MATHEMATICS OF OPERATIONS RESEARCH
Vol. 30, No. 3, August 2005, pp. 545–561
ISSN 0364-765X | EISSN 1526-5471 | 05 | 3003 | 0545

DOI 10.1287/moor.1050.0148

On the Empirical State-Action Frequencies in Markov Decision Processes Under General Policies

Shie Mannor
Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Québec, Canada H3A 2A7, shie@ece.mcgill.ca, www.ece.mcgill.ca/~shie/

John N. Tsitsiklis
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, jnt@mit.edu, web.mit.edu/~jnt/www/home.html

We consider the empirical state-action frequencies and the empirical reward in weakly communicating finite-state Markov decision processes under general policies. We define a certain polytope and establish that every element of this polytope is the limit of the empirical frequency vector, under some policy, in a strong sense. Furthermore, we show that the probability of exceeding a given distance between the empirical frequency vector and the polytope decays exponentially with time under every policy. We provide similar results for vector-valued empirical rewards.

Key words: Markov decision processes; state-action frequencies; large deviations; empirical measure
MSC2000 subject classification: Primary: 90C40; secondary: 60F99, 60B10
OR/MS subject classification: Primary: probability, Markov processes; secondary: dynamic programming/optimal control, Markov finite state
History: Received March 6, 2003; revised April 6, 2004.

1. Introduction. We consider a Markov decision process (MDP) that satisfies a weak communication assumption and describe a polytope of possible state-action frequency vectors. We show that for every point in the polytope, there exists a policy that gets "very close" to that point. More accurately, for every point in the polytope, we specify a policy that guarantees that the empirical state-action frequency vector converges to that point, with probability one. Moreover, we show that under the prescribed policy, the probability of a large distance between the point and the empirical state-action frequency vector decays exponentially with time. On the other hand, we show that no policy can "get far" from this polytope even without the weak communication assumption. Specifically, we show that the probability of a large distance between the empirical state-action frequency vector and the polytope decays exponentially with time, uniformly over all admissible policies.

While the emphasis of this work is on bounds on the empirical frequencies, we also derive some apparently new results on state-action frequency polytopes. Under the weak communication assumption, our results establish that the polytope we consider is the same as the set of possible limits (both in expectation and almost surely) of the empirical frequency vector under different policies. This extends results in Derman [7] and Puterman [15], which assumed a unichain structure. These references also showed that every point in the polytope can be achieved by a stationary policy. In contrast, for the more general case that we consider, nonstationary policies may be necessary. We note that in Kallenberg [12], a related polytope was defined for *every* Markov decision process, without any communication assumptions. However, the framework of Kallenberg [12] is too general to be useful for our purposes. In particular, some communication assumption is necessary in order to establish that every point in the polytope is a possible limit of the empirical frequency vector.

The primary motivation for this work arises in the fields of adaptive control and reinforcement learning (e.g., Kumar and Varaiya [13], Bertsekas and Tsitsiklis [4], Sutton and Barto [17]). The policies used by learning algorithms are typically nonstationary. For this reason, it is useful to have a complete characterization of the possible behaviors of empirical state-action frequencies under general (not necessarily stationary) policies. For instance, there are certain bounds on the probability of a large distance between the empirical frequencies and their limit, under the assumption that such a limit exists (Altman and

Zeitouni [2]). Our results indicate that similar bounds apply to the case of general, nonstationary policies.

Another motivation comes from the context of exploration in dynamic environments. Suppose that we wish to visit at least k times every state of a controlled Markov chain with known transition probabilities, where k is a large number. This may be the case if we desire to take a large number of measurements at each state, or in a "needle in a haystack" problem, where each state needs to be examined several times in order to identify whether something unique happens at that state. Under an appropriate accessibility assumption, it can be shown that the best possible expected time for achieving this goal is of the form $\eta k + o(k)$, where η is a positive constant that can be computed in terms of the transition probabilities. Using the results in this paper, a stronger property is obtained, namely, that there exists a policy under which the time it takes, T_k, satisfies $\mathbf{P}(T_k \geq k(1+\varepsilon)\eta) \leq ce^{-d\varepsilon^2 k}$, for every $\varepsilon > 0$, where c and d are some positive constants. Moreover, for every policy, $\mathbf{P}(T_k \leq k(1-\varepsilon)\eta) \leq ce^{-d\varepsilon^2 k}$, that is, no policy can sample more efficiently.

Yet another motivation arises from the connection between the average rewards per unit time in finite and infinite horizon problems. An important question, for a finite-horizon problem, is whether one can gain substantially by using a time-dependent policy rather than a stationary one. Our results indicate that the probability of a substantial gain is exponentially small in the time horizon.

Regarding related research, let us mention that there are large deviations results for the empirical state-action frequency vector in finite-state Markov processes (see, e.g., Dembo and Zeitouni [6]). These results were extended to Markov decision processes in Altman and Zeitouni [2], which obtained uniform convergence rates over the class of *stationary* policies. The case of nonstationary policies that have a limit was also considered to some extent in Altman and Zeitouni [2].

The question of achievable rates of convergence for controlled processes was considered in Shimkin [16]. The model therein is essentially a single-state decision process in which a decision maker may choose between sampling several stationary reward populations. Lower and upper bounds were provided on the probabilities of rare events under arbitrary policies. Of a somewhat different flavor is a Hoeffding-type inequality for bounded functions of uniformly ergodic Markov chains, which was derived in Glynn and Ormoneit [9]. We note that this reference provides an error exponent that is tighter than ours, but these results are essentially dependent on the stationary nature of the underlying policy.

The rest of the paper is organized as follows. In §2, we start by defining the model of interest. In §3, we introduce state-action frequency polytopes. In §4, we show that for every element of the polytope, there exists a policy under which the empirical state-action frequency vector converges to that point, in a strong sense. In §5, we derive a large deviations bound for the distance of the empirical state-action frequency vector from the polytope. In §6, we generalize and obtain bounds on the probability of large deviations of an empirical vector-valued reward. In §7, we provide some brief concluding remarks. The appendix contains the proofs of some of lemmas used in our development.

2. Problem definition. We consider a Markov decision process (MDP) with finite state and action spaces. The MDP is formally defined by a triplet $(\mathcal{S}, \mathcal{A}, P)$, where

(a) $\mathcal{S} = \{1, \ldots, S\}$ is a finite set of states.

(b) $\mathcal{A} = \{1, \ldots, A\}$ is a finite set of actions which is assumed, for simplicity, to be the same for all states.

(c) P is the conditional probability law. Namely, $P(s' \mid s, a)$ is the probability that the next state is s', given that the current state is s and that action a was taken.

At every time epoch t, the decision maker observes the current state s_t and chooses an action a_t. Then the next state s_{t+1} is chosen, according to $P(\cdot \mid s_t, a_t)$. For a finite set $\mathcal{B}$, we will use $\Delta(\mathcal{B})$ to denote the set of all probability distributions on $\mathcal{B}$. A *policy* is a mapping

from the set of possible past histories to the set $\Delta(\mathcal{A})$, which prescribes the probability of any particular action for every given history. A *stationary policy* is a policy that depends only on the current state.

Given a history[1] $(s_1, a_1, s_2, a_2, \ldots, s_t, a_t, s_{t+1})$, we define the *empirical state-action-state frequencies* by

$$\hat{q}_t(s, a, s') = \frac{1}{t} \sum_{\tau=1}^{t} I_{\{s_\tau = s,\, a_\tau = a,\, s_{\tau+1} = s'\}}, \tag{1}$$

where I_E stands for the indicator function of an event E. Note that the empirical frequency vector $\hat{q}_t$, with components $\hat{q}_t(s, a, s')$, is a vector in $\Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$.

Without any stationarity assumption on the underlying policy, we cannot expect $\hat{q}_t$ to have a limit. Our main objective is to show that $\hat{q}_t$ has to be close (with high probability) to the set of expected frequency vectors that can be attained using stationary policies. We will therefore start by characterizing the latter set, which is the subject of the next section.

3. State-action polytopes. In this section, we introduce a polytope and characterize it as the set of feasible limiting expected state-action-state frequencies. This characterization is used in §4 to show that the elements of this polytope are feasible empirical frequencies in a rather strong sense.

Given an MDP, the *state-action polytope*, X, is defined as the set of vectors x in $\Delta(\mathcal{S} \times \mathcal{A})$ that satisfy

$$\sum_s \sum_a P(s' \mid s, a) x(s, a) = \sum_{a'} x(s', a'), \quad \forall\, s'. \tag{2}$$

We let (as in Puterman [15]) $x^{\pi,\alpha} \in \Delta(\mathcal{S} \times \mathcal{A})$ be the limiting expected state-action frequency vector, if it exists, under policy π, starting from an initial state distribution $\alpha \in \Delta(\mathcal{S})$, under a general policy π (possibly randomized, nonstationary, or non-Markovian). That is, we define

$$x^{\pi,\alpha}(s, a) \triangleq \lim_{t \to \infty} \mathbb{E}^{\pi,\alpha} \left[\frac{1}{t} \sum_{\tau=1}^{t} I_{\{s_\tau = s,\, a_\tau = a\}} \right], \tag{3}$$

if the limit exists, and let X_Π^α be the set

$$X_\Pi^\alpha \triangleq \{x \in \Delta(\mathcal{S} \times \mathcal{A}) \text{: there exists a policy } \pi \text{ s.t. the limit (3) exists and } x = x^{\pi,\alpha}\}.$$

It is known (see Puterman [15]) that under a unichain assumption, we have $X_\Pi^\alpha = X$. We will show (Theorem 3.1) that the same is true under a less restrictive weak communication assumption. For a different approach that works under *any* assumptions but is less useful for our purposes, see Kallenberg [12]. The following proposition holds for *every* MDP.

PROPOSITION 3.1. *For every MDP, and every $\alpha \in \Delta(\mathcal{S})$, we have $X_\Pi^\alpha \subseteq X$.*

PROOF. Let Π_D denote the set of stationary deterministic policies. Using Theorem 8.9.3 from Puterman [15], we know that X_Π^α equals the convex hull of the limiting expected state-action frequency vectors associated with stationary deterministic policies. That is,

$$X_\Pi^\alpha = \text{co}(\{x^{\pi,\alpha} \mid \pi \in \Pi_D\}),$$

where $\text{co}(B)$ stands for the convex hull of a finite set B. Now, fix a stationary and deterministic policy π, specified in terms of a function $\mu\colon \mathcal{S} \mapsto \mathcal{A}$, and consider the resulting Markov chain, with transition probabilities $P(s' \mid s) = P(s' \mid s, \mu(s))$. Because the resulting chain is stationary, the limits $x^{\pi,\alpha}(s, a)$ exist and satisfy the balance equations

$$\sum_s P(s' \mid s, \mu(s)) x^{\pi,\alpha}(s, \mu(s)) = x^{\pi,\alpha}(s', \mu(s')), \quad \forall\, s'.$$

Because, in addition, $x^{\pi,\alpha}(s, a) = 0$ for $a \neq \mu(s)$, we see that the vector $x^{\pi,\alpha}$ satisfies (2), so that $x^{\pi,\alpha} \in X$. The result follows from the convexity of X. □

[1] Note that, to streamline notation, we include s_{t+1} in the history.

We now recall a definition used in Puterman [15]. An MDP is called *weakly communicating* if the set of states can be partitioned into a set of states that are accessible from each other (i.e., for any two states s and s' in that set, there exists a policy under which there is a positive probability path from s to s'), and a set of states that are transient under *all* policies.

THEOREM 3.1. *If the MDP is weakly communicating, then $X = X_\Pi^\alpha$ for all $\alpha \in \Delta(\mathcal{S})$. Furthermore, for every $x \in X$, there exists some $z \in \Delta(\mathcal{S})$ and a stationary policy π, such that $x^{\pi,z} = x$.*

PROOF. Using Proposition 3.1, is suffices to prove that $X \subseteq X_\Pi^\alpha$ for all α. We first show that X_Π^α is independent of α. For any given $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, consider an average reward MDP with reward $r(s,a)$ at state s and action a, and initial state distribution α. The corresponding optimal average reward is given by $\max_{z\in X_\Pi^\alpha} r^\top z$. Because the optimal average reward in a weakly communicating MDP is independent of the initial state (Puterman [15], p. 352), it follows that for every r the quantity $\max_{z\in X_\Pi^\alpha} r^\top z$ is the same for all α. This implies that the polytopes X_Π^α are the same for all α. Indeed, if there existed some x such that $x \in X_\Pi^{\alpha_1}$ and $x \notin X_\Pi^{\alpha_2}$, we could use the separating hyperplane theorem to obtain a vector r for which

$$\max_{z\in X_\Pi^{\alpha_1}} r^\top z \geq r^\top x > \max_{z\in X_\Pi^{\alpha_2}} r^\top z,$$

which is a contradiction.

Let us now fix some $x \in X$. We proceed to show that $x \in X_\Pi^\alpha$ for some initial state distribution α. Let $z \in \Delta(\mathcal{S})$ denote the state frequency vector associated with x, i.e., $z(s) = \sum_a x(s,a)$. One can rewrite Equation (2) in terms of the state frequency vector in the form

$$\sum_s P_x(s' \mid s) z(s) = z(s'), \tag{4}$$

where

$$P_x(s' \mid s) = \begin{cases} \dfrac{\sum_a P(s' \mid s,a) x(s,a)}{\sum_a x(s,a)}, & \text{if } z(s) > 0, \\ P(s' \mid s, 1), & \text{if } z(s) = 0. \end{cases}$$

Note that P_x corresponds to the transition probabilities for our MDP under a particular policy π: It is the policy that always chooses action 1 at states s for which $z(s) = 0$, while at other states s chooses action a with probability $x(s,a)/z(s)$. Equation (4) shows that z solves the balance equations for the Markov process governed by P_x. In particular, if this Markov process starts with z as the initial state distribution, then the state at any future time is also distributed according to z. It follows that the limiting expected state-action frequency vector under that policy, $x^{\pi,z}$, is equal to x, so that $x \in X_\Pi^z$. The result follows because X_Π^α is independent of α. □

REMARK 3.1. In the absence of the weak communication assumption, X_Π^α may be a proper subset of X. This can be seen from a simple example involving two disconnected absorbing states and no control. Here, $X = \{(\beta, 1-\beta) \mid 0 \leq \beta \leq 1\}$, while for every $\beta \in [0,1]$, we have $X_\Pi^\beta = \{(\beta, 1-\beta)\}$.

REMARK 3.2. For weakly communicating MDPs, the relative interior of X can be attained by randomized stationary policies, and the extreme points of X can be attained by deterministic stationary policies, but some boundary points of X may require either nonstationary policies or a random initial state, as the next example demonstrates.

EXAMPLE 3.1. Consider a deterministic MDP with two states ($\mathcal{S} = \{1,2\}$) and two actions ($\mathcal{A} = \{1,2\}$), in which the action determines the next state. In particular, the transition probabilities satisfy $P(1 \mid 1,1) = P(1 \mid 2,1) = P(2 \mid 1,2) = P(2 \mid 2,2) = 1$. The state-action polytope X includes a point x^* satisfying $x^*(1,1) = x^*(2,2) = 1/2$ and

$x^*(1,2) = x^*(2,1) = 0$. However, starting from an initial state distribution $\alpha = (1,0)$, no stationary policy can have x^* as a limit point. The point x^* can be attained either by using the initial distribution $\alpha = (1/2, 1/2)$, or starting from $\alpha = (1,0)$ by using a nonstationary policy like $(a_1, a_2, \dots) = (1,2,2,1,1,1,2,2,2,2,\dots)$, under which the switches between the states become less frequent with time while the expected frequency of each of the states approaches $1/2$. Furthermore, note that even when $\alpha = (1/2, 1/2)$, a stationary policy can only make the expected frequencies converge to x^*. However, there is no stationary policy that results in almost sure convergence of the empirical frequencies to x^*.

Because we will be interested in the empirical frequencies of the various possible transitions, we also define the *state-action-state* frequency polytope, Q, as the set of vectors in $\Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ that satisfy

$$q(s,a,s') = P(s' \mid s,a) \sum_{s''} q(s,a,s''), \quad \forall\, s,a,s', \tag{5}$$

$$\sum_s \sum_a q(s,a,s') = \sum_{a'} \sum_{s''} q(s',a',s''), \quad \forall\, s'. \tag{6}$$

For a loose interpretation, think of $x(s,a)$ as the frequency with which state s is visited and action a is applied, and think of $q(s,a,s')$ as the frequency with which state s is visited, action a is applied, and the next state is s'. Equation (5) requires the relative frequencies of the various transitions to conform to the transition probabilities, whereas Equation (6) is a flow conservation requirement. Equation (2) combines these two requirements in a single equation. As expected, these two polytopes, X and Q, are closely related.

LEMMA 3.1. *If $x \in X$ and if we let $q(s,a,s') = x(s,a)P(s' \mid s,a)$, then $q \in Q$. Furthermore, every element of Q can be generated in this manner from some element of X.*

PROOF. Suppose that $x \in X$ and that $q(s,a,s') = x(s,a)P(s' \mid s,a)$. We will show that $q \in Q$. Because $x \in \Delta(\mathcal{S} \times \mathcal{A})$, it is easily verified that $q \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$. Furthermore,

$$\begin{aligned} q(s,a,s') &= x(s,a)P(s' \mid s,a) \\ &= x(s,a)\left(\sum_{s''} P(s'' \mid s,a)\right)P(s' \mid s,a) \\ &= P(s' \mid s,a) \sum_{s''} q(s,a,s''), \end{aligned}$$

so that Equation (5) is satisfied. It remains to verify Equation (6). Indeed, for any s', we have

$$\begin{aligned} \sum_a \sum_s q(s,a,s') &= \sum_s \sum_a x(s,a)P(s' \mid a,s) \\ &= \sum_{a'} x(s',a') \\ &= \sum_{a'} x(s',a') \sum_{s''} P(s'' \mid s',a') \\ &= \sum_{a'} \sum_{s''} q(s',a',s''), \end{aligned}$$

as desired. Here, the first equality uses our assumption that $q(s,a,s') = x(s,a)P(s' \mid s,a)$, the second equality comes from Equation (2), the third uses the fact that transition probabilities sum to one, and the fourth uses once more the assumption $q(s,a,s') = x(s,a) \cdot P(s' \mid s,a)$.

To prove the second statement, consider an element of Q, and define x by letting $x(s,a) = \sum_{s'} q(s,a,s')$. Because $q \in Q$, Equation (5) implies that $q(s,a,s') = x(s,a)P(s' \mid s,a)$,

so that q can be indeed generated from some x. It remains to show that $x \in X$. The fact $x \in \Delta(\mathcal{S} \times \mathcal{A})$ is an immediate consequence of q belonging to $\Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$. Finally,

$$\begin{aligned}\sum_s \sum_a P(s' \mid s, a)x(s, a) &= \sum_s \sum_a q(s, a, s') \\ &= \sum_{a'} \sum_{s''} q(s', a', s'') \\ &= \sum_{a'} x(s', a'),\end{aligned}$$

where the second equality follows from Equation (6). Thus, Equation (2) is satisfied and $x \in X$, as desired. □

We now turn our attention to the feasible limiting expected frequencies. We let $q^{\pi,\alpha} \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ be the limiting expected state-action-state frequency vector, if it exists, under policy π, starting from an initial state distribution $\alpha \in \Delta(\mathcal{S})$, under a general policy π (possibly randomized, nonstationary, or non-Markovian). That is,

$$q^{\pi,\alpha}(s, a, s') = \lim_{t\to\infty} \mathbb{E}^{\pi,\alpha}\left[\frac{1}{t}\sum_{\tau=1}^{t} I_{\{s_\tau=s,\, a_\tau=a,\, s_{\tau+1}=s'\}}\right], \tag{7}$$

where $\mathbb{E}^{\pi,\alpha}$ is the expectation under policy π, given that the initial state is distributed according to α. For every $\alpha \in \Delta(\mathcal{S})$, we let Q_Π^α be the set

$$Q_\Pi^\alpha \triangleq \{q \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})\text{: there exists a policy } \pi \text{ s.t. the limit (7) exists and } q = q^{\pi,\alpha}\}.$$

We note an elementary counterpart of Lemma 3.1.

LEMMA 3.2. *If $x \in X_\Pi^\alpha$ and if we let $q(s, a, s') = x(s, a)P(s' \mid s, a)$, then $q \in Q_\Pi^\alpha$. Furthermore, every element of Q_Π^α can be generated in this manner from some element of X_Π^α.*

PROOF. Fix some policy π and some $\alpha \in \Delta(\mathcal{S})$. We have

$$\mathbb{E}^{\pi,\alpha}\left[I_{\{s_\tau=s,\, a_\tau=a,\, s_{\tau+1}=s'\}}\right] = P(s' \mid s, a)\mathbb{E}^{\pi,\alpha}\left[I_{\{s_\tau=s,\, a_\tau=a\}}\right],$$

from which it follows that the limit in the definition of $q^{\pi,\alpha}$ exists if and only if the limit in the definition of $x^{\pi,\alpha}$ exists, and in that case, $q^{\pi,\alpha}(s, a, s') = x^{\pi,\alpha}(s, a)P(s' \mid s, a)$. □

Our results so far refer to the state-action polytope X and its relation with X_Π^α. We now extend the results to the state-action-state polytope Q, and the corresponding sets Q_Π^α of limiting expected state-action-state frequencies. Once more, the containment $Q_\Pi^\alpha \subseteq Q$ holds for every MDP. However, the inclusion might be proper in the absence of some communication assumptions.

PROPOSITION 3.2. *If the MDP is weakly communicating, then for every initial state distribution α we have $Q = Q_\Pi^\alpha$. Furthermore, for every $q \in Q$, there exists some $z \in \Delta(\mathcal{S})$ and a stationary policy π, such that $q^{\pi,z} = q$.*

PROOF. By Lemmas 3.1 and 3.2, the set Q_Π^α can be constructed from the set X_Π^α using the same formula as in the construction of the set Q from the set X. Because $X = X_\Pi^\alpha$, it follows that $Q = Q_\Pi^\alpha$. Furthermore, given some $q \in Q$, Lemma 3.1 and Theorem 3.1 imply that there exists some $z \in \Delta(S)$ and some stationary policy π such that $q(s, a, s') = x^{\pi,z}(s, a)P(s' \mid s, a)$, for all (s, a, s'). It then follows that $q = q^{\pi,z}$. □

In the next two sections, we relate Q to the possible limits of the empirical frequency vector. We start in §4 with a positive result that states that for every $q \in Q$, there is a policy under which $\hat{q}_t$ converges to q almost surely. Moreover, the probability of a large distance between $\hat{q}_t$ and q decays exponentially. In §5 we provide a converse result: We show that for every policy, the probability of a large distance between $\hat{q}_t$ and Q decays exponentially.

4. Convergence to feasible points. In this section, we show that for every $q \in Q$, there exists a policy under which $\hat{q}_t$ converges to q. Moreover, the probability of a given positive distance between $\hat{q}_t$ and q decays exponentially. The proof is constructive and provides a specific policy with these properties. If q belongs to the relative interior of Q, the existence of a stationary policy with the required properties is straightforward (following Miller [14] or Glynn and Ormoneit [9]). The difficult case is when q is on the boundary of Q and the MDP has a multichain structure. In that case, there need not exist a stationary policy that guarantees convergence of $\hat{q}_t$ to q; see, e.g., Example 3.1. For this reason, we have to introduce an appropriate nonstationary policy. In the sequel, we will use $\|\cdot\|$ to denote the Euclidean norm in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$.

THEOREM 4.1. *Suppose that the MDP is weakly communicating. Then, for every initial state and for every $q \in Q$, there exists a policy π^*, and positive constants c_0, c_1, under which*

$$\mathbf{P}(\|\hat{q}_t - q\| \geq \varepsilon) \leq c_0 e^{-c_1 \varepsilon^2 t}, \quad \forall\, t \geq 1, \quad \forall\, \varepsilon > 0.$$

Furthermore, $\hat{q}_t$ converges to q, with probability 1.

PROOF. Let us fix some $q \in Q$. Then, by Proposition 3.2, there exists a stationary policy π and some $z \in \Delta(\mathcal{S})$ such that $q^{\pi,z} = q$. Consider the resulting stationary Markov chain. The standard "ergodic decomposition" (see, e.g., Puterman [15]) shows that the state space $\mathcal{S}$ can be partitioned into disjoint sets $\mathcal{S}_0, \ldots, \mathcal{S}_l$, where $\mathcal{S}_0$ is the set of states that are transient under π and where each $\mathcal{S}_i$, for $i \neq 0$ is an irreducible class of recurrent states under π. For $i \neq 0$, let $\alpha^i = \sum_{s\in\mathcal{S}_i} z(s)$, and let $q^i(s,a,s')$ be the steady-state probability of a transition from s to s' under action a, if the chain is initialized within the class $\mathcal{S}_i$. We then have

$$q = \sum_{i=1}^{l} \alpha^i q^i. \tag{8}$$

We assume without loss of generality that $\alpha^i > 0$ for all i, and define $\underline{\alpha} = \min_i \alpha^i$. (If $\alpha^i = 0$ for some i, we can work with a reduced chain from which the irreducible class $\mathcal{S}_i$ has been eliminated.) For every $i \neq 0$, let us fix a special "starting" state $s_1^i \in \mathcal{S}_i$, and let $\hat{q}_t^i$ be the resulting empirical frequency vector if the stationary policy π is used for t consecutive transitions. We will be using the following result, which is a special case of the results in Glynn and Ormoneit [9]. (The ergodicity conditions in that reference are satisfied because the Markov chain is confined to the single recurrent class $\mathcal{S}_i$.)

LEMMA 4.1. *There exist positive constants c_2 and c_3, such that for $i = 1, \ldots, l$,*

$$\mathbf{P}(\|\hat{q}_t^i - q^i\| \geq \varepsilon) \leq c_2 e^{-c_3 \varepsilon^2 t}, \quad \forall\, t \geq 1, \quad \forall\, \varepsilon > 0.$$

The main idea in the rest of the proof is as follows. For each $i \in \{1, \ldots, l\}$, we consider the infinite trajectory in $\mathcal{S}_i$ obtained by starting at s_1^i and using the stationary policy π. We break these trajectories into intervals and interleave them so that the resulting process mimics the ith trajectory for a fraction α^i of the time. This will result in

$$\hat{q}_t \approx \sum_{i=1}^{l} \hat{q}^i_{\lceil \alpha^i t \rceil} \to \sum_{i=1}^{l} \alpha^i q^i = q.$$

An exponential bound will be obtained by applying Lemma 4.1. However, such an interleaving requires that some time be spent switching from one subset $\mathcal{S}_i$ to another. To facilitate the analysis, the interleaving is arranged so that the last state of the kth interval in $\mathcal{S}_i$ is the same as the first state of the $(k+1)$st interval in $\mathcal{S}_i$. We start by characterizing the statistics of the required switching, we continue with a precise description of the interleaving, and we conclude with a rigorous version of the above outlined heuristic argument.

Consider a state $s' \in \mathcal{S}_i$, for some $i \neq 0$. We have assumed that the MDP is weakly communicating. In particular, there is a set of states that are transient under every policy. The latter states are transient under the policy π, and therefore belong to $\mathcal{S}_0$. Because the state s' is not transient under π, the weak communication assumption implies that there exists a policy under which state s' is eventually reached. Let $\rho_{s'}$ be a policy with this property, under which the expected time to reach s' is minimized, for every initial state. Standard dynamic programming results imply that $\rho_{s'}$ can be taken stationary and deterministic. Furthermore, the probability that state s' is not reached within t time steps decays exponentially with t. Let $\tau_{s,s'}$ be the time to reach s' starting from s. That is,

$$\tau_{s,s'} = \inf\{t > 0 \colon s_t = s'\}.$$

We summarize this discussion in the following result.

LEMMA 4.2. *There are positive constants c and β such that for every $s' \notin \mathcal{S}_0$, there exists a stationary policy $\rho_{s'}$ under which the random time $\tau_{s,s'}$ it takes for s' to be reached, starting from s, satisfies*

$$\mathbf{P}(\tau_{s,s'} \geq t) \leq ce^{-\beta t}, \quad \forall\, t \geq 1, \quad \forall\, s \in \mathcal{S}.$$

An immediate consequence of Lemma 4.2, which will be used later, is the following. For every $s \in \mathcal{S}$ and $s' \notin \mathcal{S}_0$, we have

$$\mathbf{P}(\tau_{s,s'} \geq t) \leq \min\{1, ce^{-\beta t}\} = \min\{1, e^{-\beta(t-c_4)}\} = \mathbf{P}(Z + c_4 \geq t), \tag{9}$$

where c_4 is such that $c = \beta e^{c_4}$, and where Z is an exponentially distributed random variable with parameter β.

We now specify the interleaved policy π^*. The policy starts in an arbitrary state s_0 and proceeds in rounds. For each round k and for each $i \neq 0$, there is a time interval consisting of

$$t_k^i = \lceil \alpha^i k(k+1) \rceil - \lceil \alpha^i (k-1)k \rceil$$

transitions during which the state lies in $\mathcal{S}_i$ and policy π is followed. (Note that t_k^i is approximately equal to $2\alpha^i k$.) For any $i \neq 0$, the initial state of the ith interval in the $(k+1)$st round, denoted by s_{k+1}^i, will be set to be the same as the final state of the ith interval in the kth round.

A precise description is as follows.

1. Initialization: $k = 1$; initial state $s_0 \in \mathcal{S}$; states $s_1^i \in S_i$, for $i = 1, \ldots, l$.
   For every round $k = 1, 2, \ldots$, do the following.
2. Let s_k^* be the state at the end of the previous round. (For $k = 1$, let s_1^* be the initial state s_0.)
3. Use policy $\rho_{s_k^1}$, until state $s_k^1 \in \mathcal{S}_1$ is reached.
4. Use policy π for t_k^1 transitions. Let s_{k+1}^1 be the final state.
5. For $i = 2$ to l, do the following.
   5a. Use policy $\rho_{s_k^i}$, until state $s_k^i \in \mathcal{S}_i$ is reached.
   5b. Use policy π for t_k^i transitions. Let s_{k+1}^i be the final state.

For any $t \geq 1$, consider the first t transitions under policy π^*. Out of these, there is a (random) number $\tau^0(t)$ of transitions during which some policy $\rho_{s'}$ is used (i.e., time spent in steps 3 or 5a). Furthermore, for $i = 1, \ldots, l$, there is a (random) number $\tau^i(t)$ of transitions, during which the policy π is used within the set $\mathcal{S}_i$ (in steps 4 or 5b). Note that, by definition, $\sum_{i=0}^{l} \tau^i(t) = t$. We have the following lemma, which is proved in the appendix.

LEMMA 4.3. *There exist constants c_5, c_6, and c_7 such that*:
(a) *For every $t \geq 1$ and $i \in \{1, \ldots, l\}$,*

$$\alpha^i(t - \tau^0(t)) - c_5\sqrt{t} \leq \tau^i(t) \leq \alpha^i t + c_5\sqrt{t}.$$

(b) *For every* $t \geq 1$ *and* $\varepsilon > 0$,

$$\mathbf{P}(\tau^0(t) \geq \varepsilon t) \leq \mathbf{P}\big(\tau^0(t) + c_5\sqrt{t} + 1 \geq \varepsilon t\big) \leq c_6 e^{-c_7\varepsilon^2 t}.$$

Let us consider the first t transitions. Out of these, there are $\tau^i(t)$ transitions that occur while using policy π within the set $\mathcal{S}_i$. The number of such transitions that involve a particular triplet (s, a, s') has the same distribution as the number of such transitions that would be observed if we were using this policy for a number $\tau^i(t)$ of contiguous time steps. The latter number equals $\tau^i(t)\hat{q}^i_{\tau^i(t)}(s, a, s')$, where $\hat{q}^i_t(s, a, s')$ is a component of the empirical frequency vector $\hat{q}^i_t$ in Lemma 4.1. In addition, there is a number $n_{\tau^0(t)}(s, a, s')$ of such transitions that occur while using one of the switching policies. Let $n_{\tau^0(t)}$ be the vector with components $n_{\tau^0(t)}(s, a, s')$. Then, using vector notation and omitting the (s, a, s') index, $\hat{q}_t$ has the same distribution as

$$\frac{n_{\tau^0(t)} + \sum_{i=1}^{l} \tau^i(t)\hat{q}^i_{\tau^i(t)}}{t}.$$

Using the representation $q = \sum_{i=1}^{l} \alpha^i q^i$ (cf. Equation (8)), we see that $\hat{q}_t - q$ has the same distribution as

$$\frac{n_{\tau^0(t)}}{t} + \sum_{i=1}^{l}\left(\frac{\hat{q}^i_{\tau^i(t)}\tau^i(t) - \hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil}{t}\right) + \sum_{i=1}^{l}\left(\frac{\hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil}{t} - \alpha^i\hat{q}^i_{\lceil\alpha^i t\rceil}\right) + \sum_{i=1}^{l}\big(\alpha^i\hat{q}^i_{\lceil\alpha^i t\rceil} - \alpha^i q^i\big). \tag{10}$$

We will now bound the tail probabilities of the norm of each one of the terms in Equation (10).

Note that the sum of the components of $n_{\tau^0(t)}$ is $\tau^0(t)$, so that $\|n_{\tau^0(t)}\| \leq \tau^0(t)$. Thus, using Lemma 4.3(b),

$$\mathbf{P}(\|n_{\tau^0(t)}/t\| \geq \varepsilon/4) \leq \mathbf{P}(\tau^0(t) \geq \varepsilon t/4) \leq c_6 e^{-c_7\varepsilon^2 t/16}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon > 0.$$

We now consider the second term in Equation (10). We note that the (s, a, s') component of the summand $q^i_{\tau^i(t)}\tau^i(t) - \hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil$ counts the number of times that a particular transition (s, a, s') is observed at times between $\lceil\alpha^i t\rceil$ and $\tau^i(t)$. Thus,

$$\big\|q^i_{\tau^i(t)}\tau^i(t) - \hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil\big\| \leq |\tau^i(t) - \lceil\alpha^i t\rceil| \leq \alpha^i\tau^0(t) + c_5\sqrt{t} + 1 \leq \tau^0(t) + c_5\sqrt{t} + 1,$$

where the second inequality follows from Lemma 4.3(a). Therefore, using Lemma 4.3(b),

$$\begin{aligned}\mathbf{P}\big(\big\|q^i_{\tau^i(t)}\tau^i(t) - \hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil\big\| \geq \varepsilon t/4l\big) &\leq \mathbf{P}\big(\tau^0(t) + c_5\sqrt{t} + 1 \geq \varepsilon t/4l\big)\\ &\leq c_6 e^{-c_7\varepsilon^2 t/16l^2}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon > 0.\end{aligned}$$

The third term in Equation (10) can be bounded by noticing that

$$\left\|\sum_{i=1}^{l}\left(\frac{\hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil}{t} - \alpha^i\hat{q}^i_{\lceil\alpha^i t\rceil}\right)\right\| \leq \sum_{i=1}^{l}\big\|\hat{q}^i_{\lceil\alpha^i t\rceil}\big\|\left(\frac{\lceil\alpha^i t\rceil}{t} - \alpha^i\right) \leq \frac{l}{t}.$$

It follows that

$$\mathbf{P}\left(\left\|\sum_{i=1}^{l}\left(\frac{\hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil}{t} - \alpha^i\hat{q}^i_{\lceil\alpha^i t\rceil}\right)\right\| \geq \frac{\varepsilon}{4}\right) \leq e \cdot e^{-\varepsilon t/4l}$$

Because $\mathbf{P}(\|\hat{q}_t - q\| > \varepsilon) = 0$ for $\varepsilon > 2$, we can assume that $\varepsilon \leq 2$, and we have

$$\mathbf{P}\left(\left\|\sum_{i=1}^{l}\left(\frac{\hat{q}^i_{\lceil\alpha^i t\rceil}\lceil\alpha^i t\rceil}{t} - \alpha^i\hat{q}^i_{\lceil\alpha^i t\rceil}\right)\right\| \geq \frac{\varepsilon}{4}\right) \leq e \cdot e^{-\varepsilon^2 t/16l}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon \in (0, 2].$$

As for the last term in Equation (10), because $\sum_{i=1}^{l} \alpha^i = 1$, its norm will exceed $\varepsilon/4$ only if the norm of $\widehat{q}^i_{\lceil \alpha^i t \rceil} - q^i$ exceeds $\varepsilon/4$ for some i. The probability of this event is bounded by

$$\sum_{i=1}^{l} \mathbf{P}\left(\left\| \widehat{q}^i_{\lceil \alpha^i t \rceil} - q^i \right\| \geq \varepsilon/4\right) \leq l c_2 e^{-c_3 \varepsilon^2 \underline{\alpha} t/16}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon > 0,$$

where we have made use of Lemma 4.1.

Putting all the above bounds together, we obtain the desired probability bound, with $c_0 = c_6 + l c_6 + e + l c_2$, and $c_1 = \min\{c_7/16, c_7/16l^2, 1/16l, c_3/16\}$. Using the Borel-Cantelli lemma, the event $\{\|\widehat{q}^i_t - q^i\| \geq \varepsilon\}$ can occur only a finite number of times, which implies that $\widehat{q}_t$ converges to q, with probability 1. □

5. A bound on the large deviation probabilities of empirical frequencies. We saw in the last section that for weakly communicating MDPs, every element of Q is a possible limit point of the empirical frequency vector. In this section, we prove a converse result, namely, that the probability that $\widehat{q}_t$ is at a given positive distance from Q decays exponentially. For any $y \in \mathbb{R}^k$ and $W \subseteq \mathbb{R}^k$, we will be using the notation $\|y - W\|$ to denote the distance of y from W, i.e., $\|y - W\| = \inf_{w \in W} \|y - w\|$.

THEOREM 5.1. *For every MDP, there exist positive constants c_0 and c_1 such that under any policy π,*

$$\mathbf{P}(\|\widehat{q}_t - Q\| \geq \varepsilon) \leq c_0 e^{-c_1 \varepsilon^2 t}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon > 0.$$

PROOF. The proof relies on the following geometric lemma that relates the Euclidean point-to-polytope distance with the amount by which the inequalities defining the polytope are violated. Its proof is given in the appendix.

LEMMA 5.1. *Suppose that a nonempty set $W \subset \mathbb{R}^k$ is defined by a set of linear inequalities, i.e., $W = \{w\colon Aw \leq b\}$ where A is an $m \times k$ matrix and $b \in \mathbb{R}^m$. Then, there exists a constant c such that for every $y \notin W$, we have $\|y - W\| \leq c\|(Ay - b)^+\|_\infty$, where $(y)^+$ is the componentwise maximum of y and the zero vector, and $\|y\|_\infty = \max_i |y_i|$.*

We apply Lemma 5.1 to the polytope Q (in place of W) and let c denote the value of the constant whose existence is asserted by the lemma. Let us fix some $\varepsilon > 0$ and some $t > c/\varepsilon$ and suppose that the event $\|\widehat{q}_t - Q\| \geq \varepsilon$ has occurred. Then, Lemma 5.1 implies that at least one of the constraints that define Q is violated by at least ε/c. Note that the simplex constraints are automatically satisfied because $\widehat{q}_t \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$, with probability 1, by construction. We also note that the constraints of the form (6) can be violated by at most $1/t$. Indeed, the number of times a state is entered can differ by at most one from the number of times a state is exited, so that the corresponding frequencies can differ by at most $1/t$. Because $1/t < \varepsilon/c$, it must be that at least one of the constraints (5) is violated by at least ε/c. Using the union bound, we obtain

$$\begin{aligned} \mathbf{P}(\|\widehat{q}_t - Q\| \geq \varepsilon) &\leq \sum_{s,a,s'} \mathbf{P}\left(|\widehat{q}_t(s,a,s') - \widehat{q}_t(s,a) P(s' \mid s,a)| \geq \frac{\varepsilon}{c} \right) \\ &= \sum_{s,a,s'} \mathbf{P}\left(|n_t(s,a,s') - n_t(s,a) P(s' \mid s,a)| \geq \frac{\varepsilon t}{c} \right), \end{aligned} \tag{11}$$

where $\widehat{q}_t(s,a) = n_t(s,a)/t$, and

$$n_t(s,a) = \sum_{\tau=1}^{t} I_{\{s_\tau = s,\, a_\tau = a\}}, \qquad n_t(s,a,s') = \sum_{\tau=1}^{t} I_{\{s_\tau = s,\, a_\tau = a,\, s_{\tau+1} = s'\}}. \tag{12}$$

We will now reason in terms of a single probability space on which the controlled process can be defined under any policy. Such a probability space can involve a countable collection of independent uniform random variables (that are used to generate actions under randomized policies), as well a collection of independent $\mathcal{S}$-valued random variables $h(s,a,\tau)$ (for

$s \in \mathcal{S}$, $a \in \mathcal{A}$, $\tau \geq 1$) that take the value s' with probability $P(s' \mid s, a)$. With these random variables at hand, the process can be constructed as follows: If the state happens to be s for the τth time, and the policy chooses action a, the random variable $h(s, a, \tau)$ is used to determine the next state.

Consider the first l transitions for which $(s_\tau, a_\tau) = (s, a)$, and let $b_l(s, a, s')$ be the number of these transitions that lead to s'. Formally, for every $l \geq 1$, we define $b_l(s, a, s')$ as the cardinality of the set $\{\tau \mid h(s, a, \tau) = s', 1 \leq \tau \leq l\}$. We observe that $b_l(s, a, s')$ is a binomial random variable with parameters l and $P(s' \mid s, a)$. Note also that if $n_t(s, a) = l$, then $n_t(s, a, s') = b_l(s, a, s')$.

Let $\delta = \varepsilon/c$, and consider the events

$$E = \{|n_t(s, a, s') - n_t(s, a)P(s' \mid s, a)| \geq \delta t\},$$
$$B = \left\{\max_{1 \leq l \leq t} |b_l(s, a, s') - lP(s' \mid s, a)| \geq \delta t\right\}.$$

We observe that $E \subseteq B$, and $\mathbf{P}(E) \leq \mathbf{P}(B)$.

To bound the probability of B, we use the following lemma, which provides large deviations bounds for the maximum of a random walk, and is in the same spirit as other large deviations bounds that can be found in Hajek [10] or Gallager [8]. Related results that apply to an asymptotic regime can be found in Hircsh [11] and references therein. The proof is again deferred to the appendix.

LEMMA 5.2. *Let $X_1, \ldots, X_t$ be independent identically distributed zero-mean random variables. Assume that the log-moment generating function $\rho(s) = \log \mathbb{E}[e^{sX_1}]$ is finite in a neighborhood of zero. We define the rate function $f(\delta) = \sup_s(\delta s - \rho(s))$. Let $\overline{X}_i = \sum_{j=1}^{i} X_j$. Then, for every $\delta \neq 0$ we have $f(\delta) > 0$, and*

$$\mathbf{P}\left(\max_{1 \leq i \leq t} |\overline{X}_i| \geq \delta t\right) \leq e^{-tf(\delta)} + e^{-tf(-\delta)}. \tag{13}$$

Let $X_1, \ldots, X_t$ be independent identically distributed Bernoulli random variables with mean p; then Equation (13) becomes

$$\mathbf{P}\left(\max_{1 \leq i \leq t} |\overline{X}_i - ip| \geq \delta t\right) \leq 2e^{-2t\delta^2}. \tag{14}$$

We apply Lemma 5.2 by identifying X_i with the shifted Bernoulli random variable $b_i(s, a, s') - b_{i-1}(s, a, s') - P(s' \mid s, a)$, so that $\overline{X}_i = b_l(s, a, s') - lP(s' \mid s, a)$. It follows that

$$\mathbf{P}(E) \leq \mathbf{P}(B) \leq 2e^{-2\delta^2 t}.$$

Substituting in Equation (11), we obtain

$$\mathbf{P}(\|\hat{q}_t - Q\| \geq \varepsilon) \leq 2S^2 A e^{-2\varepsilon^2 t/c^2}.$$

We have been assuming so far that $t > c/\varepsilon$. We now verify that the result remains valid without that assumption. Indeed, if $\varepsilon \geq 2$, we have $\mathbf{P}(\|\hat{q}_t - Q\| \geq \varepsilon) = 0$. Furthermore, if $\varepsilon \leq 2$ and $t \leq c/\varepsilon$, we have $2t\varepsilon^2/c^2 \leq 4/c$, so that

$$\mathbf{P}(\|\hat{q}_t - Q\| \geq \varepsilon) \leq 1 \leq e^{4/c} e^{-2\varepsilon^2 t/c^2}.$$

This establishes the desired result with $c_0 = \max\{2S^2A, e^{4/c}\}$ and $c_1 = 2/c^2$. □

6. A bound on the large deviation probabilities of the average reward. In this section, instead of the empirical frequencies, we focus on a vector-valued reward and show that with high probability the empirical average reward is close to a polytope of achievable limiting expected reward vectors. The motivation behind this setting comes from multicrit-

era MDPs in which one is interested in the simultaneous control of several performance measures.

We start by describing the model for the rewards. For every state-action-state frequency triplet (s, a, s'), we assume that there is a corresponding reward process, i.e., a sequence of k-dimensional random vectors $m^{\tau}_{s,a,s'}$, $\tau = 1, 2, \ldots$, and that the reward vector $m^{\tau}_{s,a,s'}$ is realized when a transition from s to s', under action a, occurs for the τth time. We assume that the reward processes are independent from the state processes. This model includes the standard case, where the random variables $m^{\tau}_{s,a,s'}$ are all independent, and for any given triplet (s, a, s'), identically distributed. But it also allows for more general reward process $m^{\tau}_{s,a,s'}$, possibly driven by additional (possibly Markov) exogenous dynamics, as long as they obey the large deviations bounds in the assumption that follows.

ASSUMPTION 6.1. *For every (s, a, s'), the limit*

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{1}{t}\sum_{\tau=1}^{t} m^{\tau}_{s,a,s'}\right]$$

exists and will be denoted by $m(s, a, s')$. Also, there exists a function $f\colon (0,\infty) \mapsto (0,\infty]$ and positive constants c_1, c_2, such that

$$\mathbf{P}\left(\left\|\frac{1}{t}\sum_{\tau=1}^{t} m^{\tau}_{s,a,s'} - m(s,a,s')\right\| > \varepsilon\right) \le c_1 e^{-c_2 f(\varepsilon) t}, \quad \forall\, \varepsilon > 0, \quad \forall\, t \ge 1.$$

Finally, there exist positive constants c, η, and ε_0, such that $f(\varepsilon) \ge c\varepsilon^{\eta}$ for all $\varepsilon \in (0, \varepsilon_0)$.

We introduce some more notation. Under a given policy and initial state distribution, we use m_t to denote the reward vector obtained at time t, and $\widehat{m}_t$ to denote the corresponding empirical average reward $(1/t)\sum_{\tau=1}^{t} m_\tau$. We also define a polytope M by

$$M = \left\{ m \in \mathbb{R}^k \colon \exists q \in Q \text{ such that } m = \sum_{s,a,s'} m(s,a,s') q(s,a,s') \right\},$$

which is the image of Q under a linear mapping. It can be shown that for weakly communicating MDPs, under Assumption 6.1, no matter what the initial state distribution is and for every point in M, there is a policy under which the sequence of empirical average rewards $\widehat{m}_t$ converges to that point (cf. Theorem 4.1). The result that follows provides a converse, namely, that the probability of a substantial deviation from M decays exponentially with time.

THEOREM 6.1. *Suppose that Assumption 6.1 holds. Then there exist positive constants c_1, c_2, and ε_0, such that under every policy π,*

$$\mathbf{P}(\|\widehat{m}_t - M\| \ge \varepsilon) \le c_1 t \exp\bigl(-c_2 t^{\min(1,\eta)} \min(\varepsilon^2, \varepsilon^{\eta})\bigr), \quad \forall\, t \ge 1, \ \forall\, \varepsilon \in (0, \varepsilon_0).$$

PROOF. Let $\widetilde{m}_t = \sum_{s,a,s'} m(s,a,s')\hat{q}_t(s,a,s')$. It follows that

$$\mathbf{P}(\|\widehat{m}_t - M\| \ge \varepsilon) \le \mathbf{P}(\|\widehat{m}_t - \widetilde{m}_t\| \ge \varepsilon/2) + \mathbf{P}(\|\widetilde{m}_t - M\| \ge \varepsilon/2). \tag{15}$$

We now proceed to bound the two terms in Equation (15), starting with the first one. We have

$$\begin{aligned}\widehat{m}_t - \widetilde{m}_t &= \frac{1}{t}\sum_{s,a,s'}\sum_{\tau=1}^{t} I_{\{s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}}(m_\tau - m(s,a,s')) \\ &= \sum_{s,a,s'} \frac{1}{t}\sum_{\tau=1}^{n_t(s,a,s')} \bigl(m^{\tau}_{s,a,s'} - m(s,a,s')\bigr),\end{aligned}$$

where $n_t(s,a,s')$ is defined in Equation (12). Thus,

$$\mathbf{P}(\|\widehat{m}_t - \widetilde{m}_t\| \ge \varepsilon/2) \le \sum_{s,a,s'} \mathbf{P}\left(\left\|\sum_{\tau=1}^{n_t(s,a,s')} \bigl(m^{\tau}_{s,a,s'} - m(s,a,s')\bigr)\right\| \ge \frac{\varepsilon t}{2S^2A}\right). \tag{16}$$

Let $\delta = \varepsilon/2S^2A$, and note that the event

$$E = \left\{ \left\| \sum_{\tau=1}^{n_t(s,a,s')} \left(m^{\tau}_{s,a,s'} - m(s,a,s') \right) \right\| \geq \delta t \right\} \tag{17}$$

is a subset of the event $\bigcup_{j=0}^{t} B_j$, where

$$B_j = \left\{ \left\| \sum_{\tau=1}^{j} \left(m^{\tau}_{s,a,s'} - m(s,a,s') \right) \right\| \geq \delta t \right\}.$$

From Assumption 6.1, we obtain

$$\mathbf{P}(B_j) \leq c_1 e^{-c_2 f(\delta t/j) j}.$$

Using the union bound, we obtain

$$\begin{aligned} \mathbf{P}(E) &\leq \sum_{j=0}^{t} \mathbf{P}(B_j) \\ &\leq c_1 \sum_{j=1}^{t} \exp(-c_2 f(\delta t/j) j) \\ &\leq c_1 \sum_{j=1}^{t} \exp(-c_2 c \delta^{\eta} t^{\eta} j^{1-\eta}), \end{aligned}$$

where the first inequality follows from the union bound and the third one from Assumption 6.1. If $\eta \geq 1$, we have $t^{\eta} j^{1-\eta} \geq t$; and if $\eta < 1$, we have $t^{\eta} j^{1-\eta} \geq t^{\eta}$. Thus,

$$\mathbf{P}(E) \leq c_1 t \exp\left(-c_2 c \delta^{\eta} t^{\min(1,\eta)}\right).$$

Substituting in Equation (16) and using the definition $\delta = \varepsilon/2S^2A$, we obtain

$$\mathbf{P}(\|\widehat{m}_t - \widetilde{m}_t\| \geq \varepsilon/2) \leq c_1' t \exp\left(-c_2' \varepsilon^{\eta} t^{\min(1,\eta)}\right),$$

for some new constants c_1', c_2'.

We now obtain a bound on the second term in Equation (15). According to our definitions, there is a linear transformation that maps $\widehat{q}_t$ to $\widetilde{m}_t$ and Q to M. It follows that $\|\widetilde{m}_t - M\| \leq c_3 \|\widehat{q}_t - Q\|$, for some constant c_3, so that $\mathbf{P}(\|\widetilde{m}_t - M\| \geq \varepsilon/2) \leq \mathbf{P}(\|\widehat{q}_t - Q\| \geq \varepsilon/2c_3)$. Using Theorem 5.1, we obtain $\mathbf{P}(\|\widetilde{m}_t - M\| \geq \varepsilon/2) \leq c_1'' \exp(-c_2'' \varepsilon^2 t)$, for some new positive constants c_1'' and c_2''. The result follows. □

Our last result is similar to Theorem 6.1 but pertains to the standard case of i.i.d. reward processes. The detailed proof is omitted because it is virtually identical to the proof of Theorem 6.1. For the purposes of the theorem, recall that the moment generating function of a vector-valued random variable Z taking values in $\mathbb{R}^m$ is defined to be $\mathbb{E}[e^{\langle s, Z \rangle}]$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^m$.

THEOREM 6.2. *Suppose that for every state-action-state triplet (s, a, s'), the corresponding rewards $m^{\tau}_{s,a,s'}$ are independent and identically distributed, and that the corresponding moment generating function is finite in a neigborhood of the origin. Then there exists a function λ: $(0, \infty) \mapsto (0, \infty]$ and a positive constant c_0, such that under any policy π,*

$$\mathbf{P}(\|\widehat{m}_t - M\| \geq \varepsilon) \leq c_0 e^{-\lambda(\varepsilon) t}, \quad \forall\, t \geq 1, \quad \forall\, \varepsilon > 0.$$

PROOF. (Outline) The only difference from the proof of Theorem 6.1 is in the bounds for $\mathbf{P}(E)$, where E is the event defined in Equation (17). Instead of using the union bound, one resorts to Lemma 5.2, as in the proof of Theorem 5.1, suitably modified to cover the vector case. □

7. Conclusions and future directions. We have provided a comprehensive characterization of the behavior of the empirical state-action frequency vectors in Markov decision

processes. We have specified a polytope of state-action frequency vectors, and we have shown that, under weak communicating assumptions, every point in the polytope is the (almost sure) limit of the empirical state-action frequency vector under some policy. We have further shown that, regardless of structural assumptions, the empirical state-action frequency vector converges to the polytope under every policy. Similar results were provided for the vector-valued reward case under some rather broad assumptions. We also note that the results of §3 are of independent interest because the available results of this type were limited to the unichain case.

There are several issues that call for further study. First, this work concerns only finite state and action spaces. It would be interesting to extend the results to the infinite case, under some strong ergodicity assumptions, as in Balaji and Meyn [3] and Glynn and Ormoneit [9], using, perhaps, ideas from Altman and Shwartz [1]. Second, the bounds in §4 involve a different exponent for every $q \in Q$. It should be possible to show that the exponent can be bounded away from zero, uniformly over all $q \in Q$, although the argument may become overly involved. Third, we did not make an attempt to optimize the exponents. It seems that optimizing the exponents (and making sure that the exponents in Theorems 5.1 and 4.1 match) is a difficult technical task.

Appendix

PROOF OF LEMMA 4.3. (a) Let t_k be the time at the end of the kth round. By that time, there is a total of $\tau^i(t_k) = \lceil \alpha^i k(k+1) \rceil$ transitions according to policy π, within each set $\mathcal{S}_i$, and $\tau^0(t_k)$ transitions according to the various switching policies $\rho_{s'}$. It follows that

$$\alpha^i(t_k - \tau^0(t_k)) - l \le \tau^i(t_k) \le \alpha^i(t_k - \tau^0(t_k)) + 1.$$

Suppose that $t_k \le t < t_{k+1}$. By our choice of the interval lengths, we have

$$t_{k+1} - \tau^0(t_{k+1}) \le t_k - \tau^0(t_k) + c'\sqrt{t},$$

for some constant c'. It follows that

$$\tau^i(t) \le \tau^i(t_{k+1}) \le \alpha^i(t_{k+1} - \tau^0(t_{k+1})) + 1 \le \alpha^i\bigl(t_k - \tau^0(t_k) + c''\sqrt{t}\bigr)$$

for some new constant c''. Because $t_k \le t$, we obtain $\tau^i(t) \le \alpha^i t + \alpha^i c''\sqrt{t}$, which proves the upper bound as long as $c_5 \ge \alpha^i c''$.

For the lower bound, note that

$$\tau^i(t) \ge \tau^i(t_k) \ge \alpha^i(t_k - \tau^0(t_k)) - l \ge \alpha^i\bigl(t_{k+1} - \tau^0(t_{k+1}) - c'''\sqrt{t}\bigr)$$

for some new constant c'''. Because $t - \tau^0(t)$ is nondecreasing in t, it follows that $\tau^i(t) \ge \alpha^i(t - \tau^0(t) - c'''\sqrt{t})$, which is of the desired form, as long as $c_5 \ge \alpha^i(1 + c''')$.

(b) Let $c_5' = c_5 + 1$. Suppose that $\varepsilon > c_5' + 1$. Then, the fact $\tau^0(t) \le t$ implies that

$$\mathbf{P}\bigl(\tau^0(t) + c_5\sqrt{t} + 1 \ge \varepsilon t\bigr) \le \mathbf{P}\bigl(\tau^0(t) + c_5'\sqrt{t} \ge \varepsilon t\bigr) \le \mathbf{P}(\tau^0(t) > t) = 0.$$

Thus, we need only to consider the case $\varepsilon \le c_5' + 1$.

It is easily checked that $t_k \ge k^2$. Thus, in t time steps, there can be at most $\lceil \sqrt{t} \rceil$ completed rounds. Thus, the total time $\tau^0(t)$ spent in switching from one set $\mathcal{S}_i$ to another consists of at most $l\lceil \sqrt{t} \rceil + l \le l'\lceil \sqrt{t} \rceil$ switching times, for a new constant l'. Using Lemma 4.2 and Equation (9), $\tau^0(t)$ is stochastically dominated by the sum of $l'\lceil \sqrt{t} \rceil$ random variables of the form $Z_j + c_4$, where the Z_j are independent exponentially distributed random variables with parameter β. Thus,

$$\mathbf{P}\bigl(\tau^0(t) + c_5'\sqrt{t} \ge \varepsilon t\bigr) \le \mathbf{P}\left(\sum_{j=1}^{l'\lceil \sqrt{t} \rceil} (Z_j + c_4 + c_5') \ge \varepsilon t\right).$$

Suppose first that

$$\varepsilon t \ge \left(c_4 + c_5' + \frac{1}{\beta}\right) l'\lceil \sqrt{t} \rceil + \frac{\varepsilon}{2} t. \tag{18}$$

The Chernoff bound yields

$$\mathbf{P}\left(\sum_{j=1}^{l'\lceil\sqrt{t}\rceil}(Z_j + c_4 + c_5') \geq \varepsilon t\right) \leq \mathbf{P}\left(\sum_{j=1}^{l'\lceil\sqrt{t}\rceil}\left(Z_j - \frac{1}{\beta}\right) \geq \frac{\varepsilon t}{2}\right) \leq e^{-l'\lceil\sqrt{t}\rceil f(\varepsilon t/(2l'\lceil\sqrt{t}\rceil))}.$$

Here, $f(\varepsilon) = \sup_s(\varepsilon s - \rho(s))$, where $\rho(s) = \log \mathbb{E}[e^{sZ_j - (s/\beta)}]$. We have $f(\varepsilon) = \beta\varepsilon - \log(1+\beta\varepsilon)$. Now, $f(0) = 0$ and f is convex, so that $f(ax) \geq af(x)$ for $a \geq 1$, and we get

$$\mathbf{P}\left(\sum_{j=1}^{l'\lceil\sqrt{t}\rceil}(Z_j + c_4 + c_5') \geq \varepsilon t\right) \leq e^{-l'tf(\varepsilon/2l')} \leq e^{-c_7\varepsilon^2 t}, \quad \forall\, t \geq 1, \ \ \forall\, \varepsilon \in [0, c_5' + 1].$$

The last inequality follows since the second derivative of $f(\varepsilon/2l')$ is $\beta^2/((2l')(1+\beta\varepsilon/2l')^2)$, and is bounded from below for $\varepsilon \in [0, c_5' + 1]$.

Finally, if Equation (18) does not hold, we have $c_7\varepsilon^2 t \leq c$, for some constant c, so that

$$\mathbf{P}\left(\tau^0(t) + c_5'\sqrt{t} \geq \varepsilon t\right) \leq 1 \leq c_8 e^{-c_7\varepsilon^2 t},$$

where $c_8 = e^c$. □

Proof of Lemma 5.1. Fix a point $y \notin W$. We have

$$\|y - W\| = \inf_{z\in\mathbb{R}^k}\left(\sum_{i=1}^k z_i^2\right)^{1/2} \quad \text{s.t.} \quad A(y+z) \leq b.$$

It follows that

$$\begin{aligned}\|y - W\| &\leq \min_{z\in\mathbb{R}^k} \sum_{i=1}^k |z_i| \\ &\quad \text{s.t.} \quad A(y+z) \leq b \\ &= \min_{z^+, z^-\in\mathbb{R}^k} e^\top z^+ + e^\top z^- \\ &\quad \text{s.t.} \quad Az^+ - Az^- \leq b - Ay \\ &\qquad z^+ \geq 0 \\ &\qquad z^- \geq 0,\end{aligned}$$

where $e \in \mathbb{R}^k$ is a vector with all components equal to 1, and where we rewrote z as a sum of positive and negative elements, that is, $z = z^+ - z^-$, where $z^+ = (z)^+$ and $z^- = (-z)^+$. Note that the right-hand side is the optimal cost in a linear programming problem. The optimal cost is finite (because it is bounded below by zero), and is attained. By the duality theorem for linear programming (Bertsimas and Tsitsiklis [5]) we can replace with the dual problem and obtain

$$\begin{aligned}\|y - W\| &\leq \max_{p\in\mathbb{R}^m} p^\top(b - Ay) \\ &\quad \text{s.t.} \quad p \leq 0 \\ &\qquad p^\top A \leq e^\top \\ &\qquad -p^\top A \leq e^\top,\end{aligned}$$

where the right-hand side is again finite. Let P denote the feasible set of the dual problem, i.e., $P = \{p \in \mathbb{R}^m \mid p \leq 0, -e^\top \leq p^\top A \leq e^\top\}$. The polyhedron P has at least one extreme

point (the point 0). Let $p_1, \ldots, p_l$ be the finite and nonempty set of extreme points of P. By the fundamental theorem of linear programming [5], the maximum over P is attained at some extreme point, so that

$$\|y - W\| \le \max_{1\le i\le l} p_i^\top (b - Ay) = \max_{1\le i\le l} (-p_i)^\top (Ay - b) \le \max_{1\le i\le l} (-p_i)^\top (Ay - b)^+,$$

where the last inequality follows because $-p_i \ge 0$. Let δ be a positive vector such that $-p_i \le \delta$ for all i, which exists because there are only finitely many vectors p_i. It follows that

$$\|y - W\| \le \delta^\top (Ay - b)^+ \le c\|(Ay - b)^+\|_\infty,$$

where $c = m\|\delta\|_\infty$. □

Proof of Lemma 5.2. Fix $\delta > 0$ and consider the sequence of random variables defined by

$$S_{k+1} = \begin{cases} S_k + X_{k+1}, & \text{if } S_k < \delta t, \\ S_k, & \text{if } S_k \ge \delta t. \end{cases}$$

Note that the following two events are identical:

$$\left\{\max_{1\le i\le t} \overline{X}_i \ge \delta t\right\} = \{S_t \ge \delta t\}.$$

We will bound the probability of the second event.

Note that $\rho(0) = 0$. Using Jensen's inequality, we have $\mathbb{E}[e^{sX_1}] \ge 1$ and therefore, $\rho(s) \ge 0$ for every $s \ne 0$. Let us fix some s for which $\rho(s)$ is finite, and consider the random variable

$$Y_k = \exp(sS_k - \min(k, N)\rho(s)),$$

where N is the stopping time defined by

$$N = \begin{cases} t, & \text{if } \overline{X}_n < \delta t \text{ for all } n = 1, \ldots, t, \\ \min\{n\colon \overline{X}_n \ge \delta t\}, & \text{otherwise.} \end{cases}$$

We claim that Y_k is a martingale. Indeed, if $k \ge N$, then $Y_{k+1} = Y_k$. If $N > k$, then

$$\begin{aligned} \mathbb{E}[Y_{k+1} \mid X_1, \ldots, X_k] &= \mathbb{E}[\exp(sS_{k+1} - (k+1)\rho(s)) \mid X_1, \ldots, X_k] \\ &= \exp(s\overline{X}_k - k\rho(s)) \cdot \mathbb{E}[\exp(sX_{k+1} - \rho(s))] \\ &= \exp(s\overline{X}_k - k\rho(s)) \\ &= Y_k. \end{aligned}$$

It follows that $\mathbb{E}[Y_t] = 1$. Because $N \le t$ and $\rho(s) \ge 0$, we have

$$\mathbb{E}[\exp(sS_t - t\rho(s))] \le \mathbb{E}[\exp(sS_t - \min(t, N)\rho(s))] = \mathbb{E}[Y_t] = 1. \tag{19}$$

From the Markov inequality, we obtain

$$\mathbf{P}(S_t \ge \delta t) = \mathbf{P}(e^{sS_t} \ge e^{s\delta t}) \le \frac{\mathbb{E}[e^{sS_t}]}{e^{s\delta t}},$$

and Equation (19) leads to

$$\mathbf{P}(S_t \ge \delta t) \le e^{-(s\delta - \rho(s))t}.$$

Because this is true for every s for which $\rho(s)$ is finite, we can take the infimum of the right-hand side over all s, which yields

$$\mathbf{P}(S_t \ge \delta t) \le e^{-f(\delta)t}.$$

(The last step rests on the observation that $\sup_s(\delta s - \rho(s))$ is not affected by restricting to those s for which $\rho(s)$ is finite.) Using a symmetrical argument, we also obtain that $\mathbf{P}(S_t \le -\delta t) \le e^{-f(-\delta)t}$. The fact that $f(\delta) > 0$ for $\delta \ne 0$, under our assumptions on $\rho(s)$, is well-known (see, e.g., Dembo and Zeitouni [6]).

We now consider the case of Bernoulli random variables. Assume that $X_1, \ldots, X_t$ are independent Bernoulli random variables with common mean p. Let $f(\delta; p) = \sup_s(\delta s - \rho(s))$, where $\rho(s) = \log \mathbb{E}[e^{s(X_1 - p)}]$. It suffices to prove that $f(\delta; p) \ge 2\delta^2$, for all δ and p. Indeed, a straightforward calculation shows that the rate function $f(\delta; p)$ is given by (see, e.g., Dembo and Zeitouni [6], p. 35):

$$f(\delta; p) = \begin{cases} (p+\delta)\log\left(\dfrac{p+\delta}{p}\right) + (1-p-\delta)\log\left(\dfrac{1-p-\delta}{1-p}\right), & \text{if } 0 \le \delta + p \le 1, \\ \infty, & \text{otherwise,} \end{cases}$$

with the convention that $0 \log 0 = 0 \log \infty = 0$. Let $f'(\delta; p)$ and $f''(\delta; p)$ be the first and second derivative, respectively, of $f(\delta; p)$, with respect to δ. We have $f(0; p) = f'(0; p) = 0$, and

$$f''(\delta; p) = \frac{1}{p+\delta} + \frac{1}{1-p-\delta} \ge 4, \quad \text{if } \delta \in (-p, 1-p).$$

It follows that $f(\delta; p) \ge 2\delta^2$, for $\delta \in [-p, 1-p]$. Outside that range, we have $f(\delta; p) = \infty$, which completes the proof of the claim. □

Acknowledgments. The authors are grateful to M. Puterman for helpful discussions and to two anonymous reviewers for helpful suggestions. This work was performed while the first author was with the Laboratory for Information and Decision Systems, MIT. This research was supported by the National Science Foundation under Grant ECS-0312921.

References

[1] Altman, E., A. Shwartz. 1991. Markov decision problems and state-action frequencies. *SIAM J. Control Optim.* **29**(4) 786–809.
[2] Altman, E., O. Zeitouni. 1994. Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality. *Math. Oper. Res.* **19**(4) 955–974.
[3] Balaji, S., S. P. Meyn. 2000. Multiplicative ergodicity and large deviations for an irreducible Markov chain. *Stochastic Processes Their Appl.* **90** 123–144.
[4] Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.
[5] Bertsimas, D., J. N. Tsitsiklis. 1999. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.
[6] Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York.
[7] Derman, C. 1970. *Finite State Markovian Decision Processes*. Academic Press, Orlando, FL.
[8] Gallager, R. G. 1996. *Discrete Stochastic Processes*. Kluwer Academics, Norwell, MA.
[9] Glynn, P. W., D. Ormoneit. 2002. Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.* **56** 143–146.
[10] Hajek, B. 1982. Hitting and occupation time bounds implied by drift analysis and applications. *Adv. Appl. Probab.* **14** 502–525.
[11] Hirsch, W. M. 1965. A strong law for the maximum cumulative sum of independent random variables. *Comm. Pure Appl. Math.* **18** 109–127.
[12] Kallenberg, L. C. M. 1983. Linear programming and finite Markovian control problems. Technical Report 148, Mathematics Centrum Tract, Amsterdam, The Netherlands.
[13] Kumar, P. R., P. Varaiya. 1986. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, Englewood Cliffs, NJ.
[14] Miller, H. D. 1961. A convexity property in the theory of random variables defined by a finite Markov chain. *Ann. Math. Statist.* **32**(4) 1260–1270.
[15] Puterman, M. 1994. *Markov Decision Processes*. Wiley-Interscience, New York.
[16] Shimkin, N. 1993. Extremal large deviations in controlled I.I.D. processes with applications to hypothesis testing. *Adv. Appl. Probab.* **25** 875–894.
[17] Sutton, R. S., A. G. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, MA.