\bar{x}\}. \tag{3.3}$$

Let us consider an absorbing MDP. Recall that the reward vector-function r is bounded and $r(\bar{x}, \bar{a}) = 0$. In view of Definition 3.3(ii), (iii), the expected total rewards v^π are well-defined for all policies π, and

$$\begin{aligned} v^\pi &= \lim_{n\to\infty} E^\pi \sum_{t=0}^{n-1} r(x_t, a_t) = E^\pi \sum_{t=0}^{\infty} r(x_t, a_t) = E^\pi \sum_{t=0}^{\infty} r(x_t, a_t) I\{x_t \in \mathbb{X}\} \\ &= E^\pi \sum_{t=0}^{T^{\bar{x}}-1} r(x_t, a_t), \end{aligned} \tag{3.4}$$

where the first two equalities follow from (3.1) and the last two follow from Definition 3.3(ii). For $\Delta \subset \Pi$, the set of performance vectors generated by policies from Δ is $\mathcal{V}^\Delta := \{v^\pi : \pi \in \Delta\}$. We also use the notation

$$\mathcal{V} := \mathcal{V}^\Pi.$$

For an absorbing MDP, the monotone convergence theorem implies that for every policy π

$$\lim_{n\to\infty} E^{\pi} \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = 0.$$

Definition 3.6 states the stronger equality. Recall that $\mathbb{M}$ is the set of all nonrandomized Markov policies and the initial measure μ is fixed.

Definition 3.6. *An absorbing MDP is called uniformly absorbing if*

$$\lim_{n\to\infty} \sup_{\pi\in\mathbb{M}} E^{\pi} \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = 0. \tag{3.5}$$

Example 3.13 describes an absorbing MDP which is not uniformly absorbing. We remark that the supremum in (3.5) is equal to the same supremum over the set of all policies $\pi \in \Pi$; see Feinberg [12, Theorem 3]. Recall that $E^{\pi} I\{t < T^{\bar{x}}\} = P^{\pi}\{T^{\bar{x}} > t\}$ and $E^{\pi} T^{\bar{x}} = \sum_{t=0}^{\infty} P^{\pi}\{T^{\bar{x}} > t\}$. Since $E^{\pi} \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = E^{\pi} T^{\bar{x}} - E^{\pi} \sum_{t=0}^{n-1} I\{t < T^{\bar{x}}\}$, assumption (3.5) means that the MDP is absorbing and the convergence $E^{\pi} \sum_{t=0}^{n-1} I\{t < T^{\bar{x}}\} \uparrow E^{\pi} T^{\bar{x}}$ as $n \to \infty$ takes place uniformly in $\pi \in \Pi$. Since the vector-function r is bounded, the convergence in (3.1) is uniform in $\pi \in \Pi$ for a uniformly absorbing MDP.

Definition 3.7. *An absorbing MDP is called atomless if* $\mu(x) = 0$ *and* $p(y|x,a) = 0$ *for all* $x, y \in \mathbb{X}$ *and* $a \in A(x)$.

In some sense, Definition 3.7 means that the state $\bar{x}$ is considered to be outside of the state space. Of course, a uniformly absorbing MDP is absorbing, and Definition 3.7 applies to uniformly absorbing MDPs, too.

As explained later in this section, the following theorem, which is the main result of this paper, generalizes Theorem 2.3, which states a similar statement for discounted MDPs.

Theorem 3.8. *For a uniformly absorbing atomless MDP,* $\mathcal{V}^{\mathbb{F}} = \mathcal{V}$.

The following corollary is an equivalent formulation of Theorem 3.8.

Corollary 3.9. *For a uniformly absorbing atomless MDP, for every policy* $\pi \in \Pi$ *there exists a deterministic policy* ϕ *such that* $v^{\phi} = v^{\pi}$.

For total-reward MDPs, the performance set $\mathcal{V}$ is convex. This simple fact follows from the convexity of the set of strategic measures; see Dynkin and Yushkevich [10, section 5.5] or, for absorbing MDPs, see Lemma 4.1 below. This fact and Theorem 3.8 imply the following corollary.

Corollary 3.10. *For a uniformly absorbing atomless MDP, the set* $\mathcal{V}^{\mathbb{F}}$ *is convex.*

Let us show that Theorem 3.8 is more general than Theorem 2.3. Recall that if an initial probability distribution ν is concentrated at one state $x \in \mathbb{X}$, then, according to Remark 2.2, we usually write E_x^{π} instead of E_{ν}^{π}. The following lemma provides a natural sufficient condition under which an absorbing MDP is uniformly absorbing.

Lemma 3.11. *Consider an MDP with a standard Borel state space* $\bar{\mathbb{X}}$ *and with a state* $\bar{x} \in \bar{\mathbb{X}}$ *such that* $A(\bar{x})$ *is a singleton and* $p(\bar{x}|\bar{x},\bar{a}) = 1$, $r(\bar{x},\bar{a}) = 0$, *where* $A(\bar{x}) = \{\bar{a}\}$. *If there is a finite constant* L *such that* $E_x^{\phi} T^{\bar{x}} < L$ *for all* $x \in \mathbb{X} = \bar{\mathbb{X}} \setminus \{\bar{x}\}$ *and for all* $\phi \in \mathbb{F}$, *then this MDP is uniformly absorbing for all initial state distributions* μ *on* $\mathbb{X}$.

Proof. Let us fix an arbitrary initial probability distribution μ on $\mathbb{X}$. As is mentioned after Definition 3.3, $\sup_{\pi\in\Pi} E_x^\pi T^{\bar{x}} = \sup_{\phi\in\mathbb{F}} E_x^\phi T^{\bar{x}}$ for all $x \in \mathbb{X}$. Therefore, $E_x^\pi T^{\bar{x}} \le L$ for all $x \in \mathbb{X}$ and for all $\pi \in \Pi$. This implies that $E^\pi T^{\bar{x}} \le L$ for all $\pi \in \Pi$. In view of Markov's inequality, for an arbitrary policy $\pi \in \Pi$ and for $n = 0, 1, \ldots,$

$$P^\pi\{T^{\bar{x}} > n\} \le (n+1)^{-1} E^\pi T^{\bar{x}} \le (n+1)^{-1} L. \tag{3.6}$$

For an arbitrary nonrandomized Markov policy $\phi = (\phi_0, \phi_1, \ldots)$ and for $n = 0, 1, \ldots,$ let us define by ϕ^{+n} the shifted nonrandomized Markov policy $\phi^{+n} = (\phi^n, \phi^{n+1}, \ldots)$. Then

$$E^\phi \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = E^\phi E_{x_n}^{\phi^{+n}} \sum_{t=0}^{\infty} I\{t < T^{\bar{x}}\} = E^\phi E_{x_n}^{\phi^{+n}} T^{\bar{x}} \le E^\phi I\{n < T^{\bar{x}}\} L$$
$$= L P^\phi\{T^{\bar{x}} > n\} \le (n+1)^{-1} L^2,$$

which implies (3.5), where the first inequality follows from $\{x_n \in \mathbb{X}\} = \{n < T^{\bar{x}}\}$ P^π-a.s. and $E_x^\pi T^{\bar{x}} \le L$ for all $\pi \in \Pi$ and all $x \in \mathbb{X}$, and the last inequality follows from (3.6). □

LEMMA 3.12. *Theorem* 3.8 *implies Theorem* 2.3.

Proof. Consider a discounted MDP. The following transformation into an absorbing MDP is well-known; see, e.g., Altman [1, p. 137]. Let us add an additional point $\bar{x}$ to the state space $\mathbb{X}$ and consider the new transition probability $\bar{p}$ defined by

$$\bar{p}(Y|x,a) := \begin{cases} \beta p(Y|x,a) & \text{if } x \in \mathbb{X},\ Y \in \mathcal{X}, \\ 1-\beta & \text{if } x \in \mathbb{X},\ Y = \{\bar{x}\}, \\ 1 & \text{if } x = \bar{x} \in Y. \end{cases}$$

Then $\mathcal{V}^{\mathbb{F}} = \mathcal{V}_\beta^{\mathbb{F}}$ and $\mathcal{V} = \mathcal{V}_\beta$. The new MDP is absorbing. It is atomless if and only if the original discounted MDP is atomless. Since $E_x^\pi T^{\bar{x}} = (1-\beta)^{-1}$, Lemma 3.11 implies that the new model is uniformly absorbing. □

Of course, the transformation of a discounted MDP into an absorbing one is trivial. However, under certain conditions it is also possible to transform an absorbing MDP into a discounted one; see Feinberg and Huang [14, 15].

The following example describes an absorbing MDP, which is not uniformly absorbing.

Example 3.13. Let $\mathbb{X} := \{(i,j) : i = 0, 1, \ldots,\ j = 0, 1, \ldots, 2^i - 1\}$, $\bar{x} := 0$, A={c,s}, where c stands for "continue" and s stands for "stop," and

$$A(x) := \begin{cases} \{c,s\} & \text{if } x = (i,0),\ i = 0, 1, \ldots, \\ \{s\} & \text{otherwise,} \end{cases}$$

and for $i = 0, 1, \ldots$

$$p(y|x,a) = \begin{cases} 0.5 & \text{if } a = c,\ x = (i,0),\ y = 0,\ \text{or } y = (i+1,0), \\ 1 & \text{if } a = s \text{ and either } x = (i,j),\ j = 0, \ldots, 2^i - 2,\ y = (i, j+1) \\ & \text{or } x = (i, 2^i - 1),\ y = 0. \end{cases}$$

In addition, $\mu(0,0) = 1$. In this example, the process starts at the state $(0,0)$. At each state $(i,0)$, $i = 0, 1 \ldots,$ the decision maker can either continue or stop the process.

If the process is continued at state $(i, 0)$, then it moves with probability 0.5 either to state $(i+1, 0)$ or to state $\bar{x}$. If the process is stopped at state $(i, 0)$, then it makes 2^i additional deterministic moves until it hits the absorbing state $\bar{x} = 0$ and stops. Let ϕ^∞ be the deterministic policy that always chooses an action c at the states $(i, 0)$, $i = 0, 1, \ldots.$ Under this policy, $T^{\bar{x}}$ has the geometric distribution with the success probability 0.5 at each step. Therefore, $E^{\phi^\infty} T^{\bar{x}} = 2$. Now let ϕ^n be a deterministic policy choosing the action s at the state $(n, 0)$ and the action c at the states $(i, 0)$ with $i = 0, 1, \ldots, n-1$, where $n = 0, 1, \ldots.$ Then $E^{\phi^n} T^{\bar{x}} = E^{\phi^n} [\sum_{t=0}^{n-1} I\{t < T^{\bar{x}}\} + 2^n I\{t < T^{\bar{x}}\}] = \sum_{t=0}^{n-1} 2^{-t} + 2^n 2^{-n} = 3 - 2^{-n+1}$. Thus, $E^\phi T^{\bar{x}} \le 3$ for all $\phi \in \mathbb{F}$. So, this MDP is absorbing. However, $\lim_{n\to\infty} \sup_{\pi\in\mathbb{M}} E^\pi \sum_{t=n}^\infty I\{t < T^{\bar{x}}\} \ge \lim_{n\to\infty} E^{\phi^n} \sum_{t=n}^\infty I\{t < T^{\bar{x}}\} = \lim_{n\to\infty} 2^{-n} 2^n = 1$. Thus, this MDP is not uniformly absorbing.

4. Occupancy measures and their properties.

For an absorbing MDP, a policy π, and an initial state distribution μ on $\mathbb{X}$, the finite occupancy measure $Q^\pi(\cdot)$ on $\mathbb{X} \times \mathbb{A}$ is defined by

$$Q^\pi(Y \times B) := E^\pi \sum_{t=0}^{T^{\bar{x}}-1} I\{x_t \in Y, a_t \in B\} = \sum_{t=0}^\infty P^\pi\{x_t \in Y, a_t \in B\}, \quad Y \in \mathcal{X},\ B \in \mathcal{A}.$$

Let $q^\pi(Y) := Q^\pi(Y \times \mathbb{A})$, where $Y \in \mathcal{X}$. Observe that $q^\pi(\mathbb{X}) = E^\pi T^{\bar{x}} \le L$. In addition,

$$v^\pi = \int_{\mathbb{X}} \int_A r(x, a) Q^\pi(dxda). \tag{4.1}$$

The set of occupancy measures for the initial distribution μ and for all policies from $\Delta \subset \Pi$ is

$$\mathcal{M}^\Delta := \{Q^\pi(\cdot) : \pi \in \Delta\}.$$

We set $\mathcal{M} := \mathcal{M}^\Pi$. For an arbitrary policy π there exists a stationary policy $\sigma \in \mathbb{S}$ such that

$$Q^\pi(Y \times B) = \int_Y \sigma(B|x) q^\pi(dx), \qquad Y \in \mathcal{X},\ B \in \mathcal{A}, \tag{4.2}$$

and (4.2) implies that

$$Q^\sigma(\cdot) = Q^\pi(\cdot); \tag{4.3}$$

see [21, Lemmas 4.1 and 4.2]. Therefore,

$$\mathcal{M}^{\mathbb{S}} = \mathcal{M}, \tag{4.4}$$

and this set is convex; see [21, Corollary 4.3]. These properties imply the corresponding properties of performance sets stated in the following lemma. Recall that the initial distribution μ is fixed.

Lemma 4.1. *For an absorbing MDP the equality $\mathcal{V}^{\mathbb{S}} = \mathcal{V}$ holds, and this set is convex.*

Proof. The lemma follows from (4.1), (4.4), and the convexity of $\mathcal{M}$. ∎

For an absorbing MDP with the initial state distribution μ, for $\pi \in \Pi$, and for $Y \in \mathcal{X}$, define

$$q_n^\pi(Y) := P^\pi\{x_n \in Y\}, \qquad n = 0, 1, \ldots.$$

Then

$$q^\pi(Y) := \sum_{n=0}^{\infty} E^\pi I\{x_n \in Y\} = \sum_{n=0}^{\infty} P^\pi(x_n \in Y) = \sum_{n=0}^{\infty} q_n^\pi(Y). \tag{4.5}$$

We observe that $q_0^\pi(Y) = \mu(Y)$ and

$$q_n^\pi(Y) = \int_{\mathbb{X}} P_x^\pi\{x_n \in Y\}\mu(dx), \quad Y \in \mathcal{X},\ \pi \in \Pi,\ n = 0, 1, \dots.$$

In particular, $q^\pi(Y) = 0$ if and only if $q_n^\pi(Y) = 0$ for all $n = 0, 1, \dots,$ $Y \in \mathcal{X}$. This implies that $q^\sigma \ll q^\pi$ for policies π and σ, if $q_n^\sigma \ll q_n^\pi$ for all $n = 0, 1, \dots,$ where the symbol $\ll$ means absolute continuity.

Observe that for a stationary policy $\pi \in \mathbb{S}$, $n = 0, 1, \dots,$ and $Y \in \mathcal{X}$,

$$q_{n+1}^\pi(Y) = \int_{\mathbb{X}} \int_{\mathbb{A}} p(Y|x,a)\pi(da|x) q_n^\pi(dx). \tag{4.6}$$

Formulae (4.5) and (4.6) imply that for $\pi \in \mathbb{S}$

$$q^\pi(Y) = \mu(Y) + \int_{\mathbb{X}} \int_{\mathbb{A}} p(Y|x,a)\pi(da|x) q^\pi(dx). \tag{4.7}$$

LEMMA 4.2. *For two stationary policies π and σ, if $\sigma(\cdot|x) \ll \pi(\cdot|x)$ for all $x \in \mathbb{X}$, then $q_n^\sigma \ll q_n^\pi$ for all $n = 0, 1, \dots,$ and therefore $q^\sigma \ll q^\pi$.*

Proof. For $n = 0$ the statement is obvious since $q_0^\pi = q_0^\sigma = \mu$. Assume that $q_n^\sigma \ll q_n^\pi$ for some $n = 0, 1, \dots$. Consider a measurable subset Y of $\mathbb{X}$ such that $q_{n+1}^\pi(Y) = 0$. In view of (4.6), this means that

$$\int_{\mathbb{A}} p(Y|x,a)\pi(da|x) = 0 \qquad (q_n^\pi\text{-a.e.}).$$

Since $\sigma(\cdot|x) \ll \pi(\cdot|x)$ for all $x \in \mathbb{X}$, as follows from the last equality,

$$\int_{\mathbb{A}} p(Y|x,a)\sigma(da|x) = 0 \qquad (q_n^\pi\text{-a.e.}).$$

Since the integral in the left-hand part of the last equation is nonnegative and $q_n^\sigma \ll q_n^\pi$,

$$\int_{\mathbb{A}} p(Y|x,a)\sigma(da|x) = 0 \qquad (q_n^\sigma\text{-a.e.}),$$

which yields

$$q_{n+1}^\sigma(Y) = \int_{\mathbb{X}} \int_{\mathbb{A}} p(Y|x,a)\sigma(da|x) q_n^\sigma(dx) = 0.$$

Thus $q_n^\sigma \ll q_n^\pi$ for all $n = 0, 1, \dots,$ which implies $q^\sigma \ll q^\pi$, as explained before formula (4.6). ◻

LEMMA 4.3. *For an atomless absorbing MDP, every occupancy measure $q^\pi(dx)$, where $\pi \in \Pi$, is atomless.*

Proof. In view of (4.4), it is sufficient to prove the lemma for stationary policies π. Let $\pi \in \mathbb{S}$. Then $q_0^\pi = \mu$ is an atomless measure. If q_n^π is atomless for some $n = 0, 1, \dots,$ then formula (4.6) implies that the measure q_{n+1}^π is atomless. Thus, all the measures q_n^π, $n = 0, 1, \dots,$ are atomless. Formula (4.5) implies that q^π is atomless. ◻

The following theorem implies that $\mathcal{M}^{\mathbb{S}} = \mathcal{M}$ and $\mathcal{V}^{\mathbb{S}} = \mathcal{V}$ for an absorbing MDP. For discounted MDPs this result was discovered by Borkar [6]; see Borkar [7] and Piunovskiy [30] for additional references.

Theorem 4.4 (see Feinberg and Rothblum [21, Lemma 4.2]). *Let π be an arbitrary policy for an absorbing MDP. Consider a stationary policy σ such that $\sigma(B|x) = \frac{Q^\pi(dx,B)}{Q^\pi(dx,\mathbb{A})}$ for each $B \in \mathcal{A}$. Then the measures Q^σ and Q^π coincide, and therefore $v^\sigma = v^\pi$.*

5. Sufficient conditions for compactness of performance sets.

We start this section with formulating sufficient conditions for compactness of the set of strategic measures $\mathcal{S} := \{P^\pi : \pi \in \Pi\}$ defined on the set of all trajectories $\mathbb{H}_\infty$ for the given initial distribution μ. Since $\mathbb{H}_\infty$ is a countable product of standard Borel spaces, it is a standard Borel space. Let $\mathcal{P}(\mathbb{H}_\infty)$ be the set of all probability measures on $\mathbb{H}_\infty$. If $\mathbb{A}$ is a Borel subset of a Polish space, let us consider the ws^∞-topology on $\mathcal{P}(\mathbb{H}_\infty)$, which is the coarsest topology in which all the mappings $P \mapsto \int f(x_0, a_0, x_1, \dots, x_t)P(dx_0 da_0 dx_1, \dots, dx_t)$ are continuous for all bounded Borel functions $f : \mathbb{H}_t \mapsto \mathbb{R}$, which are continuous in $(a_0, a_1, \dots, a_t)$, $t = 0, 1, \dots$. Let us consider the following version of a condition introduced by Schäl [31].

Condition (S).

(S1) The set $\mathbb{A}$ is a Borel subset of a Polish space, and the sets $A(x)$ are compact for all $x \in \mathbb{X}$.

(S2) The transition probability $p(\cdot|x, a)$ is setwise continuous in $a \in A(x)$; that is, for each bounded Borel function $f : \mathbb{X} \mapsto \mathbb{R}$ and for each $x \in \mathbb{X}$, the function $a \mapsto \int_{\mathbb{X}} f(y)p(dy|x, a)$ is continuous on $A(x)$.

(S3) For each $x \in \mathbb{X}$ and $i = 1, \dots, N$, the reward function $r^{(i)}(x, a)$ is continuous in $a \in A(x)$.

Theorem 5.1 (see Balder [2], Nowak [28], and Schäl [31]). *If assumptions* (S1) *and* (S2) *hold, then the set of strategic measures $\mathcal{S} = \{P^\pi : \pi \in \Pi\}$ is a compact subset of $\mathcal{P}(\mathbb{H}_\infty)$ endowed with the ws^∞-topology.*

Corollary 5.2. *Consider a uniformly absorbing MDP. If Condition* (S) *holds, then the performance set $\mathcal{V}$ is compact.*

Proof. Let the ws^∞-topology be fixed on $\mathcal{P}(\mathbb{H}_\infty)$. Since $\mathcal{V} = V(\mathcal{S})$, where $V : \mathcal{S} \mapsto \mathbb{R}^N$ with $V(P^\pi) := v^\pi$ for all $\pi \in \Pi$, the corollary follows from the continuity of V, which is established in the rest of this proof.

Let us set $r^{(i)}(\bar{x}, \bar{a}) = 0$ for all $i = 1, \dots, N$. This change affects neither the values of v^π nor the validity of (S3). Let $v^{(i),\pi}$ be the ith coordinate of the performance vector v^π, $i = 1, 2, \dots, N$,

$$v^{(i),\pi} = E^\pi \sum_{t=0}^{T^{\bar{x}}-1} r^{(i)}(x_t, a_t) = E^\pi \sum_{t=0}^{\infty} r^{(i)}(x_t, a_t),$$

where the second equality holds because the state $\bar{x}$ is absorbing and $r^{(i)}(\bar{x}, \bar{a}) = 0$. Define

$$v_n^{(i),\pi} := E^\pi \sum_{t=0}^{n-1} r^{(i)}(x_t, a_t), \qquad n = 1, 2, \dots.$$

Since the MDP is uniformly absorbing, $v_n^{(i),\pi} \to v^{(i),\pi}$ uniformly in π as $n \to \infty$.

According to Yushkevich [33, Theorem 2], each function $r^{(i)}$, $i = 1, \dots, N$, can be extended from $\mathrm{Gr}_{\mathbb{X}}(A)$ to $\mathbb{X} \times \mathbb{A}$ in such a way that the extension is a bounded measurable function which is continuous in $a \in \mathbb{A}$. By the definition of the ws^∞-topology, the functions $V_n^{(i)}(P^\pi) := v_n^{(i),\pi}$ are continuous on $\mathcal{S}$. Let $V^{(i)}(P^\pi)$ denote the ith coordinate of the vector $V(P^\pi)$. Since $V_n^{(i)}(P) \to V^{(i)}(P)$ uniformly for all $P \in \mathcal{S}$ and for all $i = 1, \ldots, N$, the mapping V is continuous. □

COROLLARY 5.3. *Consider a uniformly absorbing MDP. If each set $A(x), x \in \mathbb{X}$, is finite, then the performance set $\mathcal{V}$ is compact.*

Proof. If $\mathbb{A}$ is a Borel subset of a Polish space, then the conclusion of the corollary follows from Corollary 5.2 since Condition (S) holds. The corollary follows from this fact since a standard Borel space is isomorphic to a Borel subset of a Polish space. Indeed, let $\tilde{\mathbb{A}}$ be a Borel subset of a Polish space isomorphic to $\mathbb{A}$, and let $g : \tilde{\mathbb{A}} \mapsto \mathbb{A}$ be the corresponding isomorphism. Let us consider the MDP with the state space $\mathbb{X}$, the action space $\mathbb{A}$ replaced with the isomorphic set $\tilde{\mathbb{A}}$, the sets of available actions $\tilde{A}(x) := g^{-1}(A(x))$, one-step reward vectors $\tilde{r}(x,a) := r(x, g(a))$, and transition probabilities $\tilde{p}(\cdot|x,a) = p(\cdot|x, g(a))$, where $x \in \mathbb{X}$ and $a \in \tilde{\mathbb{A}}$. The performance sets $\mathcal{V}$ for the new and original models coincide. The set $\mathcal{V}$ is compact since $\tilde{\mathbb{A}}$ is a Borel subset of a Polish space. □

6. Submodels and dimensionality reduction.

DEFINITION 6.1. An MDP $\{\tilde{\mathbb{X}}, \tilde{\mathbb{A}}, \tilde{A}(\cdot), \tilde{p}, \tilde{r}\}$ is called a *submodel* of the MDP $\{\mathbb{X}, \mathbb{A}, A(\cdot), p, r\}$ if $\tilde{\mathbb{X}} = \mathbb{X}$, $\tilde{\mathbb{A}} = \mathbb{A}$, $\tilde{p} = p$, $\tilde{r} = r$, and $\tilde{A}(x) \subset A(x)$ for all $x \in \mathbb{X}$.

We say that a submodel is well-defined if the set $\mathrm{Gr}_{\tilde{\mathbb{X}}}(\tilde{A})$ is a Borel subset of $\tilde{\mathbb{X}} \times \tilde{\mathbb{A}}$ and there exists at least one deterministic policy (selector) in the submodel. The existence of a selector usually follows from measurable selection theorems. According to the Arsenin–Kunugui selection theorem (Kechris [26, Theorem 18.18]), a measurable selector $\phi : \tilde{\mathbb{X}} \mapsto \tilde{\mathbb{A}}$, such that $\phi(s) \in \tilde{A}(x)$ for all $x \in \tilde{\mathbb{X}}$, exists if $\tilde{\mathbb{A}}$ is a Borel subset of a Polish space, the set $\mathrm{Gr}_{\tilde{\mathbb{X}}}(\tilde{A})$ is a Borel subset of $\tilde{\mathbb{X}} \times \tilde{\mathbb{A}}$, and each set $\tilde{A}(x)$ is a union of a countable number of nonempty compact subsets of $\tilde{\mathbb{A}}$. In addition, this theorem claims that under these assumptions the projection of any Borel subset of $\mathrm{Gr}_{\tilde{\mathbb{X}}}(\tilde{A})$ onto $\tilde{\mathbb{X}}$ is a Borel subset of $\tilde{\mathbb{X}}$. If $\mathrm{Gr}_{\tilde{\mathbb{X}}}(\tilde{A})$ is a Borel subset of $\tilde{\mathbb{X}} \times \tilde{\mathbb{A}}$ and each set $\tilde{A}(x)$, $x \in \tilde{\mathbb{X}}$, is nonempty and finite or countable, then the Arsenin–Kunugui theorem implies that the submodel is well-defined and the projection of any Borel subset of $\mathrm{Gr}_{\tilde{\mathbb{X}}}(\tilde{A})$ onto $\tilde{\mathbb{X}}$ is a Borel subset of $\tilde{\mathbb{X}}$.

It is obvious that a submodel inherits many properties of the MDP including atomless, absorbing, and uniformly absorbing properties. In addition, $\tilde{\mathcal{V}} \subset \mathcal{V}$, where $\tilde{\mathcal{V}}$ is the performance set for the submodel.

LEMMA 6.2. *Consider an absorbing atomless MDP. Then for every $v \in \mathcal{V}$ there exists a submodel with finite or countable action sets $\tilde{A}(x)$, $x \in \mathbb{X}$, such that for some stationary policy π for this submodel, $v^\pi = v$ and $\pi(a|x) > 0$ for all $x \in \mathbb{X}$ and all $a \in \tilde{A}(x)$.*

Proof. According to Feinberg and Piunovskiy [19, Theorem 2.1], there exists a nonrandomized Markov policy $\phi = (\phi_0, \phi_1, \ldots)$ such that $v^\phi = v$. Let us define the nonempty sets $A_\phi(x) := \cup_{n=0}^\infty \{\phi_n(x)\}$, which are either countable or finite. Observe that the set $\mathrm{Gr}_{\mathbb{X}}(A_\phi) = \cup_{n=0}^\infty \mathrm{Gr}_{\mathbb{X}}(\phi_n)$ is Borel because the graph of a Borel function ϕ_n is a Borel set; see, e.g., Bertsekas and Shreve [3, Corollary 7.14.1].

In view of Theorem 4.4, there is a stationary policy π such that $\pi(\cdot|x)$ is concentrated on $A_\phi(x)$ and $v^\pi = v^\phi = v$. Let $\tilde{A}(x) = \{a \in A_\phi(x) : \pi(a|x) > 0\}$, $x \in \mathbb{X}$. Since $\pi(A_\phi(x)|x) = 1$, then $\tilde{A}(x) \neq \emptyset$ for all $x \in \mathbb{X}$. The set $\mathrm{Gr}_{\mathbb{X}}(\tilde{A})$ is Borel because

$\mathrm{Gr}_{\mathbb{X}}(\tilde{A}) = \{(x,a) \in \mathbb{X} \times \mathbb{A} : \ G(x,a) > 0\}$, where $G(x,a) = \sum_{n=0}^{\infty} \pi(\phi_n(x)|x) I\{(x,a) \in \mathrm{Gr}_{\mathbb{X}}(\phi_n)\}$, and because the functions $I\{(x,a) \in \mathrm{Gr}_{\mathbb{X}}(\phi_n)\}$ and $\pi(\phi_n(x)|x)$ are Borel-measurable, where the measurability of the function $I\{(x,a) \in \mathrm{Gr}_{\mathbb{X}}(\phi_n)\}$ follows from the measurability of the sets $\mathrm{Gr}_{\mathbb{X}}(\phi_n) \subset \mathbb{X} \times \mathbb{A}$, and the measurability of the function $\pi(\phi_n(x)|x)$ follows from Bertsekas and Shreve [3, Corollary 7.26.1]. □

THEOREM 6.3. *Consider a uniformly absorbing atomless MDP. Suppose that* $N = 1$ *and there exists a stationary policy* σ^* *such that* $v^{\sigma^*} = \sup_{\sigma \in \mathbb{S}} v^{\sigma}$. *For* $v := v^{\sigma^*} \in \mathcal{V}$ *consider a stationary policy* π *and a submodel with action sets* $\tilde{A}(\cdot)$, *whose existence is stated in Lemma* 6.2. *Then* $v^{\pi^*} = v^{\sigma^*}$ *for each policy* π^* *in this submodel.*

Proof. Let $N = 1$. In view of Theorem 4.4, for an absorbing MDP $\sup_{\tilde{\pi} \in \mathbb{S}} v^{\tilde{\pi}} = \sup_{\tilde{\pi} \in \Pi} v^{\tilde{\pi}}$ and $v^{\sigma} = \sup_{\tilde{\pi} \in \Pi} v^{\tilde{\pi}}$ for some policy σ if and only if $v^{\sigma^*} = \sup_{\tilde{\pi} \in \mathbb{S}} v^{\tilde{\pi}}$ for some stationary policy σ^*. Recall that $\sup_{\tilde{\pi} \in \mathbb{S}} v^{\tilde{\pi}} = \sup_{\phi \in \mathbb{F}} v^{\phi}$; see Feinberg [13].

For an arbitrary policy $\sigma \in \Pi$, let X^{σ} be the set of initial states $x \in \mathbb{X}$ for which the expected initial rewards $v^{\sigma}(x)$ are well-defined, that is,

$$X^{\sigma} = \{x \in \mathbb{X} : \ v_+^{\sigma}(x) < +\infty\} \cap \{x \in \mathbb{X} : \ v_-^{\sigma}(x) < +\infty\}. \tag{6.1}$$

In view of the Ionescu Tulcea theorem [27, section V.1], the functions $v_+^{\sigma}(x)$ and $v_-^{\sigma}(x)$ are Borel-measurable. Therefore, the set X^{σ} is Borel as the union of two Borel sets.

For $x \in \mathbb{X}$, $a \in \tilde{A}(x)$, and for a Borel function $f : \mathbb{X} \mapsto \mathbb{R}^1$, let us denote

$$\mathbf{T}^a f(x) := r(x,a) + \int_{\mathbb{X}} f(y) p(dy|x,a), \qquad x \in \mathbb{X}, \ a \in \tilde{A}(x).$$

This value is well-defined if either $\int_{\mathbb{X}} f^+(y) p(dy|x,a) < +\infty$ or $\int_{\mathbb{X}} f^-(y) p(dy|x,a) < +\infty$.

Let σ be a stationary policy in the submodel with action sets $\tilde{A}(\cdot)$. Then $\mathbf{T}^a v^{\sigma}(x)$ is well-defined for $x \in X^{\sigma}$ and $a \in \tilde{A}(x)$, where the Borel set X^{σ} is defined in (6.1). Indeed,

$$v_+^{\sigma}(x) = \sum_{a \in \tilde{A}(x)} \sigma(a|x) \left\{ r^+(x,a) + \int_{\mathbb{X}} v_+^{\sigma}(y) p(dy|x,a) \right\} < +\infty, \qquad x \in X^{\sigma},$$

and

$$v_-^{\sigma}(x) = \sum_{a \in \tilde{A}(x)} \sigma(a|x) \left\{ r^-(x,a) + \int_{\mathbb{X}} v_-^{\sigma}(y) p(dy|x,a) \right\} < +\infty, \qquad x \in X^{\sigma}.$$

Therefore,

$$v^{\sigma}(x) = v_+^{\sigma}(x) - v_-^{\sigma}(x) = \sum_{a \in \tilde{A}(x)} \sigma(a|x) \mathbf{T}^a v^{\sigma}(x), \qquad x \in X^{\sigma}, \tag{6.2}$$

and $\mathbf{T}^a v^{\sigma}(x)$ is well-defined for $x \in X^{\sigma}$ and $a \in \tilde{A}(x)$ if $\sigma(a|x) > 0$.

Observe that for an absorbing MDP $q^{\sigma}(\mathbb{X} \setminus X^{\sigma}) = 0$, which is equivalent to $q^{\sigma}(\mathbb{X}) = q^{\sigma}(X^{\sigma})$. Indeed, if $q^{\sigma}(\mathbb{X} \setminus X^{\sigma}) > 0$, then, in view of (4.5), $P^{\sigma}\{x_n \in \mathbb{X} \setminus X^{\sigma}\} > 0$ for some $n = 0, 1, \ldots$. This implies that either $v_+^{\sigma} = +\infty$ or $v_-^{\sigma} = +\infty$. This conclusion contradicts the assumptions that the MDP is absorbing and the reward function r is bounded.

In particular, for $\sigma = \pi$, where the policy π is defined in Lemma 6.2,

$$q^{\pi}(\mathbb{X} \setminus X^{\pi}) = 0. \tag{6.3}$$

By Lemma 6.2, $v^{\pi} = v = v^{\sigma^*}$. Consider the sets

$$X^{>} := \{x \in X^{\pi} : \mathbf{T}^a v^{\pi}(x) > v^{\pi}(x) \text{ for some } a \in \tilde{A}(x)\},$$

$$X^{<} := \{x \in X^{\pi} : \mathbf{T}^a v^{\pi}(x) < v^{\pi}(x) \text{ for some } a \in \tilde{A}(x)\},$$

$$X^{=} := \{x \in X^{\pi} : \mathbf{T}^a v^{\pi}(x) = v^{\pi}(x) \text{ for all } a \in \tilde{A}(x)\}.$$

The sets $X^{>}$, $X^{<}$, and $X^{=}$ are Borel. Indeed, the set $X^{>}$ is a projection of the Borel set $Y(\pi) := \{(x,a) \in \mathrm{Gr}_{X^{\pi}}(\tilde{A}) : \mathbf{T}^a v^{\pi}(x) > v^{\pi}(x)\}$ onto X^{π}. In addition, each action set $\tilde{A}(x)$, $x \in \mathbb{X}$, is finite or countable. Therefore, in view of the Arsenin–Kunugui theorem, the set $X^{>}$ is Borel and there exists a Borel mapping $\varphi^* : X^{>} \mapsto \mathbb{A}$ such that $\varphi^*(x) \in \tilde{A}(x)$ and $\mathbf{T}^{\varphi^*(x)} v^{\pi}(x) > v^{\pi}(x)$ for all $x \in X^{>}$. The set $X^{<}$ is Borel because it is a projection of the Borel set $\{(x,a) \in \mathrm{Gr}_{X^{\pi}}(\tilde{A}) : \mathbf{T}^a v^{\pi}(x) < v^{\pi}(x)\}$ onto $\mathbb{X}$. Thus, $X^{=} = X^{\pi} \setminus (X^{>} \cup X^{<})$ is a Borel set, too.

Observe that

$$q^{\pi}(X^{<}) = q^{\pi}(X^{>}) = 0. \tag{6.4}$$

To prove the second equality in (6.4), suppose that $q^{\pi}(X^{>}) > 0$. Therefore, $q^{\pi}_n(X^{>}) = P^{\pi}\{x_n \in X^{>}\} > 0$ for some $n = 0, 1, \ldots$. For the Borel mapping φ^* described in the previous paragraph, consider a randomized Markov policy π':

$$\pi'_t(B|x) = \begin{cases} I\{\varphi^*(x) \in B\} & \text{if } t = n \text{ and } x \in X^{>}, \\ \pi(B|x) & \text{otherwise,} \end{cases}$$

where $B \in \mathcal{A}$ and $t = 0, 1, \ldots$. Straightforward calculations imply that

$$v^{\pi'} - v^{\pi} = \int_{X^{>}} [\mathbf{T}^{\varphi^*(x)} v^{\pi}(x) - v^{\pi}(x)] q^{\pi}_n(dx) > 0,$$

which contradicts $v^{\pi} = v^{\sigma^*} = \sup_{\sigma \in \mathbb{S}} v^{\sigma} = \sup_{\sigma \in \Pi} v^{\sigma} \geq v^{\pi'}$, where the last equality follows from Theorem 4.4. Thus, the second equality in (6.4) is proved.

The equality $q^{\pi}(X^{<}) = 0$ holds because the inequality $q^{\pi}(X^{<}) > 0$ is impossible. Indeed, if $q^{\pi}(X^{<}) > 0$, then $q^{\pi}(X^{<} \setminus X^{>}) = q^{\pi}(X^{<}) > 0$ because $q^{\pi}(X^{>}) = 0$. Therefore,

$$0 = \int_{X^{<} \setminus X^{>}} (v^{\pi}(x) - v^{\pi}(x)) q^{\pi}(dx) = \int_{X^{<} \setminus X^{>}} \sum_{a \in \tilde{A}(x)} \pi(a|x)(\mathbf{T}^a v^{\pi}(x) - v^{\pi}(x)) q^{\pi}(dx) < 0,$$

where the second equality follows from (6.2) and the inequality holds because an integral of a negative function on a set with a positive measure is negative. The function is negative because $\pi(a|x) > 0$ for all $a \in \tilde{A}(x)$, the difference in the second integral is nonpositive for all $a \in \tilde{A}(x)$, and this difference is negative for some $a \in \tilde{A}(x)$, where $x \in X^{<} \setminus X^{>}$. Equalities (6.4) are proved.

The equality $v^{\pi^*} = v^{\sigma^*}$ holds for every policy π^* in the submodel with action sets $\tilde{A}(\cdot)$ if and only if $v^{\sigma} = v^{\pi}$ for every stationary policy σ in this submodel. This is true in view of Theorem 4.4 and because $v^{\pi} = v^{\sigma^*} = v$. Let σ be a stationary policy for the submodel with action sets $\tilde{A}(\cdot)$. To complete the proof, we show in the rest of the proof that $v^\sigma = v^\pi$.

Since $\sigma(\cdot|x) \ll \pi(\cdot|x)$ for all $x \in \mathbb{X}$, Lemma 4.2 and formulae (6.3), (6.4) imply that $q^\sigma(\mathbb{X} \setminus X^=) = 0$. Let $\sigma^{n,\pi}$ be the policy that follows σ at times $t = 0, 1, \ldots, n-1$ and follows π at $t = n, n+1, \ldots$. In particular, $\sigma^{0,\pi} = \pi$. Induction arguments imply that

$$v^{\sigma^{n,\pi}} = v^\pi, \qquad n = 0, 1, \ldots. \tag{6.5}$$

Indeed, for $n = 0$ formula (6.5) holds because $\sigma^{0,\pi} = \pi$. If (6.5) holds for some $n = 0, 1, \ldots$, then

$$v^{\sigma^{n+1,\pi}}(x) = \sum_{a \in \tilde{A}(x)} \sigma(a|x) T^a v^\pi(x) = v^\pi(x), \qquad x \in X^=,$$

and

$$v^{\sigma^{n+1,\pi}} = \int_{\mathbb{X}} v^{\sigma^{n+1,\pi}}(x)\mu(dx) = \int_{X^=} v^\pi(x)\mu(dx) = \int_{\mathbb{X}} v^\pi(x)\mu(dx) = v^\pi,$$

where the last equalities hold because $\mu(\mathbb{X} \setminus X^=) = 0$ since $\mu \ll q^\pi$, and $q^\pi(\mathbb{X} \setminus X^=) = 0$ in view of (6.3) and (6.4). Formula (6.5) is proved.

Since the MDP is uniformly absorbing,

$$\lim_{n\to\infty} E^{\sigma^{n,\pi}} \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = \lim_{n\to\infty} \sup_{\tilde{\pi}\in M} E^{\tilde{\pi}} \sum_{t=n}^{\infty} I\{t < T^{\bar{x}}\} = 0.$$

Since the reward function r is bounded,

$$\lim_{n\to\infty} E^{\sigma^{n,\pi}} \sum_{t=n}^{\infty} r(x_t, a_t) = 0.$$

Therefore,

$$\begin{aligned} v^\sigma &= \lim_{n\to\infty} E^\sigma \sum_{t=0}^{n-1} r(x_t, a_t) = \lim_{n\to\infty} E^\sigma \sum_{t=0}^{n-1} r(x_t, a_t) + \lim_{n\to\infty} E^{\sigma^{n,\pi}} \sum_{t=n}^{\infty} r(x_t, a_t) \\ &= \lim_{n\to\infty} v^{\sigma^{n,\pi}} = v^\pi, \end{aligned}$$

where the last equality follows from (6.5). □

The following lemma is correct without the assumption that the MDP is atomless. However, we need it only for an atomless MDP in this paper, and for an atomless MDP the proof follows directly from Theorem 6.3.

COROLLARY 6.4. *Consider a uniformly absorbing atomless MDP with $N = 1$. For every extreme point $v \in \mathcal{V}$ of the set $\mathcal{V}$ there exists a deterministic policy ϕ such that $v^\phi = v$.*

Proof. Since $N = 1$, the closure of the convex set $\mathcal{V}$ is a bounded interval on the line. Therefore, there could be at most two extreme points $v_* := \inf_{\pi\in\Pi} v^\pi$ and $v^* := \sup_{\pi\in\Pi} v^\pi$. Let us consider $v = v^*$. Theorem 4.4 implies that $v = \sup_{\pi\in\mathbb{S}} v^\pi$. According to Theorem 6.3, $v^\phi = v$ for every deterministic policy ϕ in the submodel, whose existence is stated in Lemma 6.2. The change $r := -r$ reduces the case $v = v_*$ to the case $v = v^*$. □

For $i = 1, \dots, N$, let us denote by b_{-i} the projection of $b \in \mathbb{R}^N$ to $\mathbb{R}^{N-1}$ obtained by removing the ith coordinate of the vector b. Also, $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors.

Definition 6.5. We say that *a point $v \in \mathcal{V}$ allows the dimensionality reduction* if there are a coordinate $i = 1, 2, \dots, N$, a vector $b \in \mathbb{R}^{N-1}$, a constant d, and a submodel $\{\mathbb{X}, \mathbb{A}, \tilde{A}(\cdot), p, r\}$ of the original MDP such that $v \in \tilde{\mathcal{V}}$, where $\tilde{\mathcal{V}}$ is the performance set for all policies in the submodel, and

$$\tilde{v}^{(i)} = d + \langle b, \tilde{v}_{-i} \rangle \qquad \text{for all} \qquad \hat{v} \in \tilde{\mathcal{V}}. \tag{6.6}$$

The following theorem plays an important role in the proof of Theorem 3.8. Recall that $\partial(C)$ is the boundary of a bounded convex set $C \in \mathbb{R}^n$, $n = 1, 2, \dots$.

Theorem 6.6 (dimensionality reduction). *For a uniformly absorbing atomless MDP, each point on the boundary of $\mathcal{V}$ allows the dimensionality reduction.*

Proof. Let $v^* \in \partial(\mathcal{V})$. Let $\langle \tilde{b}, v \rangle = \tilde{d}$ be a supporting hyperplane at the point v^* to the convex set $\mathcal{V}$ such that $\langle \tilde{b}, v \rangle \le \tilde{d}$ for all $v \in \mathcal{V}$ and $\langle \tilde{b}, v^* \rangle = \tilde{d}$, where $\tilde{b}^{(i)} \neq 0$ for at least one $i = 1, \dots, N$. Let us define the one-step reward function

$$\tilde{r}(x, a) := \langle \tilde{b}, r(x, a) \rangle, \qquad\qquad x \in \mathbb{X},\ a \in A(x).$$

Let $\tilde{v}^\sigma$ be the expected total rewards for this reward function, initial distribution μ, and a policy σ. Then $\tilde{v}^\sigma = \langle \tilde{b}, v^\sigma \rangle$.

Since $v^* \in \mathcal{V}$, then $v^* = v^{\sigma^*}$ for a stationary policy $\sigma^* \in \mathbb{S}$. Using Lemma 6.2, consider the corresponding submodel with finite or countable action sets $\tilde{A}(\cdot)$ and a stationary policy π for this submodel, where $\tilde{\mathcal{V}}$ is the performance set for the submodel. In particular, $v^\pi = v^* \in \tilde{\mathcal{V}}$. Note that $\tilde{v}^\pi = \tilde{d} = \sup_{v \in \mathcal{V}} v = \sup_{\sigma \in \mathbb{S}} \tilde{v}^\sigma$. In view of Theorem 6.3,

$$\tilde{v}^\pi = \langle \tilde{b}, \hat{v} \rangle \qquad \text{for all} \qquad \hat{v} \in \tilde{\mathcal{V}}. \tag{6.7}$$

Formula (6.7) implies (6.6) with $d := \langle \tilde{b}, v^* \rangle / \tilde{b}^{(i)}$ and $b := -\tilde{b}_{-i}/\tilde{b}^{(i)}$, where $i = 1, \dots, N$ with $\tilde{b}^{(i)} \neq 0$ and $\tilde{b}^{(i)}$ is the ith coordinate of the vector $\tilde{b}$. ◻

7. An MDP defined by two deterministic policies.

Let ϕ^0 and ϕ^1 be two deterministic policies. These two policies are considered to be fixed within this section. Let us define action sets $A^*(x) := \{\phi^0(x), \phi^1(x)\}$ and consider an MDP which is the submodel obtained from the original MDP by narrowing the action sets $A(x)$ to $A^*(x)$ for all $x \in \mathbb{X}$. We say that this MDP is defined by the deterministic policies ϕ^0 and ϕ^1.

Consider the stationary policy π^*:

$$\pi^*(B|x) := \frac{1}{2}[I\{\phi^0(x) \in B\} + I\{\phi^1(x) \in B\}], \qquad B \in \mathcal{A},\ x \in \mathbb{X}, \tag{7.1}$$

which averages the deterministic policies ϕ^0 and ϕ^1. We denote by q the occupancy measure q^{π^*} on $\mathbb{X}$,

$$q(Y) := q^{\pi^*}(Y), \qquad Y \in \mathcal{X}. \tag{7.2}$$

Lemma 7.1. *$q^\gamma \ll q$ for every stationary policy γ for the MDP defined by two deterministic policies ϕ^0 and ϕ^1.*

Proof. This lemma follows from Lemma 4.2 since $\gamma(\cdot|x) \ll \pi^*(\cdot|x)$, $x \in \mathbb{X}$, for each stationary policy γ for the MDP defined by two deterministic policies ϕ^0 and ϕ^1. ◻

The following lemma provides a useful inequality.

LEMMA 7.2. *For every stationary policy γ for the MDP defined by two deterministic policies ϕ^0 and ϕ^1, the inequality $E^\gamma f(x_t) \le 2^t E^{\pi^*} f(x_t)$ holds for an arbitrary nonnegative measurable function f and for each $t = 0, 1, \dots$.*

Proof. The proof is based on the induction in t. Since $E^\gamma f(x_0) = \int_{\mathbb{X}} f(x)\mu(dx)$ for every stationary policy γ, the inequality holds for $t = 0$ in the form of the equality. Let this inequality hold for some $t = 0, 1, \dots$. Then

$$E^\gamma[f(x_{t+1})|x_t] = \int_{\mathbb{X}} f(x) \sum_{i=0}^{1} \gamma(\phi^i(x_t)|x_t) p(dx|x_t, \phi^i(x_t)) \le \int_{\mathbb{X}} f(x) \sum_{i=0}^{1} p(dx|x_t, \phi^i(x_t))$$
$$= 2\int_{\mathbb{X}} f(x) \sum_{i=0}^{1} \frac{1}{2} p(dx|x_t, \phi^i(x_t)) = 2E^{\pi^*}[f(x_{t+1})|x_t], \tag{7.3}$$

where the first and the last equalities follow from the definitions of strategic measures, and the inequality and the second equality are obvious. Therefore, $E^\gamma f(x_{t+1}) = E^\gamma E^\gamma[f(x_{t+1})|x_t] \le 2E^\gamma E^{\pi^*}[f(x_{t+1})|x_t] \le 2^{t+1} E^{\pi^*} E^{\pi^*}[f(x_{t+1})|x_t] = 2^{t+1} E^{\pi^*} f(x_{t+1})$, where the first and the last equalities follow from the definition of a conditional expectation, the first inequality follows from (7.3), and the second inequality follows from the induction assumption. ◻

COROLLARY 7.3. *For $t = 0, 1, \dots$ and for every $Y \in \mathcal{X}$, the inequality $q_t^\gamma(Y) \le 2^t q_t(Y)$ holds for every stationary policy for the MDP defined by two deterministic policies ϕ^0 and ϕ^1.*

Proof. The corollary follows from Lemma 7.2 applied to the function $f(x) = I\{x \in Y\}$, $x \in \mathbb{X}$. ◻

For two stationary policies π and σ for the MDP defined by two deterministic policies ϕ^0 and ϕ^1, let

$$X(\pi, \sigma) := \{x \in \mathbb{X} : \pi(\cdot|x) = \sigma(\cdot|x)\} = \{x \in \mathbb{X} : \pi(\phi^0(x)|x) = \sigma(\phi^0(x)|x)\} \tag{7.4}$$

be the set of states on which π and σ choose the same decisions. In view of the last equality, this set is measurable.

LEMMA 7.4. *Consider a uniformly absorbing MDP. If $q(\mathbb{X} \setminus X(\pi, \sigma)) = 0$, then $q^\pi = q^\sigma$, where π and σ are arbitrary stationary policies in the MDP defined by two deterministic policies ϕ^0 and ϕ^1.*

Proof. As follows from (7.1), $\pi(\cdot|x) \ll \pi^*(\cdot|x)$ and $\sigma(\cdot|x) \ll \pi^*(\cdot|x)$ for all $x \in \mathbb{X}$. Lemma 4.2 implies that $q^\pi \ll q$ and $q^\sigma \ll q$. Therefore, $q^\pi(\mathbb{X} \setminus X(\pi, \sigma)) = 0$ and $q^\sigma(\mathbb{X} \setminus X(\pi, \sigma)) = 0$ if $q(\mathbb{X} \setminus X(\pi, \sigma)) = 0$. Thus, the set of states, on which the stationary policies π and σ make different decisions, will be visited with zero probability when each of these policies is used. ◻

Let $d_{TV}(\eta_1, \eta_2)$ denote the distance in total variation between two finite measures defined on the same measurable space; see, e.g., [16, section 2] or [17] for details on definitions and properties of distances in total variation for finite measures. Since $q^\pi(dx) = Q^\pi(dx, \mathbb{A})$ for an arbitrary policy π, then $d_{TV}(q^\pi, q^\sigma) \le d_{TV}(Q^\pi, Q^\sigma)$ for two policies π and σ. As follows from Lemma 7.4, $q(\mathbb{X} \setminus X(\pi,\sigma)) = 0$ implies that $q^\pi = q^\sigma$. The following theorem, which is the main result of this section, demonstrates that the value of $q(\mathbb{X} \setminus X(\pi,\sigma))$ characterizes how close the measures Q^π and Q^σ are.

THEOREM 7.5. *Consider a uniformly absorbing MDP. Let π and σ be two stationary policies for the MDP defined by two deterministic policies ϕ^0 and ϕ^1. Then for every $\epsilon > 0$ there exists $\delta > 0$ such that if $q(\mathbb{X} \setminus X(\pi,\sigma)) \le \delta$, then $d_{TV}(Q^\pi, Q^\sigma) \le \epsilon$.*

Proof. Let us fix an arbitrary $\epsilon > 0$. In this proof γ is always a policy that is equal either to π or to σ. In other words, $\gamma \in \{\pi, \sigma\}$.

We prove first the existence of $\delta > 0$ such that if $q(\mathbb{X} \setminus X(\pi,\sigma)) \le \delta$, then $d_{TV}(q^\pi, q^\sigma) \le \epsilon$. This claim follows from the following fact. There exist a constant $\delta > 0$ and measures $\bar{q}^\gamma$ and $\hat{q}^\gamma$ on $(\mathbb{X}, \mathcal{X})$ such that the inequality $q(\mathbb{X} \setminus X(\pi,\sigma)) \le \delta$ implies the correctness of the following statements: (i) $q^\gamma = \bar{q}^\gamma + \hat{q}^\gamma$, (ii) $\hat{q}^\gamma(\mathbb{X}) \le \epsilon/2$, and (iii) $\bar{q}^\pi = \bar{q}^\sigma$. If this is true, then $d_{TV}(q^\pi, q^\sigma) = d_{TV}(\hat{q}^\pi, \hat{q}^\sigma) \le \epsilon$.

Let us construct a positive constant δ and measures $\bar{q}^\gamma$ and $\hat{q}^\gamma$ on $(\mathbb{X}, \mathcal{X})$ satisfying properties (i)–(iii). We denote by $\bar{T}^Y := \min\{t = 0, 1, \dots : x_t \notin Y\}$ the first time the process leaves the set $Y \in \mathcal{X}$ and define the measure

$$\bar{q}^\gamma(C) = E^\gamma \sum_{t=0}^{\infty} I\{x_t \in C\} I\{\bar{T}^{X(\pi,\sigma)} > t\}, \qquad C \in \mathcal{X}.$$

Since the stationary policies π and γ coincide on the set $X(\pi,\sigma)$,

$$\bar{q}^\pi = \bar{q}^\sigma.$$

Thus, (iii) holds. Since the MDP is uniformly absorbing, there exist $\ell = 1, 2, \dots$ such that for every stationary policy π'

$$D_1^{\pi'} := E^{\pi'} \sum_{t=\ell}^{\infty} I\{x_t \in \mathbb{X}\} \le \epsilon/4. \tag{7.5}$$

In particular, (7.5) holds for $\pi' = \gamma$.

Our next step is to show that there exists $\delta > 0$ such that if $q(\mathbb{X} \setminus X(\pi,\sigma)) \le \delta$, then

$$D_2^\gamma := E^\gamma \sum_{t=0}^{\ell-1} I\{x_t \in \mathbb{X}\} I\{\bar{T}^{X(\pi,\sigma)} \le t\} \le \epsilon/4. \tag{7.6}$$

Indeed, by exchanging the summation and expectation in (7.6), we have

$$D_2^\gamma = \sum_{t=0}^{\ell-1} P^\gamma\{x_t \in \mathbb{X}, \bar{T}^{X(\pi,\sigma)} \le t\}. \tag{7.7}$$

Observe that for $t = 0, 1, \dots$

$$P^\gamma\{x_t \in \mathbb{X}, \bar{T}^{X(\pi,\sigma)} \le t\} \le \sum_{s=0}^{t} P^\gamma\{x_t \in \mathbb{X}, x_s \in \mathbb{X} \setminus X(\pi,\sigma)\} \le \sum_{s=0}^{t} q_s^\gamma(\mathbb{X} \setminus X(\pi,\sigma)). \tag{7.8}$$

In view of Corollary 7.3,

$$\sum_{s=0}^{t} q_s^\gamma(\mathbb{X} \setminus X(\pi,\sigma)) \le \sum_{s=0}^{t} 2^s q_s(\mathbb{X} \setminus X(\pi,\sigma)) \le 2^t \sum_{s=0}^{t} q_s(\mathbb{X} \setminus X(\pi,\sigma)) \le 2^t q(\mathbb{X} \setminus X(\pi,\sigma)). \tag{7.9}$$

Formulae (7.7)–(7.9) imply that $D_2^\gamma \leq 2^\ell q(\mathbb{X} \setminus X(\pi,\sigma))$. Thus, (7.6) holds with $\delta = 2^{-(\ell+2)}\epsilon$.

Let us define the measures

$$\hat{q}^\gamma(C) = E^\gamma \sum_{t=0}^{\infty} I\{x_t \in C\} I\{\bar{T}^{X(\pi,\sigma)} \leq t\}, \qquad C \in \mathcal{X}.$$

Then $q^\gamma = \bar{q}^\gamma + \hat{q}^\gamma$. Thus, (i) holds. Let $\delta = 2^{-(\ell+2)}\epsilon$. For $\gamma \in \{\pi,\sigma\}$

$$\begin{aligned} \hat{q}^\gamma(\mathbb{X}) &= E^\gamma \sum_{t=0}^{\infty} I\{x_t \in \mathbb{X}\} I\{\bar{T}^{X(\pi,\sigma)} \leq t\} \\ &\leq E^\gamma \sum_{t=0}^{\ell-1} I\{x_t \in \mathbb{X}\} I\{\bar{T}^{X(\pi,\sigma)} \leq t\} + E^\gamma \sum_{t=\ell}^{\infty} I\{x_t \in \mathbb{X}\} \leq \epsilon/2, \end{aligned}$$

where the last inequality follows from (7.6) and (7.5). Thus, (ii) holds. In view of (i)–(iii), $d_{TV}(q^\pi, q^\sigma) \leq \epsilon$.

Let us prove the inequality $d_{TV}(Q^\pi, Q^\sigma) \leq \epsilon$. To do this, we consider the measures $\bar{Q}^\gamma$ and $\hat{Q}^\gamma$ on $(\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$ defined by

$$\bar{Q}^\gamma(C \times B) = E^\gamma \sum_{t=0}^{\infty} I\{x_t \in C, a_t \in B\} I\{\bar{T}^{X(\pi,\sigma)} > t\}, \qquad C \in \mathcal{X},\ B \in \mathcal{A},$$

$$\hat{Q}^\gamma(C \times B) = E^\gamma \sum_{t=0}^{\infty} I\{x_t \in C, a_t \in B\} I\{\bar{T}^{X(\pi,\sigma)} \leq t\}, \qquad C \in \mathcal{X},\ B \in \mathcal{A}.$$

These two measures obviously satisfy the following properties: (i*) $Q^\gamma = \bar{Q}^\gamma + \hat{Q}^\gamma$, (ii*) $\hat{Q}^\gamma(\mathbb{X} \times \mathbb{A}) = \hat{q}^\gamma(\mathbb{X}) \leq \epsilon/2$, and (iii*) $\bar{Q}^\pi = \bar{Q}^\sigma$. Properties (i*)–(iii*) imply $d_{TV}(Q^\pi, Q^\sigma) \leq \epsilon$. ∎

Let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^N$. The following corollary follows from Theorem 7.5.

COROLLARY 7.6. *Let π and σ be two stationary policies in the MDP defined by two deterministic policies ϕ^0 and ϕ^1. Then for every $\epsilon > 0$ there exists $\delta > 0$ such that the inequality $q(\mathbb{X} \setminus X(\pi,\sigma)) \leq \delta$ implies that $\|v^\pi - v^\sigma\| \leq \epsilon$.*

Proof. Let K be a finite positive constant satisfying $K \geq |r^{(n)}(x,a)|$ for all $n = 1,\ldots,N$, $x \in \mathbb{X}$, and $a \in A(x)$. Then the corollary follows from Theorem 7.5 applied to the constant $\epsilon_1 := \epsilon/(KN^{\frac{1}{2}})$ instead of ϵ. ∎

8. Path connectedness of the set of occupancy measures generated by deterministic policies.

We recall that a subset E of a topological space is called path-connected if for every two points $e_0, e_1 \in E$ there exists a continuous function $g : [0,1] \mapsto E$ such that $g(0) = e_0$ and $g(1) = e_1$. A set is called connected if it cannot be partitioned into two nonempty subsets which are open in the relative topology induced on the set. Of course, the validity of these properties may depend on the topology chosen on the space. A subset of the Euclidean space $\mathbb{R}^N$ is connected if and only if it is path-connected.

DEFINITION 8.1. *A subset E of the set of finite measures on a measurable space is called path-connected in total variation if this set is path-connected when the set of finite measures is endowed with the metric equal to the distance in total variation.*

A sequence $\{\nu_n\}_{n=1,2,\dots}$ of finite measures on a measurable space $(\Omega,\mathcal{F})$ converges setwise to a measure ν on $(\Omega,\mathcal{F})$ if for every bounded measurable function $f:\Omega\mapsto\mathbb{R}$ $\int_\Omega f(\omega)\nu_n(d\omega)\mapsto\int_\Omega f(\omega)\nu(d\omega)$. Setwise convergence defines the topology of setwise convergence of measures; see, e.g., Bogachev [5, p. 291].

Definition 8.2. *A subset E of the space of finite measures on a measurable space is called setwise path-connected if this set is path-connected when the space of finite measures is endowed with the topology of setwise convergence of measures.*

In particular, a sequence of occupancy measures $\{Q_n\}_{n=1,2,\dots}$ converges setwise to an occupancy measure Q if for every bounded measurable function $f:\mathbb{X}\times\mathbb{A}\mapsto\mathbb{R}$

$$\int_{\mathbb{X}}\int_{\mathbb{A}} f(x,a)Q_n(dx,da)\to\int_{\mathbb{X}}\int_{\mathbb{A}} f(x,a)Q(dx,da). \tag{8.1}$$

In view of (8.1), the set $\mathcal{M}^{\mathbb{F}}$ is setwise path-connected if and only if for every two deterministic policies ϕ^0 and ϕ^1 there exists a map $g:[0,1]\mapsto\mathcal{M}^{\mathbb{F}}$ such that $g(0)=Q^{\phi^0}$, $g(1)=Q^{\phi^1}$, and the function

$$\zeta(\alpha):=\int_{\mathbb{X}}\int_{\mathbb{A}} f(x,a)g(\alpha)(dx,da) \tag{8.2}$$

is continuous for every bounded measurable function $f:\mathbb{X}\times\mathbb{A}\mapsto\mathbb{R}$.

Theorem 8.3. *For a uniformly absorbing atomless MDP, the set $\mathcal{M}^{\mathbb{F}}$ is path-connected in total variation and therefore it is setwise path-connected.*

Proof. Let ϕ^0 and ϕ^1 be two deterministic policies. Consider the stationary policy π^* defined in (7.1) and the measure q on $\mathbb{X}$ defined in (7.2). The measure q is atomless in view of Lemma 4.3. So, $q(x)=0$ for all $x\in\mathbb{X}$.

Let ψ be an isomorphic map of $\mathbb{X}$ onto the closed interval $[0,1]$; that is, ψ is a one-to-one measurable mapping of $(\mathbb{X},\mathcal{X})$ onto $([0,1],\mathcal{B}([0,1]))$. Observe that the function ψ can be viewed as a nonnegative random variable on the measurable space $(\mathbb{X},\mathcal{X})$ with the distribution function

$$F_\psi(b):=\frac{q(\{x\in\mathbb{X}:\ \psi(x)\le b\})}{q(\mathbb{X})}.$$

In particular, $F_\psi(0)=q(\{\psi^{-1}(0)\})=0$, and the second equality holds because $\{\psi^{-1}(0)\}$ is a singleton and the measure q is atomless. In addition, $F_\psi(1)=1$ because $\{x\in\mathbb{X}:\ \psi(x)\le 1\}=\mathbb{X}$.

The distribution function F_ψ is continuous. Indeed, first observe that $F_\psi(b)=0$ for $b\le 0$ and $F_\psi(b)=1$ for $b\ge 1$. Second, consider $b\in[0,1]$ and observe that $F_\psi(b-)=q(\{x\in\mathbb{X}:\ \psi(x)<b\})/q(\mathbb{X})$, $b\in\mathbb{R}$. Then $F_\psi(b)-F_\psi(b-)=q(\{\psi^{-1}(b)\})=0$, where the last inequality holds because the set $\{\psi^{-1}(b)\}$ is a singleton and the measure q is atomless.

The continuity of the function F_ψ implies the following two equalities for $\alpha\in[0,1]$:

$$F_\psi^{-1}(\alpha)=[b_{min}(\alpha),b_{max}(\alpha)],$$

where $b_{min}(\alpha):=\inf\{b\ge 0:\ F_\psi(b)=\alpha\}$ and $b_{max}(\alpha):=\sup\{b\le 1:\ F_\psi(b)=\alpha\}$, and

$$q(\psi^{-1}(F_\psi^{-1}(\alpha)))=0. \tag{8.3}$$

We observe that $b_{min}(\alpha) = \inf\{b : \ F_\psi(b) \geq \alpha\}$, and this function is well-studied in the literature under the names of the value-at-risk and quantile function. The function $b_{min}(\alpha)$ is nondecreasing and left continuous on $[0,1]$; see, e.g., Embrechts and Hofert [11, Proposition 1(2)]. Therefore, it is lower semicontinuous. Since F_ψ is a continuous function, the function $b_{min}(\alpha)$ is strictly increasing; see, e.g., [11, Proposition 1(7)].

Let us consider the collection of increasing subsets $\mathbb{X}_\alpha \subset \mathbb{X}$ and $\bar{\mathbb{X}}_\alpha \subset \mathbb{X}$:

$$\begin{aligned} &\mathbb{X}_\alpha := \{x \in \mathbb{X} : \ \psi(x) < b_{min}(\alpha)\}, && \alpha \in [0,1], \\ &\bar{\mathbb{X}}_\alpha := \{x \in \mathbb{X} : \ \psi(x) \leq b_{max}(\alpha)\} = \mathbb{X}_\alpha \cup F_\psi^{-1}(\alpha), && \alpha \in [0,1], \end{aligned} \tag{8.4}$$

and define the deterministic policies φ_α and $\bar{\varphi}_\alpha$:

$$\varphi_\alpha(x) := \begin{cases} \phi^1(x) & \text{if } x \in \mathbb{X}_\alpha, \\ \phi^0(x) & \text{if } x \in \mathbb{X} \setminus \mathbb{X}_\alpha, \end{cases} \qquad \bar{\varphi}_\alpha(x) := \begin{cases} \phi^1(x) & \text{if } x \in \bar{\mathbb{X}}_\alpha, \\ \phi^0(x) & \text{if } x \in \mathbb{X} \setminus \bar{\mathbb{X}}_\alpha. \end{cases} \tag{8.5}$$

Observe that $q(\bar{\mathbb{X}}_\alpha) = q(\mathbb{X})F_\psi(b_{\max}(\alpha)) = q(\mathbb{X})\alpha$, as follows from the definition of $\bar{\mathbb{X}}_\alpha$. According to (8.3),

$$q(\mathbb{X}_\alpha) = q(\bar{\mathbb{X}}_\alpha) = q(\mathbb{X})\alpha. \tag{8.6}$$

Recall that $X(\varphi_\alpha, \bar{\varphi}_\alpha)$ is the set of states on which φ_α and $\bar{\varphi}_\alpha$ make the same decisions; see (7.4). Since $\mathbb{X} \setminus X(\varphi_\alpha, \bar{\varphi}_\alpha) \subset F_\psi^{-1}(\alpha)$, equality (8.3) and Lemma 7.4 imply that $q^{\varphi_\alpha} = q^{\bar{\varphi}_\alpha}$ for all $\alpha \in [0,1]$. By definition, $\phi^0 = \varphi_0$ and $\phi^1 = \bar{\varphi}_1$. Thus, $q^{\phi^0} = q^{\varphi_0}$ and $q^{\phi^1} = q^{\varphi_1}$.

Observe that

$$q(\mathbb{X} \setminus X(\varphi_\alpha, \varphi_{\alpha+\Delta})) = q(\mathbb{X} \setminus X(\bar{\varphi}_\alpha, \bar{\varphi}_{\alpha+\Delta})) = q(\mathbb{X})|\Delta|, \qquad \alpha, \alpha+\Delta \in [0,1], \tag{8.7}$$

where the last equality holds because

$$q(\mathbb{X} \setminus X(\bar{\varphi}_\alpha, \bar{\varphi}_{\alpha+\Delta})) = q(\bar{\mathbb{X}}_\alpha \triangle \bar{\mathbb{X}}_{\alpha+\Delta}) = q(\mathbb{X})|F_\psi(b_{min}(\alpha+\Delta)) - F_\psi(b_{min}(\alpha))|,$$

where $\bar{\mathbb{X}}_\alpha \triangle \bar{\mathbb{X}}_{\alpha+\Delta} := (\bar{\mathbb{X}}_\alpha \cup \bar{\mathbb{X}}_{\alpha+\Delta}) \setminus (\bar{\mathbb{X}}_\alpha \cap \bar{\mathbb{X}}_{\alpha+\Delta})$ is the symmetric difference. Let us define the mapping g,

$$g(\alpha) := Q^{\varphi_\alpha}, \qquad \alpha \in [0,1].$$

As shown above, $g(0) = Q^{\phi^0}$ and $g(1) = Q^{\phi^1}$. Formula (8.7) and Theorem 7.5 imply that this mapping is continuous in total variation. ∎

COROLLARY 8.4. *For a uniformly absorbing atomless MDP the performance set $\mathcal{V}^{\mathbb{F}}$ is connected.*

Proof. Let ϕ^0 and ϕ^1 be deterministic policies. Let us consider the function $g : [0,1] \mapsto \mathcal{M}^{\mathbb{F}}$ satisfying (8.2) for all bounded measurable functions f. The existence of such a function follows from Theorem 8.3. Then the vector-function $\tilde{\zeta}(\alpha) := \int_{\mathbb{X}}\int_{\mathbb{A}} r(x,a)g(\alpha)(dx,da)$ defines a path connecting v^{ϕ^0} and v^{ϕ^1} in $\mathbb{R}^N$. ∎

COROLLARY 8.5. *If $N = 1$, then the set $\mathcal{V}^{\mathbb{F}}$ is convex for a uniformly absorbing atomless MDP.*

Proof. Corollary 8.4 and the mean value theorem imply that the bounded one-dimensional set $\mathcal{V}^{\mathbb{F}}$ is convex. ∎

Corollary 8.6. *If $N = 1$, then $\mathcal{V}^{\mathbb{F}} = \mathcal{V}$ for a uniformly absorbing atomless MDP.*

Proof. Let $v_* := \inf_{\pi\in\mathbb{S}} v^\pi$ and $v^* := \sup_{\pi\in\mathbb{S}} v^\pi$. Then $-\infty < v_* < v^* < +\infty$, where the first and the last inequality hold since the MDP is absorbing and the reward function r is bounded. According to Feinberg [13], $\inf_{\phi\in\mathbb{F}} v^\phi = v_*$ and $\sup_{\phi\in\mathbb{F}} v^\phi = v^*$. These equalities imply that the closures of the one-dimensional convex sets $\mathcal{V}^{\mathbb{F}}$ and $\mathcal{V}$ are both equal to the closed bounded interval $[v_*, v^*]$. In addition, according to Corollary 6.4, if $v \in \{v_*, v^*\} \cap \mathcal{V}$, then $v \in \mathcal{V}^{\mathbb{F}}$. Therefore, $\mathcal{V}^{\mathbb{F}} \supset \mathcal{V}$ and, by definition, $\mathcal{V}^{\mathbb{F}} \subset \mathcal{V}$. ∎

9. Proof of Theorem 3.8.

For the performance set of deterministic policies $\mathcal{V}^{\mathbb{F}}$, consider its closure $\bar{\mathcal{V}}^{\mathbb{F}}$. Since the set $\mathcal{V}^{\mathbb{F}}$ is bounded, $\bar{\mathcal{V}}^{\mathbb{F}}$ is compact.

Lemma 9.1. *Under the assumptions of Theorem* 3.8*, if the set $\mathcal{V}^{\mathbb{F}}$ is convex, then $\mathcal{V} \subset \bar{\mathcal{V}}^{\mathbb{F}}$.*

Proof. Suppose that $\mathcal{V} \not\subset \bar{\mathcal{V}}^{\mathbb{F}}$. Then there exists a stationary policy π such that $v^\pi \notin \bar{\mathcal{V}}^{\mathbb{F}}$. Therefore, there exists a hyperplane in $\mathbb{R}^N$ separating the point v^π and the convex compact set $\bar{\mathcal{V}}^{\mathbb{F}}$. Let $\langle b, v\rangle + d = 0$ be such a hyperplane, and let $\langle b, v^\pi\rangle + d > 0$ and $\langle b, v\rangle + d \le 0$ for all $v \in \mathcal{V}^{\mathbb{F}}$, where $b \in \mathbb{R}^N$ and $d \in \mathbb{R}$. Thus

$$\sup_{\phi\in\mathbb{F}} \langle b, v^\phi\rangle = \sup_{v\in\mathcal{V}^{\mathbb{F}}} \langle b, v\rangle < \langle b, v^\pi\rangle. \tag{9.1}$$

Let us consider the reward function $\tilde{r}(x,a) := \langle b, r(x,a)\rangle$, where $x \in \mathbb{X}$ and $a \in A(x)$. The expected total rewards for this reward function, a policy σ, and the initial state distribution μ is denoted by $\tilde{v}^\sigma$, and $\tilde{v}^\sigma = \langle b, v^\sigma\rangle$ for all $\sigma \in \mathbb{S}$.

Supremums of the expected total rewards are equal for deterministic and stationary policies; see Feinberg [13]. Therefore, $\sup_{v\in\mathcal{V}^{\mathbb{F}}} \langle b, v\rangle = \sup_{\phi\in\mathbb{F}} \tilde{v}^\phi \ge \tilde{v}^\pi = \langle b, v^\pi\rangle$. This contradicts (9.1). ∎

Lemma 9.2. *Let the statement of Theorem* 3.8 *be correct for $N = 1, 2, \dots$ criteria. Then, under the assumptions of Theorem* 3.8*, the set $\mathcal{V}^{\mathbb{F}}$ is convex for the case of $(N+1)$ criteria.*

Proof. Let the lemma be correct for N-dimensional vector-functions r, where $N = 1, 2, \dots$. We shall prove that the set $\mathcal{V}^{\mathbb{F}}$ is convex for $(N+1)$-dimensional vector-functions r. Let ϕ^0 and ϕ^1 be two deterministic policies, and let $\lambda \in (0,1)$. Our goal is to show that there exists a deterministic policy ϕ_λ such that $v^{\phi_\lambda} := \lambda v^{\phi^0} + (1-\lambda) v^{\phi^1}$. Let us consider the stationary policy π^* defined in (7.1), the measure q on $\mathbb{X}$ defined in (7.2), and the family of expanding sets $\mathbb{X}_\alpha \subset \mathbb{X}$ defined in (8.4). For each $\alpha \in [0,1]$ we consider the submodel with the action sets reduced to the sets

$$A^\alpha(x) = \begin{cases} \{\phi^1(x)\} & \text{if } x \in \mathbb{X}_\alpha, \\ \{\phi^0(x), \phi^1(x)\} & \text{if } x \in \mathbb{X} \setminus \mathbb{X}_\alpha. \end{cases}$$

Let $\mathcal{V}(\alpha)$ be the set of all performance vectors for the submodel with the action sets $A^\alpha(\cdot)$. According to Lemma 4.1 and Corollary 5.3, each set $\mathcal{V}(\alpha)$ is convex and compact. In addition,

$$\mathcal{V}(\alpha) \subset \mathcal{V}(\beta) \qquad \text{if } 0 \le \beta \le \alpha \le 1. \tag{9.2}$$

In view of the definition in (8.4), $\mathbb{X}_0 = \emptyset$, which implies

$$A^0(x) = \{\phi^0(x), \phi^1(x)\}, \qquad x \in \mathbb{X}.$$

Therefore $\mathcal{V}(0)$ is the performance set for the MDP defined by the deterministic policies ϕ^0 and ϕ^1. Thus, $v^{\phi^0}, v^{\phi^1} \in \mathcal{V}(0)$.

Observe that $\mathcal{V}(1) = \{v^{\phi^1}\}$. Indeed, let φ be a deterministic policy for the MDP with the action sets $A^1(\cdot)$. Then $\varphi(x) = \phi^1$ when $x \in \mathbb{X}_1 \subset \mathbb{X}$. In view of (8.6), $q(\mathbb{X} \backslash \mathbb{X}_1) = 0$. Since $\mathbb{X} \backslash X(\varphi, \phi^1) \subset \mathbb{X} \backslash \mathbb{X}_1$, we have that $q(\mathbb{X} \backslash X(\varphi, \phi^1)) = 0$. Lemma 7.4 implies that $q^\varphi = q^{\phi^1}$. Therefore, $v^\varphi = \int_{\mathbb{X}} r(x, \varphi(x)) q^\varphi(dx) = \int_{\mathbb{X}} r(x, \phi^1(x)) q^{\phi^1}(dx) = v^{\phi^1}$, where the first and the last equalities follow from the definitions of expected total rewards, occupancy measures, and deterministic policies; the equality in the middle follows from $q^\varphi = q^{\phi^1}$ and $\varphi(x) = \phi^1(x)$ for q^{ϕ^1}-almost all $x \in \mathbb{X}$.

Since the set $\mathcal{V}(0)$ is convex and $v^{\phi^0}, v^{\phi^1} \in \mathcal{V}(0)$, we have that $\lambda v^{\phi^0} + (1-\lambda) v^{\phi^1} \in \mathcal{V}(0)$. Consider an arbitrary point $\hat{v} \in \mathcal{V}(0)$. We shall prove that $v^\phi = \hat{v}$ for some deterministic policy ϕ for the submodel with action sets $A^0(x)$, $x \in \mathbb{X}$.

To do this, we'll show that $\hat{v} \in \partial(\mathcal{V}(\hat{\alpha}))$ for some $\hat{\alpha} \in [0,1]$, where $\partial(G)$ is the boundary of the convex compact subset G of $\mathbb{R}^{N+1}$. For a point $e \in \mathbb{R}^{N+1}$ and a closed set $E \subset \mathbb{R}^{N+1}$, we denote by $d(e, E) := \min\{\|e - z\| : z \in E\}$ the distance between e and E. Since E is closed, $d(e, E) = 0$ if and only if $e \in E$. If $E_1 \subset E_2$ for two closed subsets of $\mathbb{R}^{N+1}$, then $d(e, E_2) \le d(e, E_1)$.

As follows from (9.2), the function

$$G(\alpha) := d(\hat{v}, \mathcal{V}(\alpha)), \qquad \alpha \in [0,1],$$

is nondecreasing in $\alpha \in [0,1]$ and $G(0) = d(\hat{v}, \mathcal{V}(0)) = 0$. Let us prove that this function is continuous. To do this, we choose an arbitrary $\alpha \in [0,1)$ and $\Delta > 0$ such that $\alpha + \Delta \le 1$. We also choose an arbitrary point $v \in \mathcal{V}(\alpha)$. Let π be a stationary policy in the submodel with the action sets $A^\alpha(x)$, $x \in \mathbb{X}$, such that $v^\pi = v$. Let σ be the stationary policy in the model with the action sets $A^{\alpha+\Delta}(x)$, $x \in \mathbb{X}$, defined by

$$\sigma(\phi^1(x)|x) := \begin{cases} 1 & \text{if } x \in \mathbb{X}_{\alpha+\Delta} \setminus \mathbb{X}_\alpha, \\ \pi(\phi^1(x)|x) & \text{if } x \in \mathbb{X} \setminus (\mathbb{X}_{\alpha+\Delta} \setminus \mathbb{X}_\alpha). \end{cases}$$

Then $\mathbb{X} \setminus (\mathbb{X}_{\alpha+\Delta} \setminus \mathbb{X}_\alpha) \subset X(\pi, \sigma)$, which implies $\mathbb{X} \setminus X(\pi, \sigma) \subset \mathbb{X}_{\alpha+\Delta} \setminus \mathbb{X}_\alpha$. As follows from (8.6), $q(\mathbb{X} \setminus X(\pi, \sigma)) \le q(\mathbb{X})\Delta$. According to Theorem 7.5, for every $\epsilon_1 > 0$ there exists $\delta > 0$ such that $d_{TV}(Q^\pi, Q^\sigma) \le \epsilon_1$ if $\Delta \le \delta$. This implies that $\|v^\pi - v^\sigma\| \le K(N+1)^{\frac{1}{2}} \epsilon_1$, where the positive constant K is an upper bound of $|r^{(n)}(x, a)|$ for $x \in \mathbb{X}$, $a \in \mathbb{A}$, and $n = 1, 2, \ldots, N+1$. So, if we choose an arbitrary $\epsilon > 0$, set $\epsilon_1 = \epsilon / (K(N+1)^{\frac{1}{2}})$, and choose $\Delta \le \delta$, then $\|v^\pi - v^\sigma\| \le \epsilon$. This implies that if $\Delta \le \delta$, $\alpha \in [0,1)$, and $\alpha + \delta \le 1$, then

$$d(v, \mathcal{V}(\alpha + \Delta)) \le \epsilon \qquad \text{for all } v \in \mathcal{V}(\alpha). \tag{9.3}$$

Let us consider two cases: (i) $G(\alpha) > 0$ and (ii) $G(\alpha) = 0$.

(i) In this case, $\hat{v} \notin \mathcal{V}(\alpha)$. We denote by $\hat{v}_\alpha$ the projection of the point $\hat{v}$ onto the convex compact set $\mathcal{V}(\alpha)$; that is, $\hat{v}_\alpha$ is the unique point in $\mathcal{V}(\alpha)$ satisfying $\|\hat{v} - \hat{v}_\alpha\| = d(\hat{v}, \mathcal{V}(\alpha))$. Let $\hat{v}_{\alpha+\Delta} \in \mathcal{V}(\alpha+\Delta)$ be the projection of $\hat{v}_\alpha$ onto the compact set $\mathcal{V}(\alpha+\Delta)$. Then, according to the triangle inequality

$$d(\hat{v}, \mathcal{V}(\alpha)) + d(\hat{v}_\alpha, V(\alpha+\Delta)) = \|\hat{v} - \hat{v}_\alpha\| + \|\hat{v}_\alpha - \hat{v}_{\alpha+\Delta}\| \ge \|\hat{v} - \hat{v}_{\alpha+\Delta}\| \ge d(\hat{v}, \mathcal{V}(\alpha+\Delta)).$$

Since $0 < d(\hat{v}_\alpha, V(\alpha + \Delta)) < \epsilon$ and the nonnegative function $G(\alpha)$ is nondecreasing, the last formula implies

$$0 \le G(\alpha + \Delta) - G(\alpha) \le \epsilon. \tag{9.4}$$

(ii) The equality $G(\alpha) = 0$ means that $\hat{v} \in \mathcal{V}(\alpha)$. Therefore, (9.3) for $v = \hat{v}$ implies $0 \le G(\alpha + \Delta) - G(\alpha) = G(\alpha + \Delta) \le \epsilon$. So, (9.4) holds.

Since (9.4) holds for the both cases, this implies continuity of the function $G(\alpha)$ on $[0, 1]$. Let us define

$$\hat{\alpha} := \max\{\alpha \in [0,1] : d(\hat{v}, \mathcal{V}(\alpha)) = 0\}.$$

This point exists because $d(\hat{v}, \mathcal{V}(0)) = 0$ and the continuous function $G(\alpha) = d(\hat{v}, \mathcal{V}(\alpha))$ is nondecreasing in α. Since $d(\hat{v}, \mathcal{V}(\hat{\alpha})) = 0$, we have that $\hat{v} \in \mathcal{V}(\hat{\alpha})$. If $\hat{\alpha} = 1$, then $\hat{v} = v^{\phi^1} \in \mathcal{V}(1) = \partial\mathcal{V}(1)$ since $\mathcal{V}(1) = \{v^{\phi^1}\}$.

So, we need to consider the case $\hat{\alpha} \in [0, 1)$. In this case we shall prove that $\hat{v} \in \partial(\mathcal{V}(\hat{\alpha}))$.

Since $\hat{v} \in \mathcal{V}(\hat{\alpha})$, in order to prove that $\hat{v} \in \partial(\mathcal{V}(\hat{\alpha}))$, it is sufficient to show that $\hat{v}$ cannot be an interior point of $\mathcal{V}(\hat{\alpha})$. Indeed, let $\hat{v}$ be an interior point of $\mathcal{V}(\hat{\alpha})$. Then there exists $\epsilon > 0$ such that $d(\hat{v}, \partial(\mathcal{V}(\hat{\alpha}))) \ge \epsilon$. In view of (9.3) for $\alpha = \hat{\alpha}$, there exists $\Delta > 0$ such that $\hat{\alpha} + \Delta \le 1$ and $d(v, \mathcal{V}(\hat{\alpha} + \Delta)) \le \epsilon/2$ for all $v \in \mathcal{V}(\hat{\alpha})$. Thus, $d(\hat{v}, \mathcal{V}(\hat{\alpha} + \Delta)) \le \epsilon/2$. The definition of $\hat{\alpha}$ implies that $d(\hat{v}, \mathcal{V}(\hat{\alpha} + \Delta)) > 0$. Let $\hat{v}_p$ be the projection of the point $\hat{v}$ onto the convex set $\mathcal{V}(\hat{\alpha} + \Delta)$. Observe that $\hat{v}_p$ is an interior point of $\mathcal{V}(\hat{\alpha})$ because $\|\hat{v} - \hat{v}_p\| = d(\hat{v}, \mathcal{V}(\hat{\alpha} + \Delta)) \le \epsilon/2$. Since $\hat{v}$ and $\hat{v}_p$ are interior points of $\mathcal{V}(\hat{\alpha})$, there is a point $v \in \partial(\mathcal{V}(\hat{\alpha}))$ such that v belongs to the line projecting $\hat{v}$ to $\mathcal{V}(\hat{\alpha} + \Delta)$, and $\hat{v}$ is located between $\hat{v}_p$ and v. This is illustrated in Figure 9.1. Therefore $d(v, \mathcal{V}(\hat{\alpha}+\Delta)) = \|v - \hat{v}_p\| \ge \|v - \hat{v}\| \ge d(\hat{v}, \partial(\mathcal{V}(\hat{\alpha}))) \ge \epsilon$, where the first inequality holds because $\hat{v}$ is between v and $\hat{v}_p$, the second inequality follows from $v \in \partial(\mathcal{V}(\hat{\alpha}))$, and the last one follows from the choice of ϵ. This conclusion is a contradiction to $d(v, \mathcal{V}(\hat{\alpha} + \Delta)) \le \epsilon/2$. Therefore, $v \in \partial(\mathcal{V}(\hat{\alpha}))$.

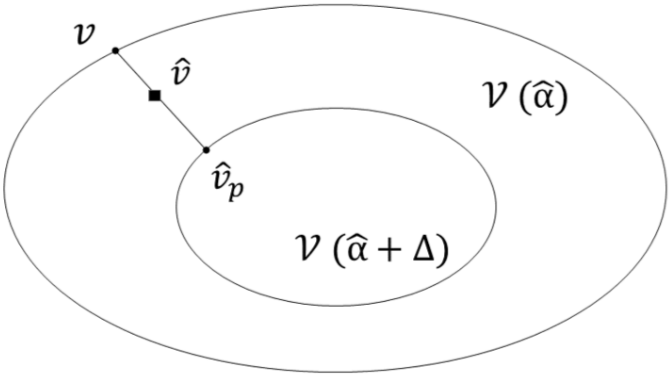
v
$\hat{v}$
$\mathcal{V}(\hat{\alpha})$
$\hat{v}_p$
$\mathcal{V}(\hat{\alpha} + \Delta)$

FIG. 9.1. *$\hat{v}$ cannot be an interior point of $\mathcal{V}(\hat{\alpha})$: Otherwise, $d(v, \mathcal{V}(\hat{\alpha} + \Delta)) = \|v - \hat{v}_p\| \ge \|v - \hat{v}\| \ge d(\hat{v}, \partial(\mathcal{V}(\hat{\alpha}))) \ge \epsilon$, and $d(v, \mathcal{V}(\hat{\alpha} + \Delta)) \le \epsilon/2$ (a contradiction).*

Since $\hat{v} \in \partial(\mathcal{V}(\hat{\alpha}))$, by Theorem 6.6 there is a coordinate $i = 1, \ldots, N + 1$ such that $\hat{v}_{-i}$ is a performance vector in a submodel of the MDP with action sets $A^{\hat{\alpha}}(\cdot)$ and the value of $v^{(i)}$ is completely defined by the vector $\hat{v}_{-i}$ according to formula (6.6). The vector $\hat{v}_{-i}$ has N coordinates. By the induction assumption, there is a deterministic policy ϕ such that $v^{\phi}_{-i} = \hat{v}_{-i}$. Thus, $v^{\phi} = \hat{v}$. □

Proof of Theorem 3.8. According to Corollary 8.6, the statement of the theorem is correct for $N = 1$. Suppose the statement of Theorem 3.8 is correct for N criteria, where $N = 1, 2, \ldots$. Let us prove that it is correct for the case of $(N + 1)$ criteria.

Consider the case of $(N+1)$ criteria. By Lemma 9.2, the set $\mathcal{V}^{\mathbb{F}}$ is convex. Therefore, Lemma 9.1 and $\mathcal{V}^{\mathbb{F}} \subset \mathcal{V}$ imply that if $v \in \mathcal{V} \setminus \partial(\mathcal{V})$, then $v \in \mathcal{V}^{\mathbb{F}}$. Let $v \in \partial(\mathcal{V})$. Theorem 6.6 implies that there exist a coordinate $i = 1, \ldots, N+1$, a vector $b \in \mathbb{R}^N$, a constant d, and a submodel with the performance set $\tilde{\mathcal{V}}$ such that $v \in \tilde{\mathcal{V}}$ and $\tilde{v}^{(i)} = d + \langle b, \tilde{v}_{-i} \rangle$ for all $\tilde{v} \in \tilde{\mathcal{V}}$, where for $w \in \mathbb{R}^N$ the following notation is used: $w^{(i)}$ is the ith coordinate of the vector w and w_{-i} is the projection of w onto $\mathbb{R}^N$ obtained by removing the ith coordinate from w. As follows from the induction assumption, there is a deterministic policy ϕ in the submodel such that $v^{\phi}_{-i} = v_{-i}$ and $v^{(i),\phi} = d + \langle b, v^{\phi}_{-i} \rangle = d + \langle b, v_{-i} \rangle = v^{(i)}$. Thus, $v^{\phi} = v$. □

10. Unbounded rewards. This section describes extensions to unbounded reward vector-functions r. These extensions are based on the standard weighted norm transformation of an MDP with unbounded rewards to an MDP with bounded rewards.

Let us consider an MDP with the expected total rewards and with a standard Borel state space $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, where $\bar{x} \notin \mathbb{X}$, a standard Borel action space $\mathbb{A}$, sets of available actions $A(x)$, where $A(\bar{x}) = \{\bar{a}\}$, with $\bar{a}$ being an arbitrary point in $\mathbb{A}$, transition probabilities p such that $p(\bar{x}|\bar{x}, \bar{a}) = 1$, a reward vector-function r with values in $\mathbb{R}^N$ such that $r^{(n)}(\bar{x}, \bar{a}) = 0$, $n = 1, 2, \ldots, N$, and an initial probability distribution μ such that $\mu(\mathbb{X}) = 1$. Let there exist a positive measurable function $w : \mathbb{X} \mapsto (0, +\infty)$ for which the following conditions hold:

(a) $\sup_{x \in \mathbb{X}} \sup_{a \in A(x)} \frac{1}{w(x)} \int_{\mathbb{X}} w(y) p(dy|x, a) \leq 1$,

(b) $\int_{\mathbb{X}} w(x) \mu(dx) < +\infty$,

(c) $\sup_{x \in \mathbb{X}} \sup_{a \in A(x)} \frac{|r^{(n)}(x,a)|}{w(x)} < +\infty$, $n = 1, 2, \ldots, N$.

Let us consider an MDP with state space $\bar{\mathbb{X}}$, action space $\mathbb{A}$, sets of available action $A(x)$, $x \in \bar{\mathbb{X}}$, transition probability $\tilde{p}$, where $\tilde{p}(\bar{x}|\bar{x}, \bar{a}) := 1$,

$$\tilde{p}(Y|x, a) := \frac{1}{w(x)} \int_Y w(y) p(dy|x, a), \qquad Y \in \mathcal{X},\ x \in \mathbb{X},\ a \in A(x),$$

and

$$\tilde{p}(\bar{x}|x, a) := 1 - \frac{1}{w(x)} \int_{\mathbb{X}} w(y) p(dy|x, a), \qquad x \in \mathbb{X},\ a \in A(x),$$

reward function $\tilde{r}$, where $\tilde{r}^{(n)}(\bar{x}, \bar{a}) = 0$ and, for $n = 1, 2, \ldots, N$,

$$\tilde{r}^{(n)}(x, a) = \frac{r^{(n)}(x, a)}{w(x)} \int_{\mathbb{X}} w(y) \mu(dy), \qquad x \in \mathbb{X},\ a \in A(x),$$

and the initial probability distribution $\tilde{\mu}$ with

$$\tilde{\mu}(Y) := \frac{\int_Y w(x) \mu(dx)}{\int_{\mathbb{X}} w(y) \mu(dy)}, \qquad Y \in \mathcal{X}, \tag{10.1}$$

and $\mu(\bar{x}) = 0$. If $\mu(x) = 0$, then $\tilde{\mu}(x) = 0$, $x \in \mathbb{X}$. Let $\tilde{v}^{\pi}$ be the vector of the total expected rewards in the MDP with the transition probabilities $\tilde{p}$ and rewards $\tilde{r}$ controlled by a policy π, when the initial state distribution is $\tilde{\mu}$.

We say that the defined MDP is uniformly absorbing if equality (3.5) holds for this MDP with the initial distribution $\tilde{\mu}$ instead of μ and the transition probability $\tilde{p}$ instead of p. This definition is consistent with Definition 3.6 because the assumptions in Definition 3.3 also hold for this MDP with the fixed initial state distribution $\tilde{\mu}$. In addition, the function $\tilde{r}$ is bounded. The following statement follows from Theorem 3.8.

Corollary 10.1. *Consider an MDP with the state space $\bar{\mathbb{X}}$ satisfying conditions* (a)–(c) *and such that $r(\bar{x},\bar{a}) = 0$ and $p(\bar{x}|\bar{x},\bar{a}) = 1$ for the state $\bar{x}$ and action $\bar{a}$ defined above. Then $\tilde{v}^\pi = v^\pi$ for all policies π. Furthermore, if the MDP with the transition probabilities $\tilde{p}$ is uniformly absorbing and atomless, then $\mathcal{V} = \mathcal{V}^{\mathbb{F}}$ for the initial MDP and this set is convex.*

Proof. Let $\tilde{E}$ and $\tilde{P}$ denote the expectations and probabilities for the MDP with the transition probabilities $\tilde{p}$ and the initial distribution $\tilde{\mu}$. $P^\pi(dx_0da_0,\dots,dx_tda_t)$ and $\tilde{P}^\pi_{\tilde{\mu}}(dx_0da_0,\dots,dx_tda_t)$ are probability distributions on the standard Borel space $(\bar{\mathbb{X}}\times\mathbb{A})^{t+1}$, where $t = 0,1,\dots$. The standard straightforward arguments imply that for $(x_0,a_0,\dots,x_t,a_t) \in (\mathbb{X}\times\mathbb{A})^{t+1}$, $t = 0,1,\dots,$

$$\tilde{P}^\pi_{\tilde{\mu}}(dx_0da_0,\dots,x_ta_t) = \frac{w(x_t)P^\pi(dx_0da_0,\dots,x_ta_t)}{\int_{\mathbb{X}} w(y)\mu(dy)}. \tag{10.2}$$

Since $p(\bar{x}|\bar{x},\bar{a}) = \tilde{p}(\bar{x}|\bar{x},\bar{a}) = 1$ and $r(\bar{x},\bar{a}) = \tilde{r}(\bar{x},\bar{a}) = 0$, equality (10.2) and the definition of the reward function $\tilde{r}$ imply that $E^\pi r(x_t,a_t) = \tilde{E}^\pi_{\tilde{\mu}}\tilde{r}(x_t,a_t)$ for all $t = 0,1,\dots$. This equality implies that $\tilde{v}^\pi = v^\pi$ for an arbitrary policy π. This implies that $\mathcal{V}^{\mathbb{G}} = \{\tilde{v}^\pi : \pi \in \mathbb{G}\}$ for every set of policies $\mathbb{G} \subset \Pi$. Since the MDP with the transition probabilities $\tilde{p}$ is uniformly absorbing and the reward vector-function $\tilde{r}$ is bounded, Theorem 3.8 implies $\{\tilde{v}^\phi : \phi \in \mathbb{F}\} = \{\tilde{v}^\pi : \pi \in \Pi\}$. Therefore, $\mathcal{V} = \mathcal{V}^{\mathbb{F}}$. □

Now let us consider a discounted MDP with the state space $\mathbb{X}$ introduced in section 2 without assuming that the reward vector-function r is bounded. Let us consider the following assumption:

(d) There exists a positive measurable function $w : \mathbb{X} \mapsto (0,+\infty)$ satisfying assumptions (b) and (c), and there exists a constant $\tilde{\beta} \in (0,1)$ such that

$$\beta \sup_{x\in\mathbb{X}} \sup_{a\in A(x)} \frac{1}{w(x)} \int_{\mathbb{X}} w(y)p(dy|x,a) \le \tilde{\beta}.$$

Then the following corollary from Theorem 2.3 holds.

Corollary 10.2. *If an atomless discounted MDP with a possibly unbounded reward vector-function r satisfies assumption* (d)*, then $\mathcal{V}_\beta = \mathcal{V}^{\mathbb{F}}_\beta$ and this set is convex.*

Proof. Let us add an isolated point $\bar{x}$ to the standard Borel space $\mathbb{X}$ and set $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$. Let us consider a discounted MDP with the action set $\mathbb{A}$, sets of available actions $A(x)$, $x \in \mathbb{X}$, reward vector-function $\tilde{r}$, initial state distribution $\tilde{\mu}$, and the discount factor $\tilde{\beta}$ described and defined above. However, instead of $\tilde{p}$, the transition probability for this MDP is $\hat{p}$, where $\hat{p}(\bar{x}|\bar{x},\bar{a}) := 1$,

$$\hat{p}(Y|x,a) := \frac{\beta}{\tilde{\beta}w(x)} \int_Y w(y)p(dy|x,a), \qquad Y \in \mathcal{X},\ x \in \mathbb{X},\ a \in A(x),$$

and

$$\hat{p}(\bar{x}|x,a) := 1 - \frac{\beta}{\tilde{\beta}w(x)} \int_{\mathbb{X}} w(y)p(dy|x,a), \qquad x \in \mathbb{X},\ a \in A(x).$$

Let $\hat{E}$ and $\hat{P}$ denote the expectations and probabilities for the defined MDP with the state space $\bar{\mathbb{X}}$ and transition probabilities $\hat{p}$. In particular, $\tilde{P}^\pi_{\tilde{\mu}}(dx_0da_0,\dots,dx_tda_t)$ is a probability distribution on the standard Borel space $(\bar{\mathbb{X}}\times\mathbb{A})^{t+1}$, where $t = 0,1,\dots$. The following formula is similar to (10.2): For $t = 0,1,\dots$ and $(x_0,a_0,\dots,x_t,a_t) \in$

$(\mathbb{X}\times\mathbb{A})^{t+1}$,

$$\tilde{\beta}^t\hat{P}^{\pi}_{\tilde{\mu}}(dx_0da_0,\dots,dx_tda_t)=\frac{\beta^t w(x_t)P^{\pi}(dx_0da_0,\dots,dx_tda_t)}{\int_{\mathbb{X}}w(y)\mu(dy)}. \tag{10.3}$$

Since $\hat{p}(\bar{x}|\bar{x},\bar{a})=1$ and $\tilde{r}(\bar{x},\bar{a})=0$, equality (10.3) and the definition of the reward function $\tilde{r}$ imply that $\beta^tE^{\pi}r(x_t,a_t)=\tilde{\beta}^t\tilde{E}^{\pi}_{\tilde{\mu}}\tilde{r}(x_t,a_t)$ for all $t=0,1,\dots$. This equality implies that $\tilde{v}^{\pi}_{\tilde{\beta}}=v^{\pi}_{\beta}$ for an arbitrary policy π, where $\tilde{v}^{\pi}_{\tilde{\beta}}$ is the vector of the total discounted expected rewards in the MDP with the transition probabilities $\hat{p}$ and discount factor $\tilde{\beta}$, when a policy π is chosen and the initial state distribution is $\tilde{\mu}$. This implies that $\mathcal{V}^{\mathbb{G}}_{\beta}=\{\tilde{v}^{\pi}_{\tilde{\beta}}:\pi\in\mathbb{G}\}$ for every set of policies $\mathbb{G}\subset\Pi$. Since the reward vector-function $\tilde{r}$ is bounded, Theorem 2.3 implies that $\{\tilde{v}^{\phi}_{\tilde{\beta}}:\phi\in\mathbb{F}\}=\{\tilde{v}^{\pi}_{\tilde{\beta}}:\pi\in\Pi\}$. Therefore, $\mathcal{V}_{\beta}=\mathcal{V}^{\mathbb{F}}_{\beta}$. □

Corollary 10.2 can also be proved by reducing discounted MDPs with discounted factors β and $\tilde{\beta}$ to undiscounted MDPs, as is done in the proof of Lemma 3.12, and by applying Corollary 10.1.

11. Compactness of performance sets and Lyapunov's convexity theorem.

In this section we describe sufficient conditions for the compactness of the sets $\mathcal{V}$ and $\mathcal{V}^{\mathbb{F}}$ and discuss the relation of our results to Lyapunov's convexity theorem. From an intuitive point of view, it is clear that the set of the ranges of vector-measures is a particular case of the sets $\mathcal{V}$ and $\mathcal{V}^{\mathbb{F}}$, when a one-step problem is considered. We demonstrate this in Example 11.2. The following example shows that the set $\mathcal{V}$ may be noncompact.

Example 11.1. Let $\mathbb{X}:=[0,1]$, $A(x):=\mathbb{A}:=(0,1)$, $r(x,a)=a$, and let μ be the Lebesgue measure on $[0,1]$. Under every decision the process moves from every state $x\in\mathbb{X}$ to an absorbing state. For every deterministic policy ϕ, we have that $v^{\phi}=\int_0^1\phi(x)dx$, where $\phi:[0,1]\mapsto(0,1)$ is an arbitrary Borel function. In this example, $\mathcal{V}^{\mathbb{F}}=(0,1)$. Since this MDP is uniformly absorbing and atomless, $\mathcal{V}=\mathcal{V}^{\mathbb{F}}=(0,1)$. By changing the action sets to $(0,1]$, $[0,1)$, and $[0,1]$, we obtain MDPs with performance sets $(0,1]$, $[0,1)$, and $[0,1]$, respectively.

As stated in Corollary 5.2, Condition (S) from section 5 is sufficient for the compactness of $\mathcal{V}$. For example, in Example 11.1 this condition holds when $A(x)=\mathbb{A}=[0,1]$, $x\in\mathbb{X}$. Condition (S) always holds when all the action sets $A(x)$ are finite. Another sufficient condition (W) for the compactness of the set of strategic measures was introduced by Schäl [31]. This condition assumes weak continuity of transition probabilities. Being combined with continuity of the bounded reward vector-functions $r:\mathbb{X}\times\mathbb{A}\mapsto\mathbb{R}^N$, this weak continuity condition implies compactness of the performance set $\mathcal{V}$. Condition (W) was used in Feinberg and Piunovskiy [18]. We neither use nor consider weak continuity condition (W) in this paper. In general, a measure ν is called atomless if for any measurable set E with $\nu(E)>0$ there exists a measurable subset E' of E such that $\nu(E)>\nu(E')>0$. A vector-measure is called atomless if each of its coordinates is an atomless measure.

Lyapunov's convexity theorem states that the range of a finite atomless vector-measure is convex and compact. In other words, if $(\mathbb{X},\mathcal{X})$ is a measurable space and ν is a finite atomless vector-measure with values in $\mathbb{R}^N$, then the set $\mathcal{W}:=\{\nu(B):B\in\mathcal{X}\}$ is a compact and convex subset of $\mathbb{R}^N$.

One of the equivalent formulations of this version of Lyapunov's convexity theorem (see, e.g., Blackwell [4]) states that if μ is a finite atomless measure on a measurable space $(\mathbb{X},\mathcal{X})$ and $r:(\mathbb{X},\mathcal{X})\mapsto(\mathbb{R}^N,\mathcal{B}(\mathbb{R}^N))$ is a measurable vector-function whose coordinates are nonnegative functions satisfying $\int_{\mathbb{X}} r^{(n)}(x)\mu(dx)<+\infty$, where $n=1,\dots,N$, then the set $\mathcal{W}^*:=\{\int_B r(x)\mu(dx): B\in\mathcal{X}\}$ is a compact and convex subset of $\mathbb{R}^N$.

To see that the classic Lyapunov convexity theorem is equivalent to this statement, for an atomless vector-measure $\nu=(\nu^{(1)},\dots,\nu^{(N)})$, define the atomless measure $\mu=\sum_{n=1}^N\nu^{(n)}$. Since $\nu^{(n)}\ll\mu$, there are Radon–Nikodym derivatives $r^{(n)}:=d\nu^{(n)}/d\mu$, $n=1,\dots,N$. Therefore, $\nu(B)=\int_B r(x)\mu(dx)$ for all $\in\mathcal{X}$, and $\mathcal{W}=\mathcal{W}^*$. Conversely, for an atomless finite measure μ and the vector-function r described in the previous paragraph, $\nu(B)=\int_B r(x)\mu(dx)$, where $B\in\mathcal{X}$ is the atomless vector-measure, and $\mathcal{W}^*=\mathcal{W}$ is its range.

The following example demonstrates that Theorem 3.8 and Corollaries 5.3 and 10.1 imply Lyapunov's convexity theorem for the case when an atomless measure is defined on a standard Borel space.

Example 11.2. Let us consider an MDP with a state space $\bar{\mathbb{X}}=\mathbb{X}\cup\{\bar{x}\}$, where $\mathbb{X}$ is a standard Borel space, action sets $A(x):=\mathbb{A}:=\{0,1\}$ and $A(\bar{x})=\{0\}$, rewards $r:\mathbb{X}\times\mathbb{A}\mapsto\mathbb{R}^N$, and let μ be an atomless initial probability measure on $\mathbb{X}$. We also set $p(\bar{x}|x,a)=1$ for all $x\in\bar{X}$ and $a\in A(x)$. That is, from each state x the process moves to the absorbing state $\bar{x}$. We also set $r(x,0):=\bar{0}$ for all $x\in\bar{X}$, where $\bar{0}$ is the zero-vector in $\mathbb{R}^N$, and $r(x,1):=r(x)$, $x\in X$, where $r=(r^{(1)},\dots,r^{(N)})$ is a Borel vector-function such that each coordinate function $r^{(n)}$ is nonnegative and $\int_{\mathbb{X}} r^{(n)}(x)\mu(dx)<+\infty$ for all $n=1,\dots,N$.

Every deterministic policy $\phi\in\mathbb{F}$ is defined by the set $B^\phi:=\{x\in\mathbb{X}:\phi(x)=1\}$. Observe that $v^\phi=\int_{B^\phi} r(x)\mu(dx)$. In addition, $\{B^\phi:\phi\in\mathbb{F}\}$ is the Borel σ-algebra on $\mathbb{X}$. Thus, we are in the framework of the equivalent formulation of Lyapunov's convexity theorem, and $\mathcal{W}^*=\mathcal{V}^{\mathbb{F}}$. Since the function r can be unbounded, we define the weight function $w(x):=1+\sum_{n=1}^N|r^{(n)}(x)|$, $x\in\mathbb{X}$.

Then $v^\phi=\int_{B^\phi}\tilde{r}(x)\tilde{\mu}(dx)$, where the measure $\tilde{\mu}$ is defined in (10.1) and the vector-function $\tilde{r}(x):=r(x)(w(x))^{-1}\int_{\mathbb{X}} w(y)\mu(dy)$, $x\in\mathbb{X}$, is bounded. Therefore, in view of Corollary 10.1, $\mathcal{V}^{\mathbb{F}}=\mathcal{V}$ and this set is closed and compact. The compactness of the set $\mathcal{V}$ follows from Corollaries 5.3 and 10.1. The set $\mathcal{W}^*=\mathcal{V}^{\mathbb{F}}$ is convex and compact. Thus, Lyapunov's convexity theorem for a standard Borel space $\mathbb{X}$ is a particular example of an application of Corollary 10.1, which follows from Theorem 3.8.

REFERENCES

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

[2] E. J. Balder, *On compactness of the space of policies in dynamic programming*, Stochastic Proc. Appl., 32 (1989), pp. 141–150.

[3] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control*, Athena Scientific, Belmont, MA, 1996.

[4] D. Blackwell, *On a theorem of Lyapunov*, Ann. Math. Statistics, 22 (1951), pp. 112–114.

[5] V. I. Bogachev, *Measure Theory*, Vol. I, Springer, Berlin, 2007.

[6] V. S. Borkar, *A convex analytic approach to Markov decision processes*, Probab. Theory Related Fields, 79 (1988), pp. 642–657.

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

[8] A. Dvoretzky, A. Wald, and J. Wolfowitz, *Elimination of randomization in certain problems of statistics and of the theory of games*, Proc. Natl. Acad. Sci. USA, 36 (1950), pp. 256–260.

[9] A. Dvoretzky, A. Wald, and J. Wolfowitz, *Elimination of randomization in certain statistical procedures and zero-sum two-person games*, Ann. Math. Statist., 22 (1951), pp. 1–21.

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

[11] P. Embrechts and M. Hofert, *A note on generalized inverses*, Math. Methods Oper. Res., 77 (2013), pp. 423–432.

[12] E. A. Feinberg, *Nonrandomized Markov and semi-Markov strategies in dynamic programming*, Theory Probab. Appl., 27 (1982), pp. 116–126.

[13] E. A. Feinberg, *On stationary strategies for Borel dynamic programming*, Math. Oper. Res., 17 (1992), pp. 392–397.

[14] E. A. Feinberg and J. Huang, *On the reduction of total-cost and average-cost MDPs to discounted MDPs*, Naval Res. Logist., (2017), https://doi.org/10.1002/nav.21743.

[15] E. A. Feinberg and J. Huang, *Reduction of total-cost and average-cost MDPs with weakly continuous transition probabilities to discounted MDPs*, Oper. Res. Lett., 46 (2018), pp. 179–184.

[16] E. A. Feinberg, P. O. Kasyanov, and M. Z. Zgurovsky, *Convergence of probability measures and Markov decision models with incomplete information*, Proc. Steklov Inst. Math., 287 (2014), pp. 96–117.

[17] E. A. Feinberg, P. O. Kasyanov, and M. Z. Zgurovsky, *Uniform Fatou's lemma*, J. Math. Anal. Appl., 444 (2016), pp. 550–567.

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

[19] E. A. Feinberg and A. B. Piunovskiy, *Nonatomic total rewards Markov decision processes with multiple criteria*, J. Math. Anal. Appl., 273 (2002), pp. 93–111.

[20] E. A. Feinberg and A. B. Piunovskiy, *On the Dvoretzky–Wald–Wolfowitz theorem on nonrandomized statistical decisions*, Theory Probab. Appl., 50 (2006), pp. 463–466.

[21] E. A. Feinberg and U. G. Rothblum, *Splitting randomized stationary policies in total-reward Markov decision processes*, Math. Oper. Res., 37 (2012), pp. 129–153.

[22] E. A. Feinberg and A. Shwartz, *Constrained discounted dynamic programming*, Math. Oper. Res., 21 (1996), pp. 922–945.

[23] E. A. Feinberg and I. M. Sonin, *Notes on equivalent stationary policies in Markov decision processes with total rewards*, Math. Methods Oper. Res., 44 (1996), pp. 205–221.

[24] O. Hernández-Lerma and J. González-Hernández, *Constrained Markov control processes in Borel spaces: The discounted case*, Math. Methods Oper. Res., 52 (2000), pp. 271–285.

[25] A. Jaśkiewicz and A. S. Nowak, *On a generalization of the Dvoretzky-Wald-Wolfowitz theorem with an application to a robust optimization problem*, J. Math. Anal. Appl., 469 (2019), pp. 126–135.

[26] A. S. Kechris, *Classical Descriptive Set Theory*, Springer-Verlag, New York, 1995.

[27] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.

[28] A. S. Nowak, *On the weak topology in the space of probability measures induced by policies*, Bull. Pol. Acad. Sci. Math., 36 (1988), pp. 181–186.

[29] A. B. Piunovskiy, *Optimal Control of Random Sequences in Problems with Constraints*, Kluwer Academic Publishers, Dordrecht, 1997.

[30] A. B. Piunovskiy, *Controlled random sequences: Methods of convex analysis and problems with functional constraints*, Russian Math. Surveys, 53 (1998), pp. 1233–1293.

[31] M. Schäl, *On dynamic programming: Compactness of the space of policies*, Stochastic Process. Appl., 3 (1975), pp. 345–364.

[32] S. M. Srivastava, *A Course on Borel Sets*, Springer-Verlag, New York, 1998.

[33] A. A. Yushkevich, *The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions*, Math. Oper. Res., 22 (1997), pp. 458–467.