Operations Research Letters 37 (2009) 123–126

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Fast convergence to state-action frequency polytopes for MDPs

Mathieu Tracol

LRI, University Paris-Sud, France

ARTICLE INFO

Article history:
Received 9 November 2008
Accepted 19 December 2008
Available online 30 January 2009

Keywords:
Markov Decision Processes
State-action frequencies
Polytopes
Deviation bounds

ABSTRACT

In the context of finite weakly communicating Markov Decision Processes, we tackle the problem of fast convergence of state-action frequency vectors to the polytope of stationary distributions on state-action frequencies. Using unichain policies, we derive bounds on the speed of convergence which are independent of the limit points.

1. Introduction

Consider a Markov Decision Process (MDP) with a finite state space S and a finite action set A. Given an execution run, i.e. a sequence of alternative states and actions, one is interested in the fraction of time a couple (s, a) of a state s and an action a is observed. It turns out that in many cases, when we consider reward MDPs, the associated "average state-action frequency" vectors give information for average costs on the system. The set of accumulation points of such vectors is a polytope H in $\mathbb{R}^{|S\times A|}$. Under some communication assumptions between the states of the system, H is characterized by a finite set of linear constraints [1]. The question of the almost sure convergence of empirical state-action frequency vectors to this polytope has been raised, as well as questions on the speed of convergence.

Under a weak communication assumption, we prove that we can find two positive constants C_1, C_2 such that, for any $x \in H$ and $\epsilon > 0$, we can find a policy on the system such that the probability that the associated empirical state-action frequency vector at time n is ϵ-far from x is smaller than $C_1 e^{-C_2\cdot\epsilon^2\cdot n}$. In order to achieve this result we consider *unichain policies*, which we link to the *end-components* of de Alfaro [2]. We follow the construction of [1] for the general policy. The bounds we derive come from a slight generalization of the work of Glynn and Ormoneit in [3], which gives a classical Chernoff bound in the context of mixing Markov chains.

Altman and Zeitouni, in [4], have given asymptotic large deviation bounds in a more restrictive context, and in particular for stationary policies. In this paper we give an answer to a problem posed in [1], where the authors derive a non-uniform bound on the speed of convergence: it depends on the point of the polytope.

In Section 2 we present the basic concepts for the study of state-action frequency vectors on an MDP. In Section 3 we present a particular kind of policy called a unichain policy. We give a bound on the speed of convergence of the empirical state-action frequency vector to its limit for unichain policies. In Section 4 we build a non stationary policy on an MDP such that the associated empirical state-action frequency vector converges quickly to a given point of the state-action frequency polytope.

2. Preliminaries

We consider finite-state, finite-action MDPs, which consist of a triple $\mathcal{S} = (S, A, P)$. $S = \{s_1, \ldots, s_n\}$ is the set of states, $A = \{a_1, \ldots a_l\}$ is a finite set of actions which is assumed, for simplicity, to be the same for all states. P is the conditional probability law: $P(s'|s, a)$ is the probability that the next state is s', given that the current state is s and that action a was taken. If X is a finite set, $\Delta(X)$ is the set of probability distributions on X. If E is an event, I_E is the associated characteristic function. $\|\ \|$ is a norm.

A *history* is a sequence, finite or infinite, of state and actions, which starts with a state (the initial state), and ends at a state if the history is finite. Ω is the set of infinite histories of our system. A *policy* is a mapping from the set of finite histories to the set $\Delta(A)$. A policy σ on $\mathcal{S}$ and an initial distribution α naturally induce a probability distribution $\mathbb{P}^{\sigma,\alpha}$ on the standard sigma-field Λ on the set of histories (See [5]). $X_i, Y_i, i \geq 0$ are the random variables which associate to a history h of length greater than i, respectively the state and the action it takes at time i. We will use several classes of policy: we write *HR* for the class of history dependent and randomized policies, *HD* for history dependent and deterministic, *SR* for stationary and randomized, and *SD* for

E-mail address: tracol@lri.fr.

0167-6377/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2008.12.003

stationary and deterministic policies. A stationary policy induces in a natural way a Markov chain on the state space of the system. In the future a finite Markov chain which is aperiodic and irreducible will be called *regular*.

Following [6], we can classify MDPs according to the structure of the set of Markov chains induced by stationary policies:

An MDP is *unichain* if the transition matrix corresponding to *every* deterministic stationary policy is unichain, that is, it consists of a single recurrent class of states plus a possibly empty set of transient states.

An MDP is *weakly communicating*, see [1], 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 to reach s' from s), and a set of states which are transient under *all* policies.

In the study of MDPs, and more generally in the study of controlled probabilistic processes, there has been research aiming to differentiate the respective powers of the different classes of policy. In this paper we consider the possible *state-action frequency vectors* that the different classes of policy can generate.

Definition 1 (*State-action Frequency Vectors*). Given a policy σ on $\mathcal{S}$, an initial distribution α, and $t \geq 1$, the *empirical state-action frequency vector* $\hat{x}_t$ at time t is the random vector on the probability space $(\Omega, \Lambda, P^{\sigma,\alpha})$ taking values in $\Delta(S \times A)$ with coordinates:

$$\hat{x}_t(s,a) = \frac{1}{t}\sum_{i=1}^{t} I_{s_i=s \wedge a_i=a}, \quad \forall\, s \in S, a \in A.$$

The *expected state-action frequency vector* at time t is the expectation of the empirical state-action frequency vector:

$$x_t^{\sigma,\alpha}(s,a) = \mathbb{E}^{\sigma,\alpha}\left[\frac{1}{t}\sum_{i=1}^{t} I_{s_i=s \wedge a_i=a}\right], \quad \forall\, s, \in S, a \in A.$$

If it exists, the associated *limit expected state-action frequency vector* is the limit $x^{\sigma,\alpha}$ of $x_t^{\sigma,\alpha}$ as t goes to infinity.

Remark 1. Given a Markov chain induced on S by a *stationary* policy σ, we know from [5], that for any initial distribution α, $x_t^{\sigma,\alpha}$ has a limit $x^{\sigma,\alpha}$ as t goes to infinity. Moreover, if the chain is regular, we know that $\hat{x}_t^{\sigma,\alpha}$ converges to $x^{\sigma,\alpha}$ as t goes to infinity with probability one. However, we do not have general results for the rate of this convergence. In fact, for a general Markov chain, $\hat{x}_t^{\sigma,\alpha}$ will not converge with probability one to a unique point. (Consider a process such that $X_n = X_2, \forall n \geq 2$).

The next couple of results are standard.

Proposition 1.
- *Suppose the probabilistic process induced on S by σ has only one recurrent class. Then there exists some x^σ such that $x_t^{\sigma,\alpha}$ converges to x^σ for any initial distribution α.*
- *Let σ be a stationary policy on $\mathcal{S}$, and α an initial distribution on S. Let $S = S_0 \cup S_1 \cup \cdots \cup S_l$ be the ergodic decomposition of S for σ: S_0 is the set of states of S transient under the policy σ, and the S_i are the irreducible classes of recurrent states. For all $i \in [1; l]$, σ induces an irreducible Markov chain on S_i, with associated limit state-action frequency vector x_i. For all $i \in [1; l]$, let $\lambda_i = \mathbb{P}^{\sigma,\alpha}(Reach(S_i))$, where $Reach(S_i)$ is the set of histories which reach (and stay in) S_i ultimately. Then we have the following:*

$$x^{\sigma,\alpha} = \sum_{i=1}^{l} \lambda_i \cdot x_i.$$

Proof. For the first point see [5]. For the second point: First, since the S_i, $i \in [1; l]$ form a partition of the set of recursive states, the λ_i sum to one. Given $(s,a) \in S \times A$, let Z_n be the random vector $\frac{\sum_{i=1}^{n} I_{X_i=s \wedge Y_i=a}}{n}$. By definition, $x^{\sigma,\alpha} = \lim_{n\to\infty} \mathbb{E}^{\sigma,\alpha}(Z_n)$. Let $J = \{i \in [1; l] \lambda_i > 0\}$. Let τ be the random time defined as the first time X_n enters in $\bigcup_{i\in J} S_i$. Let $N \in \mathbb{N}$, large enough such that for all $i \in J$ we have $\mathbb{P}^{\sigma,\alpha}(X_N \in S_i) > 0$. For $i \in J$, let

$$\mathbb{E}^{\sigma,\alpha}[Z_n | \{X_N \in S_i\}] = \frac{\mathbb{E}^{\sigma,\alpha}[Z_n \cdot I_{X_N \in S_i}]}{\mathbb{P}^{\sigma,\alpha}(X_N \in S_i)}.$$

Then $\mathbb{E}^{\sigma,\alpha}[Z_n | \{X_N \in S_i\}] \to x_i$ as n goes to infinity. Indeed, once X_n has entered S_i, it stays in it forever, and behaves as a regular Markov chain. Now,

$$\mathbb{E}^{\sigma,\alpha}[Z_n] = \mathbb{E}^{\sigma,\alpha}[Z_n \cdot I_{\{\tau>N\}}] + \sum_{i\in J} \mathbb{E}^{\sigma,\alpha}[Z_n \cdot I_{\{X_N \in S_i\}}].$$

The states in $S \setminus \bigcup_{i=1}^{l} S_i$ are transient, hence $\mathbb{P}^{\sigma,\alpha}(\{\tau > N\})$ goes to zero as N goes to infinity. Since $\lambda_i = \lim_{N\to\infty} \mathbb{P}^{\sigma,\alpha}(\{X_N \in S_i\})$ by definition, we get that $\mathbb{E}^{\sigma,\alpha}[Z_n]$ goes to $\sum_{i=1}^{l} \lambda_i \cdot x_i$ as n goes to infinity. □

Given a class of policy Π and an initial distribution α, we define the set of the possible limit points which can be reached using these policies. That is, for a class Π of policy,

$$X_\Pi^\alpha = \{x \in \Delta(S \times A) | \exists \sigma \in \Pi \text{ s.t. } x = x^{\sigma,\alpha}\}.$$

3. Unichain policies

In this section we define a more restrictive class of policy, *SDU*, which is rich enough for the associated X_{SDU} to contain the extreme points of X_{HR}. We present first the notion of *end components* introduced in [2].

Definition 2 (*Sub-MDP, End Component. [2]*). A *sub-MDP* of $\mathcal{S}$ is a pair (C, D) where $C \subseteq S$ and D is a function that associates to each $s \in C$ a set $D(s) \subseteq A$ of actions. A sub-MDP (C, D) induces a relation $\rho_{(C,D)}$ on $C \times S$, as:

$$\rho_{(C,D)} = \{(s, s') \in C \times S | \exists a \in D(s) \text{ s.t. } P(s'|s,a) > 0\}.$$

An *end component* on $\mathcal{S}$ is a sub-MDP such that:

- $\text{succ}(s,a) \subseteq C$ for all $s \in C$ and $a \in D(s)$.
- The graph $(C, \rho_{(C,D)})$ is strongly connected.

An end component (C, D) of $\mathcal{S}$ is a *deterministic end component* if for all $s \in C$ we have $|D(s)| = 1$.

Definition 3 (*Unichain Policy*). A *stationary* policy σ on $\mathcal{S}$ is *unichain* if the Markov chain it induces on the state space of $\mathcal{S}$ is unichain. We write SDU for the set of stationary, deterministic and unichain policies.

By Proposition 1, for a unichain policy σ on $\mathcal{S}$, $x^{\sigma,\alpha}$ is independent of the initial distribution α.

Chernoff bound analogues for Markov chains rely heavily on mixing conditions for the chain. In this paper we generalize a theorem of [3] to the context of unichain policies. [3] requires a mixing condition, closely related to the assumptions of uniform ergodicity of [5]:
A Markov chain X_n, $n \geq 0$ on S is *uniformly ergodic* if there exists constants λ, m and a probability measure ϕ on S, such that:

$$\mathbb{P}_x(X_m \in \cdot) \geq \lambda\phi(\cdot), \quad \text{for each } x \in S.$$

Here $\mathbb{P}_x(\cdot)$ denotes the conditional probability $\mathbb{P}(\cdot|X_1 = x)$, and more generally in the future $\mathbb{P}_{X_i=x}$ stands for $\mathbb{P}(\cdot|X_i = x)$. A finite regular Markov chain is uniformly ergodic [5]. However a general Markov chain, non-irreducible or non-periodic, is not (consider a cycle for instance).

Theorem 1 ([3]). *Let X_n, $n \geq 1$ be a Markov chain on S. Let $f : S \to \mathbb{R}$ bounded by $\|f\|$, set $Y_i \equiv f(X_i)$ and $S_n = \frac{\sum_{i=1}^n Y_i}{n}$. Suppose that X_n is uniformly ergodic. Then we have:*

$$\mathbb{P}_x(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq c_1 \cdot \mathrm{e}^{-c_2 \cdot \epsilon^2 \cdot n}$$

for $n > \frac{2\|f\|m}{\lambda \cdot \epsilon}$, where $c_1 = \mathrm{e}^{\frac{2\lambda\epsilon}{\|f\|m}}$, $c_2 = \frac{\lambda^2}{2\|f\|^2 m^2}$.

This theorem does not apply to general Markov chains, but we prove now that it remains valid for unichain Markov chains. (This extension was stated in Lemma 4.1 of [1] without proof). Our method works as follows: first we bound the time to enter the class of recurrent states, and next we use a union bound to deal with the possible period.

Proposition 2. *Let X_n, $n \geq 1$ be a unichain Markov chain on S. Let f, Y_i, and S_n be as before. Then we can find two positive constants $C_1, C_2 \in \mathbb{R}$ and $N \in \mathbb{N}$ such that:*

$$\forall x \in S,\ \forall \epsilon > 0, \forall n > N, \quad \mathbb{P}_x(|S_n - \mathbb{E}[S_n]| \geq \epsilon) \leq C_1 \cdot \mathrm{e}^{-C_2 \cdot n \cdot \epsilon^2}.$$

Proof. We decompose the state space S into disjoint sets S_1 and S_2 such that all states of S_1 are transient, and $S_2 = \{s_1, \ldots, s_l\}$ is a single class of recurrent states.

Once X_n has entered S_2, it behaves like an irreducible Markov chain on S_2. Let d be its period. Let $N \in \mathbb{N}$. Suppose $X_N \in S_2$. Then for any $j \in [0; d-1]$, $(X_{N+j+kd})_{k \geq 0}$ is a regular Markov chain on S_2. For $j \in [0; d-1]$, let $S_n^j = \frac{\sum_{i=1}^n Y_i \cdot \chi_{d\mathbb{N}+j}(i)}{n}$, where $\chi_{d\mathbb{N}+j}(i)$ is 1 iff i is of the form $d \cdot k + j$ with $k \in \mathbb{N}$. Let τ be the random time defined as the first n when $X_n \in S_2$. Then:

$$\mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon\right) \leq \mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon \wedge \tau \leq N\right) + \mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon \wedge \tau > N\right).$$

X_n has the Markov property, and if $\tau \leq N$, $X_N \in S_2$. Hence

$$\mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon\right) \leq \mathbb{P}_x(\tau > N) + \sum_{i=1}^{l} \mathbb{P}\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon | X_N = s_i\right). \tag{1}$$

Now we bound $\mathbb{P}_x(\tau \geq N)$ and

$$\mathbb{P}\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon | X_N = s_i\right), \quad \text{for } i \in [1; l].$$

Let $i \in [1, l]$. By the Theorem 1, we can find c_1^i, c_2^i, N_i such that $\mathbb{P}(\frac{S_n^j - \mathbb{E}[S_n^j]}{n/d} \geq \epsilon | X_0 = s_i) \leq c_1^i \cdot \mathrm{e}^{-c_2^i \epsilon^2 n}$ for $n \geq N_i$. It implies that for $N \geq N_i$ such that $d/N \leq \epsilon$,

$$\mathbb{P}\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n/d} \geq \epsilon | X_N = s_i\right) \leq c_3^i \cdot \mathrm{e}^{-4 \cdot c_2^i \epsilon^2 n} \tag{2}$$

with $c_3^i = 2c_1^i \cdot \mathrm{e}^{c_2^i \epsilon^2 N}$.

Now for $\mathbb{P}_x(\tau \geq N)$. Let $\delta = \mathrm{Min}_{x \in S} \mathbb{P}_x(X_{|S|} \in S_2)$. Then $\delta > 0$, and for all $n \geq 1$ and $x \in S$ we have $\mathbb{P}_x(X_n \in S_1) \leq (1-\delta)^{n/|S|}$. That is,

$$\mathbb{P}_x(\tau \geq n) \leq \mathrm{e}^{-C \cdot n} \tag{3}$$

where $C = -\frac{\ln(1-\delta)}{|S|}$

Finally, gathering (1)–(3),

$$\mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon\right) \leq 2c_1 \mathrm{e}^{c_2^i \epsilon^2 \cdot N} \cdot \mathrm{e}^{-4 \cdot c_2 \epsilon^2 \cdot n} + \mathrm{e}^{-C \cdot N}$$

for all $N \geq \max(N_i, d/\epsilon)$, $n \geq N$. Taking $N \geq n\epsilon^2 \cdot c_2/C$ we get:

$$\mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \epsilon\right) \leq c_5^i \cdot \mathrm{e}^{-c_6^i \cdot n \cdot \epsilon^2}$$

where $c_5^i = \max(3c_3^i, 1)$ and $c_6^i = \min(C/2, \epsilon^2 \cdot c_2^i)$.

Finally, $(S_n - \mathbb{E}[S_n]) = \sum_{j=0}^{d-1} S_n^j \left(\frac{S_n^j - \mathbb{E}(S_n^j)}{n}\right)$, so $\mathbb{P}_x((S_n - \mathbb{E}[S_n]) \geq \epsilon) \leq \mathrm{Max}_{j=0}^d \mathbb{P}_x\left(\frac{S_n^j - \mathbb{E}[S_n^j]}{n} \geq \frac{\epsilon}{d}\right)$, which proves the result, using the symmetry of the metric. □

Corollary 1. *Let σ be a unichain policy on $\mathcal{S}$. Let x^σ be the unique limit state-action frequency vector induced by σ. Then we can find two positive constants $C_1, C_2 \in \mathbb{R}$ and $N \in \mathbb{N}$ such that*

$$\mathbb{P}_\sigma^\alpha(\|\hat{x}_{\sigma,\alpha}^t - x^\sigma\| \geq \epsilon) \leq C_1 \mathrm{e}^{-C_2 \epsilon^2 t},$$
$$\forall\, t \geq N, \quad \forall\, \epsilon \in [0; 1], \forall \alpha \in \Delta(S).$$

The core of our method will lie in the following lemma.

Lemma 1. *Suppose $\mathcal{S}$ is weakly communicating. Let (C, D) be a deterministic end component of $\mathcal{S}$. Then there exists a policy $\sigma \in SDU$ such that:*

- *For all $s \in C$, $\sigma(s) = D(s)$.*
- *C is the irreducible class of recurrent states under σ.*

Proof. Since $\mathcal{S}$ is weakly communicating, we can partition S in two subsets S_c and S_t. S_c is a set of communicating states. (for each pair $(s, t) \in S_c$ there exist a deterministic scheduler on $\mathcal{S}$ which connects them). S_t is the set of states which are transient for all deterministic scheduler on $\mathcal{S}$. C is then a subset of S_c.
We construct recursively a growing sequence of sets of states $(B_i)_{i \geq 0}$ on which we define σ:

- $B_0 = C$. $\forall s \in C$, let $\sigma(s) = D(s)$.
- Given B_i, $i \geq 0$, B_{i+1} is the set of states s in $S - B_i$ such that there exist an action $a \in A(s)$ with $P(B_i|s, a) > 0$. For all $s \in B_{i+1}$ we fix such an action a_s. We define, for all $s \in B_{i+1}$, $\sigma(s) = \delta_{a_s}$.

This construction ends: in fact if $B_i \neq S$, then $B_{i+1} \geq |B_i| + 1$. Indeed, since $\mathcal{S}$ is weakly communicating it is not difficult to see that if $B_i \neq S$ then there exist $t \in S$ at distance one from B_i, that is such that $\exists a \in A$ s.t. $P(B_i|t, a) > 0$. Hence $\exists t \in S \setminus B_i$ s.t. $t \in B_{i+1}$. This proves in particular $B_{|S|} = S$. σ is clearly stationary and deterministic. Consider the Markov chain σ induces on S. By construction of σ, initiated on any $s \in S$, we have a positive probability of being in C after $|S|$ steps. C is an irreducible set of recurrent states of the chain by the irreducibility criterion in the definition of an end component, hence C is the unique irreducible class of recurrent states of the unichain policy σ on $\mathcal{S}$. □

By [6,7,1], we know that if $\mathcal{S}$ is weakly communicating, then the associated X_{HR}^α is independent of α and a closed convex set in $\mathbb{R}^{|S \times A|}$. Moreover $X_{HR}^\alpha = \mathrm{Conv}\,(X_{SD}^\alpha)$, where $\mathrm{Conv}\,(X_{SD}^\alpha)$ is the closed convex hull of X_{SD}^α. In the following we drop α in the notations when we deal with sets independent of the initial distribution.

Proposition 3. *Let $\mathcal{S}$ be a weakly communicating MDP. Then X^{α}_{SDU} is independent of α, and $X_{HR} = \mathrm{Conv}(X_{SDU})$. More precisely, X_{SDU} contains all the extremal points of X_{HR}.*

Proof. X^{α}_{SDU} is clearly independent of α, since $x^{\sigma,\alpha}$ is independent of α for any $\sigma \in SDU$. We just have to prove that for all $\sigma \in SD$ there exists $\sigma_1, \ldots, \sigma_l \in SDU$ and $\lambda_1, \ldots, \lambda_l \in [0, 1]$ such that $\sum_{i=1}^{l} \lambda_i = 1$ and $x^{\sigma,\alpha} = \sum_{i=1}^{l} \lambda_i \cdot x^{\sigma_i,\alpha}$.

Let $\sigma \in SD$. Let $S_0, \ldots, S_l$ be the associated ergodic decomposition of S. Let $\lambda_i, x_i, i \in [1; l]$ be as in Proposition 1. Then we know $x^{\sigma,\alpha} = \sum_{i=1}^{l} \lambda_i \cdot x^{\sigma_i,\alpha}$. According to Lemma 1, we know that for all $i \neq 0$ there exist a scheduler $\sigma \in SDU(\mathcal{S})$ whose recurrent class is S_i and which coincide with σ on S_i. By Proposition 1 we know that the limit $x^{\sigma_i,\alpha}$ associated to σ_i is independent of α, hence is equal to x_i. This proves the result. □

4. A uniform bound

This section is devoted to the proof of the bound on the speed of convergence to the limit points in the polytope. We can use the same construction as in [1] for the policy, using rounds. However this time we will get uniform bounds on the speed of convergence, independent of the point we choose in the polytope. In the following $\mathcal{S}$ is a weakly communicating MDP. Let $S = S_0 \cup S_1$ be the decomposition of the state space of $\mathcal{S}$ in a set of transient states (S_0), and a set of communicating set (S_1). Given $s, s' \in S$ and a stationary policy σ on $\mathcal{S}$, we define the random time $\tau^{\sigma}_{s,s'}$ to be the number of steps it takes for s' to be reached, starting from s, when policy σ is followed. The following lemma is the analogue of Lemma 4.2 of [1], in our context. It can be proved using the same tools as for the proof of Proposition 2.

Lemma 2 (*See [1]*). *We can find two positive constants C_1 and C_2 such that, for every $s' \in S_1$, there exists a deterministic and stationary policy $\sigma_{s'}$ such that for all $s \in S$, $t \geq 1$, $\tau^{\sigma_{s'}}_{s,s'}$ satisfies $\mathbb{P}^{\sigma_{s'}}(\tau^{\sigma_{s'}}_{s,s'} \geq t) \leq C_1 \cdot \mathrm{e}^{-C_2 \cdot t}$.*

We know that the extremal points of the polytope X_{HR} can be reached using unichain policies. There are a finite number of unichain policies on $\mathcal{S}$, hence a finite number of unichain policies σ whose associated x^{σ} is an extremal point of X_{HR}. We call such policies *extremal unichain* policies. By Corollary 1, we can find three positive constants c_1, c_2, N such that for any extremal unichain policy σ, for any initial distribution α on $\mathcal{S}$,

$$\mathbb{P}^{\sigma,\alpha}(\|\hat{x}^{\sigma,\alpha}_t - x^{\sigma}\| \geq \epsilon) \leq c_1 \cdot \mathrm{e}^{-c_2 \epsilon^2 \cdot t}, \quad \forall\, t \geq N.$$

Now we can conclude using a construction analogue of the construction in [1].

Theorem 2. *Suppose that $\mathcal{S}$ is a weakly communicating MDP. Then there exist two positive constants C_0, C_1, such that for all $x \in X_{HR}$ there exists a policy $\sigma \in HD(\mathcal{S})$ under which, $\forall t \geq 0$, $\forall \epsilon > 0$:*

$$\mathbb{P}^{\sigma,\alpha}(\|\hat{x}^{\sigma,\alpha}_t - x\| \geq \epsilon) \leq C_0 \mathrm{e}^{-C_1 \epsilon^2 t}, \quad \forall\, \alpha \in \mathrm{Dist}(S).$$

Proof. Let $\mathcal{S}$ and $x \in X_{HR}$ be as presented. By Proposition 3 we know that we can find $l \in \mathbb{N}$ such that for all $i \in [1, l]$ there exist $\lambda_i \geq 0$ and σ_i an extremal unichain policy such that if x_i is the unique stationary distribution on S for the Markov chain associated to σ_i, then $x = \sum_{i=1}^{l} \lambda_i \cdot x_i$. For each $i \in [1; l]$, let $S_i \subseteq S$ be the set of states recurrent for the Markov chain induced on S by σ_i. It is not difficult to see that the construction of [1] can be paralleled. We rewrite it to keep a complete procedure.First we define some time indexes. For $k \in \mathbb{N}$ and $i \in [1; l]$, let $t^i_k = \ulcorner \lambda_i k(k+1) \urcorner - \ulcorner \lambda_i (k-1) k \urcorner$. The interleaved policy σ is constructed by rounds:
Initialization. Initial state $s_0 \in S$, states $s^i_1 \in S_i$ for all $i \in [1; l]$.
For every round $k = 1, 2, \ldots$ do the following:

- Let s^*_k be the state at the end of the previous round. (for $k = 1$, $s^*_1 = s_0$).
- Use policy $\sigma_{s^1_k}$ until state $s^1_k \in S_1$ is reached.
- use policy σ_1 for t^1_k transitions. Let s^1_{k+1} be the final state.
- For $i = 2$ to l do the following:
  - Use policy $\sigma_{s^i_k}$ until state s^i_k is reached.
  - use policy σ_i for t^i_k transitions; let s^i_{k+1} be the final state.

Then the bounds derived in the tricky proof of [1] can be derived in the same way in our context. The difference being that the constants for the rate of convergence (in particular c_1 and c_2, which correspond to constant c_2, c_3 in [1]), which depended on x in [1], depend now only on the system in our case. Explicit c_1, c_2 can be derived from Proposition 2. Using the results of [1], explicit bounds can be given for the speed of convergence. □

References

[1] S. Mannor, J.N. Tsitsiklis, On the empirical state-action frequencies in Markov decision processes under general policies, Mathematics of Operations Research 30 (3) (2005) 545.
[2] L. Alfaro, Formal verification of probabilistic systems, Ph.D. Thesis, Stanford University, Stanford, CA, USA, 1997.
[3] P.W. Glynn, D. Ormoneit, Hoeffding's inequality for uniformly ergodic Markov chains, Statistics and Probability Letters 56 (2) (2002) 143–146.
[4] E. Altman, O. Zeitouni, Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality, Mathematics of Operations Research 19 (1994) 955.
[5] J.G. Kemeny, J.L. Snell, A.W. Knapp, Denumerable Markov Chains, Springer, 1976.
[6] Martin L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley, 1994.
[7] C. Derman, Finite State Markovian Decision Processes, Academic Press, Inc., Orlando, FL, USA, 1970.