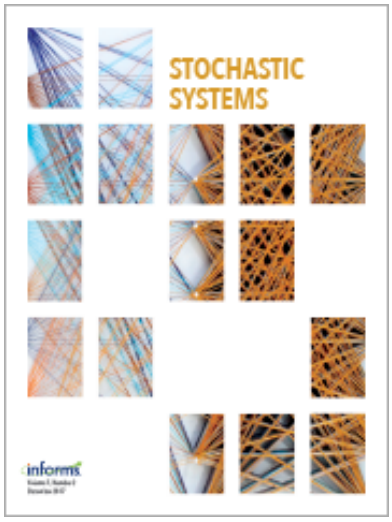

Stochastic Systems

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

A Concentration Bound for TD(0) with Function Approximation

Siddharth Chandak, Vivek S. Borkar

To cite this article:
Siddharth Chandak, Vivek S. Borkar (2026) A Concentration Bound for TD(0) with Function Approximation. Stochastic Systems 16(1):44-60. https://doi.org/10.1287/stsy.2023.0055

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

STOCHASTIC SYSTEMS
Vol. 16, No. 1, March 2026, pp. 44–60
ISSN 1946-5238 (online)

A Concentration Bound for TD(0) with Function Approximation

Siddharth Chandak,[a,*] **Vivek S. Borkar**[b]

[a] Department of Electrical Engineering, Stanford University, Stanford, California 94305; [b] Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra 400076, India

*Corresponding author

Contact: chandaks@stanford.edu, https://orcid.org/0000-0003-3237-7729 (SC); borkar.vs@gmail.com, https://orcid.org/0000-0003-0756-5402 (VSB)

Received: December 16, 2023
Revised: October 29, 2024; May 26, 2025
Accepted: October 24, 2025
Published Online in Articles in Advance: December 26, 2025

https://doi.org/10.1287/stsy.2023.0055

Abstract. We derive uniform all-time concentration bound of the type 'for all $n \geq n_0$ for some n_0' for TD(0) with linear function approximation. We work with online TD learning with samples from a single sample path of the underlying Markov chain. This makes our analysis significantly different from offline TD learning or TD learning with access to independent samples from the stationary distribution of the Markov chain. We treat TD(0) as a contractive stochastic approximation algorithm with both martingale and Markov noises. Markov noise is handled using the Poisson equation, and the lack of almost-sure guarantees on boundedness of iterates is handled using the concept of relaxed concentration inequalities.

Funding: The work of V. S. Borkar was supported in part by the S. S. Bhatnagar Fellowship from the Government of India.

Keywords: TD(0) • reinforcement learning • stochastic approximation • concentration bound • temporal difference learning • all time bound

1. Introduction

TD(0) is one of the most popular reinforcement learning (RL) algorithms for policy evaluation (Tsitsiklis and Van Roy 1997). Given a fixed policy, the algorithm is an iterative method to obtain the value function for each state under the long-term discounted reward framework. To mitigate the issues of large state spaces, the value function is often approximated using a linear combination of feature vectors. This algorithm is referred to as TD(0) with linear function approximation. In this paper, we work with online TD(0) with a single sample path of the underlying Markov chain. Our goal in this paper is to obtain a concentration bound of the form *from some time on* or, more precisely, *for all $n \geq n_0$ for a suitably chosen n_0* for this algorithm.

A bound of this form was published for TD(0) as a section in our paper titled "Concentration of Contractive Stochastic Approximation and Reinforcement Learning" (Chandak et al. 2022). This paper established an all-time bound for contractive stochastic approximation with Markov noise and applied the bound to asynchronous Q-learning and TD(0). Although the main theorem and its application to asynchronous Q-learning are correct, TD(0) does not satisfy a key assumption for the main theorem,[1] and hence, the theorem was incorrectly applied to TD(0). We remove the need for that assumption in this version, giving a completely different proof tailored to the TD(0) algorithm.

The previous paper required the iterates of the stochastic approximation iteration to be bounded by a constant with probability 1. This assumption is not known to be true for the iterates of online TD(0) with function approximation for a single sample path. In fact, a common method to alleviate this issue is to project the iterates back into a ball centered around the origin (Bhandari et al. 2018, Patil et al. 2023). The key difficulty caused by the lack of this assumption is in applying martingale inequalities, which often require some restrictions on the increments of the martingale that are often not easy to verify. We do not modify the algorithm and instead adapt relaxed concentration inequalities (Chung and Lu 2006, section 8) for our problem. These bounds have an extra term given by the probability of increments going above a certain threshold (Tao and Vu 2015, proposition 34). Although the proof in this paper is restricted to TD(0), the underlying idea of using relaxed concentration inequalities is broadly applicable to other algorithms that face similar challenges because of unboundedness.

1.1. Related Works

There has been growing interest in analyzing the finite-time performance of reinforcement learning (RL) algorithms. Existing results can broadly be categorized by the type of bounds they establish. The most extensive body of work concerns expectation or mean square bounds (see, e.g., Chen et al. 2020, 2021). Another prominent line of research focuses on regret bounds, which characterize how the cumulative error grows over time—typically through almost-sure or expected regret guarantees (see, e.g., Azar et al. 2017, Jin et al. 2018, Yang and Wang 2019, Yang et al. 2020). A third class comprises high-probability or concentration bounds (see, e.g., Even-Dar and Mansour 2003, Qu and Wierman 2020, Li et al. 2023). Our work falls within this category but differs from conventional analyses that establish high-probability guarantees only for sufficiently large time n. In contrast, we derive uniform all-time bounds, that is, bounds that hold for all $n \geq n_0$ with probability at least $1 - \delta$.

Specifically for TD(0), moment bounds have been established in Bhandari et al. (2018), Srikant and Ying (2019), and Chen et al. (2021). High-probability bounds have been established under various modifications of the TD(0) algorithm. These include uniform sampling from data sets (Prashanth et al. 2021), projection and tail averaging (Patil et al. 2023), and oracle access to independent and identically distributed (i.i.d.) samples of state–action–next state triplets (s, a, s') (Dalal et al. 2018, Chen et al. 2025).

One of us considered stochastic approximation involving contractive maps and martingale noise and derived maximal concentration bounds for this class of algorithms (Borkar 2022). This covered, in particular, synchronous Q-learning for discounted cost and some related schemes. In Chandak et al. (2022), we extended this work to cover "Markov noise" (Meerkov 1972) in the stochastic approximation scheme, allowing us to give bounds for the asynchronous case. As mentioned before, this work assumed almost-sure boundedness of iterates, which is not satisfied by the TD(0) algorithm. We remove the need for this assumption in our current work. Other articles aiming at bounds of these forms, such as the one we provide, are found in Chandak et al. (2023) for the LSPE algorithm and Borkar (2002), Kamal (2010), and Thoppe and Borkar (2019) for abstract stochastic approximation schemes. A recent work (Chen et al. 2025) considers all-time bounds for iterates without almost-sure boundedness, but they only consider additive and multiplicative noise and not Markovian noise as considered in this paper. Their proof technique relies on Moreau envelopes and a bootstrapping technique, which differs significantly from our work.

1.2. Outline and Notation

The rest of the paper is structured as follows. Section 2 gives a background to the TD(0) algorithm, along with the required assumptions and the stochastic approximation formulation. Section 3 states the main result and provides some insights into the result. The result is proved in Section 4. A concluding section highlights some future directions. Appendix A states a martingale inequality used in our proof, and Appendix B gives proofs for some technical lemmas. which are used to prove the main theorem.

Throughout this work, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^d$. θ denotes the zero vector in $\mathbb{R}^d$. The ℓth component of a vector x and a vector-valued function $h(\cdot)$ are denoted by $x(\ell)$ and $h^\ell(\cdot)$, respectively.

2. Background on TD(0)

TD(0) is an algorithm for policy evaluation, that is, for learning the performance of a fixed policy, and not for optimizing over policies. Hence, a stationary policy is fixed a priori, giving us a time-homogeneous uncontrolled Markov chain $\{Y_n\}$ over a finite state space $\mathcal{S}$. The transition probabilities are given by $p(\cdot|\cdot)$, where the dependence on the policy is suppressed. Assume that the chain is aperiodic irreducible with the stationary distribution $\pi = [\pi(1), \ldots, \pi(S)], S = |\mathcal{S}|$. Let D denote the $S \times S$ diagonal matrix whose sth diagonal entry is $\pi(s)$. Reward $r(s)$ is received when a transition from state s takes place. Note that this reward can be stochastic as well, and the additional noise thereof can be combined with other noise terms without affecting our concentration result. For simplicity, we assume that we receive a deterministic $r(s)$. The objective is to evaluate the long-term discounted reward for each state given by the value function

$$V(s) = E\left[\sum_{m=0}^{\infty} \gamma^m r(X_m) \,|\, X_0 = s\right], s \in \mathcal{S}.$$

Here, $0 < \gamma < 1$ is the discount factor. The dynamic programming equation for evaluating the same is

$$V(s) = r(s) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s) V(s'), \, s \in \mathcal{S}.$$

This can be written as the following vector equation:

$$V = r + \gamma PV,$$

for $r = [r(1), \ldots, r(S)]^T$ and $P = [[p(s'|s)]]_{s',s\in\mathcal{S}} \in \mathbb{R}^{S\times S}$.

The state space can often be large ($S \gg 1$), and to alleviate this "curse of dimensionality," V is often approximated using a linear combination of d linearly independent basis functions (feature vectors) $\phi_i \in \mathbb{R}^S, 1 \le i \le d$, with $S >> d \ge 1$. Also, let $\varphi(s) = [\phi_1(s), \ldots, \phi_d(s)]^T$ for $s \in \mathcal{S}$ denote the vector comprising components corresponding to state s in each feature. Thus, $V(s) \approx \sum_{i=1}^{d} x(i)\phi_i(s)$; that is, $V(s) \approx x^T\varphi(s)$ and $V \approx \Phi x$, where $x = [x(1), \ldots, x(d)]^T$, and Φ is an $S \times d$ matrix whose ith column is ϕ_i. Here, x denotes the learnable weights for the linear function approximator. Because $\{\phi_i\}$ are linearly independent, Φ is full rank. Substituting this approximation into the dynamic programming equation above leads to

$$\Phi x \approx r + \gamma P\Phi x.$$

But the right-hand side (RHS) may not belong to the range of Φ. So we use the following fixed point equation:

$$\Phi x = \Pi(r + \gamma P\Phi x) := H(\Phi x), \tag{1}$$

where Π denotes the projection to Range(Φ) with respect to a suitable norm. It turns out to be convenient to take a projection with respect to the weighted norm $\|y\|_D := \sqrt{y^T D y} = (\sum_{s\in\mathcal{S}} \pi(s)(y(s))^2)^{1/2}$ for $y \in \mathbb{R}^S$. The projection map with respect to this norm is

$$\Pi y := \Phi(\Phi^T D\Phi)^{-1}\Phi^T Dy.$$

The invertibility of $\Phi^T D\Phi$ is guaranteed by the fact that Φ is full rank. Finally, the TD(0) algorithm is given by the recursion

$$x_{n+1} = x_n + a(n)\varphi(Y_n)(r(Y_n) + \gamma\varphi(Y_{n+1})^T x_n - \varphi(Y_n)^T x_n). \tag{2}$$

Here, $a(n)$ denotes the positive step-size sequence. At the end of this section, we explain how this iteration can be expected to converge to the required fixed point from (1).

2.1. Assumptions

We impose two assumptions on the algorithm. The first is about the feature vectors, and as we explain next, it does not restrict the algorithm. The second specifies the class of step sizes $a(n)$ considered, which is standard in the analysis of RL. In fact, our results hold for a broader class of step sizes than those typically required in stochastic approximation frameworks.

1. For the assumption on Φ, define $\Psi := \Phi^T\sqrt{D}$, and let λ_M be the largest singular value of Ψ, that is, the square of the largest eigenvalue of $\Psi\Psi^T$ and, equivalently, of $\Psi^T\Psi$. Assume that

$$\lambda_M < \frac{\sqrt{2(1-\gamma)}}{(1+\gamma)}. \tag{3}$$

Because the feature vectors can be scaled without affecting the algorithm (the weights $x(i)$ get scaled accordingly), this assumption does not restrict the algorithm. Alternatively, the step size can be appropriately scaled as well; that is, $a(n) = b(n)c$, where $b(n)$ acts as the effective step size, and $c\varphi(\cdot)$ act as the effective feature vectors that satisfy the above assumption. This assumption can be replaced with the following assumption on the basis vectors:

$$\|\varphi(s)\| \le \frac{\sqrt{2(1-\gamma)}}{1+\gamma} \ \forall s \in \mathcal{S};$$

that is, the ℓ_2 norm of each row of Φ is bounded by $\sqrt{2(1-\gamma)}/(1+\gamma)$. To see this, note that

$$\frac{\|\Psi x\|_2}{\|x\|_2} = \frac{\|\Phi x\|_D}{\|x\|_2} \le \frac{\|\Phi x\|_\infty}{\|x\|_2}.$$

Now, $\max_{x\neq\theta} \frac{\|\Phi x\|_\infty}{\|x\|_2}$ is the operator norm defined with ℓ_2 norm for domain and ℓ_∞ for codomain, which is equal to the maximum ℓ_2 norm of a row. Hence,

$$\lambda_M = \max_{x\neq\theta} \frac{\|\Psi x\|_2}{\|x\|_2} \le \max_{x\neq\theta} \frac{\|\Phi x\|_\infty}{\|x\|_2} = \max_{s\in\mathcal{S}} \|\varphi(s)\|_2.$$

2. $\{a(n)\}$ is a sequence of nonnegative step sizes satisfying the conditions

$$a(n) \to 0, \sum_n a(n) = \infty$$

and is assumed to be nonincreasing; that is, $a(n+1) \leq a(n)\ \forall n$. We also assume that $a(n) < 1$ for all n. We further assume that $\frac{d_1}{n+1} \leq a(n) \leq d_3\left(\frac{1}{n+1}\right)^{d_2}$, $\forall n$, where $d_1 > 0$ and $0 < d_2 \leq 1$. Larger values of d_1 and d_2 and smaller values of d_3 improve the main result presented below. The role this assumption plays in our bounds will become clear later. Observe that we do not require the classical square-summability condition in stochastic approximation, namely, $\sum_n a(n)^2 < \infty$. This is because the contractive nature of our iterates (Lemma 1) gives us an additional handle on errors by putting less weight on past errors. A similar effect was observed in Chandak et al. (2022). The above assumptions on the step-size sequence can be weakened so as to hold only after some $N > 1$ without any changes in the analysis.

2.2. Formulation as a Stochastic Approximation Iteration

We next rearrange Algorithm (2) to separate the martingale noise and the Markov noise, and we write it as a stochastic approximation iteration:

$$\begin{aligned} x_{n+1} &= x_n + a(n)\varphi(Y_n)(r(Y_n) + \gamma\varphi(Y_{n+1})^T x_n - \varphi(Y_n)^T x_n) \\ &= x_n + a(n)(F(x_n, Y_n) - x_n + M_{n+1}x_n) \\ &= x_n + a(n)\left(\sum_{s\in\mathcal{S}} \pi(s)F(x_n, s) - x_n\right) + \underbrace{a(n)M_{n+1}x_n}_{\tau_1} + \underbrace{a(n)\left(F(x_n, Y_n) - \sum_{s\in\mathcal{S}\pi}(s)F(x_n, s)\right)}_{\tau_2}, \end{aligned}$$

where

$$F(x, Y) = \varphi(Y)r(Y) + \gamma\varphi(Y)\sum_{s'\in\mathcal{S}} p(s'|Y)\varphi(s')^T x - \varphi(Y)\varphi(Y)^T x + x,$$

and

$$M_{n+1} = \gamma\varphi(Y_n)\left(\varphi(Y_{n+1})^T - \sum_{s'\in\mathcal{S}} p(s'|Y_n)\varphi(s')^T\right).$$

Define the family of σ-fields $\mathcal{F}_n := \sigma(x_0, Y_m, m \leq n)$, $n \geq 0$. Then, $\{M_n x_{n-1}\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$; that is,

$$E[M_{n+1}x_n|\mathcal{F}_n] = \theta, a.s.\ \forall n,$$

where θ denotes the zero vector. The term τ_1 denotes the error because of the martingale noise term, and term τ_2 denotes the error because of the Markov noise $\{Y_n\}$.

The following lemma shows that the function $\sum_{s\in\mathcal{S}}\pi(s)F(\cdot, s)$ is a contraction. Let $\langle x, x'\rangle = x^T x'$ and $\langle x, z\rangle_D = x^T Dz$. Whereas the contraction property of the TD(0) algorithm is well-known (Tsitsiklis and Van Roy 1997), we obtain an explicit expression for the contraction factor.

Lemma 1. *For any $x, z \in \mathbb{R}^d$,*

$$\left\|\sum_{s\in\mathcal{S}} \pi(s)(F(x, s) - F(z, s))\right\| \leq \alpha\|x - z\|,$$

where

$$\alpha = \sqrt{1 - \min_{x\neq\theta} \frac{\|\Phi x\|_D^2}{\|x\|^2}(2(1-\gamma) - \lambda_M^2(1+\gamma)^2)}.$$

Moreover, $0 < \alpha < 1$, and hence, the function $\sum_{s\in\mathcal{S}}\pi(s)F(\cdot, s)$ is a contraction.

The proof appears in Appendix B. The Banach contraction mapping theorem implies that $\sum_{s\in\mathcal{S}}\pi(s)F(\cdot, s)$ has a unique fixed point x^*; that is, there exists a unique point x^* such that $\sum_{s\in\mathcal{S}}\pi(s)F(x^*, s) = x^*$. We next show that the fixed point x^* is the required fixed point we wish to converge to. Before that, we first observe that

$$\sum_{s\in\mathcal{S}} \pi(s)F(x, s) = (\Phi^T Dr + \gamma\Phi^T DP\Phi - \Phi^T D\Phi + I)x.$$

Then,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \pi(s)F(x^*,s) = (\Phi^T D r + \gamma \Phi^T D P \Phi - \Phi^T D \Phi + I)x^* = x^* \\
&\quad \Rightarrow (\Phi^T D r + \gamma \Phi^T D P \Phi)x^* = \Phi^T D \Phi x^* \\
&\quad \Rightarrow (\Phi^T D \Phi)^{-1}(\Phi^T D r + \gamma \Phi^T D P \Phi)x^* = x^* \\
&\quad \Rightarrow \Phi(\Phi^T D \Phi)^{-1}\Phi^T D(r + \gamma P \Phi)x^* = \Phi x^* \\
&\quad \Rightarrow H(\Phi x^*) = \Phi x^*.
\end{aligned}$$

So Φx^* is the required fixed point of (1).

3. Main Result

Before stating the main result, we define the following two sequences. For $n \geq 0$,

$$b_k(n) = \sum_{m=k}^{n} a(m),\ 0 \leq k \leq n < \infty,$$

$$\beta_k(n) = \begin{cases} \frac{1}{k^{d_2-d_1} n^{d_1}}, & \text{if } d_1 \leq d_2 \\ \frac{1}{n^{d_2}}, & \text{otherwise.} \end{cases}$$

Our main result is as follows:

Theorem 1. *There exist finite positive constants c_1, c_2 and D such that for $0 < \delta \leq 1, 0 < \epsilon \leq 1$, $n_0 > 0$ large enough to satisfy $\alpha + a(n_0)c_1 < 1$, and $n \geq n_0$,*

a. *the inequality*

$$\|x_m - x^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\epsilon + \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1},\ \ \forall n_0 \leq m \leq n,$$

holds with probability exceeding

$$1 - 2d \sum_{m=n_0+1}^{n} e^{-D\delta^2/\beta_{n_0}(m-1)} - P(\|x_{n_0} - x^*\| > \epsilon).$$

b. *In particular,*

$$\|x_m - x^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\epsilon + \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1},\ \ \forall m \geq n_0,$$

holds with probability exceeding

$$1 - 2d \sum_{m\geq n_0+1} e^{-D\delta^2/\beta_{n_0}(m-1n)} - P(\|x_{n_0} - x^*\| > \epsilon).$$

The following are some remarks about the theorem and the proof that follows.

Remark 1. The assumption that $\delta \leq 1$ and $\epsilon \leq 1$ is used in the proof for Lemma 3 and has been made only for simplicity. These can be taken as any positive values, with changes required only in the constant D.

Remark 2. The term $P(\|x_{n_0} - x^*\| > \epsilon)$ captures the unavoidable contribution of the initial condition at n_0. This can be bounded by combining moment bounds (Bhandari et al. 2018, Srikant and Ying 2019, Chen et al. 2021) with Markov's inequality.

Remark 3. In Chen et al. (2025), an all-time bound is obtained, which goes to zero as $m \uparrow \infty$. In our bound, the term $\frac{a(n_0)(c_2+c_1\epsilon)+\delta}{1-\alpha-a(n_0)c_1}$ remains constant as m is increased. Here, δ can be modified to $\delta(m)$ (similar to the treatment in Chandak et al. 2022, corollary 1). But the term $a(n_0)(c_2 + c_1\epsilon)$ arises from our treatment of Markov noise using the Poisson equation. Note that Chen et al. (2025) do not consider the case of Markov noise, but only consider the case of additive and multiplicative noise (i.i.d. samples of state-action-next state triplets). We leave incorporating their ideas into our approach to get a bound decaying with m for Markov noise as future work.

Remark 4. For the special case of $a(n) = \frac{d_1}{n+1}$, we combine our result with Chen et al. (2021, theorem 2.1), a mean square error bound, to get the following corollary.

Corollary 1. *Let $a(n) = d_1/(n+1)$ with sufficiently large d_1. Let n_0 be large enough to satisfy assumptions of Theorem 1 and Chen et al. (2021, theorem 2.1). Then, with probability at least $1-\varepsilon_1-\varepsilon_2$, we have, for all $m \geq n_0$, that*

$$\|x_m - x^*\| = \mathcal{O}\left(\frac{1}{\sqrt{n_0}}\log^{1/2}\left(\frac{1}{\varepsilon_1}\right) + \sqrt{\frac{\log(n_0)}{n_0}}\frac{1}{\sqrt{\varepsilon_2}}\left(\frac{n_0}{m}+\frac{1}{n_0}\right)\right).$$

The proof for this corollary has been presented at the end of Appendix B. The first term here corresponds to the term δ in Theorem 1. This term has a $\sqrt{1/n_0}$ decay rate and an exponentially small tail. The second term is the contribution of the initial condition at n_0. We have a polynomial tail in this case, but the dependence on m and n_0 is stronger, as the term $(\sqrt{\log(n_0)}/\sqrt{n_0}) \times (n_0/m)$ decays with m, and the other term decays as $\log^{1/2}(n_0)n_0^{-3/2}$.

4. Proof of the Main Result

We present the proof of the main theorem in this section. The key martingale concentration inequality used in our proof is stated in Appendix A, and proofs for the technical lemmas used in the proof are presented in Appendix B.

Proof of Theorem 1. Define z_n for $n \geq n_0$ by

$$z_{n+1} = z_n + a(n)\left(\sum_s \pi(s)F(z_n,s) - z_n\right),$$

where $z_{n_0} = x_{n_0}$. Note that $\|x_n - x^*\| \leq \|x_n - z_n\| + \|z_n - x^*\|$. To bound the second term, note that

$$\begin{aligned} z_{n+1} - x^* &= (1-a(n))(z_n - x^*) + a(n)\left(\sum_{s\in\mathcal{S}} \pi(s)F(z_n,s) - x^*\right) \\ &= (1-a(n))(z_n - x^*) + a(n)\sum_{s\in\mathcal{S}} \pi(s)(F(z_n,s) - F(x^*,s)). \end{aligned}$$

The second equality follows from the fact that x^* is a fixed point for $\sum_{s\in\mathcal{S}}\pi(s)F(\cdot,s)$. Then,

$$\begin{aligned} \|z_{n+1} - x^*\| &\leq (1-a(n))\|z_n - x^*\| + a(n)\left\|\sum_{s\in\mathcal{S}} \pi(s)(F(z_n,s) - F(x^*,s))\right\| \\ &\leq (1-(1-\alpha)a(n))\|z_n - x^*\|. \end{aligned}$$

which finally gives us

$$\|z_n - x^*\| \leq \prod_{k=n_0}^{n-1}(1-(1-\alpha)a(k))\|x_{n_0} - x^*\| \leq e^{-(1-\alpha)b_{n_0}(n-1)}\|x_{n_0} - x^*\|, \tag{4}$$

This also implies that for all $n \geq n_0$,

$$\|z_n\| \leq \|x_{n_0} - x^*\| + \|x^*\|. \tag{5}$$

Next, we give a probabilistic bound on the term $\|x_n - z_n\|$. Note that

$$\begin{aligned} x_{n+1} - z_{n+1} &= (1-a(n))(x_n - z_n) + a(n)M_{n+1}x_n + a(n)\left(F(x_n,Y_n) - \sum_s \pi(s)F(z_n,s)\right) \\ &= (1-a(n))(x_n - z_n) + a(n)M_{n+1}x_n + a(n)\left(\sum_{s\in\mathcal{S}} \pi(s)(F(x_n,s) - F(z_n,s))\right) \\ &\quad + a(n)\left(F(x_n,Y_n) - \sum_{s\in\mathcal{S}} \pi(s)F(x_n,s)\right). \end{aligned}$$

For $n, m \geq 0$, let $\chi(n,m) = \prod_{k=m}^{n}(1-a(k))$ if $n \geq m$, and one otherwise. For some $n \geq n_0$, we iterate the above for $n_0 \leq m \leq n$ to obtain

$$\begin{aligned} x_{m+1} - z_{m+1} = &\sum_{k=n_0}^{m} \chi(m,k+1)a(k)M_{k+1}x_k \\ &+ \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(\sum_{s\in\mathcal{S}} \pi(s)(F(x_k,s) - F(z_k,s))\right) \\ &+ \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(F(x_k,Y_k) - \sum_{s\in\mathcal{S}} \pi(s)F(x_k,s)\right). \end{aligned} \tag{6}$$

Here, we use the definition that $x_{n_0} = z_{n_0}$. We first simplify the third term above. For simplicity, we define $F(x,Y) = F_1(Y) + F_2(Y)x + x$, where

$$F_1(Y) = \varphi(Y)r(Y) \in \mathbb{R}^d \text{ and } F_2(Y) = \left(\gamma\varphi(Y)\sum_{s'\in\mathcal{S}} p(s'|Y)\varphi(s')^T - \varphi(Y)\varphi(Y)^T\right) \in \mathbb{R}^{d\times d}.$$

We define $U : \mathcal{S} \longmapsto \mathbb{R}^d$ to be a solution of the Poisson equation

$$U(s) = F_1(s) - \sum_{s'\in\mathcal{S}} \pi(s')F_1(s') + \sum_{s'\in\mathcal{S}} p(s'|s)U(s'),\ s \in \mathcal{S}. \tag{7}$$

For $s_0 \in \mathcal{S}$, $\tau := \min\{n > 0 : Y_n = s_0\}$, and $E_s[\cdots] = E[\cdots | Y_0 = s]$; we know that

$$U'(s) = E_s\left[\sum_{m=0}^{\tau-1}(F_1(Y_m) - \sum_{s'\in\mathcal{S}} \pi(s')F_1(s'))\right],\ s \in S \tag{8}$$

is one particular solution to the Poisson equation (see, e.g., Borkar 1991, pp. 85–91, section VI.4, lemma 4.2 and theorem 4.2). Thus, $\|U'(s)\|_\infty \leq 2\max_{s\in\mathcal{S}}\|F_1(s)\|_\infty E_s[\tau]$. For an irreducible Markov chain with a finite state space, $E_s[\tau]$ is finite for all s, and hence, the solution $U'(s)$ is bounded for all s. For each ℓ, the Poisson equation specifies $U^\ell(\cdot)$ uniquely only up to an additive constant. Along with the additional constraint that $U(s_0) = 0$ for a prescribed $s_0 \in S$, the system of equations given by (7) has a unique solution. Henceforth, U refers to the unique solution of the Poisson equation with $U(s_0) = 0$. Similarly, let $W : \mathcal{S} \longmapsto \mathbb{R}^{d\times d}$ be the unique solution of the Poisson equation

$$W(s) = F_2(s) - \sum_{s'} \pi(s')F_2(s') + \sum_{s'} p(s'|s)W(s),\ s \in S, \tag{9}$$

with the additional constraint that $W(s_0) = 0$ for a prescribed $s_0 \in S$ as above. □

The following lemma gives a simplification of the third term in (6), using the solutions of the Poisson equation stated above. Before stating the lemma, we first define $x'_m = \sup_{n_0\leq k\leq m}\|x_m - z_m\|$.

Lemma 2. *There exist positive constants c_1, c_2 such that for all $n_0 \leq m \leq n$,*

$$\begin{aligned} &\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(F(x_k,Y_k) - \sum_{s\in\mathcal{S}} \pi(s)F(x_k,s)\right) \\ &\quad = \sum_{k=n_0}^{m} \chi(m,k+1)a(k)(\tilde{U}_{k+1} + \tilde{W}_{k+1}x_k) + \mu_m(n_0), \end{aligned}$$

where

$$\|\mu_m(n_0)\| \leq a(n_0)(c_2 + c_1x'_m + c_1\|x_{n_0} - x^*\|).$$

Here, $\tilde{U}_{k+1}$ and $\tilde{W}_{k+1}x_k$ are martingale difference sequences with respect to $\mathcal{F}_k$, where $\tilde{U}_{k+1} = U(Y_{k+1}) - \sum_{s'} p(s'|Y_k)U(s')$ and $\tilde{W}_{k+1} = W(Y_{k+1}) - \sum_{s'} p(s'|Y_k)W(s')$ for $k \geq n_0$ and the zero vector, respectively, or the zero matrix, otherwise.

The proof appears in Appendix B. Returning to (6), we now have

$$x_{m+1} - z_{m+1} = \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(\sum_{s\in\mathcal{S}} \pi(s)(F(x_k,s) - F(z_k,s))\right) + \sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1}) + \mu_m(n_0).$$

Now,

$$\begin{aligned}\|x_{m+1} - z_{m+1}\| &\le \left\|\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(\sum_{s\in\mathcal{S}} \pi(s)(F(x_k,s) - F(z_k,s))\right)\right\| \\ &\quad + \left\|\sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1})\right\| \\ &\quad + a(n_0)(c_2 + c_1 x'_m + c_1\|x_{n_0} - x^*\|) \\ &\le \alpha\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\|x_k - z_k\| + a(n_0)(c_2 + c_1 x'_m + c_1\|x_{n_0} - x^*\|) \\ &\quad + \left\|\sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1})\right\|. \end{aligned} \tag{10}$$

For any $0 < k \le m$,

$$\chi(m,k) + \chi(m,k+1)a(k) = \chi(m,k+1),$$

and hence,

$$\chi(m,n_0) + \sum_{k=n_0}^{m} \chi(m,k+1)a(k) = \chi(m,m+1) = 1.$$

This implies that

$$\sum_{k=n_0}^{m} \chi(m,k+1)a(k) \le 1.$$

Using the definition of x'_m, we have

$$x'_{m+1} \le (\alpha + a(n_0)c_1)x'_m + \left\|\sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1})\right\| + a(n_0)(c_2 + c_1\|x_{n_0} - x^*\|). \tag{11}$$

Next, we wish to obtain a bound on the probability

$$P(\|x_m - x^*\| \le \exp(-(1-\alpha)b_{n_0}(m-1))\epsilon + \Delta(n_0,\epsilon,\delta), \ \ \forall n_0 \le m \le n),$$

for some $\epsilon > 0$ and $\delta > 0$ (recall the assumption that $\alpha + a(n_0)c_1 < 1$). For ease of notation, here, we have defined

$$\Delta(n_0,\epsilon,\delta) := \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}.$$

From (4), recall that $\|z_n - x^*\| \le \exp(-(1-\alpha)b_{n_0}(n-1))\|x_{n_0} - x^*\|$ *a.s.*, and hence,

$$\|x_{n_0} - x^*\| \le \epsilon \Rightarrow \|z_m - x^*\| \le \exp(-(1-\alpha)b_{n_0}(m-1))\epsilon.$$

Also recall that $\sup_{n_0\le m\le n}\|x_m - z_m\| = x'_n$. Hence,

$$\{\|x_{n_0} - x^*\| \le \epsilon\} \bigcap \{x'_n \le \Delta(n_0,\epsilon,\delta)\} \subseteq \{\|x_m - x^*\| \le \exp(-(1-\alpha)b_{n_0}(m-1))\epsilon + \Delta(n_0,\epsilon,\delta), \ \ \forall n_0 \le m \le n\}.$$

This implies the following relation between the probabilities of the two sets:

$$\begin{aligned} &P(\|x_m - x^*\| \le \exp(-(1-\alpha)b_{n_0}(m-1))\|x_{n_0} - x^*\| + \Delta(n_0,\epsilon,\delta), \ \ \forall n_0 \le m \le n) \\ &\ge 1 - P(\{x'_n > \Delta(n_0,\epsilon,\delta)\} \bigcup \{\|x_{n_0} - x^*\| > \epsilon\}). \end{aligned}$$

To compensate for the lack of an almost-sure bound on the iterates $\{x_n\}$, we adapt the proof method from Tao and Vu (2015, proposition 34) (see Chung and Lu 2006, section 8, for a detailed explanation). For this, we define

$\xi = \{x_0, Y_k, k \geq 0\}$ and the "bad" set $\mathbb{B}_m$ as

$$\mathbb{B}_m = \left\{\xi \,|\, x'_m(\xi) > \Delta(n_0,\epsilon,\delta) \bigcup \|x_{n_0}(\xi) - x^*\| > \epsilon\right\}.$$

Here, the notation $x'_m(\xi)$ and $x_{n_0}(\xi)$ highlights the dependence of x'_m and x_{n_0} on the realizations of x_0 and $\{Y_k\}$. Analogous notation is used for other random variables. For $\xi \notin \mathbb{B}_{n-1}$, let us define $\overline{x}_{k,n-1}(\xi) = x_k(\xi)$ and $\overline{z}_{k,n-1}(\xi) = z_k(\xi)$ for all k. For $\xi \in \mathbb{B}_{n-1}$, we define $\overline{x}_{k,n-1}(\xi) = x^*$ and $\overline{z}_{k,n-1}(\xi) = x^*$ for all k. Also, define $\overline{x}'_{m,n-1}(\xi) = \sup_{n_0 \leq k \leq m} \|\overline{x}_{k,n-1}(\xi) - \overline{z}_{k,n-1}(\xi)\|$. Note that $\overline{x}'_{m,n-1} = 0$ when $\xi \in \mathbb{B}_{n-1}$, and $\overline{x}'_{m,n-1} = x'_m \leq \Delta(n_0,\epsilon,\delta)$ when $\xi \notin \mathbb{B}_{n-1}$. The intuition behind these definitions is that $\overline{x}'_{m,n-1}$ is always bounded by $\Delta(n_0,\epsilon,\delta)$ for all $m \leq n-1$.

Note that $\xi \notin \mathbb{B}_{n-1} \Rightarrow \overline{x}'_{n,n-1}(\xi) = x'_n(\xi)$, which implies $P(\overline{x}'_{n,n-1}(\xi) \neq x'_n(\xi)) \leq P(\mathbb{B}_{n-1})$. Henceforth, we drop ξ for ease of notation, rendering implicit the dependence of all random variables on ξ. Then,

$$\begin{aligned}
P(x'_n > \Delta(n_0,\epsilon,\delta)) &\leq P\left(x'_n > \Delta(n_0,\epsilon,\delta) \bigcup \|x_{n_0} - x^*\| > \epsilon\right) \\
&\overset{(a)}{\leq} P\left(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta) \bigcup \overline{x}'_{n,n-1} \neq x'_n \bigcup \|x_{n_0} - x^*\| > \epsilon\right) \\
&\overset{(b)}{\leq} P\left(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta)\right) + P(\overline{x}'_{n,n-1} \neq x'_n \bigcup \|x_{n_0} - x^*\| > \epsilon) \\
&\overset{(c)}{\leq} P\left(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta)\right) + P(\mathbb{B}_{n-1}) \\
&= P\left(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta)\right) + P\left(x'_{n-1} > \Delta(n_0,\epsilon,\delta) \bigcup \|x_{n_0} - x^*\| > \epsilon\right).
\end{aligned}$$

Inequality (a) here follows from the observation that

$$\left\{\overline{x}'_{n,n-1} \leq \Delta(n_0,\epsilon,\delta)\right\} \bigcap \{\overline{x}'_{n,n-1} = x'_n\} \subseteq \{x'_n \leq \Delta(n_0,\epsilon,\delta)\},$$

which implies that

$$\{x'_n > \Delta(n_0,\epsilon,\delta)\} \subseteq \{\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta)\} \bigcup \{\overline{x}'_{n,n-1} \neq x'_n\},$$

which gives us the required inequality. Inequality (b) follows from union bound, and inequality (c) follows from the observations that $\{\|x_{n_0} - x^*\| > \epsilon\} \subseteq \mathbb{B}_{n-1}$ and $\{\overline{x}'_{n,n-1} \neq x'_n\} \subseteq \mathbb{B}_{n-1}$.

Now, we obtain a bound for $P(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta))$ by induction. We first note that $\overline{x}'_{n-1,n-1}$ is bounded by $\Delta(n_0,\epsilon,\delta)$ by definition. Hence, $P(\overline{x}'_{n,n-1} > \Delta(n_0,\epsilon,\delta)) = P(\|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| > \Delta(n_0,\epsilon,\delta))$. We first restate (11) for $m = n-1$.

$$\begin{aligned}
\|x_n - z_n\| \leq x'_n \leq (\alpha + a(n_0)c_1)x'_{n-1} + a(n_0)(c_2 + c_1\|x_{n_0} - x^*\|) \\
+ \left\|\sum_{k=n_0}^{n-1} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1})\right\|.
\end{aligned}$$

Now, let $I\{\cdot\}$ denote the indicator function, which is one when $\{\cdot\}$ holds true, and zero otherwise.

$$\begin{aligned}
&\|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| \\
&= \|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| I\{\xi_{n-1} \in \mathbb{B}_{n-1}\} + \|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| I\{\xi_{n-1} \notin \mathbb{B}_{n-1}\} \\
&\overset{(a)}{=} 0 \times I\{\xi_{n-1} \in \mathbb{B}_{n-1}\} + \|x_n - z_n\| \times I\{\xi_{n-1} \notin \mathbb{B}_{n-1}\} \\
&\leq I\{\xi_{n-1} \notin \mathbb{B}_{n-1}\} \times \Bigg((\alpha + a(n_0)c_1)x'_{n-1} + a(n_0)(c_2 + c_1\|x_{n_0} - x^*\|) \\
&\quad + \left\|\sum_{k=n_0}^{n-1} \chi(m,k+1)a(k)(M_{k+1}x_k + \tilde{W}_{k+1}x_k + \tilde{U}_{k+1})\right\| \Bigg) \\
&\overset{(b)}{\leq} I\{\xi_{n-1} \notin \mathbb{B}_{n-1}\} \times \Bigg((\alpha + a(n_0)c_1)\Delta(n_0,\epsilon,\delta) + a(n_0)(c_2 + c_1\epsilon) \\
&\quad + \left\|\sum_{k=n_0}^{n-1} \chi(n-1,k+1)a(k)(M_{k+1}\overline{x}_{k,n-1} + \tilde{W}_{k+1}\overline{x}_{k,n-1} + \tilde{U}_{k+1})\right\| \Bigg).
\end{aligned}$$

Here, inequality (a) follows from our definition of $\mathbb{B}_{n-1}$ that $\|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| = 0$ when $\xi_{n-1} \in \mathbb{B}_{n-1}$ and $x_n = \overline{x}_{n,n-1}$ when $\xi_{n-1} \notin \mathbb{B}_{n-1}$. Inequality (b) follows from the fact that when $\xi_{n-1} \notin \mathbb{B}_{n-1}$, then $x_k = \overline{x}_{k,n-1}$ for all k, and $\|x_{n_0} - x^*\| \leq \epsilon$. Substituting the expression for $\Delta(n_0, \epsilon, \delta)$, we obtain the following:

$$\begin{aligned}\|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| &\leq (\alpha + a(n_0)c_1)\frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1} + a(n_0)(c_2 + c_1\epsilon) \\ &\quad + \left\| \sum_{k=n_0}^{n-1} \chi(n-1, k+1)a(k)(M_{k+1}\overline{x}_{k,n-1} + \tilde{W}_{k+1}\overline{x}_{k,n-1} + \tilde{U}_{k+1}) \right\| \\ &\leq \frac{a(n_0)(c_2 + c_1\epsilon)}{1 - \alpha - a(n_0)c_1} + \frac{\alpha + a(n_0)c_1}{1 - \alpha - a(n_0)c_1}\delta \\ &\quad + \left\| \sum_{k=n_0}^{n-1} \chi(n-1, k+1)a(k)(M_{k+1}\overline{x}_{k,n-1} + \tilde{W}_{k+1}\overline{x}_{k,n-1} + \tilde{U}_{k+1}) \right\|.\end{aligned}$$

When

$$\left\| \sum_{k=n_0}^{n-1} \chi(n-1, k+1)a(k)(M_{k+1}\overline{x}_{k,n-1} + \tilde{W}_{k+1}\overline{x}_{k,n-1} + \tilde{U}_{k+1}) \right\| \leq \delta,$$

we have

$$\|\overline{x}_{n,n-1} - \overline{z}_{n,n-1}\| \leq \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}.$$

Hence,

$$P\left(\overline{x}'_{n,n-1} > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}\right) \leq P\left(\left\| \sum_{k=n_0}^{n-1} \chi(n-1, k+1)a(k)(M_{k+1}\overline{x}_{k,n-1} + \tilde{W}_{k+1}\overline{x}_{k,n-1} + \tilde{U}_{k+1}) \right\| > \delta\right).$$

Let us denote the probability on the right side of the inequality as p_{n-1}. Then,

$$P\left(x'_n > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}\right) \leq p_{n-1} + P\left(x'_{n-1} > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1} \bigcup \|x_{n_0} - x^*\| > \epsilon\right).$$

Then, repeating the same procedure using $\mathbb{B}_{n-2}$, we obtain

$$\begin{aligned}&P\left(x'_{n-1} > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1} \bigcup \|x_{n_0} - x^*\| > \epsilon\right) \\ &\quad \leq p_{n-2} + P\left(x'_{n-2} > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1} \bigcup \|x_{n_0} - x^*\| > \epsilon\right).\end{aligned}$$

Iterating this for $n \geq m \geq n_0 + 1$, we get

$$P\left(x'_n > \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}\right) \leq \sum_{m=n_0+1}^{n} p_{m-1} + P(\|x_{n_0} - x^*\| > \epsilon).$$

The probabilities p_m can be bounded using standard martingale inequalities as the terms of the martingale difference sequence are almost surely bounded. The following lemma, proved in Appendix B, gives a bound on the probabilities p_m:

Lemma 3. *There exists positive constant D such that for $0 < \epsilon \leq 1, 0 < \delta \leq 1$,*

$$p_m \leq 2de^{-D\delta^2/\beta_{n_0}(m)}.$$

Recall that d here denotes the dimension of the iterates $\{x_n\}$.

This completes the proof for the first part of Theorem 1.

Let A_n be the set

$$\left\{\|x_m - x^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\epsilon + \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1},\ \forall n_0 \leq m \leq n\right\}.$$

Then, $\{A_n\}$ is a decreasing sequence of sets; that is, $A_{n+1} \subseteq A_n$ for all $n \geq n_0$. Now, let A be the set

$$\left\{\|x_m - x^*\| \leq e^{-(1-\alpha)b_{n_0}(m-1)}\epsilon + \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1},\ \forall m \geq n_0\right\}.$$

Then, $A = \cap_{n=n_0}^{\infty} A_n$. Hence, $P(A) = \lim_{n\uparrow\infty} P(A_n)$. This completes the proof for Theorem 1.

5. Conclusions

In conclusion, we note some future directions. The concept of relaxed martingale concentration inequalities can be used to obtain bounds of the similar flavor for algorithms that suffer from similar issues. These include TD(λ) and SSP Q-Learning. Alternatively, similar bounds can be obtained for variants of temporal difference learning (Chen et al. 2021). Another direction could be to improve the bounds in this paper to get an exponentially small tail for Markovian stochastic approximation.

Appendix A. A Martingale Inequality

Let $\{M_n\}$ be a real-valued martingale difference sequence with respect to an increasing family of σ-fields $\{\mathcal{F}_n\}$. Assume that there exist $\varepsilon, C > 0$ such that

$$E[e^{\varepsilon|M_n|}\,|\,\mathcal{F}_{n-1}] \leq C \quad \forall n \geq 1 \text{ a.s.}$$

Let $S_n := \sum_{m=1}^{n} \zeta_{m,n} M_m$, where $\zeta_{m,n}$, $m \leq n$, for each n, are a.s. bounded $\{\mathcal{F}_n\}$-previsible random variables; that is, $\zeta_{m,n}$ is $\mathcal{F}_{m-1}$-measurable $\forall m \geq 1$, and $|\zeta_{m,n}| \leq A_{m,n}$ a.s. for some constant $A_{m,n}$, $\forall m, n$. Suppose

$$\sum_{m=1}^{n} A_{m,n} \leq \gamma_1,\ \max_{1\leq m\leq n} A_{m,n} \leq \gamma_2\omega(n),$$

for some $\gamma_i, \omega(n) > 0$, $i = 1, 2; n \geq 1$. Then, we have

Theorem A.1. *There exists a constant $D > 0$ depending on $\varepsilon, C, \gamma_1, \gamma_2$ such that for $\epsilon > 0$,*

$$P(|S_n| > \epsilon) \leq 2e^{-\frac{D\epsilon^2}{\omega(n)}}, \text{ if } \epsilon \in \left(0, \frac{C\gamma_1}{\varepsilon}\right], \tag{A.1}$$

$$2e^{-\frac{D\epsilon}{\omega(n)}} \quad \textit{otherwise.} \tag{A.2}$$

This is a variant of Liu and Watbled (2009, theorem 1.1). See Thoppe and Borkar (2019, pp. 21–23, theorem A.1) for details.

Appendix B. Technical Proofs

B.1. Proof of Lemma 1

Proof.

$$\begin{aligned}\left\|\sum_{s\in\mathcal{S}} \pi(s)(F(x,s) - F(z,i))\right\|^2 &= \left\|\gamma\sum_{s\in\mathcal{S}} \pi(s)\varphi(s)\sum_{s'\in\mathcal{S}} p(s'|s)\varphi(s')^T(x-z) - \sum_{s\in\mathcal{S}} \pi(s)\varphi(s)\varphi(s)^T(x-z) + (x-z)\right\|^2 \\ &= \|(\gamma\Phi^T DP\Phi - \Phi^T D\Phi + I)(x-z)\|^2 \\ &= \|(\gamma\Phi^T DP\Phi - \Phi^T D\Phi)(x-z)\|^2 + (x-z)^T(x-z) - 2(x-z)^T\Phi^T D\Phi(x-z) \\ &\quad + (x-z)^T(\gamma\Phi^T DP\Phi + \gamma\Phi^T P^T D\Phi)(x-z). \end{aligned} \tag{B.1}$$

Now,

$$\begin{aligned}(x-z)^T(\gamma\Phi^T DP\Phi + \gamma\Phi^T P^T D\Phi)(x-z) &= (x-z)^T\gamma\Phi^T(DP + P^T D)\Phi(x-z) \\ &= \gamma\langle\Phi(x-z), P\Phi(x-z)\rangle_D + \gamma\langle P\Phi(x-z), \Phi(x-z)\rangle_D \\ &\overset{(a)}{\leq} 2\gamma\|P\Phi(x-z)\|_D\|\Phi(x-z)\|_D \\ &\overset{(b)}{\leq} 2\gamma\|\Phi(x-z)\|_D^2, \end{aligned} \tag{B.2}$$

and

$$2(x-z)^T\Phi^T D\Phi(x-z) = 2\langle\Phi(x-z), \Phi(x-z)\rangle_D = 2\|\Phi(x-z)\|_D^2. \tag{B.3}$$

Inequality (a) follows from the Cauchy–Schwarz inequality, and (b) follows from the observation that $\|Py\|_D \leq \|y\|_D$, which can be proved as follows:

$$\|Py\|_D^2 = \sum_{s\in\mathcal{S}} \pi(s)\left(\sum_{s'\in\mathcal{S}} p(s'|s)y(s')\right)^2 \leq \sum_{s\in\mathcal{S}} \pi(s)\sum_{s'\in\mathcal{S}} p(s'|s)y(s')^2 = \sum_{s'\in\mathcal{S}} \pi(s')y(s')^2 = \|y\|_D^2.$$

Here, the inequality follows from Jensen's inequality.

Combining (B.2) and (B.3) with (B.1) gives us

$$\left\|\sum_{s\in\mathcal{S}} \pi(s)(F(x,s)-F(z,s))\right\|^2 \leq \|x-z\|^2 - 2(1-\gamma)\|\Phi(x-z)\|_D^2 + \|(\gamma\Phi^T DP\Phi - \Phi^T D\Phi)(x-z)\|^2. \tag{B.4}$$

To analyze the last term in (B.4), we use the fact that the operator norm of a matrix defined as $\|M\| = \sup_{x\neq\theta} \frac{\|Mx\|}{\|x\|}$, using the Euclidean norm for vectors, is equal to the largest singular value of that matrix. Thus,

$$\begin{aligned}
\|(\gamma\Phi^T DP\Phi - \Phi^T D\Phi)(x-z)\|^2 &= \|\Phi^T\sqrt{D}(\gamma\sqrt{D}P\Phi - \sqrt{D}\Phi)(x-z)\|^2 \\
&\leq \lambda_M^2\|(\gamma\sqrt{D}P\Phi - \sqrt{D}\Phi)(x-z)\|^2 \\
&= \lambda_M^2\langle(\gamma P - I)\Phi(x-z), (\gamma P - I)\Phi(x-z)\rangle_D \\
&= \lambda_M^2\|(I-\gamma P)\Phi(x-z)\|_D^2 \\
&\leq \lambda_M^2(1+\gamma)^2\|\Phi(x-z)\|_D^2.
\end{aligned} \tag{B.5}$$

The last inequality follows from the triangle inequality. We now invoke Assumption (3) and combine (B.5) with (B.4) as follows:

$$\begin{aligned}
\left\|\sum_{s\in\mathcal{S}} \pi(s)(F(x,s)-F(z,s))\right\|^2 &\leq \|x-z\|^2 - 2(1-\gamma)\|\Phi(x-z)\|_D^2 + \lambda_M^2(1+\gamma)^2\|\Phi(x-z)\|_D^2 \\
&< \|x-z\|^2 - 2(1-\gamma)\|\Phi(x-z)\|_D^2 + \left(\frac{\sqrt{2(1-\gamma)}}{1+\gamma}\right)^2(1+\gamma)^2\|\Phi(x-z)\|_D^2 \\
&= \|x-z\|^2.
\end{aligned} \tag{B.6}$$

This gives us the required contraction property with contraction factor α for which an explicit expression can be obtained, using the first inequality in (B.6) as

$$\alpha = \sqrt{1 - \min_{x\neq\theta} \frac{\|\Phi x\|_D^2}{\|x\|^2}(2(1-\gamma) - \lambda_M^2(1+\gamma)^2)}.$$

Note that as the columns of Φ are linearly independent, $x \neq \theta \Rightarrow \Phi x \neq \theta$, and hence, $\frac{\|\Phi x\|_D}{\|x\|} > 0$ when $x \neq \theta$. Also, note that $\min_{x\neq\theta} \frac{\|\Phi x\|_D}{\|x\|} = \min_{\|x\|=1}\|\Phi x\|_D$, and hence, by the extreme value theorem, we have that $\min_{\|x\|=1}\|\Phi x\|_D$ is attained and is greater than zero. Along with Assumption (3), this implies that $\alpha < 1$. □

B.2. Proof of Lemma 2

Proof. Using definitions of $U(\cdot)$ and $W(\cdot)$, we have

$$\begin{aligned}
&\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(F(x_k,Y_k) - \sum_{s\in\mathcal{S}}\pi(s)F(x_k,s)\right) \\
&= \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(U(Y_k) - \sum_{s'\in\mathcal{S}} p(s'|Y_k)U(s')\right)
\end{aligned} \tag{B.7a}$$

$$+\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(W(Y_k) - \sum_{s'\in\mathcal{S}} p(s'|Y_k)W(s')\right)x_k. \tag{B.7b}$$

We first simplify (B.7a) as follows:

$$\begin{aligned}
&\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(U(Y_k) - \sum_{s'\in\mathcal{S}} p(s'|Y_k)U(s')\right) \\
&= \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(U(Y_{k+1}) - \sum_{s'\in\mathcal{S}} p(s'|Y_k)U(s')\right)
\end{aligned} \tag{B.8a}$$

$$+ \sum_{k=n_0+1}^{m} ((\chi(m,k+1)a(k) - \chi(m,k)a(k-1))U(Y_k)) \tag{B.8b}$$

$$+\chi(m,n_0+1)a(n_0)U(Y_{n_0}) - \chi(m,m+1)a(m)U(Y_{m+1}). \tag{B.8c}$$

For (B.8a), define $\tilde{U}_{k+1} = U(Y_{k+1}) - \sum_{s' \in \mathcal{S}} p(s'|Y_k)U(s')$ for $k \geq n_0$, and zero otherwise. This is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$.

We define $U_{max} := \max_{i \in S} \|U(i)\|$ and bound the norm of (B.8b) as follows:

$$\begin{aligned}
&\left\| \sum_{k=n_0+1}^{m} ((\chi(m,k+1)a(k) - \chi(m,k)a(k-1))U(Y_k)) \right\| \\
&\leq \left\| \sum_{k=n_0+1}^{m} ((\chi(m,k+1)a(k) - \chi(m,k+1)a(k-1))U(Y_k) \right\| \\
&\quad + \left\| \sum_{k=n_0+1}^{m} ((\chi(m,k+1)a(k-1) - \chi(m,k)a(k-1))U(Y_k) \right\| \\
&\leq \sum_{k=n_0+1}^{m} ((a(k-1) - a(k))\chi(m,k+1)U_{max}) + \sum_{k=n_0+1}^{m} ((\chi(m,k+1) - \chi(m,k))a(k-1)U_{max}) \\
&\leq \sum_{k=n_0+1}^{m} ((a(k-1) - a(k))U_{max}) + \sum_{k=n_0+1}^{m} ((\chi(m,k+1) - \chi(m,k))a(n_0)U_{max}) \\
&= (a(n_0) - a(m))U_{max} + (\chi(m,m+1) - \chi(m,n_0+1))a(n_0)U_{max} \\
&\leq 2a(n_0)U_{max}. \qquad \text{(B.9)}
\end{aligned}$$

The second and third inequalities follow from $a(k-1) - a(k) \geq 0$ because $a(k)$ is a nonincreasing sequence for $k > n_0$, and $\chi(m,k+1) - \chi(m,k)$ is positive because $1 \geq \chi(m,k+1) \geq \chi(m,k)$ for $m,k > n_0$, as $a(k) < 1$ for $k > n_0$. Note that the norm of (B.8c) is directly bounded by $2a(n_0)U_{max}$.

Now, we simplify (B.7b) as follows:

$$\begin{aligned}
&\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(W(Y_k) - \sum_{s' \in \mathcal{S}} (p(s'|Y_k)W(s')\right)x_k \qquad \text{(B.10a)} \\
&= \sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(W(Y_{k+1}) - \sum_{s' \in \mathcal{S}} (p(s'|Y_k)W(s')\right)x_k \\
&\quad + \sum_{k=n_0+1}^{m} (\chi(m,k+1)a(k) - \chi(m,k)a(k-1))W(Y_k)x_k \qquad \text{(B.10b)} \\
&\quad + \sum_{k=n_0+1}^{m} \chi(m,k)a(k-1)W(Y_k)(x_k - x_{k-1}) \qquad \text{(B.10c)} \\
&\quad + \chi(m,n_0+1)a(n_0)W(Y_{n_0})x_{n_0} - \chi(m,m+1)a(m)W(Y_{m+1})x_m. \qquad \text{(B.10d)}
\end{aligned}$$

Similar to the sequence $\tilde{U}_{k+1}$, for (B.10a), define $\tilde{W}_{k+1} = W(Y_{k+1}) - \sum_{s' \in \mathcal{S}} p(s'|Y_k)W(s')$ for $k \geq n_0$, and zero otherwise. Note that $\tilde{W}_{k+1}x_k$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$.

Define $W_{max} := \max_{i \in S} \|W(i)\|$. Note that, here, $\|W(i)\|$ denotes the operator norm of a matrix, that is, $\|W(i)\| = \sup_{x \neq \theta} \frac{\|W(i)x\|}{\|x\|}$, using the Euclidean norm for vectors. Similar to (B.8b), we bound the norm of (B.10b) as follows:

$$\begin{aligned}
&\left\| \sum_{k=n_0+1}^{m} (\chi(m,k+1)a(k) - \chi(m,k)a(k-1))W(Y_k)x_k \right\| \\
&\leq \left\| \sum_{k=n_0+1}^{m} (\chi(m,k+1)a(k) - \chi(m,k)a(k-1))W(Y_k)(x_k - z_k) \right\| \\
&\quad + \left\| \sum_{k=n_0+1}^{m} (\chi(m,k+1)a(k) - \chi(m,k)a(k-1))W(Y_k)z_k \right\| \\
&\leq 2a(n_0)W_{max}(x'_m + \|x_{n_0} - x^*\| + \|x^*\|).
\end{aligned}$$

The last inequality here follows from the definition of $x'_m = \sup_{n_0 \leq k \leq m} \|x_m - z_m\|$ and from the bound on $\|z_n\|$ (5). For (B.10c), let us first bound $\|x_k - x_{k-1}\|$.

$$\begin{aligned}
\|x_k - x_{k-1}\| &= a(k)\|\varphi(Y_{k-1})(r(Y_{k-1}) + \gamma\varphi(Y_k)^T x_{k-1} - \varphi(Y_{k-1})^T x_{k-1})\| \\
&\leq a(k)(K_1 + K_2\|x_{k-1}\|) \\
&\leq a(n_0)(K_1 + K_2(x'_m + \|x_{n_0} - x^*\| + \|x^*\|)),
\end{aligned}$$

for appropriate K_1 and K_2. Before simplifying (B.10c), we first need to repeat an important simplification from our main

proof. Note that for any $0 < k \leq m$,

$$\chi(m,k) + \chi(m,k+1)a(k) = \chi(m,k+1),$$

and hence,

$$\chi(m,n_0) + \sum_{k=n_0}^{m} \chi(m,k+1)a(k) = \chi(m,m+1) = 1.$$

This implies that

$$\sum_{k=n_0}^{m} \chi(m,k+1)a(k) \leq 1.$$

We can finally bound the norm of (B.10c):

$$\begin{aligned}
&\left\| \sum_{k=n_0+1}^{m} \chi(m,k)a(k-1)W(Y_k)(x_k - x_{k-1}) \right\| \\
&\leq \sum_{k=n_0+1}^{m} \chi(m,k)a(k-1)\|W(Y_k)(x_k - x_{k-1})\| \\
&\leq \sum_{k=n_0+1}^{m} \chi(m,k)a(k-1)a(n_0)W_{max}(K_1 + K_2(x'_m + \|x_{n_0} - x^*\| + \|x^*\|)) \\
&\leq a(n_0)W_{max}(K_1 + K_2(x'_m + \|x_{n_0} - x^*\| + \|x^*\|)).
\end{aligned}$$

Finally, the norm of (B.10d) can directly be bounded by

$$\|\chi(m,n_0+1)a(n_0)W(Y_{n_0})x_{n_0} - \chi(m,m+1)a(m)W(Y_{m+1})x_m\| \leq 2a(n_0)W_{max}(x'_m + \|x_{n_0} - x^*\| + \|x^*\|).$$

Combining the bounds above gives us

$$\begin{aligned}
&\sum_{k=n_0}^{m} \chi(m,k+1)a(k)\left(F(x_k,Y_k) - \sum_{s\in\mathcal{S}} \pi(s)F(x_k,s)\right) \\
&= \sum_{k=n_0}^{m} \chi(m,k+1)a(k)(\tilde{U}_{k+1} + \tilde{W}_{k+1}x_k) + \mu_m(n_0),
\end{aligned}$$

where

$$\|\mu_m(n_0)\| \leq 4a(n_0)U_{max} + a(n_0)W_{max}(K_1 + (4+K_2)(x'_m + \|x_{n_0} - x^*\| + \|x^*\|)).$$

Define constants $c_1 := W_{max}(4+K_2)$ and $c_2 := 4U_{max} + K_1W_{max} + c_1\|x^*\|$. This completes the proof of Lemma 2. □

B.3. Proof of Lemma 3

Proof. We first note that for $n_0 \leq k \leq m$, $\|\overline{x}_{k,m}\| \leq \|\overline{x}_{k,m} - \overline{z}_{k,m}\| + \|\overline{z}_{k,m}\| \leq \overline{x}'_{m,m} + \|\overline{z}_{k,m}\|$. The following follow from the definition of $\mathbb{B}_m$. If $\xi \in \mathbb{B}_m$, $\overline{x}'_{m,m}(\xi) = 0$ and if $\xi \notin \mathbb{B}_m$, $\overline{x}'_{m,m}(\xi) = x'_m(\xi) \leq \frac{a(n_0)(c_2+c_1\epsilon)+\delta}{1-\alpha-a(n_0)c_1}$. Hence,

$$\overline{x}'_{m,m} \leq \frac{a(n_0)(c_2 + c_1\epsilon) + \delta}{1 - \alpha - a(n_0)c_1}.$$

Using (5), we have $\|\overline{z}_{k,m}\| \leq \epsilon + \|x^*\|$. Under the condition that $\epsilon \leq 1$ and $\delta \leq 1$, we have

$$\|\overline{x}_{k,m}\| \leq 1 + \|x^*\| + \frac{a(n_0)(c_2 + c_1) + 1}{1 - \alpha - a(n_0)c_1}.$$

Let $v^{(\ell)}$ denote the ℓth component of a vector v. Then,

$$\begin{aligned}
\Gamma_m :&= \left\| \sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}\overline{x}_{k,m} + \tilde{W}_{k+1}\overline{x}_{k,m} + \tilde{U}_{k+1}) \right\| \\
&\leq \sqrt{d} \max_{1\leq\ell\leq d} \left| \sum_{k=n_0}^{m} \chi(m,k+1)a(k)(M_{k+1}\overline{x}_{k,m} + \tilde{W}_{k+1}\overline{x}_{k,m} + \tilde{U}_{k+1})^{(\ell)} \right|.
\end{aligned}$$

Recall that d here is the dimension of our iterates $\{x_n\}$. We apply Theorem A.1 from Appendix A component-wise. For this, first note that

$$(M_{k+1}\overline{x}_{k,m} + \tilde{W}_{k+1}\overline{x}_{k,m} + \tilde{U}_{k+1})^{(\ell)} \le c_3\left(2 + \|x^*\| + \frac{a(n_0)(c_2+c_1)+1}{1-\alpha-a(n_0)c_1}\right),$$

where $c_3 = \max\{M_{max} + 2W_{max}, 2U_{max}\}$. In the theorem statement, let

$$C = \sqrt{d}c_3\left(2 + \|x^*\| + \frac{a(n_0)(c_2+c_1)+1}{1-\alpha-a(n_0)c_1}\right), \zeta_{k,m} = \chi(m,k+1)a(k), \varepsilon = 1, \gamma_1 = 1.$$

Next, we choose suitable γ_2 and $\omega(m)$ such that $\max_{n_0\le k\le m}\zeta_{k,m} \le \gamma_2\omega(m)$. For this, we use our assumption that $\frac{d_1}{n+1} \le a(n) \le d_3\left(\frac{1}{n+1}\right)^{d_2}$, $\forall n \ge n_0$, to obtain

$$\begin{aligned}
\chi(m,k+1) &= \prod_{i=k+1}^{m}(1-a(i)) \le \exp\left(-\sum_{i=k+1}^{m}a(i)\right) \le \exp\left(-\sum_{i=k+1}^{m}\frac{d_1}{i+1}\right) \\
&\le \exp\left(-\int_{k+1}^{m+1}\frac{d_1}{y+1}dy\right) \le \exp(d_1(\log(k+2) - \log(m+2))) \\
&= \left(\frac{k+2}{m+2}\right)^{d_1} \\
\Rightarrow \max_{n_0\le k\le m} a(k)\chi(m,k+1) &\le \max_{n_0\le k\le m} d_3\left(\frac{1}{k}\right)^{d_2}\left(\frac{k+2}{m+2}\right)^{d_1} \le \max_{n_0\le k\le m} d_3\left(\frac{1}{k}\right)^{d_2}\left(\frac{2k}{m+2}\right)^{d_1}.
\end{aligned}$$

From the last inequality, $\gamma_2 = d_3 2^{d_1}$ and $\omega(m) = \beta_{n_0}(m)$ satisfy the required conditions. Then, there exists a constant $D > 0$ such that for $n_0 < m$ and $\delta \in (0, C]$, we have

$$P(\Gamma_m \ge \delta) \le 2de^{-D\delta^2/\beta_{n_0}(m)},$$

and for $\delta > C$,

$$P(\Gamma_m \ge \delta) \le 2de^{-D\delta/\beta_{n_0}(m)}.$$

The factor d comes from the application of union bound to bound the maximum over all components. Under the assumption that $\delta \le 1$, we have that $e^{-D\delta^2/\beta_{n_0}(m)} \ge e^{-D\delta/\beta_{n_0}(m)}$, and hence, $P(\Gamma_m \ge \delta) \le 2de^{-D\delta^2/\beta_{n_0}(m)}$. □

B.4. Proof of Corollary 1

To show Corollary 1, we first obtain values of δ and ϵ such that the probability in Theorem 1 is $1-\varepsilon_1-\varepsilon_2$. We use $\mathfrak{s}_i, i = 1,2,\dots$ to denote different constants throughout this proof. For $a(n) = d_1/(n+1)$ with a sufficiently large d_1, we have $\beta_{n_0}(m) \le 1/m$. This implies that

$$\sum_{m\ge n_0+1}\exp(-D\delta^2/\beta_{n_0}(m)) \le \sum_{m\ge n_0+1}\exp(-D\delta^2 m) \le \mathfrak{s}_1\exp(-D\delta^2 n_0).$$

Let $\varepsilon_1/(2d) = \mathfrak{s}_1\exp(-D\delta^2 n_0)$, which gives us $\delta = \mathfrak{s}_2 n_0^{-1/2}\log^{1/2}(\mathfrak{s}_3/\varepsilon_1)$ for appropriate constants $\mathfrak{s}_2$ and $\mathfrak{s}_3$. This choice of δ gives us

$$2d\sum_{m\ge n_0+1}\exp(-D\delta^2/\beta_{n_0}(m)) \le \varepsilon_1.$$

Let $\epsilon = \sqrt{E[\|x_{n_0} - x^*\|^2]}/\sqrt{\varepsilon_2}$, which implies that

$$P(\|x_{n_0} - x^*\| > \epsilon) = P\left(\|x_{n_0} - x^*\|^2 > \frac{E[\|x_{n_0} - x^*\|^2]}{\varepsilon_2}\right) \le \varepsilon_2.$$

Here, the last inequality follows from Markov's inequality. Being a linear contractive stochastic approximation iteration with an aperiodic irreducible Markov chain, our formulation satisfies the assumptions for Chen et al. (2021, theorem 2.1). To apply their result, we note the corresponding mapping between constants: the norm $\|\cdot\|_c$ is the Euclidean norm in our case, h is one, φ_2 is $1-\alpha$ in our case, and their α is d_1 in our case. For $d_1 > 1/(1-\alpha)$ and n_0 sufficiently large to satisfy the condition for Chen et al. (2021, theorem 2.1(2)), we can use their result (Chen et al. (2021, theorem 2.1(2)(b)(iii)) to obtain the following mean square bound:

$$E[\|x_{n_0} - x^*\|^2] \le \mathfrak{s}_4\left(\frac{1}{n_0+1}\right)^{(1-\alpha)d_1} + \mathfrak{s}_5\frac{\log(n_0+1)}{n_0+1} \le \mathfrak{s}_6\frac{\log(n_0+1)}{n_0+1}.$$

Substituting the values of δ and ϵ in our bound, we get, with probability greater than $1-\varepsilon_1-\varepsilon_2$,

$$\begin{aligned}\|x_m - x^*\| &\le e^{-(1-\alpha)b_{n_0}(m-1)}\frac{\sqrt{E[\|x_{n_0}-x^*\|^2]}}{\sqrt{\varepsilon_2}} \\ &\quad + \mathfrak{s}_7\left(\frac{c_2 d_1}{n_0+1} + \frac{c_1 d_1}{n_0+1}\frac{\sqrt{E[\|x_{n_0}-x^*\|^2]}}{\sqrt{\varepsilon_2}} + \mathfrak{s}_2 n_0^{-1/2}\log^{1/2}(\mathfrak{s}_3/\varepsilon_1)\right) \\ &\le e^{-(1-\alpha)b_{n_0}(m-1)}\frac{\mathfrak{s}_6}{\sqrt{\varepsilon_2}}\sqrt{\frac{\log(n_0+1)}{n_0+1}} \\ &\quad + \mathfrak{s}_7\left(\frac{c_2 d_1}{n_0+1} + \frac{c_1 d_1}{n_0+1}\frac{\mathfrak{s}_6}{\sqrt{\varepsilon_2}}\sqrt{\frac{\log(n_0+1)}{n_0+1}} + \mathfrak{s}_2 n_0^{-1/2}\log^{1/2}(\mathfrak{s}_3/\varepsilon_1)\right)\end{aligned}$$

for all $m \ge n_0$. Now,

$$\begin{aligned}&\exp(-(1-\alpha)b_{n_0}(m-1)) \\ &\le \exp\left(-\sum_{i=n_0}^{m-1}(1-\alpha)a(i)\right) \le \exp\left(-\sum_{i=n_0}^{m-1}\frac{(1-\alpha)d_1}{i+1}\right) \\ &\le \exp\left(-\int_{n_0}^{m}\frac{(1-\alpha)d_1}{y+1}dy\right) \le \exp((1-\alpha)d_1(\log(k+1)-\log(m+1))) \\ &= \left(\frac{n_0+1}{m+1}\right)^{d_1(1-\alpha)} \le \frac{n_0+1}{m+1}.\end{aligned}$$

Here, the final inequality follows from the assumption that $(1-\alpha)d_1 > 1$. Hence, we get that for sufficiently large n_0, the following holds with probability $1-\varepsilon_1-\varepsilon_2$ for all $m \ge n_0$:

$$\|x_m - x^*\| = \mathcal{O}\left(\frac{1}{\sqrt{n_0}}\log^{1/2}\left(\frac{1}{\varepsilon_1}\right) + \sqrt{\frac{\log(n_0)}{n_0}}\frac{1}{\sqrt{\varepsilon_2}}\left(\frac{n_0}{m}+\frac{1}{n_0}\right)\right).$$

Endnote

[1] In Chandak et al. (2022), TD(0) does not satisfy Assumption (6), which is required in the proof of Lemma 1 that shows almost-sure boundedness of the iterates. The authors thank Zaiwei Chen for pointing this out.

References

Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. Precup D, Teh YW, eds. *Proc. 34th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 70 (PMLR, New York), 263–272.

Bhandari J, Russo D, Singal R (2018) A finite time analysis of temporal difference learning with linear function approximation. *Proc. 31st Conf. Learn. Theory* (PMLR, New York), 1691–1692.

Borkar VS (1991) *Topics in Controlled Markov Chains*, Pitman Research Notes in Mathematics Series, vol. 240 (Longman Scientific & Technical, Harlow, Essex, UK).

Borkar VS (2002) On the lock-in probability of stochastic approximation. *Combin. Probab. Comput.* 11(1):11–20.

Borkar VS (2022) Corrigendum to "A concentration bound for contractive stochastic approximation" [Syst. Control Lett. 153 (2021) 104947]. *Systems Control Lett.* 153:104947.

Chandak S, Borkar VS, Dodhia P (2022) Concentration of contractive stochastic approximation and reinforcement learning. *Stochastic Systems* 12(4):411–430.

Chandak S, Borkar VS, Dolhare H (2023) A concentration bound for LSPE(λ). *Systems Control Lett.* 171:105418.

Chen Z, Maguluri ST, Zubeldia M (2025) Concentration of contractive stochastic approximation: Additive and multiplicative noise. *Ann. Appl. Probab.* 35(2):1298–1352.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2020) Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *NIPS'20: Proc. 34th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 8223–8234.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2021) A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. Preprint, submitted February 2, https://arxiv.org/abs/2102.01567.

Chung F, Lu L (2006) Concentration inequalities and martingale inequalities: A survey. *Internet Math.* 3(1):79–127.

Dalal G, Szörényi B, Thoppe G, Mannor S (2018) Finite sample analyses for TD(0) with function approximation. *Proc. AAAI Conf. Artificial Intelligence*, vol. 32 (AAAI Press, Palo Alto, CA).

Even-Dar E, Mansour Y (2003) Learning rates for Q-learning. *J. Machine Learn. Res.* 5:1–25.

Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, eds. *NIPS'18: Proc. 32nd Internat. Conf. Neural Inform. Processing Systems*, vol. 31 (Curran Associates, Inc., Red Hook, NY), 4868–4878.

Kamal S (2010) On the convergence, lock-in probability, and sample complexity of stochastic approximation. *SIAM J. Control Optim.* 48(8):5178–5192.

Li G, Cai C, Chen Y, Wei Y, Chi Y (2023) Is Q-learning minimax optimal? A tight sample complexity analysis. *Oper. Res.* 72(1):222–236.

Liu Q, Watbled F (2009) Exponential inequalities for martingales and asymptotic properties of the free energy of directed polymers in a random environment. *Stochastic Processes Their Appl.* 119(10):3101–3132.

Meerkov SM (1972) Simplified description of slow Markov walks. Part II. *Automation Remote Control* 33(5):761.

Patil G, Prashanth LA, Nagaraj D, Precup D (2023) Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. *Proc. 26th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 206 (PMLR, New York), 5438–5448.

Prashanth LA, Korda N, Munos R (2021) Concentration bounds for temporal difference learning with linear function approximation: The case of batch data and uniform sampling. *Machine Learn.* 110(3):559–618.

Qu G, Wierman A (2020) Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Proc. 33rd Conf. Learn. Theory* (PMLR, New York), 3185–3205.

Srikant R, Ying L (2019) Finite-time error bounds for linear stochastic approximation and TD learning. *Proc. 32nd Conf. Learn. Theory* (PMLR, New York), 2803–2830.

Tao T, Vu V (2015) Random matrices: Universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.* 43(2):782 –874.

Thoppe G, Borkar V (2019) A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems* 9(1):1–26.

Tsitsiklis J, Van Roy B (1997) An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automatic Control* 42(5):674–690.

Yang L, Wang M (2019) Sample-optimal parametric Q-learning using linearly additive features. *Proc. 36th Internat. Conf. Machine Learn.* (PMLR, New York), 6995–7004.

Yang Z, Jin C, Wang Z, Wang M, Jordan M (2020) On function approximation in reinforcement learning: Optimism in the face of large state spaces. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *NIPS'20: Proc. 34th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 13903–13916.