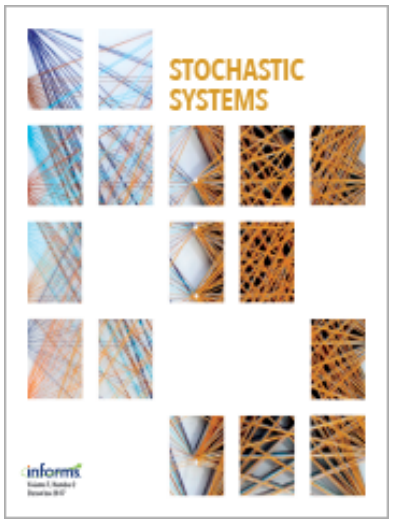


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Empirical Q-Value Iteration

Dileep Kalathil, Vivek S. Borkar, Rahul Jain

To cite this article:
Dileep Kalathil, Vivek S. Borkar, Rahul Jain (2020) Empirical Q-Value Iteration. Stochastic Systems

Published online in Articles in Advance 09 Oct 2020

. https://doi.org/10.1287/stsy.2019.0062

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–18
ISSN 1946-5238 (online)

Empirical Q-Value Iteration

Dileep Kalathil,[a] Vivek S. Borkar,[b] Rahul Jain[c]

[a] Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas 77843; [b] Department of Electrical Engineering, Indian Institute of Technology Mumbai, Mumbai 400076, India; [c] Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089

Contact: dileep.kalathil@tamu.edu, https://orcid.org/0000-0002-7403-4006 (DK); borkar.vs@gmail.com (VSB); rahul.jain@usc.edu (RJ)

Received: April 28, 2020
Accepted: April 29, 2020
Published Online in Articles in Advance: October 9, 2020

https://doi.org/10.1287/stsy.2019.0062

Abstract. We propose a new simple and natural algorithm for learning the optimal Q-value function of a discounted-cost Markov decision process (MDP) when the transition kernels are unknown. Unlike the classical learning algorithms for MDPs, such as Q-learning and actor-critic algorithms, this algorithm does not depend on a stochastic approximation-based method. We show that our algorithm, which we call the *empirical Q-value iteration* algorithm, converges to the optimal Q-value function. We also give a rate of convergence or a nonasymptotic sample complexity bound and show that an asynchronous (or online) version of the algorithm will also work. Preliminary experimental results suggest a faster rate of convergence to a ballpark estimate for our algorithm compared with stochastic approximation-based algorithms.

Funding: V. S. Borkar's work was supported in part by a J. C. Bose Fellowship and a grant for "Distributed Computation for Optimization over Large Networks and High Dimensional Data Analysis" from the Department of Science and Technology, Government of India. R. Jain and D. Kalathil's research was supported by the Office of Naval Research Young Investigator Award [N000141210766] and the National Science Foundation CAREER Award [0954116].

Keywords: dynamic programming • empirical methods • simulation • stochastic approximations

1. Introduction

The Q-learning algorithm of Watkins (Watkins 1989, Watkins and Dayan 1992) has been an early and one of the most popular and widely used algorithms for approximate dynamic programming for Markov decision processes. An important feature of this and other algorithms of this ilk (actor-critic, temporal difference(λ), least squares temporal difference, least squares policy evaluation, natural gradient, etc.) has been that they are *stochastic approximations*, that is, recursive schemes that update a vector *incrementally* based on observed payoffs (Jaakkola et al. 1994). This is achieved by using step sizes that are either decreasing slowly in a precise sense or equal a small positive constant. In either case, this induces a slower time scale for the iteration compared with the natural time scale on which the underlying stochastic phenomena evolve. Thus, the two time scale effects such as averaging kick in, ensuring that the algorithm effectively follows an averaged dynamics, that is, its original dynamics averaged out over the random processes affecting it on the natural time scale. The iterations are designed such that this averaged dynamics has the desired convergence properties. This extends even when the algorithm is asynchronous, for example, Q-learning (Tsitsiklis 1994). In fact, it can be generalized to stochastic approximations for general nonexpansive maps (Abounadi et al. 2002, Yu and Bertsekas 2013).

What we propose here is an alternative scheme for Q-learning that is *not* incremental and therefore evolves on the natural time scale. It does the usual Q-value iteration with the proviso that the conditional averaging with respect to the actual transition kernel of the underlying controlled Markov chain is replaced by a simulation-based empirical surrogate. One obvious advantage one might expect from this is that if it works, it will have much faster convergence. Indeed, this was observed earlier in Kearns and Singh (1999), who called it a phased-Q learning algorithm. A sample complexity result was provided via some back-of-the-envelope calculations although convergence is not implied. Our contribution is to provide a rigorous proof that it indeed works and provide simulation evidence that the expected fast convergence to a ballpark estimate is indeed a reality, although the theoretically predicted convergence is much slower. We first show that with fixed number of samples iterates almost surely converge to a random vector and then show that it coincides with the optimal Q-value function. Then, we obtain the rate of convergence and sample complexity bounds via a random operator analysis technique based on stochastic dominance.

The proof technique we use should be of independent interest as it is based on the constructs borrowed from the celebrated backward coupling scheme for exact simulation (Propp and Wilson 1996; see also Diaconis and Freedman 1999 for a discussion of the scheme and other related dynamics). In hindsight, this need not be surprising, as value and Q-value iterations in finite time yield finite horizon values/Q-values looking backward with the initial guess as the terminal cost.

There is an enormous amount of literature on reinforcement learning for approximate dynamic programming, and there is no point in even attempting a bird's eye view here. We refer the reader instead to the classic (Bertsekas and Tsitsiklis 1996) and its update in chapters 6 and 7 of Bertsekas (2012). Other related expositions are Sutton and Barto (1998), Szepesvári (2010), and Powell (2007).

We set up the framework and state the main result in the next section, followed by its proof in Section 3. Section 4 presents rate of convergence analysis, a nonasymptotic sample complexity bound, and its asynchronous and online extensions. Section 5 presents some simulation results, and Section 6 concludes with pointers to future possibilities.

2. Preliminaries and Main Result

2.1. Markov Decision Process

Consider a Markov decision process (MDP) on a finite state space $\mathbb{S}$ and a finite action space $\mathbb{A}$. Let $\mathcal{P}(\mathbb{A})$ denote the space of all probability measures on $\mathbb{A}$. Also given is a transition kernel

$$p : (s, a, s') \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \mapsto p(s'|s, a) \in [0, 1]$$

satisfying $\sum_{s' \in \mathbb{S}} p(s'|s, a) = 1$. Let $c : \mathbb{S} \times \mathbb{A} \to \mathbb{R}_+$ denote the cost function that depends on the state-action pair.

An MDP is a controlled Markov chain $\{X_t\}$ on the set $\mathbb{S}$ controlled by an $\mathbb{A}$-valued control process $\{Z_t\}$ such that $P(X_{t+1} = s|X_r, Z_r, r \le t) = p(s|X_t, Z_t)$. Define Π to be the class of *stationary randomized policies*: mappings $\pi : \mathbb{S} \to \mathcal{P}(\mathbb{A})$ such that $\pi(X_t)$ is the conditional distribution of Z_t given $\{X_r, Z_r, r < t; X_t\}$ for all t. Our objective is to minimize over all admissible $\{Z_t\}$ the infinite horizon discounted cost $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(X_t, Z_t)]$ where $\gamma \in (0, 1)$ is the discount factor. It is well known that Π contains an optimal policy that minimizes the infinite horizon discounted cost (Puterman 2005). Also, let Σ denote the set of nonstationary policies $\{\sigma_t\}$ with $\sigma_t : \mathbb{S} \to \mathcal{P}(\mathbb{A})$, that is, $\sigma_t(X_t)$ is the conditional distribution of Z_t given $\{X_r, Z_r, r < t; X_t\}$ for each t.

For any $\pi \in \Pi$, we define the transition probability matrix P^π as

$$P^\pi(s, s') := \sum_{a \in \mathbb{A}} p(s'|s, a)\pi(s, a). \tag{1}$$

We make the following assumption.

Assumption 1. *For any $\pi \in \Pi$, the Markov chain defined by the transition probability matrix P^π is irreducible and aperiodic.*

Remark 1. By Assumption 1, for any $\pi \in \Pi$, there exists a positive integer r_π such that, $(P^\pi)^{r_\pi}(s, s') > 0, \forall s, s' \in \mathbb{S}$, where $(P^\pi)^{r_\pi}(s, s')$ denotes the (s, s')th element of the matrix $(P^\pi)^{r_\pi}$ (Levin et al. 2009, proposition 1.7, p. 8).

Define the optimal value function $V^* : \mathbb{S} \to \mathbb{R}_+$ as

$$V^*(s) = \inf_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(X_t, \pi(X_t)) \middle| X_0 = s\right]. \tag{2}$$

Also define the Bellman operator $T : \mathbb{R}_+^{|\mathbb{S}|} \to \mathbb{R}_+^{|\mathbb{S}|}$ as

$$T(V)(s) := \min_{a \in \mathbb{A}} \left[c(s, a) + \gamma \sum_{s'} p(s'|s, a)V(s')\right]. \tag{3}$$

The Bellman operator is a contraction mapping, that is, $\|T(V) - T(V')\|_\infty \le \gamma\|V - V'\|_\infty$, and the optimal value function V^* is the unique fixed point of $T(\cdot)$. Given the optimal value function, an optimal policy π^* can be calculated as (Puterman 2005)

$$\pi^*(s) \in \arg\min_{a \in \mathbb{A}} \left[c(s, a) + \gamma \sum_{s'} p(s'|s, a)V^*(s')\right]. \tag{4}$$

2.2. Value Iteration, *Q*-Value Iteration

A standard scheme for finding the optimal value function (and hence an optimal policy) is *value iteration*. One starts with an arbitrary function V_0. At the kth iteration, given the current iterate V_k, we calculate $V_{k+1} = TV_k$. Because $T(\cdot)$ is a contraction mapping, by Banach fixed point theorem, $V_k \to V^*$.

Another way to find the optimal value function is via Q-value iteration. Although this requires more computation than the value iteration, Q-value iteration is extremely useful in developing learning algorithms for MDPs.

Define the Q-value operator $G : \mathbb{R}_+^d \to \mathbb{R}_+^d$ as

$$G(Q)(s,a) := c(s,a) + \gamma \sum_{s' \in \mathbb{S}} p(s'|s,a) \min_b Q(s',b), \tag{5}$$

where $d = |\mathbb{S}||\mathbb{A}|$. Similar to the Bellman operator T, Q-value operator G is also a contraction mapping, that is, $\|G(Q) - G(Q')\|_\infty \le \gamma \|Q - Q'\|_\infty$. Let Q^* be the unique fixed point of $G(\cdot)$, that is,

$$Q^*(s,a) = c(s,a) + \gamma \sum_{s' \in \mathbb{S}} p(s'|s,a) \min_b Q^*(s',b).$$

This Q^* is called the optimal Q-value. By the uniqueness of V^*, it is clear that $V^* = \min_{a \in \mathbb{A}} Q^*(s,a)$. Thus, given Q^*, one can compute V^* and hence an optimal policy π^*.

The standard method to compute Q^* is Q-value iteration. We start with an arbitrary Q_0 and then update $Q_{k+1} = G(Q_k)$. Because of the contraction property of G, $Q_k \to Q^*$ almost surely.

2.3. Empirical Q-Value Iteration for MDPs

The Bellman operator T and the Q-value operator G require the knowledge of the exact transition kernel $p(\cdot|\cdot,\cdot)$. In practical applications, these transition probabilities may not be readily available, but it may be possible to simulate a transition according to any of these probabilities. Without loss of generality, we assume that the MDP is driven by uniform random noise according to the simulation function

$$\psi : \mathbb{S} \times \mathbb{S} \times [0,1] \to \mathbb{S} \quad \text{such that} \quad \Pr\big(\psi(s,a,\xi) = s'\big) = p(s'|s,a), \tag{6}$$

where ξ is a random variable distributed uniformly in $[0,1]$. Using this convention, the Q-value operator can be written as

$$G(Q)(s,a) := c(s,a) + \gamma \, \mathbb{E}\left[\min_b Q\big(\psi(s,a,\xi), b\big)\right]. \tag{7}$$

In the *empirical Q-value iteration* (EQVI) algorithm, we replace the expectation in the previous equation with an empirical estimate. Given a sample of n i.i.d. random variables distributed uniformly in $[0,1]$, denoted $\{\xi_i\}_{i=1}^n$, the empirical estimate of $\mathbb{E}[\min_b Q(\psi(s,a,\xi),b)]$ is $\frac{1}{n}\sum_{i=1}^n \min_b Q(\psi(s,a,\xi_i),b)$. We summarize our *empirical Q-value iteration* algorithm here.

Algorithm 1: EQVI Algorithm

Input: $\widehat{Q}_0 \in \mathbb{R}_+^d$, sample size $n \ge 1$, maximum iterations $k_{\max}$. Set counter $k = 0$.

1. For each $(s,a) \in \mathbb{S} \times \mathbb{A}$, sample n uniformly distributed random variables $\{\xi_i^k(s,a)\}_{i=1}^n$, and compute

$$\widehat{Q}_{k+1}(s,a) = c(s,a) + \gamma \, \frac{1}{n} \sum_{i=1}^n \left(\min_b \widehat{Q}_k\big(\psi\big(s,a,\xi_i^k(s,a)\big), b\big)\right).$$

2. Increment $k \leftarrow k+1$. If $k > k_{\max}$, STOP. Else, return to Step 1.

We introduce some notation to state our results precisely. Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be the probability space of one-sided infinite sequences $\omega = (\omega_k : k \in \mathbb{Z}^*)$, where $\mathbb{Z}^*$ is the set of nonnegative integers. Each element ω_k is a vector, $\omega_k = (\xi_i^k(s,a), 1 \le i \le n, s \in \mathbb{S}, a \in \mathbb{A})$, where $\xi_i^k(s,a)$ is a random noise distributed uniformly in $[0,1]$. We assume that $\xi_i^k(s,a)$ are i.i.d. $\forall i$, $\forall (s,a) \in \mathbb{S} \times \mathbb{A}$ and $\forall k \in \mathbb{Z}^*$. $\mathbb{E}_1$ denotes expectation with respect to measure $\mathbb{P}_1$.

Our main result then is the following.

Theorem 1. *For a given $\omega \in \Omega_1$, let $\widehat{Q}_k(\omega), k \ge 0$, be the corresponding Q-value iterates as defined in Algorithm 1. Then, there exists a random variable $Q^*(\omega)$ such that $\widehat{Q}_k(\omega) \to Q^*(\omega)$, $\omega - a.s.$*

The main idea that we exploit is the fact that (exact) Q-value iteration in finite time is equal to finite horizon Q-values obtained by *looking backward* with the initial guess as the terminal cost. More precisely, when the transition kernels $p(\cdot|\cdot,\cdot)$ are known, the kth iterate Q_k of the (exact) Q-value iteration is obtained via the

iteration $Q_k = G(Q_{k-1})$ (compare with (5)) with an initial guess Q_0. One can show that this Q_k is equal to Q'_k, which is the Q-value obtained by looking backward where

$$Q'_k(s,a) = \mathbb{E}\left[\sum_{l=-k}^{-1} \gamma^{l+k} c\left(X'_l, Z'_l\right) + \gamma^k Q_0\left(X'_0, Z'_0\right) | X'_{-k} = s, Z'_{-k} = a\right].$$

Therefore, rather than showing that the forward iteration Q_k converges to the optimal Q-value function Q^*, one can also establish the convergence of the (exact) Q-value iteration by showing that the backward iterate Q'_k converges to Q^* almost surely. When the transition kernels are known, this is obviously a convoluted route because the convergence of the forward iteration $Q_{k+1} = G(Q_k)$ is immediate by the contraction property of G.

However, when the transition kernels are unknown, it is not clear if we can directly prove the convergence of the (simulation-based) forward iteration sequence $\widehat{Q}_k(\omega)$ (given in Algorithm 1 and formalized in Equation (14)) to the optimal Q-value function Q^*. To overcome this difficulty, we take the approach mentioned above and define the (simulation-based) backward iteration sequence $\widetilde{Q}_k(\omega)$ (compare with Equation (26)) similar to the Q'_k, and we rigorously show that $\widehat{Q}_k(\omega) = \widetilde{Q}_k(\omega), \forall \omega$ (compare with Proposition 2). Then, using an approach similar to the well-known Propp-Wilson backward simulation algorithm (Propp and Wilson 1996), we show that $\widetilde{Q}_k$ (and hence $\widehat{Q}_k$) converges to a random variable $Q^*(\omega)$ almost surely (compare with Proposition 1).

We can further establish that the random limit $Q^*(\omega)$ in Theorem 1 is indeed a constant almost surely.

Corollary 1. *The empirical Q-value iteration converges to the optimal Q-value function, that is,* $\widehat{Q}_k \to Q^*$ *as* $k \to \infty$ *for any fixed* n.

We also provide a rate of convergence, or a nonasymptotic sample complexity bound. This follows from methods that had been developed in Haskell et al. (2013) for empirical value and policy iteration, which, however, only provide a convergence in probability guarantee.

Let $\widehat{Q}_k^n$ be the kth iterate of EQVI when using n samples. Then,

Theorem 2. *Given* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$, *fix* $\epsilon_g = \epsilon/\eta^*$ *and select* $\delta_1, \delta_2 > 0$ *such that* $\delta_1 + 2\delta_2 \le \delta$ *where* $\eta^* = \lceil 2/(1-\gamma) \rceil$. *Select an* n *such that*

$$n \ge n(\epsilon, \delta) = \frac{(\kappa^*)^2}{2\epsilon_g^2} \log \frac{2|\mathbb{S}||\mathbb{A}|}{\delta_1},$$

where $\kappa^* = \max_{(s,a)\in\mathbb{K}} c(s,a)/(1-\gamma)$ *and select a* k *such that*

$$k \ge k(\epsilon, \delta) = \log\left(\frac{1}{\delta_2\, \mu_{n,\min}}\right).$$

Then,

$$\mathbb{P}_1\left(\|\widehat{Q}_k^n - Q^*\| \ge \epsilon\right) \le \delta.$$

Here $\mu_{n,\min} = \min_i \mu_n(i)$, *and* $\mu_n(i)$ *is given by*

$$\mu_n(\eta^*) = p_n^{N^*-\eta^*-1}, \qquad \mu_n(N^*) = \frac{1-p_n}{p_n},$$
$$\mu_n(i) = (1-p_n)p_n^{(N^*-i-1)}, \qquad \forall i = \eta^*+1, \ldots, N^*-1,$$
$$p_n = 1 - 2|\mathbb{S}||\mathbb{A}|e^{-2(\epsilon/\gamma)^2 n/((\kappa^*)^2)}, \quad N^* = \left\lceil \frac{\kappa^*}{\epsilon_g} \right\rceil.$$

2.4. Comparison with Classical *Q*-Learning

Synchronous variant of the classical Q-learning algorithm for discounted MDPs works as follows (see Bertsekas and Tsitsiklis 1996, section 5.6). For every state-action pair $(s,a) \in \mathbb{S} \times \mathbb{A}$, we maintain a Q-value function and use the update rule

$$Q_{k+1}(s,a) = Q_k(s,a) + \alpha_k\left(c(s,a) + \gamma \min_{b\in\mathbb{A}} Q_k\left(\psi\left(s,a,\xi^k(s,a)\right), b\right) - Q_k(s,a)\right), \tag{8}$$

where $\xi^k(s,a)$ is a random noise sampled uniformly from $[0,1]$ and $\{\alpha_k, k \geq 0\}$ is the standard stochastic approximation step sequence such that $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$. It can be shown that $Q_k \to Q^*$ almost surely (Bertsekas and Tsitsiklis 1996). The rate of convergence depends on the sequence $\{\alpha_k, k \geq 0\}$ (Borkar 2008). In general, the convergence is very slow.

The *empirical Q-value iteration* algorithm does not use stochastic approximation and is a nonincremental scheme. The rate of convergence will depend on the number of noise samples n.

3. Proof of Theorem 1

In the following, we first formally set the notations for the underlying probability space and define EQVI iterate $\widehat{Q}_k$ using those notations (14). Then we define the forward simulation model for controlled Markov chains (Equation (19)) and show the finite time coupling property of this simulated chain (Proposition 1). Then we define the backward simulation model for controlled Markov chain (24). Equipped with these notions, we proceed to prove Proposition 2. Finally, we will give the proof for the main results Theorem 1 and for the corollary

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be the probability space of one-sided infinite sequences ω such that $\omega = \{\omega_k : k \in \mathbb{Z}^*\}$, where $\mathbb{Z}^*$ is the set of nonnegative integers. Each element ω_k of the sequence is a vector $\omega_k = (\xi_i^k(s,a), 1 \leq i \leq n, s \in \mathbb{S}, a \in \mathbb{A})$, where $\xi_i^k(s,a)$ is a random noise distributed uniformly in $[0,1]$. We assume that $\xi_i^k(s,a)$ are i.i.d. $\forall i$, $\forall (s,a) \in \mathbb{S} \times \mathbb{A}$, and $\forall k \in \mathbb{Z}^*$. $\mathbb{E}_1$ denotes expectation with respect to measure $\mathbb{P}_1$.

For each $k \in \mathbb{Z}^*$, θ_k denotes the left shift operator, that is,

$$\theta_k \omega := \{\omega_{\tau+k} : \tau \geq 0\}. \tag{9}$$

Also, let Γ be the projection operator such that $\Gamma(\theta_k \omega) = \omega_k, \forall k \in \mathbb{Z}^*$, $\forall \omega \in \Omega_1$. Recall that ψ is the simulation function defined in Equation (6) such that

$$\mathbb{P}_1\big(\psi\big(s,a,\xi_i^k(s,a)\big) = s'\big) = p(s'|s,a), \quad \forall i,k. \tag{10}$$

Using ψ, for each $\omega \in \Omega_1$, we define a sequence of empirical transition kernels $\widehat{p}(\omega) = (\widehat{p}_k(\omega_k))_{k \geq 0}$ as

$$\widehat{p}_k(s'|s,a) := \frac{1}{n} \sum_{i=1}^{n} I\{\psi\big(s,a,\xi_i^k(s,a)\big) = s'\}. \tag{11}$$

We dropped ω_k from the previous definition for ease of notation. For any $\pi \in \Pi$, we also define the transition probability matrix $\widehat{P}_k^\pi$ as

$$\widehat{P}_k^\pi(s,s') := \sum_{a \in \mathbb{A}} \widehat{p}_k(s'|s,a)\pi(s,a). \tag{12}$$

The rows of $\widehat{P}_k^\pi$ are independent because of the independence assumption on the elements of the vector ω_k. Also, $\widehat{P}_k^\pi$ are independent $\forall k$.

We define the *empirical Q-value operator* $\widehat{G} : \Omega_1 \times \mathbb{R}_+^d \to \mathbb{R}_+^d$ as

$$\begin{aligned} \widehat{G}(\theta_k\omega, Q)(s,a) &:= \check{G}_n(\Gamma(\theta_k\omega), Q)(s,a) \\ &:= c(s,a) + \gamma \frac{1}{n} \sum_{i=1}^{n} \min_b Q\big(\psi\big(s,a,\xi_i^k(s,a)\big), b\big) \\ &= c(s,a) + \gamma \sum_{s'} \widehat{p}_k(s'|s,a) \min_b Q(s',b). \end{aligned} \tag{13}$$

Then, the empirical Q-value iteration given in Algorithm 1 can be succinctly represented as

$$\widehat{Q}_{k+1}(\omega) = \widehat{G}\Big(\theta_k\omega, \widehat{Q}_k(\omega)\Big). \tag{14}$$

We drop ω from the notation of $\widehat{Q}_k$ whenever it is not necessary.

From Equations (10) and (13), for any fixed Q,

$$\mathbb{E}_1\Big[\widehat{G}(\theta_k\omega, Q)\Big] = G(Q), \ \forall k \in \mathbb{Z}^*, \tag{15}$$

where G is the Q-value operator defined in Equation (5).

We define another probability space $(\Omega_2 = \Omega_2' \times \Omega_2', \mathcal{F}_2, \mathbb{P}_2)$ of one-sided infinite sequences μ such that $\mu = \{(\nu_k, \tilde{\nu}_k), k \in \mathbb{Z}^*\}$. Here $\nu = \{\nu_k : k \in \mathbb{Z}^*\} \in \Omega_2'$. Each element ν_k of the sequence ν is a $|\mathbb{S}||\mathbb{A}|$-dimensional vector, $\nu_k = (\nu_k(s,a), s \in \mathbb{S}, a \in \mathbb{A})$ where $\nu_k(s,a)$ is a random variable distributed uniformly in $[0,1]$. We assume that $\nu_k(s,a)$ are i.i.d. $\forall (s,a) \in \mathbb{S} \times \mathbb{A}$ and $\forall k \in \mathbb{Z}^*$. Likewise, let $\tilde{\nu} = \{\tilde{\nu}_k : k \in \mathbb{Z}^*\} \in \Omega_2'$. Each element $\tilde{\nu}_k$ of the sequence $\tilde{\nu}$ is a $|\mathbb{S}|$-dimensional vector, $\tilde{\nu}_k = (\tilde{\nu}_k(s), s \in \mathbb{S})$, where $\tilde{\nu}_k(s)$ is a random variable distributed uniformly in $[0,1]$. We assume that $\tilde{\nu}_k(s)$ are i.i.d., independent of ν, $\forall s \in \mathbb{S}$ and $\forall k \in \mathbb{Z}^*$. $\mathbb{E}_2$ denotes the expectation with respect to $\mathbb{P}_2$. Let $\mathbb{P}$ be the product measure, $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$ and let $\mathbb{E}$ denote the expectation with respect to $\mathbb{P}$.

For each $\omega \in \Omega_1$, that is, for each sequence of transition kernels $\widehat{p}(\omega) = (\widehat{p}_k(\omega_k))_{k \geq 0}$, we define a sequence of simulation functions $(\phi_k = (\phi_k^1, \phi_k^2))_{k \geq 0}$ as

$$\phi_k^1 : \ \mathbb{S} \times \mathbb{A} \times \Omega_2' \to \mathbb{S}, \tag{16}$$

$$\phi_k^2 : \ \mathbb{S} \times \Omega_2' \to \mathbb{A}, \tag{17}$$

such that

$$\mathbb{P}_2\big(\phi_k^1(s,a,\nu_k(s,a)) = s'\big) = \widehat{p}_k(s'|s,a), \tag{18}$$

and ϕ_k^2 is the (randomized) control strategy that maps the output of the function ϕ_k^1 to an action space-valued random variable $\phi_k^2(\phi_k^1(s,a,\nu_k(s,a)), \tilde{\nu}_k(\phi_k^1(s,a,\nu_k(s,a))))$. We note that the control strategy can be identified with an element π, respectively, σ, of the set Π or Σ, when, respectively,

$$\mathbb{P}_2\big(\phi_k^2(s,\tilde{\nu}(s)) = a\big) = \pi(s,a) \ \text{ or } \mathbb{P}_2\big(\phi_k^2(s,\tilde{\nu}(s)) = a\big) = \sigma_k(s,a).$$

In such a case, we write $\phi_k^2 \approx \pi$ or $\phi_k^2 \approx \sigma_k$ as the case may be.

For $k_2 > k_1$, define the composition function $\Phi_{k_1}^{k_2}$ as

$$\Phi_{k_1}^{k_2} := \phi_{k_2-1} \circ \phi_{k_2-2} \circ \cdots \circ \phi_{k_1}. \tag{19}$$

Given an $\omega \in \Omega_1, \nu \in \Omega_2$ and an initial condition (s_0, a_0), we can *simulate* an MDP with state-action sequence $(X_k(\omega,\nu), Z_k(\omega,\nu))_{k \geq 0}$ as follows:

$$(X_k(\omega,\nu), Z_k(\omega,\nu)) = \Phi_0^k(s_0,a_0) \ \text{ and } \ (X_{k+1}(\omega,\nu), Z_{k+1}(\omega,\nu)) = \phi_k \circ \Phi_0^k(s_0,a_0). \tag{20}$$

We call this simulation method as *forward simulation*. The dependence on the control strategy ϕ_k^2 is implicit and is not used in the notation. Whenever not necessary, we also drop ω and ν from the notation and denote the simulated chain by (X_k, Z_k). Because

$$\mathbb{P}_2(X_{k+1}|X_m, Z_m, m \leq k) = \widehat{p}_k(X_{k+1}|X_k, Z_k),$$

the sequence $(X_k(\omega,\nu))_{k \geq 0}$ is a *controlled Markov chain*.

Consider two controlled Markov chains $X_k^1(\omega,\nu), X_k^2(\omega',\nu'), k \geq k_0$, with different initial conditions, defined on $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathbb{P} \times \mathbb{P}')$ where $(\Omega', \mathcal{F}', \mathbb{P}')$ is another copy of $(\Omega, \mathcal{F}, \mathbb{P})$. Define the *coupling time*, $\widetilde{\tau}_{\omega^*,\nu^*}$, for $\omega^* := (\omega,\omega'), \nu^* := (\nu,\nu')$, as $\widetilde{\tau}_{\omega^*,\nu^*}(s_0^1, s_0^2) := :$

$$\min\Big\{m \geq 0 : X_{k_0+m}^1(\omega,\nu) = X_{k_0+m}^2(\omega',\nu'), X_{k_0}^1(\omega,\nu) = s_0^1, X_{k_0}^2(\omega',\nu') = s_0^2\Big\}. \tag{21}$$

We prove that the expected value of the coupling time is finite.

Proposition 1. *Let $(X_k^1(\omega,\nu), Z_k^1)_{k \geq k_0}, (X_k^2(\omega',\nu'), Z_k^2)_{k \geq k_0}$ be two sequences of state-action pairs for an MDP simulated according to (20) using an arbitrary control strategy $\phi_k^2 \approx \sigma_k$. Let $\widetilde{\tau}_{\omega^*,\nu^*}$ be the coupling time as defined in Equation (21). Then,*

$$\mathbb{E}\Big[\widetilde{\tau}_{\omega^*,\nu^*}\big(s_0^1, s_0^2\big)\Big] < \infty, \ \ \forall s_0^1, s_0^2 \in \mathbb{S}.$$

Proof is given in the Appendix.

We now consider the *backward simulation* of an MDP. This is similar to the *coupling from the past* idea introduced in Propp and Wilson (1996). For us, this is a proof technique, a *thought experiment*, and not the actual algorithm.

Given $\omega \in \Omega_1, \nu \in \Omega_2$, the sequence of simulation functions $(\phi_k = (\phi_k^1, \phi_k^2))$, a $k_0 > 0$, and an initial condition $\widetilde{X}_{-k_0}(\omega,\nu) = s_0, \widetilde{Z}_{-k_0}(\omega,\nu) = a_0$, we simulate a controlled Markov chain $(\widetilde{X}_m(\omega,\nu))_{m=-k_0}^0$ of length k_0+1 using the backward simulation. As a first step, we do an offline computation of all possible simulation trajectories as follows:

1. Input k_0. Initialize $m = -1$.
2. Compute $\widetilde{\phi}_m^1(s,a,\nu_{-m}(s,a)) := \phi_{-m}^1(s,a,\nu_{-m}(s,a)),\ \forall (s,a) \in \mathbb{S}\times\mathbb{A}$.
3. $m \leftarrow m-1$. If $m < -k_0$, stop. Else, return to Step 2.

Then we simulate $(\widetilde{X}_m(\omega,\nu))_{m=-k_0+1}^0$ as

$$\widetilde{X}_m = \widetilde{\phi}_m^1\left(\widetilde{X}_{m-1}, \widetilde{Z}_{m-1}, \nu_{-(m-1)}\left(\widetilde{X}_{m-1}, \widetilde{Z}_{m-1}\right)\right), \tag{22}$$

$$\widetilde{Z}_m = \widetilde{\phi}_m^2\left(\widetilde{X}_m, \tilde{\nu}_{-m}\left(\widetilde{X}_m\right)\right) := \phi_{-m}^2\left(\widetilde{X}_m, \tilde{\nu}_{-m}\left(\widetilde{X}_m\right)\right), \tag{23}$$

starting from the initial condition $\widetilde{X}_{-k_0}(\omega,\nu) = s_0, \widetilde{Z}_{-k_0}(\omega,\nu) = a_0$. We define the composition function as

$$\widetilde{\Phi}_{-k_0}^0 := \widetilde{\phi}_0 \circ \widetilde{\phi}_{-1} \circ \cdots \circ \widetilde{\phi}_{-k_0+2} \circ \widetilde{\phi}_{-k_0+1}, \tag{24}$$

where $\widetilde{\phi}_m = (\widetilde{\phi}_m^1, \widetilde{\phi}_m^2)$.

Recall that (20) in the forward simulation starting from $k = 0$, we go from a path of length k_0 to a path of length k_0+1 by taking the composition $\phi_{k_0} \circ \Phi_0^{k_0}(s_0,a_0)$. In backward simulation, we do this by taking the composition $\widetilde{\Phi}_{-k_0+1}^0 \circ \widetilde{\phi}_{-k_0}(s_0,a_0)$. Therefore, forward simulation is done by forward composition of simulation functions, whereas the backward simulation is done by backward composition of the simulation functions. Furthermore, in forward simulation, we can successively generate consecutive states of a single controlled Markov chain trajectory one transition at a time, whereas in backward simulation, one is obliged to generate one transition per state and any trajectory from $-k_0$ to 0 has to be traced out of this collection by choosing contiguous state transitions at each successive time. This feature is familiar from the Propp-Wilson backward simulation algorithm mentioned previously.

In the following, we fix the control strategy ϕ_k^2 as

$$\phi_k^2(s) = \arg\min \widetilde{Q}_k(s,\cdot), \forall s, \tag{25}$$

where $\widetilde{Q}_k$ is defined as

$$\widetilde{Q}_k(s,a) := \mathbb{E}_2\left[\sum_{l=-k}^{-1} \gamma^{l+k} c\left(\widetilde{X}_l, \widetilde{Z}_l\right) + \gamma^k \widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) | \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right] \tag{26}$$

and $\widetilde{Q}_0(\cdot,\cdot) = h(\cdot,\cdot)$ for any bounded function $h : \mathbb{S}\times\mathbb{A} \to \mathbb{R}_+$. The expectation in this equation is with respect to the measure $\mathbb{P}_2$ for a given ω (i.e., for a given sequence of transition kernels $(\widehat{p}_k(\omega_k))_{k\geq 0}$).

We now show an important connection between the $\widetilde{Q}_k$ iterate defined previously and the empirical Q-value iterate $\widehat{Q}_k$.

Proposition 2. *Let $\widehat{Q}_0(\cdot,\cdot) = \widetilde{Q}_0(\cdot,\cdot) = h(\cdot,\cdot)$ for any bounded function $h : \mathbb{S}\times\mathbb{A} \to \mathbb{R}_+$. Then, $\widehat{Q}_k = \widetilde{Q}_k$ for all $k \geq 0$.*

We prove this by induction. First, by the definition of $\widehat{Q}_k$ given in Equation (14), for all $(s,a) \in \mathbb{S}\times\mathbb{A}$, we get

$$\widehat{Q}_0(s,a) = h(s,a),$$
$$\widehat{Q}_1(s,a) = c(s,a) + \gamma \sum_{s'} \widehat{p}_0(s'|s,a) \min_b h(s',b).$$

Now, by the definition in Equation (26),

$$\widetilde{Q}_0(s,a) = \mathbb{E}_2\left[h\left(\widetilde{X}_0, \widetilde{Z}_0\right) | \widetilde{X}_0 = s, \widetilde{Z}_0 = a\right] = h(s,a),$$
$$\widetilde{Q}_1(s,a) = \mathbb{E}_2\left[c\left(\widetilde{X}_{-1}, \widetilde{Z}_{-1}\right) + \gamma\ \widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) | \widetilde{X}_{-1} = s, \widetilde{Z}_{-1} = a\right]$$
$$= c(s,a) + \gamma \mathbb{E}_2\left[\widetilde{Q}_0\left(\phi_0\left(\widetilde{X}_{-1}, \widetilde{Z}_{-1}, \nu_1\right), \widetilde{Z}_0\right) | \widetilde{X}_{-1} = s, \widetilde{Z}_{-1} = a\right],$$

where $\widetilde{Z}_0 = \arg\min \widetilde{Q}_0(\phi_0(\widetilde{X}_{-1}, \widetilde{Z}_{-1}, \nu_1), \cdot)$. Then,

$$\widetilde{Q}_1(s,a) = c(s,a) + \gamma \sum_{s'} \widehat{p}_0(s'|s,a) \min_b h(s',b),$$

where we used the fact that $\widetilde{Q}_0 = h$.

Now, assume that $\widetilde{Q}_m = \widehat{Q}_m$ for all $m \le k-1$. Then,

$$\begin{aligned}
\widetilde{Q}_k(s,a) &= \mathbb{E}_2\left[\sum_{l=-k}^{-1} \gamma^{l+k} c\left(\widetilde{X}_l, \widetilde{Z}_l\right) + \gamma^k \widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) \middle| \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right] \\
&= c(s,a) + \mathbb{E}_2\left[\sum_{l=-k+1}^{-1} \gamma^{l+k} c\left(\widetilde{X}_l, \widetilde{Z}_l\right) + \gamma^k \widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) \middle| \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right] \\
&= c(s,a) \\
&\quad + \gamma\, \mathbb{E}_2\left[\sum_{l=-k+1}^{-1} \gamma^{l+k-1} c\left(\widetilde{X}_l, \widetilde{Z}_l\right) + \gamma^{k-1} \widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) \middle| \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right] \\
&= c(s,a) + \gamma\, \mathbb{E}_2\left[\widetilde{Q}_{k-1}\left(\widetilde{X}_{-k+1}, \widetilde{Z}_{-k+1}\right) \middle| \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right] \\
&= c(s,a) + \gamma\, \mathbb{E}_2\left[\widetilde{Q}_{k-1}\left(\phi_{-k+1}\left(\widetilde{X}_{-k}, \widetilde{Z}_{-k}, \nu_k\right), \widetilde{Z}_{-k+1}\right) \middle| \widetilde{X}_{-k} = s, \widetilde{Z}_{-k} = a\right],
\end{aligned}$$

where $\widetilde{Z}_{-k+1} = \arg\min \widetilde{Q}_{k-1}(\phi_{-k+1}(\widetilde{X}_{-k}, \widetilde{Z}_{-k}, \nu_k), \cdot)$.

Then,

$$\begin{aligned}
\widetilde{Q}_k(s,a) &= c(s,a) + \gamma \sum_{s'} \widehat{p}_{k-1}(s'|s,a) \min_b \widetilde{Q}_{k-1}(s',b) \\
&= c(s,a) + \gamma \sum_{s'} \widehat{p}_{k-1}(s'|s,a) \min_b \widehat{Q}_{k-1}(s',b) \\
&= \widehat{Q}_k(s,a).
\end{aligned}$$

Now we show the following results.

Proposition 3. *For $\omega \in \Omega_1, \nu \in \Omega_2$, we trace out two MDPs with state-action sequences*

$$\left(\widetilde{X}_m(\omega,\nu), \widetilde{Z}_m(\omega,\nu)\right)_{m=-k}^{0}, \left(\widetilde{X}'_m(\omega,\nu), \widetilde{Z}'_m(\omega,\nu)\right)_{m=-k}^{0},$$

with initial conditions

$$\left(\widetilde{X}_{-k}(\omega,\nu), \widetilde{Z}_{-k}(\omega,\nu)\right) = (s,a), \left(\widetilde{X}'_{-k}(\omega,\nu), \widetilde{Z}'_{-k}(\omega,\nu)\right) = (s',a').$$

These chains couple with probability 1 as $k \to \infty$.

By construction, two Markov chain paths initiated at time $-k$ traced from the backward simulation in forward time beginning at $-k$ will merge once they hit a common state, that is, get coupled (Propp and Wilson 1996). Let $\tilde{\tau}^k_{\omega,\nu}$ be the time after which these chains couple, that is, $\widetilde{X}_{-k+\tilde{\tau}^k_{\omega,\nu}} = \widetilde{X}'_{-k+\tilde{\tau}^k_{\omega,\nu}}$ and $\widetilde{X}_{-k+l} \neq \widetilde{X}'_{-k+l}$ for all $0 \le l < \tilde{\tau}^k_{\omega,\nu}$. Because these chains are of finite length (from $-k$ to 0), we may need to define the value of $\tilde{\tau}^k_{\omega,\nu}$ arbitrarily if they don't couple during this time.

To overcome this, we let these chains run to infinity. This can be done without loss of generality as follows. For $-k \le m < 0$, simulate the chains according to the backward simulation method specified by (22) ad (23). Suppose the i.i.d. random vectors ν_m are generated for all $-\infty < m < \infty$. For $m \ge 0$, continue the simulation to generate chains $(\widetilde{X}_m(\omega,\mu), \widetilde{Z}_m(\omega,\mu))_{m=1}^{\infty}$, $(\widetilde{X}'_m(\omega,\mu), \widetilde{Z}'_m(\omega,\mu))_{m=1}^{\infty}$ as

$$\widetilde{X}_m = \phi^1_{m+k}\left(\widetilde{X}_{m-1}, \widetilde{Z}_{m-1}, \nu_{-(m-1)}\right), \tag{27}$$

$$\widetilde{Z}_m = \phi^2_{m+k}\left(\widetilde{X}_m, \tilde{\nu}_{-m}\right). \tag{28}$$

It is easy to see that the $\widetilde{\tau}^k_{\omega,\nu}$ has the same statistical properties as the coupling time defined in Equation (21). Therefore, by Proposition 1, $\mathbb{E}[\widetilde{\tau}^k_{\omega,\nu}] < \infty$. Now,

$$\sum_{n\geq 1} \mathbb{P}\left(2\widetilde{\tau}^k_{\omega,\nu} \geq n\right) = \mathbb{E}\left[2\widetilde{\tau}^k_{\omega,\nu}\right] < \infty.$$

Also, it is easy to see that $\widetilde{\tau}^k_{\omega,\nu}$s are identically distributed $\forall k$. Therefore,

$$\sum_{n\geq 1} \mathbb{P}\left(2\widetilde{\tau}^k_{\omega,\nu} \geq n\right) = \sum_{n\geq 1} \mathbb{P}\left(2\widetilde{\tau}^n_{\omega,\nu} \geq n\right) < \infty,$$

which implies

$$\sum_{n\geq 1} \mathbb{P}\left(\widetilde{\tau}^n_{\omega,\nu} - n > -\frac{n}{2}\right) < \infty.$$

Then, by the Borel-Cantelli lemma, $\widetilde{\tau}^n_{\omega,\nu} - n \to -\infty$, (ω,ν)-a.s. Thus, the chains will couple with probability 1.

We shall need the following lemma of Blackwell and Dubins (Blackwell and Dubins 1962; Borkar 1995, chapter 3, theorem 3.3.8).

Lemma 1 (Blackwell and Dubins (1962)). *Let $Y_k, k = 1, 2, \ldots, \infty$ be real random variables on a probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ such that $Y_k \to Y_\infty$ a.s. and $\mathbb{E}[\sup_k |Y_k|] < \infty$. Let $\{\mathcal{F}_k\}$ be a family of sub-σ-fields of $\mathcal{F}$ that is either increasing or decreasing, with $\mathcal{F}_\infty = \vee_k \mathcal{F}_k$ or $\cap_k \mathcal{F}_k$ accordingly. Then, $\lim_{k,j\to\infty} \mathbb{E}[Y_k|\mathcal{F}_j] = \mathbb{E}[Y_\infty|\mathcal{F}_\infty]$ a.s. and in L_1.*

We now show that $\widetilde{Q}_k(\omega)$ converges to a random variable $Q^*(\omega)$ almost surely. By the previous proposition, this will imply the almost sure convergence of $\widehat{Q}_k$ to $Q^*(\omega)$.

We now give the proof of Theorem 1.

Consider the backward simulation described previously. For $\omega \in \Omega_1, \nu \in \Omega_2$, we trace out two MDPs with state-action sequences:

$$\left(\widetilde{X}_m(\omega,\nu), \widetilde{Z}_m(\omega,\nu)\right)_{m=-k}^{0}, \left(\widetilde{X}'_m(\omega,\nu), \widetilde{Z}'_m(\omega,\nu)\right)_{m=-k}^{0},$$

with initial conditions

$$\left(\widetilde{X}_{-k}(\omega,\nu), \widetilde{Z}_{-k}(\omega,\nu)\right) = (s,a), \left(\widetilde{X}'_{-k}(\omega,\nu), \widetilde{Z}'_{-k}(\omega,\nu)\right) = (s',a').$$

By construction, two Markov chain paths initiated at time $-k$ traced from the backward simulation but in forward time beginning at $-k$ will merge once they hit a common state, that is, get coupled (Propp and Wilson 1996). Decrease $-k$ until all paths initiated at $-k$ couple. Once they couple, they follow the same sample path.

Now, by construction,

$$\begin{aligned}\widetilde{Q}_k(s,a) - \widetilde{Q}_k(s',a') = \mathbb{E}_2\Bigg[&\sum_{l=-k}^{(-k+\tau^k_{\omega,\nu}-1)\wedge(-1)} \gamma^{l+k}\left(c\left(\widetilde{X}_l, \widetilde{Z}_l\right) - c\left(\widetilde{X}'_l, \widetilde{Z}'_l\right)\right)\\ &+ \gamma^{k\wedge\tau^k_{\omega,\nu}}\left(\widetilde{Q}_0\left(\widetilde{X}_0, \widetilde{Z}_0\right) - \widetilde{Q}_0\left(\widetilde{X}'_0, \widetilde{Z}'_0\right)\right)\\ &\Bigg|\left(\widetilde{X}_{-k}, \widetilde{Z}_{-k}\right) = (s,a), \left(\widetilde{X}'_{-k}, \widetilde{Z}'_{-k}\right) = (s',a')\Bigg].\end{aligned}$$

Because the chains will couple with probability 1 (according to Proposition 3), the right-hand side (RHS) of this equation will converge to a random variable $R(\omega)(s,a,s',a')$, ω-a.s. as $k \to \infty$, that is,

$$R_k(\omega)(s,a,s',a') := \widetilde{Q}_k(s,a) - \widetilde{Q}_k(s',a') \to R(\omega)(s,a,s',a'),\ \omega - \text{a.s.} \tag{29}$$

We revert to the *forward time* picture henceforth. Now,

$$\begin{aligned}\widehat{Q}_{k+1}(s,a) &= c(s,a) + \gamma \sum_{s'} \widehat{p}_k(s'|s,a) \min_b \widehat{Q}_k(s',b)\\ &= c(s,a) + \gamma \sum_{s'} \widehat{p}_k(s'|s,a) \min_b \left(\widehat{Q}_k(s',b) - \widehat{Q}_k(s,a)\right) + \gamma\ \widehat{Q}_k(s,a)\\ &= c(s,a) + \gamma \sum_{s'} \widehat{p}_k(s'|s,a) \min_b R_k(\omega)(s',b,s,a) + \gamma\ \widehat{Q}_k(s,a).\end{aligned}$$

Because $\widehat{p}_k$ depends only on ω, we can define another random variable $R'_k(\omega)(s,a)$ such that

$$R'_k(\omega)(s,a) := \sum_{s'} \widehat{p}_k(s'|s,a) \min_b R_k(\omega)(s',b,s,a),$$
$$= E\left[\min_b R_k(\omega)(s',b,s,a)|\tilde{\mathcal{F}}_{k-1}\right],$$

where $\tilde{\mathcal{F}}_{k-1} := \sigma(\xi_i^{k'}(s,a), s \in \mathbb{S}, a \in \mathbb{A}, 1 \le i \le n, k' < k)$. Because $R_k(\omega) \to R(\omega),\ \omega -$ a.s., it follows from the preceding lemma that there exists another random variable $R^*(\omega)$ such that

$$R'_k(\omega) \to R^*(\omega),\ \omega - \text{a.s.}$$

Then,

$$\begin{aligned}\widehat{Q}_{k+1}(s,a) &= c(s,a) + \gamma\ R'_k(\omega)(s,a) + \gamma\ \widehat{Q}_k(s,a)\\ &= c(s,a) + \gamma\ R'_k(\omega)(s,a) + \gamma\ c(s,a) + \gamma^2\ R'_{k-1}(\omega)(s,a) + \gamma^2\ \widehat{Q}_{k-1}(s,a)\\ &\ \ \vdots \qquad\qquad \vdots\\ &= c(s,a)\sum_{l=0}^{k}\gamma^l + \gamma\ \sum_{l=0}^{k}\gamma^l R'_{k-l}(\omega)(s,a) + \gamma^{k+1}\widehat{Q}_0(s,a).\end{aligned}$$

Clearly,

$$\widehat{Q}_k(s,a) \to Q^*(\omega) := \frac{c(s,a)}{(1-\gamma)} + \frac{\gamma R^*(\omega)(s,a)}{(1-\gamma)},\ \omega - \text{a.s.}$$

Next, we provide a proof of Corollary 1.

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be the probability space as defined before. By $\mathcal{F}_k$ denote $\sigma(\widehat{Q}_m, m \le k)$. From Proposition 1, $\widehat{Q}_k(\omega) \to \widehat{Q}^*(\omega),\ \omega -$ a.s., and hence, $\widehat{Q}_k(\omega) - \widehat{Q}_{k-1}(\omega) \to 0$. Taking conditional expectation and using Lemma 1 we get

$$\mathbb{E}\left[\widehat{Q}_k(\omega)|\mathcal{F}_{k-1}\right] - \widehat{Q}_{k-1}(\omega) \to 0.$$

Because $\widehat{Q}_k(\omega) = \widehat{G}(\theta_{k-1}\omega, \widehat{Q}_{k-1}(\omega))$, from Equation (15),

$$\mathbb{E}\left[\widehat{Q}_k(\omega)|\mathcal{F}_{k-1}\right] = G\left(\widehat{Q}_{k-1}(\omega)\right),$$

where G is the Q-value operator defined in Equation (5). This gives $G(\widehat{Q}_{k-1}(\omega)) - \widehat{Q}_{k-1}(\omega) \to 0$. Then, by the continuity of G, $G(\widehat{Q}^*(\omega)) = \widehat{Q}^*(\omega)$, which implies that $\widehat{Q}^*(\omega)$ is indeed equal to the optimal Q-function Q^*, by the uniqueness of the fixed point of G.

4. Rate of Convergence and Asynchronous EQVI

In this section, we now provide a rate of convergence or a nonasymptotic sample complexity bound. This follows from methods that had been developed in Haskell et al. (2013) for empirical value and policy iteration, which, however, only provide a convergence in probability guarantee. In the second section, we provide an argument of why asynchronous EQVI will also work. This also uses methods developed earlier in Haskell et al. (2013).

4.1. Rate of Convergence

One notable observation about Theorem 1 is that the almost sure convergence of the EQVI iterate holds for any n. However, the rate of convergence will and does depend on n, and this is confirmed by the simulation results (Section 5). Although the convergence guarantee, $\widehat{Q}_k \to Q^*$ is a strong result, rate of convergence is an important consideration in practical applications. Unfortunately, the coupling argument used in the proof of Theorem 1 does not yield a rate of convergence. However, we note that the exact Q-value operator $G(\cdot)$ is a contraction, and its empirical variant $\widehat{G}(\cdot)$ is a random contraction operator.

In Haskell et al. (2013), a technique for analyzing the rate of convergence of a random sequence resulting from iteration of a random contraction operator was developed. This was used to show the probabilistic

convergence of empirical value iteration, and explicit bounds were given on the number of simulations samples n and the number of iterations k that are needed to get an ϵ-optimal value function with a probability greater than $(1-\delta)$. We now argue that the exact Q-value operator $G(\cdot)$ is a contraction, its empirical variants $\widehat{G}(\cdot)$ satisfy assumptions 4.1–4.4 in Haskell et al. (2013), and thus a very similar methodology can be used in establishing convergence in probability of the iterates of EQVI (weaker than Theorem 1 in this paper). However, more importantly, it yields a rate of convergence and a nonasymptotic sample complexity result; that is, for any given $\epsilon > 0, \delta > 0$, we give an explicit bound on the number of simulation samples n and the number of iterations k that are needed to get an ϵ optimal Q-value with probability greater that $(1-\delta)$.

Assumptions. The classical operator G and a sequence of random operators $\{\widehat{G}_n\}$ satisfy the following:

4.1. $\mathcal{P}(\lim_{n\to\infty} \|\widehat{G}_n q - G q\| \geq \epsilon) = 0$ $\forall \epsilon > 0$ and $\forall q \in \mathbb{R}^{|\mathbb{S}|}$. Also G has a (possibly nonunique) fixed point q^* such that $Gq^* = q^*$.

4.2. There exists a $\kappa^* < \infty$ such that $\|\hat{q}_n^k\| \leq \kappa^*$ almost surely for all $k \geq 0$, $n \geq 1$. Also, $\|q^*\| \leq \kappa^*$.

4.3. $\|G q - q^*\| \leq \gamma \|q - q^*\|$ for all $q \in \mathbb{R}^{|\mathbb{S}|}$.

4.4. There is a sequence $\{p_n\}_{n\geq 1}$ such that

$$P\left(\|Gq - \widehat{G}_n q\| < \epsilon\right) > p_n(\epsilon)$$

and $p_n(\epsilon) \uparrow 1$ as $n \to \infty$ for all $v \in \overline{B_{\kappa^*}(0)}$, $\forall \epsilon > 0$.

It can be shown that the exact Q-value operator G and its empirical variants $\widehat{G}_n$ (where the index n is for number of samples) satisfy the previous assumptions. It can be argued easily by using strong law of large numbers that Assumption 4.1 is satisfied. Assumption 4.2 is satisfied easily when rewards are bounded. Assumption 4.3 is satisfied because G is a contraction operator. It can easily be checked that Assumption 4.4 is satisfied with

$$p_n = 1 - 2|\mathbb{S}||\mathbb{A}|e^{-2(\epsilon/\gamma)^2 n/\left((\kappa^*)^2\right)}. \tag{30}$$

This implies convergence in probability of the Q-value iterates (weaker than in the previous section) to the optimal Q-value. Now, following arguments and construction similar to section 5.1 in Haskell et al. (2013), we can derive a nonasymptotic sample complexity bound given in Theorem 2.

Because the details of the proofs are the same as in Haskell et al. (2013), we only give a short outline here. Readers are referred to Haskell et al. (2013) for details. The proof is based on the idea of constructing a sequence of Markov chains that stochastically dominate a discrete error process. More precisely, we are interested in the rate of convergence of the sequence $\{\|\widehat{Q}_k - Q^*\|, k \geq 0\}$ to 0. However, because the error process $\{\|\widehat{Q}_k - Q^*\|, k \geq 0\}$ is continuous valued, we first discretize it and get a discrete error process now defined on nonnegative integers. Unfortunately, this process is not Markovian. Hence, we construct a Markov chain $\{Y_k^n, k \geq 0\}$ that has the following structure:

$$Y_k^n = \begin{cases} \max\{Y_{k-1}^n, \eta^*\}, & \text{with probability } p_n, \\ N^*, & \text{with probability} 1 - p_n. \end{cases} \tag{31}$$

Note that p_n is close to 1 for sufficiently large n. The Markov chain $\{Y_k^n, k \geq 0\}$ will either move one unit closer to zero until it reaches η^*, or it will move (as far away from zero as possible) to N^* (and hence bounds are very conservative). We can show that this Markov chain stochastically dominates the discrete error process. Furthermore, as n goes to infinity, the invariant distribution of the Markov chain will concentrate at zero, which establishes convergence of the error process $\{\|\widehat{Q}_k - Q^*\|, k \geq 0\}$ to zero in probability. Now the mixing rate of the Markov chain gives the rate of convergence and the sample complexity bound for EQVI in the previous theorem.

4.2. Asynchronous EQVI

We now show that just as for empirical value iteration, the asynchronous version of EQVI works as well. That is, the Q-value function estimates converge in probability even when the updates are asynchronous, including in the *online* case when updates are done for one state at a time. We consider each state to be visited at least once to complete a full cycle, and the time for a full cycle could be random.

Let $(\sigma_k, \alpha_k)_{k\geq 0}$ be any infinite sequence of states and actions. This sequence $(\sigma_k, \alpha_k)_{k\geq 0}$ may be deterministic or stochastic, and it may even depend online on the Q-value function updates. For shorthand, denote $z = (\sigma, \alpha)$. For any $z \in \mathbb{S} \times \mathbb{A}$, we define the asynchronous Q-value operator G_z as

$$[G_z Q](s,a) = \begin{cases} c(\sigma,\alpha) + \gamma \mathbb{E}\left[\min_{b\in\mathbb{A}} Q\left(\psi(\sigma,\alpha,\xi), b\right)\right], & (s,a) = z \\ Q(s,a), & \text{otherwise.} \end{cases}$$

Also define its empirical variant with n samples as

$$\left[\widehat{G}_{z,n}(\omega)\widehat{Q}\right](s,a) = \begin{cases} c(\sigma,\alpha) + \frac{\gamma}{n}\sum_{i=1}^{n} \min_{b\in\mathbb{A}} \widehat{Q}\left(\psi(\sigma,\alpha,\xi_i), b\right), & (s,a) = z, \\ \widehat{Q}(s,a), & \text{otherwise.} \end{cases}$$

The operators G_z and $\widehat{G}_{z,n}$ only update the Q-value function for state s and action a, and leaves the other estimates unchanged. This will then produce a sequence of updates $\{Q_k\}$ and $\{\widehat{Q}_k^n\}$, respectively, starting from some initial seed Q_0.

Suppose that in some finite number of steps K_1, each state-action pair is visited at least once. Define

$$\widetilde{G} := G_{z_{K_1}} \cdots G_{z_1} G_{z_0},$$

which is a contraction with constant γ. It is well known (Borkar 2008) that if each state-action pair is visited infinitely often, the sequence produced by asynchrononus Q-value iteration, $\{Q_k\}$ will converge to Q^*, the optimal Q-value.

Now define the time of $(m+1)th$ full update

$$K_{m+1} := \inf\left\{k : k \geq K_m,\ (z_i)_{i=K_m+1}^{k} \text{ includes every state-action pair in } \mathbb{S}\times\mathbb{A}\right\},$$

with $K_0 = 0$. We can now give a slightly modified stochastic dominance argument to show that asynchronous EVI will converge in a probabilistic sense by checking the progress of the algorithm at these hitting times, that is, we look at the sequence $\{\widehat{Q}_{K_m}^n\}_{m\geq 0}$. In the simplest update scheme, each state-action pair is updated in turn and the length of a full update cycle is $|\mathbb{S}||\mathbb{A}|$.

Now, analogous to $\widetilde{G}$, we can define an operator $\widehat{G}_n$,

$$\widehat{G}_n := \widehat{G}_{z_{K_1},n} \cdots \widehat{G}_{z_1,n} \widehat{G}_{z_0,n}.$$

Each random operator in this iteration introduces an error $\epsilon/|\mathbb{S}||\mathbb{A}|$ compared with the corresponding non-random operator. This can be ensured by picking n large enough such that

$$P\left\{\|\widehat{G}_{z,n}Q - G_z Q\| \geq \epsilon/|\mathbb{S}||\mathbb{A}|\right\} \leq 2\, e^{-2\left(\epsilon/\left(\gamma|\mathbb{S}||\mathbb{A}|\right)\right)^2 n/(2\kappa^*)^2},$$

where κ^* is a constant that can be computed. This can be used now to guarantee that $P\{\|\widehat{G}_n Q - \widetilde{G}_z Q\| \geq \epsilon\}$ is upper bounded by

$$p_n = 2\,|\mathbb{S}|\,|\mathbb{A}|\, e^{-2\left(\epsilon/\left(\gamma|\mathbb{S}||\mathbb{A}|\right)\right)^2 n/(2\kappa^*)^2}.$$

Now, the stochastic dominance argument developed in Haskell et al. (2013) can be applied to obtain the following result.

Theorem 3. *If each state-action pair is visited in turn infinitely often, the iterates of asynchronous EQVI,*

$$\widehat{Q}_k^n \to Q^* \quad \text{in probability}$$

as $n, k \to \infty$.

Remark 2. We note that the online version of asynchronous EQVI is like the popular Q-learning algorithm used for reinforcement learning. As we see in the numerical results in the next section, online EQVI has a much faster convergence than Q-learning, although the theoretical guarantees are weaker, that is, convergence in probability for EQVI and almost sure for Q-learning.

5. Numerical Results

In this section, we show some numerical results comparing the classical Q-Learning (QL) algorithm (given in Equation (8)) with our EQVI. We generate a random MDP, with $|\mathbb{S}| = 500$ and $|\mathbb{A}| = 10$, where the transition matrix P and the cost $c(s,a)$ are generated randomly. We plot relative error $e_k := \|\widehat{Q}_k - Q^*\|/\|Q^*$ versus the number of iterations. The synchronous version of QL was used in which we used more than one simulation samples and updated all state-action pairs at the same time.

We can represent the update equations of both QL and EQVI using the operator $\widehat{G}$ ((13) and (8))

$$\text{QL} : Q_{k+1} = (1-\alpha_k)Q_k + \alpha_k\left(\widehat{G}(\theta_k\omega, Q_k)\right), \tag{32}$$

$$\text{EQVI} : \widehat{Q}_{k+1} = \widehat{G}\left(\theta_k\omega, \widehat{Q}_k\right). \tag{33}$$

Therefore, both EQVI and QL can be run using the same MatLab code. For EQVI, set $\alpha_k = 1, \forall k$. This does not make EQVI a stochastic approximation scheme because it does not satisfy the step-size requirement.

As you can see from Figure 1, the rate of decay of relative error is way faster in EQVI (and close to exact QVI) compared with synchronous QL. In fact, to reach 5% relative error, QVI takes about 30 iterations, and EQVI takes just a bit more (about 35), whereas synchronous QL takes more than 300. Thus, EQVI promises at least a 10 times speedup over synchronous QL. In fact, in about 35 iterations, synchronous QL has a 50% relative error. The relative error has been estimated from 50 simulation runs, and the confidence intervals are very tight. As we take more samples per iteration, we start to approach performance of exact QVI.

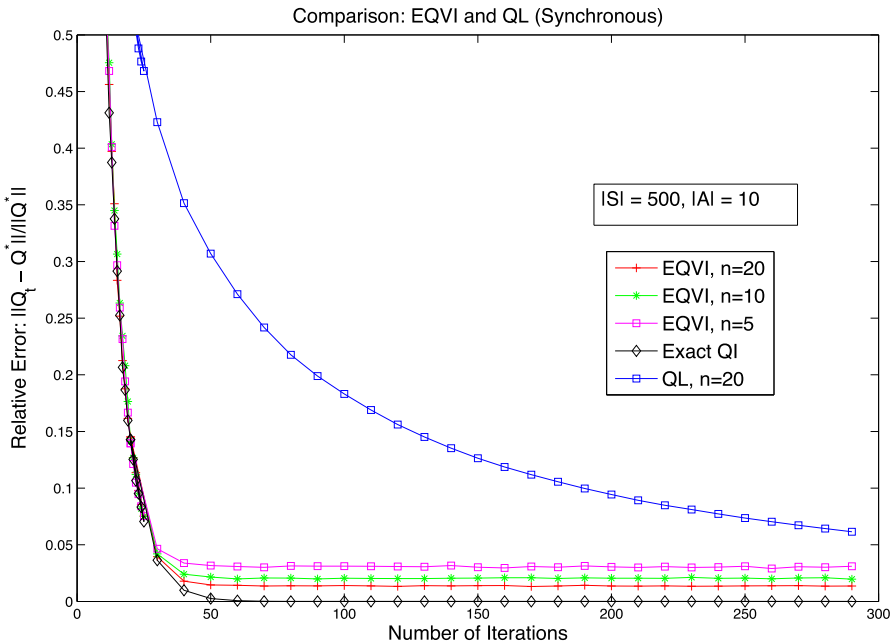
Figure 2 shows asynchronous EQVI and QL for a random MDP with 500 states and 10 actions wherein state-action pairs were chosen randomly in each iteration. As can be seen, exact QVI and EQVI get to within 5% relative error in about 500 iterations (quite remarkable because there are 5,000 state-action pairs), whereas QL in 500 iterations has a 90% relative error. In fact, (asynchronous) QL is so slow that even after 10,000 iterations, the relative error is still about 50%. As before, the relative error has been estimated from 50 simulation runs, and the confidence intervals are very tight.

From these simulations. it is clear that EQVI promises significantly faster performance than Q-learning in both synchronous and asynchronous settings.

Remark 3. As mentioned previously, we get a very fast convergence with EQVI to a ballpark estimate but then an extremely slow (in fact, imperceptible in the given time frame) movement to the exact value as guaranteed by theory. To get some intuition about why, consider the uncontrolled case. Then, the iterations are of the form

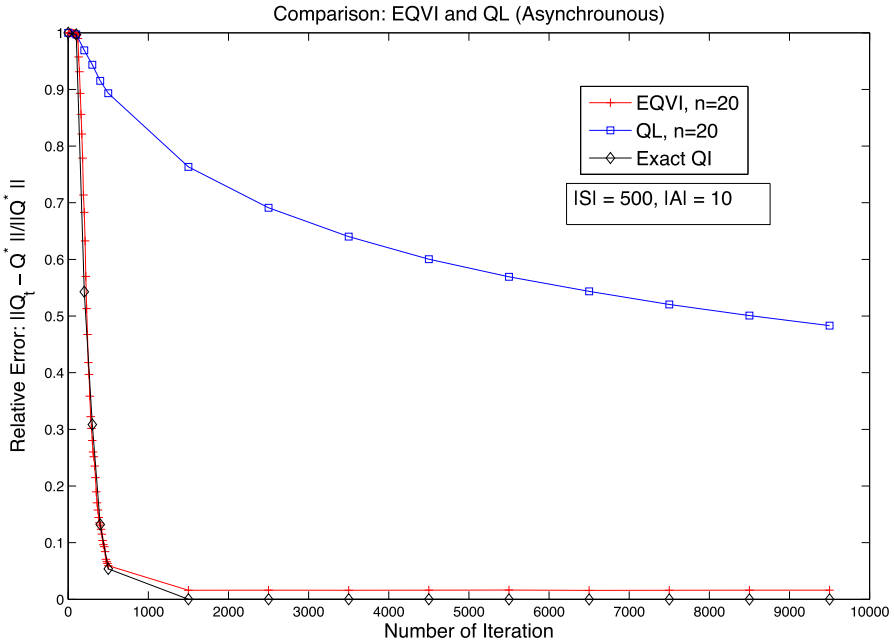
$$Q_{k+1} = \check{G}Q_k,$$

Figure 1. Comparison of Synchronous Exact QVI, EQVI, and QL for a 500 State and 10 Action Random MDP



Notes. For QL, the step size $\alpha_k = 1/k^\theta$, $\theta = 0.6$. Average is over 50 runs.

Figure 2. Comparison of Asynchronous Exact QVI, EQVI, and QL for a 500 State and 10 Action Random MDP with Multiple Samples in Each Step



Notes. For QL, the step size $\alpha_k = 1/k^{\theta}$, $\theta = 0.6$. Average is over 50 runs.

where $\check{G}$ is a random *affine* contraction. This may further be written as

$$Q_{k+1} = \check{G}Q_k + M_k = AQ_k + b + M_k,$$

where $\check{G}(x) = Ax + b$ for suitably defined A, b is a deterministic affine contraction and $\{M_k\}, M_k := \check{G}Q_k - \bar{G}Q_k$, a martingale difference sequence. Note that A in our case is γ times a stochastic matrix, hence a stable matrix. Then

$$Q_k = A^k Q_0 + \sum_{m=0}^{k-1} A^{k-m} b + \sum_{j=0}^{k-1} A^{k-j} M_j.$$

The first term on the right decays to zero, the second converges to the desired limit, and the third represents noise. If $\{M_k\}$ were i.i.d., this would converge to a stationary process and not to zero. In our case, it does converge to zero as implicit in the proof of Theorem 1. In case of stochastic approximation, M_k would be weighted by a square-summable step size that accelerates this convergence to zero. However, in our case, in the absence of such additional damping, the fluctuations can be expected to diminish only very slowly. On the other hand, the decay of dependence on initial condition and convergence of the middle term to the desired limit are no longer incremental as in the stochastic approximation counterpart and therefore very rapid. This is in tune with the well-known bias-variance tradeoff and not surprising. This does, however, suggest that a hybrid scheme that runs empirical Q-value iteration initially and then switches to conventional Q-learning will have the best of both the worlds if a faster almost sure convergence is needed. The performance of our scheme improves rapidly with increasing n. For practical problems, using EQVI until the relative error is below some threshold (e.g., 1%–5%) may be enough.

6. Conclusions

We presented a new (offline and online) Q-value iteration algorithm for discounted-cost MDPs. We have rigorously established the convergence of this algorithm to the desired limit with probability 1. Unlike the classical learning schemes for MDPs such as Q-learning and actor-critic algorithms, our algorithm or analysis does not use a stochastic approximation method and is a nonincremental scheme. Preliminary experimental results suggest a faster rate of convergence for our algorithm than currently popularly used algorithms.

A particularly interesting and useful aspect is whether distributed and asynchronous implementation of EQVI will work. We have been able to show that for the special case where each state-action pair is updated in turn. Moreover, the convergence guarantee is only probabilistic. It would be useful to show that even with randomly picked state-action pairs, as long as each one of them is picked infinitely often, we will get convergence and in the stronger almost sure sense (as for our main result for the synchronous case.)

Another useful direction will be to show that this would work with infinite (even continuous) state and action spaces. This would then make such an algorithm useful even for partially observed MDP problems. This will require combining current methods with function approximation in an appropriate space (e.g., a reproducing kernel Hilbert space).

Another useful direction would be the average reward case. Average reward MDPs are typically hard to analyze because the dynamic programming operator for average reward MDP is not a contraction mapping. There are, however, provably convergent Q-learning and actor-critic algorithms for average reward MDPs because of the powerful ODE approach to stochastic approximation (Konda and Borkar 1999, Abounadi et al. 2001). It would be interesting to see if our algorithm works for learning in MDPs with average reward criterion.

These are directions for future research.

Appendix. Proof of Proposition 1

We present this as a series of lemmas.

Given an initial time k_0 and states $s_0, s' \in \mathbb{S}$, we define the *hitting time* $\tau_{\omega,\nu}$ of the controlled Markov chain $(X_k(\omega,\nu))_{k\geq k_0}$ as

$$\tau_{\omega,\nu}(s_0,s') := \min\{m \geq 0 | X_{k_0+m}(\omega,\nu) = s', X_{k_0}(\omega,\nu) = s_0\}. \tag{A.1}$$

We first show that the expected value of the hitting time is finite when the chain is controlled by a stationary strategy, that is, $\phi_k^2 \approx \pi \in \Pi, \forall k$.

Lemma A.1. *Let $(X_k(\omega,\nu), Z_k)_{k\geq k_0}$ be the sequence of state-action pairs for the MDP simulated according to (20) using a stationary control strategy $\phi_k^2 \approx \pi \in \Pi, \forall k$. Let $\tau_{\omega,\nu}$ be the hitting time as defined in Equation (A.1). Then,*

$$\mathbb{E}\left[\tau_{\omega,\nu}(s_0,s')\right] < \infty, \ \forall s_0, s' \in \mathbb{S}.$$

Consider a sequence of states, $(s_{k_0+j})_{j=0}^r$, with $s_{k_0} = s_0$ and $s_{k_0+r} = s'$ such that $P^\pi(s_{k_0}, s_{k_0+1}) \cdots P^\pi(s_{k_0+r-1}, s_{k_0+r}) > 0$. By Remark 1, such a sequence of states exists. Furthermore, r can be picked independent of the choice of s_0, s' and we assume that it is so. Let

$$W^\pi = W^\pi\left((s_{k_0+j})_{j=0}^r\right) := \widehat{P}^\pi_{k_0}(s_{k_0}, s_{k_0+1}) \cdots \widehat{P}^\pi_{k_0+r-1}(s_{k_0+r-1}, s_{k_0+r}).$$

Using (10)–(12), $\mathbb{E}_1[\widehat{P}^\pi_k] = P^\pi \ \forall k$. Because $\widehat{P}^\pi_k$ are i.i.d.,

$$\mathbb{E}_1[W^\pi] = P^\pi(s_{k_0}, s_{k_0+1}) \cdots P^\pi(s_{k_0+r-1}, s_{k_0+r}) > 0.$$

Therefore, there exist $\epsilon > 0, \delta > 0$ such that $\mathbb{P}_1(W^\pi > \epsilon) > \delta$. Then,

$$\mathbb{P}(\tau_{\omega,\nu}(s_0,s') \leq r) \ \geq \ \mathbb{P}_2(\tau_{\omega,\nu}(s_0,s') \leq r | W^\pi > \epsilon)\, \mathbb{P}_1(W^\pi > \epsilon) \ > \ \epsilon\delta,$$

because $\mathbb{P}_2(\tau_{\omega,\nu}(s_0,s') \leq r | W^\pi) \geq \mathbb{P}_2(X_{k_0+r} = s', X_{k_0} = s_0 | W^\pi) \geq W^\pi$. Therefore,

$$\mathbb{P}(\tau_{\omega,\nu}(s_0,s') > r) \ \leq \ (1-\epsilon\delta).$$

Because of the i.i.d. nature of ω and the Markov property of $X_k(\omega,\nu)$, it is clear that the previous probability does not depend on k_0, and hence, for any $k > 0$,

$$\mathbb{P}(\tau_{\omega,\nu}(s,s') > kr) \ \leq \ (1-\epsilon\delta)^k.$$

$$\text{Then,} \quad \mathbb{E}\left[\tau_{\omega,\nu}(s,s')\right] = \sum_{t\geq 0} \mathbb{P}(\tau_{\omega,\nu}(s,s') > t) \leq \sum_{k\geq 0} r\mathbb{P}\left(\tau_{\omega,\nu}(s,s') > kr\right)$$
$$\leq r \sum_{k\geq 0} (1-\epsilon\delta)^k < \infty.$$

We next show that the expected value of the coupling time is finite when the chain is controlled by a stationary strategy, that is, $\phi_k^2 \approx \pi \in \Pi, \forall k$.

Lemma A.2. *Let $(X_k^1(\omega,\nu), Z_k^1)_{k\geq k_0}, (X_k^2(\omega',\nu'), Z_k^2)_{k\geq k_0}$ be two sequences of state-action pairs for an MDP simulated according to (20) using a stationary control strategy $\phi_k^2 \approx \pi \in \Pi, \forall k$. Let $\widetilde{\tau}_{\omega^*,\nu^*}$ be the coupling time as defined in Equation (21). Then,*

$$\mathbb{E}\left[\widetilde{\tau}_{\omega^*,\nu^*}\left(s_0^1, s_0^2\right)\right] < \infty, \forall s_0^1, s_0^2 \in \mathbb{S}.$$

Consider two sequences of states, $(s_{k_0+j}^1)_{j=0}^r$ and $(s_{k_0+j}^2)_{j=0}^r$ with $s_{k_0}^1 = s_0^1$, $s_{k_0}^2 = s_0^2, s_{k_0+r}^1 = s_{k_0+r}^2 = s$, for some $s \in \mathbb{S}$ such that

$$P^\pi\left(s_{k_0}^1, s_{k_0+1}^1\right)\cdots P^\pi\left(s_{k_0+r-1}^1, s_{k_0+r}^1\right) > 0, \quad \text{and}$$
$$P^\pi\left(s_{k_0}^2, s_{k_0+1}^2\right)\cdots P^\pi\left(s_{k_0+r-1}^2, s_{k_0+r}^2\right) > 0.$$

By Remark 1, such $(s_{k_0+j}^1)_{j=0}^r$ and $(s_{k_0+j}^2)_{j=0}^r$ exist. Using, by abuse of notation, some common notation for entities defined on the two copies of $(\Omega, \mathcal{F}, \mathbb{P})$, let

$$W_1^\pi = W_1^\pi\left(\left(s_{k_0+j}^1\right)_{j=0}^r\right) := \widehat{P}_{k_0}^\pi\left(s_{k_0}^1, s_{k_0+1}^1\right)\cdots\widehat{P}_{k_0+r-1}^\pi\left(s_{k_0+r-1}^1, s_{k_0+r}^1\right),$$
$$W_2^\pi = W_2^\pi\left(\left(s_{k_0+j}^2\right)_{j=0}^r\right) := \widehat{P}_{k_0}^\pi\left(s_{k_0}^2, s_{k_0+1}^2\right)\cdots\widehat{P}_{k_0+r-1}^\pi\left(s_{k_0+r-1}^2, s_{k_0+r}^2\right).$$

As in the proof of Lemma A.1,

$$\mathbb{E}_1\left[W_1^\pi\right] = P^\pi\left(s_{k_0}^1, s_{k_0+1}^1\right)\cdots P^\pi\left(s_{k_0+r-1}^1, s_{k_0+r}^1\right) > 0,$$
$$\mathbb{E}_1\left[W_2^\pi\right] = P^\pi\left(s_{k_0}^2, s_{k_0+1}^2\right)\cdots P^\pi\left(s_{k_0+r-1}^2, s_{k_0+r}^2\right) > 0.$$

Therefore, there exist $\epsilon > 0, \delta > 0$ such that $\mathbb{P}_1(W_1^\pi > \epsilon) > \delta$ and $\mathbb{P}_1(W_2^\pi > \epsilon) > \delta$. Moreover, because of the independence of $\widehat{P}_{k_0+j}^\pi(s_{k_0+j}^1, s_{k_0+j+1}^1)$ and $\widehat{P}_{k_0+j}^\pi(s_{k_0+j}^2, s_{k_0+j+1}^2)$,

$$\mathbb{P}_1\left(W_1^\pi > \epsilon, W_2^\pi > \epsilon\right) > \delta^2.$$

Also,

$$\mathbb{P}_2\left(X_{k_0+r}^1 = X_{k_0+r}^2, X_{k_0}^1 = s_0^1, X_{k_0}^2 = s_0^2 | W_1^\pi, W_2^\pi\right) \geq W_1^\pi W_2^\pi.$$

Then, by an argument analogous to that of Lemma A.1, we have

$$\begin{aligned}\mathbb{P}\left(\widetilde{\tau}_{\omega^*,\nu^*}\left(s_0^1, s_0^2\right) \leq r\right) &\geq \mathbb{P}_2\left(\widetilde{\tau}_{\omega,\nu}\left(s_0^1, s_0^2\right) \leq r | W_1^\pi > \epsilon, W_2^\pi > \epsilon\right)\mathbb{P}_1\left(W_1^\pi > \epsilon, W_2^\pi > \epsilon\right)\\ &\geq \epsilon^2\delta^2,\end{aligned}$$

where the ϵ, δ may be chosen independent of the choice of s_0^1, s_0^2. Hence,

$$\mathbb{P}\left(\widetilde{\tau}_{\omega^*,\nu^*}\left(s_0^1, s_0^2\right) > r\right) \leq \left(1 - \epsilon^2\delta^2\right).$$

Now the same arguments as in the proof of Lemma A.1 can be applied to get the desired conclusion.

We now extend the result of Lemmas A.1 and A.2 to nonstationary control strategies. For that, we use the following result from Borkar (1991) for a homogeneous MDP defined by the original transition kernel $p(\cdot|\cdot,\cdot)$. We include the proof for completeness.

Lemma A.3. (Borkar 1991, lemma 1.1, p. 42). *Let $(X_k, Z_k), k \geq k_0$ be the sequence of state-action pairs corresponding to the homogeneous MDP defined by an arbitrary control strategy $\sigma \in \Sigma$ and the transition kernel $p(\cdot|\cdot,\cdot)$. Then, there exist integer r^* and $\epsilon > 0$ such that*

$$\mathbb{P}\left(\tau(s,s') > r^*\right) < 1 - \epsilon, \quad \forall s, s' \in \mathbb{S}.$$

Suppose not. Then, there exists a sequence of controlled Markov chains $\{X_k^\alpha, k \geq k_0\}$, $\alpha = 1, 2, \ldots$ governed by control strategies $\{\sigma_k^\alpha, t \geq k_0\}$ (with the corresponding control sequences $\{Z_k^\alpha, k \geq k_0\}$) such that the following holds: If $\tau^\alpha(s,s') := \min\{k \geq 0 | X_{k_0+k}^\alpha = s', X_{k_0}^\alpha = s\}$, then

$$\mathbb{P}(\tau^\alpha(s,s') > \alpha) > 1 - \frac{1}{\alpha}, \quad \alpha \geq 1.$$

Because the state and action spaces are finite, the laws of $\{(X_k^\alpha, Z_k^\alpha), k \geq k_0\}, \alpha \geq 1$ are tight. By dropping to a subsequence if necessary and invoking Skorohod's theorem, we may assume that these chains are defined on a common probability space,

and there exists a controlled Markov chain $\{X_k^\infty, k \geq k_0\}$ governed by controls $Z_k^\infty, k \geq k_0$, corresponding to a control strategy σ^∞ with $X_{k_0}^\infty = s$, such that $(X_k^\alpha, Z_k^\alpha)_{k\geq 0} \to (X_k^\infty, Z_k^\infty)_{k\geq 0}$ a.s. Because

$$\mathbb{P}\left(\tau^\alpha(s,s') > j\right) = \mathbb{E}\left[\prod_{k=1}^{j} \mathbb{I}\left\{X_{k_0+k}^\alpha \neq s'\right\},\right] \quad \alpha, t = 1,2,\ldots,$$

a straightforward limiting argument leads to

$$\Pr(\tau^\infty(s,s') > \alpha) > 1 - \frac{1}{\alpha}, \quad \alpha \geq 1,$$

for $\tau^\infty(s,s') := \min\{k \geq 0 | X_{k_0+k}^\infty = s', X_{k_0}^\infty = s\}$. Then, $\tau^\infty = \infty$ a.s. This is possible only if there exists a nonempty subset H of $\mathbb{S} \setminus \{s'\}$ such that for each $i \in H$, $\max_{k\notin H} \min_{a\in\mathbb{A}} p(k|i,a) = 0$. Let a_i be the action at which the above minimum is achieved. Then the chain starting at H and governed by a stationary control strategy π such that $\pi(i) = a_i$ never leaves H. This contradicts Assumption 1 that under any stationary control strategy, $\mathbb{S}$ is irreducible. Thus, the given statement must hold.

Now we extend the result of Lemma A.1 to nonstationary control strategies.

Lemma A.4. *Let $(X_k(\omega,\nu), Z_k)_{k\geq k_0}$ be the sequence of state-action pairs for the MDP simulated according to (20) using an arbitrary control strategy $\phi_k^2 \approx \sigma_k, \forall k$. Let $\tau_{\omega,\nu}$ be the hitting time as defined in Equation (A.1). Then,*

$$\mathbb{E}\left[\tau_{\omega,\nu}(s_0,s')\right] < \infty, \ \forall s_0, s' \in \mathbb{S}.$$

The proof is similar to that of Lemma A.1. By Lemma A.3, there exists a $j^*, 0 < j^* \leq r^*$ and a sequence of states, $(s_{k_0+j})_{j=0}^{j^*}$, with $s_{k_0} = s_0$ and $s_{k_0+j^*} = s'$ such that $P^{\sigma_{k_0+1}}(s_{k_0}, s_{k_0+1}) \cdots P^{\sigma_{k_0+j^*}}(s_{k_0+r-1}, s_{k_0+j^*}) > 0$ where P^{σ_k} is defined as in (1) by replacing π with σ_k. Let

$$W^\sigma = W^\sigma\left((s_{k_0+j})_{j=0}^{j^*}\right) := \widehat{P}_{k_0}^{\sigma_{k_0}}(s_{k_0}, s_{k_0+1}) \cdots \widehat{P}_{k_0+r-1}^{\sigma_{k_0+j^*-1}}(s_{k_0+r-1}, s_{k_0+j^*}),$$

where $\widehat{P}^{\sigma_k}$ is defined as in (12) by replacing π with σ_k. As in the proof of Lemma A.1, $\mathbb{E}[\widehat{P}_k^{\sigma_k}] = P^{\sigma_k}, \forall k$ and because $\widehat{P}_k^{\sigma_k}$ are independent $\forall k$,

$$\mathbb{E}_1[W^\sigma] = P^{\sigma_{k_0+1}}(s_{k_0}, s_{k_0+1}) \cdots P^{\sigma_{k_0+j^*}}(s_{k_0+r-1}, s_{k_0+j^*}) > 0.$$

Then, there exists an $\epsilon > 0, \delta > 0$ such that $\mathbb{P}_1(W^\sigma > \epsilon) > \delta$. Then, as in the proof of Lemma A.1,

$$\mathbb{P}(\tau_{\omega,\nu}(s_0,s') > r) \ \leq \ (1 - \epsilon\delta), \quad \text{and} \quad \mathbb{E}\left[\tau_{\omega,\nu}(s_0,s')\right] < \infty.$$

Now, the proof of Proposition 1 is straightforward by combining the proofs of Lemmas A.2 and A.4.

References

Abounadi J, Bertsekas D, Borkar VS (2001) Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.* 40(3):681–698.

Abounadi J, Bertsekas DP, Borkar V (2002) Stochastic approximation for nonexpansive maps: Application to q-learning algorithms. *SIAM J. Control Optim.* 41(1):1–22.

Bertsekas DP (2012) *Dynamic Programming and Optimal Control*, vol. 2, 4th ed. (Athena Scientific, Nashua, NH).

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming* (Athena Scientific, Nashua, NH).

Blackwell D, Dubins L (1962) Merging of opinions with increasing information. *Ann. Math. Statist.* 33(3):882–886.

Borkar VS (1991) *Topics in Controlled Markov Chains* (Longman Scientific & Technical, Harlow, England).

Borkar VS (1995) *Probability Theory: An Advanced Course* (Springer, New York).

Borkar VS (2008) *Stochastic Approximation: A Dynamical Systems Viewpoint* (Cambridge University Press).

Diaconis P, Freedman D (1999) Iterated random functions. *SIAM Rev.* 41(1):45–76.

Haskell WB, Jain R, Kalathil D (2013) Empirical dynamic programming. *Math. Oper. Res.* 41(2);402–429.

Jaakkola T, Jordan MI, Singh SP (1994) On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.* 6(6):1185–1201.

Kearns MJ, Singh SP (1999) Finite-sample convergence rates for q-learning and indirect algorithms. Solla SA, Leen TK, Müller KR, eds. *Adv. Neural Inf. Processing Systems* (MIT Press, Cambridge, MA), 996–1002.

Konda VR, Borkar VS (1999) Actor-critic–type learning algorithms for Markov decision processes. *SIAM J. Control Optim.* 38(1):94–123.

Levin DA, Peres Y, Wilmer EL (2009) *Markov Chains and Mixing Times* (American Mathematical Society, Providence, RI).

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, vol. 703 (John Wiley & Sons, New York).

Propp JG, Wilson DB (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* 9(1-2):223–252.

Puterman ML (2005) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York).

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction, vol. 1 (MIT Press, Cambridge, MA).
Szepesvári C (2010) Algorithms for reinforcement learning. *Synthesis Lectures Artificial Intelligence Machine Learn.* 4(1):1–103.
Tsitsiklis JN (1994) Asynchronous stochastic approximation and q-learning. *Machine Learn.* 16(3):185–202.
Watkins CJ, Dayan P (1992) Q-learning. *Machine Learn.* 8(3-4):279–292.
Watkins CJH (1989) Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge, UK.
Yu H, Bertsekas DP (2013) On boundedness of q-learning iterates for stochastic shortest path problems. *Math. Oper. Res.* 38(2):209–227.