Statistical Inference for Policy Evaluation with Temporal Difference Learning

Weichen Wu, Gen Li, Yuting Wei, Alessandro Rinaldo

Keywords: Temporal Difference Learning, Statistical Inference, Berry-Esseen bound, Central Limit Theorem

Summary

We investigate the statistical properties of Temporal Difference (TD) learning with Polyak-Ruppert averaging, arguably one of the most widely used algorithms in reinforcement learning, for the task of estimating the parameters of the optimal linear approximation to the value function. Assuming independent samples, we make three theoretical contributions that improve upon the current state-of-the-art results: (i) we establish refined high-dimensional Berry-Esseen bounds over the class of convex sets, achieving faster rates than the best known results, and (ii) we propose and analyze a novel, computationally efficient online plug-in estimator of the asymptotic covariance matrix; (iii) we derive sharper high probability convergence guarantees that depend explicitly on the asymptotic variance and hold under weaker conditions than those adopted in the literature. These results enable the construction of confidence regions and simultaneous confidence intervals for the linear parameters of the value function approximation, with guaranteed finite-sample coverage. We demonstrate the applicability of our theoretical findings through numerical experiments.

Contribution(s)

1. We derive a new high-dimensional Berry-Esseen bound for the estimation error over the class of convex sets with an online variance estimator, achieving an $O(T^{-\frac{1}{3}})$ rate with T the number of iterations, which is the fastest in the literature.
   Context: The Gaussian approximation results we obtained deliver the sharpest rate in the literature, improving on recent contributions.
2. We develop a computationally efficient online algorithm to construct an estimator of the asymptotic variance of the TD estimator. We also show that the total variation distance between the computable estimated Gaussian distribution and the asymptotic Gaussian distribution vanishes with high probability at a rate $O(T^{-\frac{1}{3}})$, matching the order of our Berry-Esseen rate.
   Context: Existing estimators of the asymptotic variance suffer from a significantly higher time and space complexity than the online procedure we propose. Thus, our methodology is practically relevant.
3. We construct confidence regions and simultaneous confidence intervals for the linear parameters of the value function approximation, with guaranteed finite-sample coverage.
   Context: To the best of our knowledge, ours is the first procedure for formal statistical inference on the model parameters with a finite sample coverage guarantee.
4. We derive the first high-probability convergence guarantee for the TD estimator with Polyak-Ruppert averaging that matches the asymptotic variance up to logarithm factors, without requiring a projection step to control the iterative trajectory or stepsizes that depend on the number of iterations.
   Context: Prior work established similar convergence guarantees with larger variances and under stronger assumptions on the stepsizes.

Statistical Inference for Policy Evaluation with Temporal Difference Learning

Weichen Wu[1], **Gen Li**[2], **Yuting Wei**[3], **Alessandro Rinaldo**[4]
weichen.wu.1996@gmail.com, genli@cuhk.edu.hk,
ytwei@wharton.upenn.edu, alessandro.rinaldo@austin.utexas.edu

[1]**The Voleon Group, New York, NY, USA**
[2]**Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR.**
[3]**Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA.**
[4]**Department of Statistics and Data Sciences, University of Texas, Austin TX, USA.**

Abstract

We investigate the statistical properties of Temporal Difference (TD) learning with Polyak-Ruppert averaging, arguably one of the most widely used algorithms in reinforcement learning, for the task of estimating the parameters of the optimal linear approximation to the value function. Assuming independent samples, we make three theoretical contributions that improve upon the current state-of-the-art results: (i) we establish refined high-dimensional Berry-Esseen bounds over the class of convex sets, achieving faster rates than the best known results, and (ii) we propose and analyze a novel, computationally efficient online plug-in estimator of the asymptotic covariance matrix; (iii) we derive sharper high probability convergence guarantees that depend explicitly on the asymptotic variance and hold under weaker conditions than those adopted in the literature. These results enable the construction of confidence regions and simultaneous confidence intervals for the linear parameters of the value function approximation, with guaranteed finite-sample coverage. We demonstrate the applicability of our theoretical findings through numerical experiments.

1 Introduction

The feasibility and accuracy of statistical inference tasks, such as hypothesis testing or constructing confidence regions or simultaneous confidence intervals, are traditionally predicated on using estimators that are not only consistent but also satisfy some type of distributional approximation. In most cases, this is accomplished via Central Limit Theorems (CLTs), which ensure that the procedures in use admit a Gaussian approximation. Due to the high dimensionality of modern scientific and engineering applications, there has been a surge of interest in establishing new *high-dimensional Berry-Esseen bounds,* CLTs that hold in high dimensions and have a finite-sample guarantee. While efforts in this area have focused on Berry-Esseen bounds for normalized sums, few results are available to handle more complicated scenarios, where estimators are computed in a nonlinear, iterative manner. In this paper, we investigate one such scenario and study in detail the consistency, Gaussian approximations, and statistical inference for the output of the Temporal Difference (TD) learning algorithm for policy evaluation with linear function approximation in Reinforcement Learning (RL).

A critical task in RL is to evaluate with confidence the quality of a given policy, measured by its value function. As RL models are based on Markov Decision Processes (MDPs), statistical inference must accommodate the online nature of the sampling mechanism. TD learning (see, e.g., Sutton, 1988), an iterative procedure that can be framed as an instance of stochastic approximation (SA), is arguably

the most widely used and effective algorithm for policy evaluation in RL. The last several years have witnessed a flurry of activities within the statistical machine learning community to derive finite sample consistency rates of TD iterates, leading to deep results and a much improved understanding of the performance and limitations of TD learning. See, e.g., Khamaru et al. (2020); Li et al. (2024a); Bhandari et al. (2021); Dalal et al. (2018); Srikant & Ying (2019); Lakshminarayanan & Szepesvari (2018); Mou et al. (2020); Durmus et al. (2022); Li et al. (2024b); Samsonov et al. (2024b); Prashanth et al. (2013); Borkar et al. (2025) for a partial and growing list of relevant references. Though for the tabular case (finite state and action spaces) the TD algorithm is known to be minimax rate optimal (Li et al., 2024a), its properties – and, in particular, its sample complexity – in more complex and high-dimensional settings have not been fully understood.

In a more specialized line of work, concerned with establishing distributional approximations of the TD iterates, it has been shown that, using Polyak-Ruppert averaging (Polyak & Juditsky, 1992; Ruppert, 1998), the TD estimation error converges to a Gaussian distribution in the large sample limit; see, e.g., Fort (2015); Mou et al. (2020); Li et al. (2023). While these results provide a formal justification for the validity of statistical inference, the guarantees from those contributions are non-quantitative, in the sense that they are of asymptotic nature and hold only in fixed dimensions. As a result, they do not adequately capture the dependence on the sample size and other problem-related quantities, most notably the dimensionality, all of which affect the accuracy of statistical inference. In this paper, we partially address this gap. We focus on TD learning with Polyak-Ruppert averaging for estimating the coefficients of the best linear approximation of the value function, assuming i.i.d. samples. We make the following theoretical and methodological contributions.

1. *Sharper Berry-Esseen bounds.* We derive a new high-dimensional Berry-Esseen bound for the estimation error over the class of convex sets and using an estimated variance, achieving an $O(T^{-\frac{1}{3}})$ rate as a function of the number of iterations, which appears to be the fastest in the literature. While completing this work, we became aware of the recent related paper by Samsonov et al. (2024a), which, under the same settings, derived a high-dimensional Berry-Esseen bound for linear SA with Polyak-Ruppert averaging, obtaining the slower rate $O(T^{-\frac{1}{4}})$. See Section 3.1 for a discussion and detailed comparisons.
2. *Improved high-probability guarantees.* We show that, for an appropriate choice of the initial stepsize, the Euclidean norm of the estimation error for the TD learning algorithm is of order

$$O\left(\sqrt{\frac{\mathsf{Tr}(\mathbf{\Lambda}^\star)}{T}\log\frac{1}{\delta}}\right),$$

   with probability at least $1-\delta$, where T denotes the number of iterations, $\mathbf{\Lambda}^\star$ represents the asymptotic variance matrix of the estimation error, and $\mathsf{Tr}(\mathbf{\Lambda}^\star)$ its trace. To the best of our knowledge, this is the first high-probability convergence guarantee for this problem that matches the asymptotic variance, up to logarithm factors, without requiring a projection step to control the iterative trajectory. Furthermore, our assumptions are weaker than those used in the literature. Table 1 compares our high-probability convergence guarantees with the latest available results.
3. *Novel, computationally efficient variance estimator.* We develop a computationally efficient online algorithm to construct an estimator $\widehat{\mathbf{\Lambda}}_T$ of the asymptotic variance $\mathbf{\Lambda}^\star$. We also show that the total variation distance between the (computable) distribution $\mathcal{N}(\mathbf{0}, \widehat{\mathbf{\Lambda}}_T)$ and the asymptotic distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)$ vanishes with high probability at a rate $O(T^{-\frac{1}{3}})$, matching the order of our Berry-Esseen rate. As a result, estimating the variance, a crucial step for constructing confidence sets, does not lead to worse rates.
4. *Statistical inference.* Based on the Berry-Esseen bound and the variance estimator, we design an online, efficient algorithm to construct confidence regions and simultaneous confidence intervals for the linear parameters of the value function approximation. We prove that the finite-sample coverage rate of proposed confidence region converges to the target level at a rate of $O(T^{-\frac{1}{3}})$.

We verify our theoretical findings with numerical simulations, discussed in Section 4 and the appendix. As an illustration, Figure 1 shows how the empirical distribution of TD estimation error in $10,000$

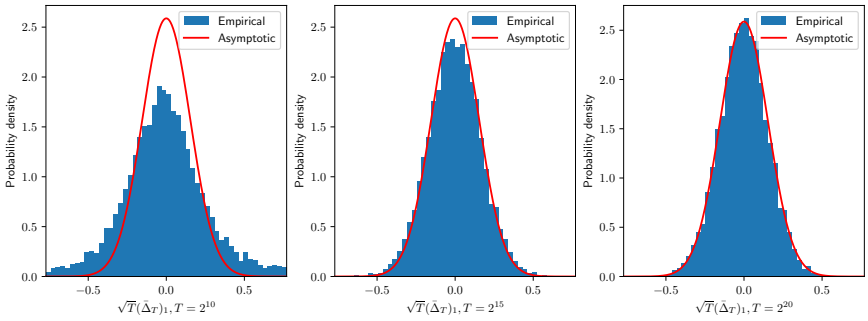

Figure 1: Emergence of the first entry of TD estimation error on an MDP with a 3-dimensional feature map.

independent trials (as shown in blue histograms), scaled by the square-root of the number of iterations T, converges to its asymptotic Gaussian distribution (shown by the red curves).

The i.i.d. condition. Our results are derived under the standard assumption of i.i.d. data, which has been widely used in the recent literature on statistical analyses of TD learning, SA and stochastic gradient descent (see, e.g. Sheshukova et al., 2025). While unrealistic, the theoretical investigation of this simplified scenario has proved challenging; the present work addresses some of the gap still existing in the literature and provide state-of-the art rates. We mention that analogous results, obtained using different techniques and yielding slower rates, have been recently obtained for the more challenging case of Markovian data: see Srikant (2024) and Wu et al. (2025).

Notation. We use boldface small letters to denote vectors and capital letters to denote matrices. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_2$ denotes its Euclidean norm; for a matrix $\boldsymbol{M}$, we use $\mathsf{Tr}(\boldsymbol{M})$, $\det(\boldsymbol{M})$, $\|\boldsymbol{M}\|_{\mathsf{F}}$ and $\|\boldsymbol{M}\|$ to denote its trace, determinant, Frobenius and spectral norm, respectively. For a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we use $\lambda_{\max}(\boldsymbol{M}) = \lambda_1(\boldsymbol{M}) \geq \ldots \lambda_d(\boldsymbol{M}) = \lambda_{\min}(\boldsymbol{M})$ to denote its eigenvalues and $\mathrm{cond}(\boldsymbol{M})$ to indicate its condition number. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ and a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, we let $\|\boldsymbol{x}\|_{\boldsymbol{M}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}$. We will use the notation $\tilde{C}$ to express a quantity independent of T, the number of iterations/sample size, but possibly dependent on other problem-related parameters. Similarly, for positive sequences $\{f_t\}_{t\in\mathbb{N}}$ and $\{g_t\}_{t\in\mathbb{N}}$, we write $f_t \lesssim g_t$, or $f_t = O(g_t)$, to signify that $f_t \leq \tilde{C} g_t$, for some $C > 0$ independent of t. The *convex distance* between two probability distributions on $\mathbb{R}^d$, say P and Q, is defined as

$$d_{\mathsf{C}}(P, Q) = \sup_{\mathcal{A}\in\mathscr{C}_d} |P(\mathcal{A}) - Q(\mathcal{A})|,$$

where $\mathscr{C}_d$ is the class of all convex subsets of $\mathbb{R}^d$. Also, we use $d_{\mathsf{TV}}(P, Q)$ to denote the total variation (TV) distance between P and Q. Notice that $d_{\mathsf{C}}(P, Q) \leq d_{\mathsf{TV}}(P, Q)$.

2 Background

Markov decision process. An infinite-horizon discounted Markov Decision Process (MDP) (Bertsekas, 2017; Sutton & Barto, 2018) is the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ represent the *state space* of the environment and the *action space* available to an agent, respectively, both assumed for simplicity to be finite (though our results hold for countable spaces). The function $P : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is a transition kernel that associates to any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ the probability distribution $P(\cdot \mid s, a)$ of the next state of the environment when the agent executes action a while the environment is in state s. The *reward function* $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ measures the immediate reward to the agent after taking action a at state s; finally, $\gamma \in [0, 1)$ is the *discount factor*, used to assess the impact of future rewards.

In choosing their actions, the agent follows a randomized action selection rule, or *policy*, represented as a (time-invariant) conditional probability distribution $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ over the action space given the state of the environment. For any policy π, its associated reward function $r : \mathcal{S} \mapsto [0, 1]$ is the expected reward of the next action selected according to the policy π, i.e. $r(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s, a)]$, for $s \in \mathcal{S}$. Any policy π induces a transition kernel P, representing the probability distribution of the next state given the current state s and the policy π, given by $P(s' \mid s) = \mathbb{E}_{a\sim\pi(\cdot|s)}P(s' \mid s, a)$ for $s, s' \in \mathcal{S}$. We assume throughout that there exists a unique stationary distribution μ for the transition matrix P; that is, $\mu(i) = \mathbb{P}_{s\sim\mu,s'\sim P(\cdot|s)}(s' = i)$, $i \in \mathcal{S}$.

The value function and the Bellman equation. The worth of a policy π is measured by its *value function* $s \in \mathcal{S} \mapsto V(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$ returning, for each initial state, the expected value of all cumulative future rewards, appropriately discounted. A fundamental, defining feature of the value function is that it arises as the *unique* solution to the *Bellman equation*, a functional fixed-point equation of the form

$$V(s) = r(s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s)}V(s') =: \mathcal{T}V(s). \tag{1}$$

where $\mathcal{T}$ is known as the *Bellman's operator*.

Linear policy approximation and the projected Bellman equation. In most practical scenarios, the value function is very complex and impossible to compute due to the prohibitively large size of the state space $\mathcal{S}$. A common way to overcome this issue is to consider a parametric approximation of V through a linear function, which is both computationally tractable and amenable to statistical estimation. In detail, for a given feature map $\boldsymbol{\phi} : \mathcal{S} \mapsto \mathbb{R}^d$, mapping each state in $\mathcal{S}$ to a centered feature vector in $\mathbb{R}^d$, where d is typically smaller than $|\mathcal{S}|$, a linear value function approximation is given by

$$s \in \mathcal{S} \mapsto V_{\boldsymbol{\theta}}(s) := \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of linear coefficients. Throughout this paper, we assume that the Euclidean norm of the features is uniformly bounded, i.e., $\|\boldsymbol{\phi}(s)\|_2 \leq 1$ for all $s \in \mathcal{S}$. Throughout, we let

$$\boldsymbol{\Sigma} := \mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}^\top(s)] \tag{3}$$

be the feature covariance matrix and use λ_0, λ_Σ to denote its smallest and largest eigenvalues, respectively. We assume $\lambda_0 > 0$.

The best approximating linear function is the one parametrized by the coefficient $\boldsymbol{\theta}^\star$ solving the *projected Bellman equation* (Tsitsiklis & Van Roy, 1997) $V_{\boldsymbol{\theta}^\star} = \Pi\mathcal{T}V_{\boldsymbol{\theta}^\star}$, with Π the orthogonal projection onto the subspace of functions that can be represented by linear combinations of the features. See, e.g., Duan et al. (2021). In turn, this (unique!) solution is given by [1]

$$\boldsymbol{\theta}^\star := \boldsymbol{A}^{-1}\boldsymbol{b}, \tag{4}$$

where

$$\boldsymbol{A} := \mathop{\mathbb{E}}_{s\sim\mu,s'\sim P(\cdot|s)}\left[\boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')\right)^\top\right] \in \mathbb{R}^{d\times d} \tag{5}$$

$$\text{and} \quad \boldsymbol{b} := \mathop{\mathbb{E}}_{s\sim\mu}[\boldsymbol{\phi}(s)r(s)] \in \mathbb{R}^d. \tag{6}$$

The Temporal Difference (TD) learning with Polyak-Ruppert averaging. Although in most real-world RL scenarios the state-action-reward tuples follow a Markov decision process, many of the most recent theoretical contributions – including those delivering the sharpest guaranties listed in Table 1 and Samsonov et al. (2024a) – have made the simplifying but technically very convenient

[1] notice that the matrix $\boldsymbol{A}$ is invertible due to our assumption that $\lambda_0 > 0$; please refer to Lemma 6.4 for details.

assumption of $i.i.d.$ data. Here we follow the same convention and analyze the properties of the TD learning algorithm applied to a sequence of T $i.i.d.$ state-action-reward tuples of the form

$$\{(s_t, a_t, s'_t, r_t)\}_{t=1}^{T}, \tag{7}$$

where $(s_1, \ldots, s_T) \overset{i.i.d.}{\sim} \mu$, the stationary distribution and, for each t, $a_t \sim \pi(s_t), r_t = r(s_t, a_t)$ and $s'_t \sim P(\cdot \mid s_t, a_t)$.

The TD learning algorithm is a stochastic approximation method to find, in an iterative manner, the solution (4). In detail, for a sequence of pre-selected step sizes $\{\eta_t\}_{t\geq 1}$, TD proceeds using the update rule

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t(\boldsymbol{A}_t\boldsymbol{\theta}_{t-1} - \boldsymbol{b}_t), \tag{8}$$

where

$$\boldsymbol{A}_t := \boldsymbol{\phi}(s_t)\left(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s'_t)\right)^\top \quad \text{and} \quad \boldsymbol{b}_t := r_t\boldsymbol{\phi}(s_t). \tag{9}$$

Rather than the final TD iterate $\boldsymbol{\theta}_T$, to estimate $\boldsymbol{\theta}^\star$ it is often preferred to use instead the average of all the iterates

$$\overline{\boldsymbol{\theta}}_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}_t. \tag{10}$$

This estimator is an instance of the Polyak-Ruppert averaging scheme (Polyak & Juditsky, 1992; Ruppert, 1998), a stochastic approximation methodology that provably reduces the variability and accelerates convergence. Furthermore, with a proper choice of the stepsizes, the averaged iterates converge to a Gaussian distribution, thus enabling statistical inference. To date, the best Berry-Esseen bounds rates for stochastic approximation methods and TD learning with Polyak-Ruppert averaging are due to Samsonov et al. (2024a), which in turn are based on general non-linear Berry-Esseen bounds by Shao & Zhang (2022). In this paper, we improve on these results by deriving novel and sharper high-probability and Berry-Esseen bounds for the estimation error of the TD algorithm

$$\overline{\boldsymbol{\Delta}}_T = \overline{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star. \tag{11}$$

We conclude by mentioning that the choice of learning rates $\{\eta_t\}_{t\geq 0}$ in the updates (8) plays a critical role in determining the statistical properties of (both the original and) the averaged TD estimator (10). Here we assume a *polynomial-decaying* stepsize schedule $\eta_t = \eta_0 t^{-\alpha}$ for $\alpha \in (1/2, 1)$.

3 Results

It is known (Polyak & Juditsky, 1992; Fort, 2015) that, in fixed-dimensional settings, the TD estimation error of (11) satisfies

$$\sqrt{T}\,\overline{\boldsymbol{\Delta}}_T \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star), \tag{12}$$

where the asymptotic covariance matrix $\boldsymbol{\Lambda}^\star$ is given by

$$\boldsymbol{\Lambda}^\star = \boldsymbol{A}^{-1}\boldsymbol{\Gamma}\boldsymbol{A}^{-\top}, \quad \text{with} \tag{13}$$

$$\boldsymbol{\Gamma} = \mathbb{E}[(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top]. \tag{14}$$

Thus, the above result implies the familiar parametric rate

$$\|\overline{\boldsymbol{\Delta}}_T\|_2 = O_P\left(\sqrt{\frac{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)}{T}}\right), \tag{15}$$

which may serve as a benchmark. Below, we will extend these guarantees to a high-dimensional scenario by demonstrating an explicit dependence on all model-related parameters including, in

particular, the dimension of the feature space. Our main goal is to show how the TD algorithm can be used to construct non-asymptotic confidence regions and simultaneous confidence intervals for $\boldsymbol{\theta}^\star$ in high dimensions. To that effect, we will establish a high-probability convergence guarantee for $\overline{\boldsymbol{\Delta}}_T$, develop high-dimensional Berry-Esseen bounds over the class of convex sets, and finally build confidence regions for $\boldsymbol{\theta}^\star$ using $\overline{\boldsymbol{\theta}}_T$ and a novel, online estimator of the asymptotic variance.

3.1 High-dimensional Berry-Esseen bounds

We start with a novel high-dimensional Berry-Esseen bounds for the TD learning estimator, certifying that, uniformly over the class of d-dimensional convex sets, $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ is close in distribution to a centered Gaussian random vector with the optimal asymptotic covariance (14), thus closely mimicking the classical, fixed-dimensional CLT guarantee in (12). We reiterate that, to the best of our knowledge, the resulting rates are the sharpest to date.

Theorem 3.1. *Consider the TD algorithm with Polyak-Ruppert averaging (8), independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$ for $\alpha \in (\frac{1}{2}, 1)$ and $\eta_0 < \frac{1}{4}$. Then,*

$$d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\mathsf{Tr}(\boldsymbol{\Gamma})\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{\alpha}{2}} + \frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{(1-\gamma)\lambda_0\eta_0}T^{\alpha-1} + \tilde{C}\, T^{-\frac{1}{2}} + \tilde{C}'\, T^{2\alpha-2}$$

where $\tilde{C}, \tilde{C}'$ are problem-related quantities independent of T.

On the final rate. By choosing $\alpha = 2/3$, we obtain a rate in T of order $O(T^{-1/3})$. In fact, in Theorem 7.3 in the Appendix we establish the more refined result that the Berry-Esseen bound (in the convex distance) between the distribution of $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ and a Gaussian distribution with (approximately) matching first two moments is of order $\sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\mathsf{Tr}(\boldsymbol{\Gamma})\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{\alpha}{2}}$. As $\alpha \to 1$, this rate approaches the optimal rate of $1/\sqrt{T}$. The additional term $\frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{(1-\gamma)\lambda_0\eta_0}T^{\alpha-1}$ in the above bound, which is ultimately the reason for the final rate $O(T^{-1/3})$, originates from replacing an approximation to the finite sample variance of $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ with its asymptotic variance $\boldsymbol{\Lambda}^\star$.

Comparison with Samsonov et al. (2024a). Prior to this work, the best Berry-Esseen bounds for stochastic approximation with polynomial-decaying stepsizes and Polyak-Ruppert averaging are those obtained by Samsonov et al. (2024a), thus covering the settings considered here. Specifically, the authors show in their Theorem 3 that when $\eta_0 \leq \min\{\lambda_0(1-\gamma), \frac{1-\gamma}{(1+\gamma)^2}\}$ and $\alpha \in [\frac{1}{2}, 1)$,

$$d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim O(T^{-\frac{\alpha}{2}} + T^{\frac{\alpha-1}{2}}).$$

Choosing $\alpha = \frac{1}{2}$ leads to a convergence rate in T of order $O(T^{-\frac{1}{4}})$. In contrast, Theorem 3.1 shows that, by selecting $\alpha = \frac{2}{3}$, we obtain the faster rate $O(T^{-\frac{1}{3}})$. As demonstrated in Appendix 7.2.3, this rate is tight for $\alpha \in (\frac{2}{3}, 1)$, since the $O(T^{\alpha-1})$ term is inevitable. See also Figure 2 below, showing the output of a numerical experiment indicating that, in fact, the choice of $\alpha = \frac{2}{3}$ leads to a smaller approximation error. Though worked out independently, the proofs of our Theorem 3.1 and of Theorem 3 of Samsonov et al. (2024a) bear similarities, as they both rely on the general Berry-Esseen bounds for non-linear statistics developed by Shao & Zhang (2022). We obtain sharper rates by a more targeted analysis of the TD settings, while, in comparison, the results of Samsonov et al. (2024a) apply more broadly to SA problems. Specifically we (i) adopt a different decomposition of the TD estimation error, (ii) handle in a more refined way certain terms appearing in the bound of Theorem 3.1 of Shao & Zhang (2022)[2] and (iii) carry out a more specialized analysis of the properties of the matrices $\boldsymbol{A}_t$ and $\boldsymbol{A}$ (see, e.g., Lemma 6.4 in the Appendix) originating from TD learning, which has allowed us to tighten problem-related coefficient, as well as to work with a substantially larger initial

[2] Specifically, the term I_3 in Appendix 7.2.1.

stepsize η_0. This non-trivial sharpening may be potentially generalized to analyze stochastic gradient algorithms applied to non-linear functions, as, e.g., in Theorem 3.4 of Shao & Zhang (2022).

After the appearance of our preprint, Butyrin et al. (2025) obtained analogous Berry–Esseen rates for linear stochastic approximation using similar techniques.

3.2 High-probability bounds

Our next result presents a high-dimensional, high-probability bound on the size of the TD estimation error.

Theorem 3.2. *Consider the TD algorithm with Polyak-Ruppert averaging (8), independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$ for $\alpha \in (\frac{1}{2}, 1)$. For any $\delta \in (0,1)$, there exists $\eta_0 = \eta_0(\delta) > 0$, such that*

$$\|\overline{\boldsymbol{\Delta}}_T\|_2 \lesssim \sqrt{\frac{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)}{T} \log \frac{1}{\delta}} + \tilde{C}(T^{-\frac{\alpha+1}{2}} + T^{\frac{\alpha}{2}-1}) \log \frac{T}{\delta}$$

with probability at least $1-\delta$, where $\tilde{C}$ is a problem-related quantity independent of T and δ.

The proof of this result can be found in Appendix 7.1, where the value of $\tilde{C}$ is given in equation (81). Throughout the paper, for the sake of readability, we have opted to express unwieldy problem-dependent quantities that do not directly depend on T in implicit form; their values can be tracked in the proofs. Accordingly, when writing our finite-sample bounds, we will first present the leading terms in T.

In a nutshell, Theorem 3.2 states that $\overline{\boldsymbol{\theta}}_T$ converges to $\boldsymbol{\theta}^*$ (in Euclidean distance) at the rate of order $O(\sqrt{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)/T})$, up to a log factor. It is worth comparing the high-probability sample complexity bound of Theorem 3.2 with those in the literature. To the best of our knowledge, the tightest results were attained by Dalal et al. (2018) for inverse-decaying stepsizes and Li et al. (2024b); Samsonov et al. (2024b) for constant stepsize as summarized in Table 1. Theorem 3.2 is the first result to achieve tight high-probability convergence guarantees up to log factors using polynomial-decaying stepsizes. Another subtle but consequential difference concerns the restriction on the step size. For our bound to hold, the choice of the initial stepsize η_0 depends on δ – specifically, η_0 is required to be sufficiently small such that $\eta_0 \sqrt{\log 1/(\delta\eta_0)} \leq \tilde{C}$, where $\tilde{C}$ is independent of T and δ but depends on the other parameters of interest (its specific form is shown in (73) and (74) in the Appendix). Though not ideal, this condition is necessary for a Chernoff-type convergence guarantee; we refer to Appendix 7.1.3 for a detailed discussion. Nonetheless, our stepsize requirements are considerably weaker than those imposed by Samsonov et al. (2024b) and Li et al. (2024b), which depend on both δ and T; see Table 1 in Appendix 7.1.4.

As it turns out, the dependence of η_0 on the target probability coverage δ can be removed in the special case of *tabular settings* in which the state space is finite. See Appendix 7.1.4, where we also discuss related (but not applicable to our problem!) high probability bounds for TD-learning in tabular settings by Chen et al. (2024) and Khodadadian & Zubeldia (2025), in which the stepsize is also independent of the choice of δ.

3.3 Online estimation of the covariance

In the last section, we obtained finite sample bounds on the convex distance between $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ and its Gaussian approximation, which features an *unknown* variance. To construct confidence regions for $\boldsymbol{\theta}^\star$, it is necessary to estimate the covariance matrix $\boldsymbol{\Lambda}^\star$. Two types of approaches have recently been proposed: *plug-in estimation* (Chen et al., 2020) and *bootstrapping* (White & White, 2010; Hanna et al., 2017; Hao et al., 2021; Ramprasad et al., 2022; Zhu et al., 2023; Samsonov et al., 2024a).

Here we propose to use a computationally feasible online plug-in estimator for $\boldsymbol{\Lambda}^\star$ with superior computational efficiency to existing methodologies and establish its non-asymptotic rate of convergence. Specifically, we consider the following

$$\overline{\boldsymbol{A}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{A}_t, \quad \text{and} \tag{16}$$

$$\widehat{\boldsymbol{\Gamma}}_T = \sum_{t=1}^{T} w_{t,T}(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)^\top \tag{17}$$

as estimators for $\boldsymbol{A}$ and $\boldsymbol{\Gamma}$ respectively, where the *weights* $\{w_{t,T}\}_{1\le t\le T}$ in $\widehat{\boldsymbol{\Gamma}}_T$ are chosen as

$$w_{t,T} := \frac{t}{\sum_{t=1}^{T} t} = \frac{2t}{T(T+1)},$$

following the intuition that as t grows larger, $\overline{\boldsymbol{\theta}}_t$ is a more accurate estimator of $\boldsymbol{\theta}^\star$ and hence the corresponding term should be assigned higher weights. We then consider

$$\widehat{\boldsymbol{\Lambda}}_T = \overline{\boldsymbol{A}}_T^{-1}\widehat{\boldsymbol{\Gamma}}_T\overline{\boldsymbol{A}}_T^{-\top} \tag{18}$$

as the final estimator of $\boldsymbol{\Lambda}^\star$.

Crucially, the estimator $\widehat{\boldsymbol{\Lambda}}_T$ can be evaluated in an online fashion. See Algorithm 1 in Appendix 9.1 for a detailed description of the procedure, which does *not* require a priori knowledge of the total round of iterations T, and has time and space complexities of order $O(Td^2 + d^3)$ and $O(d^2)$, respectively. The estimator $\widehat{\boldsymbol{\Lambda}}_T$ also enjoys nice statistical properties. This is made precise by the following Gaussian comparison result, showing that the unknown asymptotic variance $\boldsymbol{\Lambda}^\star$ in the Berry-Esseen bound of Theorem 7.4 can be replaced by $\widehat{\boldsymbol{\Lambda}}_T$ without compromising the accuracy of the approximation (i.e., without affecting the rate in T). The proof of this result can be found in Appendix 7.3.

Theorem 3.3. *Consider the TD algorithm* (8) *with independent samples and decaying stepsizes* $\eta_t = \eta_0 t^{-\alpha}$ *for all* $t > 0$*, any* $\alpha \in (1/2, 1)$ *and* $\eta_0 < \frac{1}{4}$ *, and let* $\widehat{\boldsymbol{\Lambda}}_T$*,* $\boldsymbol{\Lambda}^\star$ *be defined as in* (18)*,* (14) *respectively. Then,*

$$d_{\mathsf{TV}}(\mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Lambda}}_T), \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^\star)) \lesssim \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)}(\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{1}{3}} + o(T^{-\frac{1}{3}}).$$

with probability at least $1 - 3T^{-\frac{1}{3}} - \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)\sqrt{T}}$.

Above, the term $o(T^{-\frac{1}{3}})$ depends on all the parameters of the problem (and, in particular, on α) but, as a function of T alone, it is of order $O(T^{-\frac{1}{3}})$. See Appendix 7.3 for the proof and details.

Comparison with Chen et al. (2020). In different settings, Chen et al. (2020) considered the problem of carrying out statistical inference for the true model parameters using stochastic gradient descent and assuming a strongly convex and smooth population loss function. In their work, they consider the slightly different plug-in estimator $\widehat{\boldsymbol{\Lambda}}'_T = \overline{\boldsymbol{A}}_T^{-1}\widehat{\boldsymbol{\Gamma}}'_T\overline{\boldsymbol{A}}_T^{-\top}$, where $\widehat{\boldsymbol{\Gamma}}'_T := \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{A}_t\boldsymbol{\theta}_t - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}_t - \boldsymbol{b}_t)^\top$. The authors proved that the estimation error of $\widehat{\boldsymbol{\Lambda}}'_T$ is bounded in expectation as

$$\mathbb{E}\|\widehat{\boldsymbol{\Lambda}}'_T - \boldsymbol{\Lambda}^\star\| \le \|\boldsymbol{\Gamma}\| \cdot (\widetilde{C}_1 T^{-\frac{\alpha}{2}} + \widetilde{C}_2 T^{-\alpha}), \tag{19}$$

for problem-related constants $\widetilde{C}_1$ and $\widetilde{C}_2$. It is possible to modify the proof of Theorem 3.3 to obtain a guarantee in expectation for our problem as well, albeit in the TV metric. Specifically, we can show that

$$\mathbb{E}\left[d_{\mathsf{TV}}(\mathcal{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Lambda}}_T), \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^\star))\right] \lesssim \widetilde{C}\|\boldsymbol{\Gamma}\|T^{-\frac{1}{2}} \tag{20}$$

where $\widetilde{C}$ depends on $\lambda_0, \gamma, \eta_0, \alpha$ and $\boldsymbol{\theta}^\star$. Comparing (19) with (20), we have sharpened the rate from $O\left(T^{-\frac{\alpha}{2}}\right)$ to $O\left(T^{-\frac{1}{2}}\right)$, *independent of* the choice of α.

Comparison with Samsonov et al. (2024a). Leveraging the multiplier-bootstrap technique proposed by Zhu et al. (2023), Samsonov et al. (2024a) construct a randomized estimator $\overline{\boldsymbol{\theta}}_T^{(b)}$ based on the samples $\mathcal{S}_T := \{s_t, s_t'\}_{t=1}^T$. Their Theorems 4 and 5 yield that

$$d_{\mathsf{C}}(\sqrt{T}(\overline{\boldsymbol{\theta}}_T^{(b)} - \boldsymbol{\theta}^\star)|\mathcal{S}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim \frac{\|\boldsymbol{\Gamma}^{-1}\|}{[\lambda_0(1-\gamma)]^{\frac{3}{2}}}(2\|\boldsymbol{\theta}^\star\|_2 + 1)^2 \frac{\log T}{T^{\frac{1}{4}}} + o(T^{-\frac{1}{4}}) \tag{21}$$

Compared with this bootstrap-based procedure, our methodology is advantageous in the following ways: (i) the upper bound in Theorem 3.3 vanishes, as a function of T, at the rate of $O(T^{-\frac{1}{3}})$, which is faster than the $O(T^{-\frac{1}{4}} \log T)$ rate from the bound (21). (ii) The leading term in the upper bound in Theorem 3.3 entails a smaller problem-related coefficient than that of (21). (iii) Perhaps most importantly, in order to estimate the bootstrap distribution with an $O(T^{-\frac{1}{4}})$ error rate, it is necessary to store the sample trajectory $\mathcal{S}_T$ and generate $O(T^{\frac{1}{2}})$ perturbed trajectories. Consequently, the bootstrap procedure has a time complexity of order $O(T^{\frac{3}{2}}d^2)$ and a space complexity of order $O(T)$. This amounts to a much higher computational burden than our online Algorithm 1.

3.4 Statistical inference

Combining the results from Sections 3.1 and 3.3, we describe a general methodology for constructing convex confidence sets for $\boldsymbol{\theta}^\star$ with guaranteed finite sample coverage. Based on the TD estimator $\overline{\boldsymbol{\theta}}_T$ in (10) and the variance estimator $\widehat{\boldsymbol{\Lambda}}_T$ (18), and given any tolerance level $\delta \in (0,1)$, it is enough to pick any convex set $\mathcal{C}_\delta$ that has probability at least $1-\delta$ under the Gaussian distribution $\mathcal{N}(\overline{\boldsymbol{\theta}}_T, \frac{\widehat{\boldsymbol{\Lambda}}_T}{T})$ with random mean and covariance; that is, satisfying the condition

$$\mathbb{P}_{\boldsymbol{\theta} \sim \mathcal{N}(\overline{\boldsymbol{\theta}}_T, \frac{\widehat{\boldsymbol{\Lambda}}_T}{T})} \left(\boldsymbol{\theta} \in \mathcal{C}_\delta \middle| \overline{\boldsymbol{\theta}}_T, \widehat{\boldsymbol{\Lambda}}_T \right) \geq 1 - \delta. \tag{22}$$

Natural choices of $\mathcal{C}_\delta$ are (i) hyper-rectangles, leading to *simultaneous confidence intervals* of the form

$$\mathcal{C}_\delta = \prod_{j=1}^d [(\overline{\boldsymbol{\theta}}_T)_j - \widehat{R}/\sqrt{T}, (\overline{\boldsymbol{\theta}}_T)_j + \widehat{R}/\sqrt{T}], \tag{23}$$

where $\widehat{R} > 0$ is the (random) quantile of the ℓ_∞ norm of a centered Gaussian random vectors with covariance $(\widehat{\boldsymbol{\Lambda}}_T/T)$, which can be calculated via simulations; and (ii) *ellipsoids* or *balls*, for example

$$\mathcal{C}_\delta = \left\{ \boldsymbol{\theta} : \sqrt{T}\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_T\|^2_{\widehat{\boldsymbol{\Lambda}}_T^{-1}} \leq R \right\}, \tag{24}$$

where R is the $1-\delta$ quantile of the $\chi^2(d)$ distribution.

The next result, whose proof is in Appendix 7.4, shows that the coverage of $\mathcal{C}_\delta$ converges to the nominal level of $1-\delta$ at a rate $O(T^{-1/3})$, as a function of T alone.

Theorem 3.4. *Consider TD with Polyak-Ruppert averaging (8) with independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$, for all $t > 0$ with $\alpha = 2/3$ and $\eta_0 < \frac{1}{4}$. For a given significance level $\delta \in (0,1)$, it holds that*

$$\mathbb{P}(\boldsymbol{\theta}^\star \in \mathcal{C}_\delta) \geq 1 - \delta - \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)}(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{1}{3}} - \frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{2(1-\gamma)\lambda_0\eta_0}T^{-\frac{1}{3}} - o(T^{-\frac{1}{3}}),$$

uniformly over all convex sets $\mathcal{C}_\delta \subset \mathbb{R}^d$ satisfying (22).

We highlight two important features of the above result. First, the coverage guarantee holds *uniformly* over all choices of the convex set $\mathcal{C}_\delta$ meeting condition (22), such as those in (23) and (24). Secondly, the size (i.e., volume) of the confidence sets vanishes at the standard (and optimal) parametric rate of $O(T^{-1/2})$, keeping all the other variables constant.

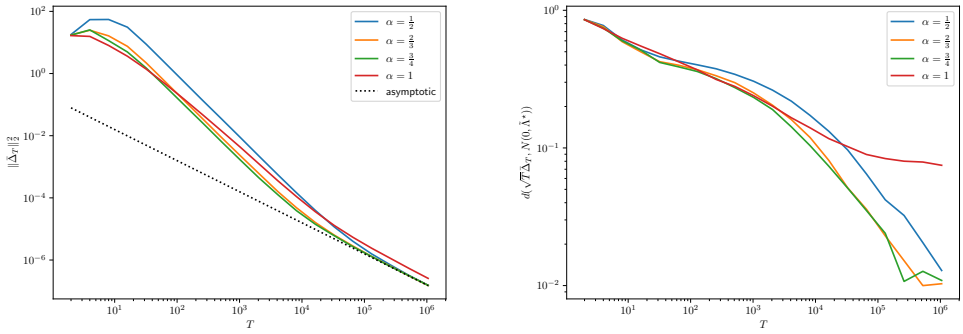

Figure 2: (a) Empirical 95% quantiles of $\|\overline{\boldsymbol{\Delta}}_T\|_2$, and (b) distance between the empirical and asymptotic distributions of $\overline{\boldsymbol{\Delta}}_T$.

4 Simulations

We corroborate our theoretical findings with numerical experiments. Towards that goal, we consider a family of MDPs that are constructed to be difficult to distinguish from each other. The specification of this family of MDPs, along with detailed descriptions of our experiments, is provided in Appendix 9. Here we summarize some of our results. We ran $10,000$ independent trials, each for $T = 2^{20}$ iterations, and recorded data at iterations $2^1, 2^2, 2^3, ..., 2^{19}, 2^{20}$. We report the rate of convergence of $\|\overline{\boldsymbol{\Delta}}_T\|_2^2$, the rate of converge of $\overline{\boldsymbol{\Delta}}_T$ to its asymptotic Gaussian approximation, and the precision of our estimator for the covariance matrix, under different choices of the parameter $\alpha = \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1$. The results of these experiments confirm that the convergence speed is in accordance with our theoretical findings. In particular, for the choice of $\alpha = \frac{2}{3}$, the statistical inference procedure described in Section 3.4 achieves the target coverage rate when T is sufficiently large.

Figure 2(a) illustrates the behavior of $\|\overline{\boldsymbol{\Delta}}_T\|_2^2$ under different choices of α. As a reference, the black dotted line represents the 95% quantile of $\{\boldsymbol{Z}_T\}$ where $\boldsymbol{Z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star/T)$. Our theory predicts that when T grows large, this should be the asymptotic line for the empirical 95% quantiles when $\alpha \in (\frac{1}{2}, 1)$. This is indeed what is shown for $\alpha = \frac{2}{3}$ and $\frac{3}{4}$. When $\alpha = 1$, however, the empirical quantiles do not appear to converge to the asymptotic quantiles. This fact is also consistent with our theory, as reflected in the $O(T^{\frac{\alpha}{2}-1})$ term in the bound of Theorem 3.2, which when $\alpha = 1$, converges at the $O(T^{-\frac{1}{2}})$ rate, just like the leading term. As for the case of $\alpha = \frac{1}{2}$, we observe that when T is small, $\|\overline{\boldsymbol{\Delta}}_T\|_2$ is relatively large and the empirical and asymptotic lines begin to overlap only when T is very large (i.e., $T > 10^5$). This phenomenon is also predicted by the $O(T^{-\frac{\alpha+1}{2}})$ term in the bound of Theorem 3.2, which vanishes at a slower rate when $\alpha = \frac{1}{2}$. Notice that for ease of computation, we use a different measure of the distributional difference than the convex distance, as specified in Eq. (120) in Appendix 9.

Figure 2(b) depicts the speed of convergence of the empirical distribution of the rescaled TD estimation error $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ to its asymptotic distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)$. When $\alpha = 1$, this distance does *not* converge to 0 as T grows large, attesting to our claim that the $O(T^{\alpha-1})$ term in the distributional difference is inevitable. On the other hand, choosing $\alpha = \frac{1}{2}$, as done in Samsonov et al. (2024a), leads to a slower convergence than with $\alpha = \frac{2}{3}$ (our choice) and $\alpha = \frac{3}{4}$.

5 Discussion

We develop high-probability convergence guarantees and Berry-Esseen bounds for TD learning with linear function approximation and $i.i.d.$ samples. We also propose an efficient way to construct a covariance estimator and derive its consistency rates. These results lead to a practicable methodology for constructing convex confidence sets for the target parameter, with its size vanishing at a parametric

rate $O(T^{-1/2})$. An important direction for future research is to investigate whether the order $O(T^{-1/3})$ for the convergence speed of coverage rate in Theorem 3.4 is tight. Another interesting extension would be to derive specialized Berry-Esseen bounds for the sub-class of hyper-rectangles, where one would expect to find a better dependence on the dimension d of the feature space, possibly of poly-logarithmic, as opposed to polynomial, order.

Our analysis and results can be generalized to cover other iterative algorithms, such as stochastic gradient descent and stochastic approximation methods. In Appendix 10, we describe conditions under which our bounds and rates apply to the output of a general linear stochastic approximation algorithm for i.i.d. data, a setting that has received attention in the recent literature (see, e.g., Sheshukova et al., 2025; Samsonov et al., 2024a).

References

Dimitri P Bertsekas. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 2017.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 35(2):936 – 982, 2025.

Bogdan Butyrin, Eric Moulines, Alexey Naumov, Sergey Samsonov, Qi-Man Shao, and Zhuo-Song Zhang. Improved central limit theorem and bootstrap approximations for linear stochastic approximation, 2025. URL https://arxiv.org/abs/2510.12375.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. 2020.

Zaiwei Chen, Siva Theja Maguluri, and Martin Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise, 2024. URL https://arxiv.org/abs/2303.15740.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Yaqi Duan, Mengdi Wang, and Martin J Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for polyak-ruppert averaged iterates of linear stochastic approximation. *arXiv preprint arXiv:2207.04475*, 2022.

Gersende Fort. Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80, 2015.

Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR, 2021.

Olav Kallenberg and Rafal Sztencel. Some dimension-free features of vector-valued martingales. *Probability Theory and Related Fields*, 88(2):215–247, 1991.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

Sajad Khodadadian and Martin Zubeldia. A general-purpose theorem for high-probability bounds of stochastic approximation with polyak averaging, 2025. URL https://arxiv.org/abs/2505.21796.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024a.

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, pp. 1–1, 2024b. DOI: 10.1109/TIT.2024.3394685.

Xiang Li, Jiadong Liang, and Zhihua Zhang. Online statistical inference for nonlinear stochastic approximation with markovian data. *arXiv preprint arXiv:2302.07690*, 2023.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pp. 2947–2997. PMLR, 2020.

Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 24724–24761. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2c15b0221da28bc6f4373a7e78b896dd-Paper-Conference.pdf.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

L. A. Prashanth, Nathaniel Korda, and Rémi Munos. Concentration bounds for temporal difference learning with linear function approximation: the case of batch data and uniform sampling. *Machine Learning*, 110:559 – 618, 2013. URL https://api.semanticscholar.org/CorpusID:210911778.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, pp. 1–14, 2022.

David Ruppert. Efficient estimators from a slowly converging robbins-monro process. Technical report, 1998.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 12408–12460. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1700ad4e6252e8f2955909f96367b34d-Paper-Conference.pdf.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4511–4547. PMLR, 30 Jun–03 Jul 2024b. URL https://proceedings.mlr.press/v247/samsonov24a.html.

Qi-Man Shao and Zhuo-Song Zhang. Berry–esseen bounds for multivariate nonlinear statistics with applications to m-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3): 1548–1576, 2022.

Marina Sheshukova, Sergey Samsonov, Denis Belomestny, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for stochastic gradient descent, 2025. URL https://arxiv.org/abs/2502.06719.

R. Srikant. Rates of convergence in the central limit theorem for markov chains, with an application to td learning, 2024.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, aug 2011. ISSN 1615-3383. DOI: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

John Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Martha White and Adam White. Interval estimation for reinforcement-learning algorithms in continuous-state domains. *Advances in Neural Information Processing Systems*, 23, 2010.

Weichen Wu, Yuting Wei, and Alessandro Rinaldo. Uncertainty quantification for markov chains with application to temporal difference learning, 2025.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

6 Preliminaries

In this section, we summarize some preliminary facts that would be useful in the proof of the theoretical results in this paper.

6.1 Useful algebraic facts

Supplementary notation. For any complex number $z \in \mathbb{C}$, we use $\mathsf{Re}(z)$ to denote its real part. If a matrix $\boldsymbol{M}$ is positive (semi-)definite, i.e. $\lambda_{\min}(\boldsymbol{M}) > (\geq) 0$, we write $\boldsymbol{M} \succ (\succeq) \mathbf{0}$; when two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy $\boldsymbol{X} - \boldsymbol{Y} \succ (\succeq) \mathbf{0}$, we write $\boldsymbol{X} \succ (\succeq) \boldsymbol{Y}$ or equivalently $\boldsymbol{Y} \prec (\preceq) \boldsymbol{X}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we use $\{x_i\}_{1 \leq i \leq d}$ to denote its entries; for a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, we use $\{X_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq n}$ to denote its entries. The *vectorization* of matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, denoted as $\mathbf{vec}(\boldsymbol{X})$, is defined as a vector in $\mathbb{R}^{mn}$ with entries $[\mathbf{vec}(\boldsymbol{X})]_{(i-1)m+j} = \boldsymbol{X}_{ij}$ for every $1 \leq i \leq m$ and $1 \leq j \leq n$.

For a matrix sequence $\{\boldsymbol{F}_t\}_{t \in \mathbb{N}}$, a number sequence $\{f_t\}_{t \in \mathbb{N}}$ and a fixed matrix $\boldsymbol{X}$, if there exists a constant $\widetilde{C}$ independent of t and $\boldsymbol{X}$ such that

$$\|\boldsymbol{F}_t\| \leq \widetilde{C} f_t \|\boldsymbol{X}\|, \quad \text{and} \quad \mathsf{Tr}(\boldsymbol{F}_t) \leq \widetilde{C} f_t \mathsf{Tr}(\boldsymbol{X}) \tag{25}$$

hold simultaneously, we denote for simplicity $\boldsymbol{F}_t = O(f_t)\boldsymbol{X}$.

The following lemma captures some basic properties in linear algebra regarding the relationship between Frobenius and operator norms of matrices.

Lemma 6.1. *The following properties hold true:*

1. For any sequence of matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{d \times d}$:

$$\left\| \prod_{i=1}^{n} \boldsymbol{X}_i \right\|_{\mathsf{F}} \leq \|\boldsymbol{X}_j\|_{\mathsf{F}} \prod_{\substack{1 \leq i \leq n \\ i \neq j}} \|\boldsymbol{X}_i\|. \tag{26}$$

2. For any matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d}$,

$$\|\boldsymbol{X}\|_{\mathsf{F}} \leq \sqrt{d} \|\boldsymbol{X}\|. \tag{27}$$

Lemma 6.2. *For any matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ and a symmetric matrix $\boldsymbol{Y} \in \mathbb{S}^{d \times d}$, if $\boldsymbol{Y} \succeq \mathbf{0}$, then*

$$\mathsf{Tr}(\boldsymbol{X}\boldsymbol{Y}) = \mathsf{Tr}(\boldsymbol{Y}\boldsymbol{X}) \leq \|\boldsymbol{X}\| \mathsf{Tr}(\boldsymbol{Y}).$$

Lemma 6.2 directly imply that for any matrix sequence $\{\boldsymbol{F}_t\}_{t \in \mathbb{N}}$ on $\mathbb{R}^{d \times d}$ and a fixed matrix $\boldsymbol{X} \succeq \mathbf{0}$, it can be guaranteed that

$$\boldsymbol{F}_t \boldsymbol{X} = O(\|\boldsymbol{F}_t\|)\boldsymbol{X}, \quad \text{and} \quad \boldsymbol{X}\boldsymbol{F}_t = O(\|\boldsymbol{F}_t\|)\boldsymbol{X}. \tag{28}$$

The following lemma, due to Chen et al. (2020)[Lemma 4.1], bounds the difference between the inverse matrices and would be useful for our analysis of the plug-in variance estimator.

Lemma 6.3. *Let $\boldsymbol{X}, \boldsymbol{Y}$ be two invertible matrices that satisfy $\|\boldsymbol{X}^{-1}(\boldsymbol{Y} - \boldsymbol{X})\| \leq \frac{1}{2}$. Then the difference between $\boldsymbol{Y}^{-1}$ and $\boldsymbol{X}^{-1}$ can be bounded by*

$$\left\|\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1}\right\| \leq 2\|\boldsymbol{Y} - \boldsymbol{X}\|\|\boldsymbol{X}^{-2}\|.$$

The following lemma reveals several critical algebraic features of the matrix $\boldsymbol{A}$.

Lemma 6.4. *Let $\boldsymbol{A}_t, \boldsymbol{A}$ be defined as in (9) and (5) respectively, and $\boldsymbol{\Sigma}$, λ_0 and λ_Σ be defined as in (3).Then the following features hold:*

$$2(1-\gamma)\boldsymbol{\Sigma} \preceq \boldsymbol{A} + \boldsymbol{A}^\top \preceq 2(1+\gamma)\boldsymbol{\Sigma}; \tag{29}$$

$$\min_{1\le i\le d} \mathsf{Re}(\lambda_i(\boldsymbol{A})) \ge (1-\gamma)\lambda_0; \tag{30}$$

$$\mathbb{E}[\boldsymbol{A}_t^\top \boldsymbol{A}_t] \preceq \boldsymbol{A} + \boldsymbol{A}^\top; \tag{31}$$

$$\boldsymbol{A}^\top \boldsymbol{A} \preceq \lambda_\Sigma(\boldsymbol{A} + \boldsymbol{A}^\top); \tag{32}$$

$$\|\boldsymbol{I} - \eta \boldsymbol{A}\| \le 1 - \frac{1-\gamma}{2}\lambda_0\eta, \quad \forall \eta \in \left(0, \frac{1}{2\lambda_\Sigma}\right); \tag{33}$$

$$\|\boldsymbol{A}^{-1}\| \le \frac{1}{\lambda_0(1-\gamma)}. \tag{34}$$

Proof: We address these properties in order.

Proof of Equation (29). By definition, $\boldsymbol{A} + \boldsymbol{A}^\top$ is featured by

$$\begin{aligned}\boldsymbol{A} + \boldsymbol{A}^\top &= \mathbb{E}_{s\sim\mu, s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s'))^\top + (\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s'))\boldsymbol{\phi}(s)^\top] \\ &= 2\mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top] - \gamma\mathbb{E}_{s\sim\mu, s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top + \boldsymbol{\phi}(s')\boldsymbol{\phi}(s)^\top].\end{aligned}$$

By Cauchy-Schwartz inequality, the second term is bounded by

$$-\left(\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top + \boldsymbol{\phi}(s')\boldsymbol{\phi}(s')^\top\right) \preceq \boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top + \boldsymbol{\phi}(s')\boldsymbol{\phi}(s)^\top \preceq \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top + \boldsymbol{\phi}(s')\boldsymbol{\phi}(s')^\top;$$

Hence, $\boldsymbol{A} + \boldsymbol{A}^\top$ is bounded above by

$$\begin{aligned}\boldsymbol{A} + \boldsymbol{A}^\top &\preceq 2\mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top] + \gamma\mathbb{E}_{s\sim\mu, s'\sim P(\cdot|s)}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top + \boldsymbol{\phi}(s')\boldsymbol{\phi}(s')^\top\right] \\ &= 2\mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top] + \gamma\left(\mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top] + \mathbb{E}_{s'\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top]\right) \\ &= (2+2\gamma)\mathbb{E}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top] = 2(1+\gamma)\boldsymbol{\Sigma},\end{aligned}$$

and similarly below by

$$\boldsymbol{A} + \boldsymbol{A}^\top \succeq 2(1-\gamma)\boldsymbol{\Sigma}.$$

Notice we applied the fact that when $s \sim \mu$ and $s' \sim P(\cdot \mid s)$, the marginal distribution of s is also μ.

Proof of Equation (30). This result is a direct corollary of (29). Specifically, let $z = \lambda + i\mu \in \mathbb{C}$ denote any eigenvalue of $\boldsymbol{A}$, where $\lambda, \mu \in \mathbb{R}$; and $\boldsymbol{x} + i\boldsymbol{y}$ denote the corresponding eigenvector in $\mathbb{C}^d$, where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ satisfy $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$. It then follows that

$$\boldsymbol{A}(\boldsymbol{x} + i\boldsymbol{y}) = (\lambda + i\mu)(\boldsymbol{x} + i\boldsymbol{y}) = (\lambda\boldsymbol{x} - \mu\boldsymbol{y}) + i(\mu\boldsymbol{x} + \lambda\boldsymbol{y}).$$

Comparing the real and imaginary parts of the left-most and right-most terms, we obtain

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} - \mu\boldsymbol{y}, \quad \text{and} \quad \boldsymbol{A}\boldsymbol{y} = \mu\boldsymbol{x} + \lambda\boldsymbol{y}.$$

Direct calculation yields

$$\begin{aligned}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} &= \lambda\|\boldsymbol{x}\|_2^2 - \mu\boldsymbol{x}^\top\boldsymbol{y} = \lambda - \mu\boldsymbol{x}^\top\boldsymbol{y}, \quad \text{and} \\ \boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{y} &= \mu\boldsymbol{y}^\top\boldsymbol{x} + \lambda\|\boldsymbol{y}\|_2^2 = \mu\boldsymbol{y}^\top\boldsymbol{x} + \lambda.\end{aligned}$$

Hence, Equation (29) implies

$$\begin{aligned}\lambda &= \frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y}) \\ &= \frac{1}{4}\boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{x} + \frac{1}{4}\boldsymbol{y}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{y} \\ &\geq \frac{1}{4}\boldsymbol{x}^\top(1-\gamma)\boldsymbol{\Sigma}\boldsymbol{x} + \frac{1}{4}\boldsymbol{y}^\top(1-\gamma)\boldsymbol{\Sigma}\boldsymbol{y} \\ &\geq \frac{1}{4}(2(1-\gamma)\lambda_0 + 2(1-\gamma)\lambda_0) = (1-\gamma)\lambda_0.\end{aligned}$$

This completes the proof of (30).

Proof of Equation (31). The left-hand-side of (31) can be bounded by

$$\begin{aligned}\mathbb{E}[\boldsymbol{A}_t^\top \boldsymbol{A}_t] &= \mathbb{E}[(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))\boldsymbol{\phi}(s_t)^\top \boldsymbol{\phi}(s_t)(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))^\top] \\ &= \mathbb{E}\|\boldsymbol{\phi}(s_t)\|_2^2[(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))^\top] \\ &\preceq \max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2^2 \mathbb{E}[(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t'))^\top] \\ &\preceq \mathbb{E}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top] - \gamma\mathbb{E}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top + \boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top] + \gamma^2\mathbb{E}[\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t')^\top] \\ &\preceq 2\mathbb{E}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top] - \gamma\mathbb{E}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top + \boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top] \\ &= \boldsymbol{A} + \boldsymbol{A}^\top,\end{aligned}$$

Notice that the inequality on the fourth line follows from the assumption that $\max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2 \leq 1$ and the inequality on the fifth line follows from the fact that since s_t and s_t' and both drawn from the stationary distribution, $\mathbb{E}[\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t')^\top] = \mathbb{E}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top]$.

Proof of Equation (32). It suffices to show that for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} \leq \lambda_\Sigma \boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{x}.$$

By definition, the left hand side of this inequality is featured by

$$\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 = \left\|\mathbb{E}[\boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s'))^\top \boldsymbol{x}]\right\|_2^2,$$

where the expectation is taken over the distribution $s \sim \mu, s' \sim P(\cdot \mid s)$. For simplicity, denote

$$f(s) = \mathbb{E}_{s'\sim P(\cdot|s)}[(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s'))^\top \boldsymbol{x}];$$

then the law of total expectations yields

$$\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \|\mathbb{E}[f(s)\boldsymbol{\phi}(s)]\|_2^2.$$

In what follows, we aim to show that

$$\|\mathbb{E}[f(s)\boldsymbol{\phi}(s)]\|_2^2 \leq \lambda_\Sigma \mathbb{E}[f^2(s)]. \tag{35}$$

Observe that for any vector $\boldsymbol{y} \in \mathbb{R}^d$, it can always be guaranteed that

$$\mathbb{E}[f(s) - \boldsymbol{\phi}^\top(s)\boldsymbol{y}]^2 \geq 0.$$

A direct expansion yields

$$\begin{aligned}0 &\leq \mathbb{E}[f^2(s)] - 2\mathbb{E}[f(s)\boldsymbol{\phi}^\top(s)\boldsymbol{y}] + \boldsymbol{y}^\top\mathbb{E}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top]\boldsymbol{y} \\ &= \mathbb{E}[f^2(s)] - 2\mathbb{E}[f(s)\boldsymbol{\phi}^\top(s)\boldsymbol{y}] + \boldsymbol{y}^\top\boldsymbol{\Sigma}\boldsymbol{y} \\ &\leq \mathbb{E}[f^2(s)] - 2\mathbb{E}[f(s)\boldsymbol{\phi}^\top(s)\boldsymbol{y}] + \lambda_\Sigma\|\boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{y} \in \mathbb{R}^d.\end{aligned}$$

By taking

$$y = \frac{\mathbb{E}[f(s)\boldsymbol{\phi}(s)]}{\lambda_\Sigma},$$

we obtain

$$0 \leq \mathbb{E}[f^2(s)] - \frac{\|\mathbb{E}[f(s)\boldsymbol{\phi}(s)]\|_2^2}{\lambda_\Sigma},$$

and (35) follows immediately. In other words, it can be guaranteed for every $\boldsymbol{x} \in \mathbb{R}^d$ that

$$\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} \leq \lambda_\Sigma \mathbb{E}[f^2(s)].$$

and we now proceed to bounding $\mathbb{E}[f^2(s)]$. The Jensen's inequality implies

$$\begin{aligned}
\mathbb{E}[f^2(s)] &= \mathbb{E}_{s\sim\mu}\left\{\boldsymbol{x}^\top \mathbb{E}_{s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')]\mathbb{E}_{s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')]^\top \boldsymbol{x}\right\} \\
&\leq \mathbb{E}_{s\sim\mu}\left\{\boldsymbol{x}^\top \mathbb{E}_{s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')][\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')]^\top \boldsymbol{x}\right\} \\
&= \boldsymbol{x}^\top \mathbb{E}\left\{[\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')][\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')]^\top\right\}\boldsymbol{x} \\
&\leq \boldsymbol{x}^\top(\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x},
\end{aligned}$$

where the last inequality follows from the same logic as in the proof of Equation (31). Equation (32) follows immediately.

Proof of Equation (33). We firstly notice that

$$(\boldsymbol{I} - \eta\boldsymbol{A})^\top(\boldsymbol{I} - \eta\boldsymbol{A}) = \boldsymbol{I} - \eta(\boldsymbol{A} + \boldsymbol{A}^\top) + \eta^2\boldsymbol{A}^\top\boldsymbol{A}.$$

By invoking (32), we obtain

$$(\boldsymbol{I} - \eta\boldsymbol{A})^\top(\boldsymbol{I} - \eta\boldsymbol{A}) \preceq \boldsymbol{I} - (\eta - \lambda_\Sigma\eta^2)(\boldsymbol{A} + \boldsymbol{A}^\top).$$

For any $\eta \in (0, \frac{1}{2\lambda_\Sigma})$, this upper bound is controlled by

$$(\boldsymbol{I} - \eta\boldsymbol{A})^\top(\boldsymbol{I} - \eta\boldsymbol{A}) \preceq \boldsymbol{I} - \frac{\eta}{2}(\boldsymbol{A} + \boldsymbol{A}^\top) \preceq (1 - (1-\gamma)\lambda_0\eta)\boldsymbol{I}.$$

Hence, the spectral norm of $\boldsymbol{I} - \eta\boldsymbol{A}$ is controlled by

$$\|\boldsymbol{I} - \eta\boldsymbol{A}\| \leq \sqrt{1 - (1-\gamma)\lambda_0\eta} \leq 1 - \frac{1-\gamma}{2}\lambda_0\eta.$$

Proof of Equation (34). For any unit vector $\boldsymbol{x} \in \mathbb{R}^d$, we observe

$$\begin{aligned}
\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} &= \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}[\boldsymbol{\phi}(s)\boldsymbol{\phi}^\top(s)]\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} - \gamma\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}[\boldsymbol{\phi}(s)\boldsymbol{\phi}^\top(s')]\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} \\
&= \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} - \gamma\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}[\boldsymbol{\phi}(s)\boldsymbol{\phi}^\top(s')]\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} \\
&= 1 - \gamma\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s))(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s'))] \\
&\leq 1 - \gamma\sqrt{\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s))^2]\mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s'))^2]} \\
&= 1 - \gamma\sqrt{(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}[\boldsymbol{\phi}(s)\boldsymbol{\phi}^\top(s)]\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}) \cdot (\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}[\boldsymbol{\phi}(s')\boldsymbol{\phi}^\top(s')]\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x})} \\
&= 1 - \gamma\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} = 1 - \gamma,
\end{aligned}$$

where the fourth line is an application of the Cauchy-Schwartz inequality. Consequently, for any unit vector $\boldsymbol{x} \in \mathbb{R}^d$, it can be guaranteed that

$$\begin{aligned}
\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\|_2 &= \|\boldsymbol{x}\|_2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\|_2 \\
&\geq \boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} \geq 1 - \gamma,
\end{aligned}$$

where the inequality, again, is an application of the Cauchy-Schwartz inequality. This essentially means that $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\| \leq \frac{1}{1-\gamma}$, and therefore

$$\|\boldsymbol{A}^{-1}\| \leq \frac{1}{1-\gamma}\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})} = \frac{1}{\lambda_0(1-\gamma)}.$$

□

Lemma 6.4 implies the following property regarding a Lyapunov equation.

Lemma 6.5. *For any matrix $\boldsymbol{E} \succeq \boldsymbol{0}$, there exists a unique positive definite matrix $\boldsymbol{X} \succeq \boldsymbol{0}$, such that*

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}^\top = \boldsymbol{E}.$$

Furthermore, it can be guaranteed that

$$\|\boldsymbol{X}\| \leq \frac{1}{2(1-\gamma)\lambda_0}\|\boldsymbol{E}\|, \quad \textit{and} \quad \mathsf{Tr}(\boldsymbol{X}) \leq \frac{1}{2(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{E}).$$

Proof: The existence of $\boldsymbol{X}$ follows from the basic property of Lyapunov equation and (30); we focus on bounding the norm and trace of $\boldsymbol{X}$.

Bounding the operator norm of $\boldsymbol{X}$. Let $\lambda = \lambda_{\max}(\boldsymbol{X})$ and $\boldsymbol{x}$ denote a corresponding eigen vector with $\|\boldsymbol{x}\|_2 = 1$. By definition, $\boldsymbol{X}\boldsymbol{x} = \lambda\boldsymbol{x}$, and therefore

$$\|\boldsymbol{E}\| \geq \boldsymbol{x}^\top\boldsymbol{E}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}^\top)\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x} \geq [2(1-\gamma)\lambda_0]\cdot\lambda,$$

where the last inequality follows from (29). The bound on $\|\boldsymbol{X}\|$ follows naturally.

Bounding the norm of $\boldsymbol{X}$. Since $\boldsymbol{X} \succeq \boldsymbol{0}$, direct calculation yields

$$\begin{aligned}
\mathsf{Tr}(\boldsymbol{E}) &= \mathsf{Tr}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}^\top) \\
&= \mathsf{Tr}(\boldsymbol{X}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{X}^{\frac{1}{2}} + \boldsymbol{X}^{\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{X}^{\frac{1}{2}}) \\
&= \mathsf{Tr}(\boldsymbol{X}^{\frac{1}{2}}(\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{X}^{\frac{1}{2}}) \\
&\geq \mathsf{Tr}(\boldsymbol{X}^{\frac{1}{2}}\cdot 2(1-\gamma)\lambda_0\cdot\boldsymbol{X}^{\frac{1}{2}}) = 2(1-\gamma)\lambda_0\mathsf{Tr}(\boldsymbol{X}),
\end{aligned}$$

where we invoked the cyclability of trace and (29) through the process. □

The following lemma further features the compressive power of the matrix $\prod_{k=1}^{t}(\boldsymbol{I} - \eta_k\boldsymbol{A})$.

Lemma 6.6. *Assume that $\beta \in (0,1)$ and $\alpha \in (\frac{1}{2}, 1)$. Let $\eta_t = \eta_0 t^{-\alpha}$ for all positive integer t, then the following inequalities hold:*

$$t^\alpha\prod_{k=1}^{t}\left(1-\beta k^{-\alpha}\right) \leq e\left(\frac{\alpha}{\beta}\right)^{\frac{\alpha}{1-\alpha}}; \tag{36}$$

$$\sum_{i=1}^{t} i^{-\nu}\prod_{k=i+1}^{t}\left(1-\beta k^{-\alpha}\right) \leq \frac{1}{\nu-1}\left(\frac{2(\nu-\alpha)}{\beta}\right)^{\frac{\nu-\alpha}{1-\alpha}} t^{\alpha-\nu}, \quad \forall\nu \in (1, \alpha+1]; \tag{37}$$

$$\sum_{i=1}^{t} i^{-2\alpha}\prod_{k=i+1}^{t}\left(1-\beta k^{-\alpha}\right) = \frac{1}{\beta}t^{-\alpha} + 5\left(\frac{2}{\beta}\right)^{\frac{2-\alpha}{1-\alpha}} t^{-1}; \tag{38}$$

$$\max_{1\leq i\leq t} i^{-\alpha}\prod_{k=i+1}^{t}\left(1-\beta k^{-\alpha}\right) \leq e\left(\frac{\alpha}{\beta}\right)^{\frac{\alpha}{1-\alpha}} t^{-\alpha}. \tag{39}$$

Proof: We prove these four inequalities in order.

Proof of Equation (36). We firstly observe that

$$t^{\alpha}\prod_{k=1}^{t}\left(1-\beta k^{-\alpha}\right)<t^{\alpha}\exp\left(-\beta\sum_{k=1}^{t}k^{-\alpha}\right),$$

in which

$$\sum_{k=1}^{t}k^{-\alpha}>\int_{1}^{t}x^{-\alpha}\mathrm{d}x=\frac{t^{1-\alpha}-1}{1-\alpha}.$$

Hence, the left hand side of (36) is upper bounded by

$$\begin{aligned}t^{\alpha}\prod_{k=1}^{t}\left(1-\beta k^{-\alpha}\right)&<t^{\alpha}\exp\left(-\beta\cdot\frac{t^{1-\alpha}-1}{1-\alpha}\right)\\&=e^{\frac{\beta}{1-\alpha}}\cdot t^{\alpha}\exp\left(-\frac{\beta}{1-\alpha}t^{1-\alpha}\right).\end{aligned}$$

By letting $u=t^{1-\alpha}$, we obtain

$$t^{\alpha}\prod_{k=1}^{t}\left(1-\beta k^{-\alpha}\right)<e^{\frac{\beta}{1-\alpha}}\cdot u^{\frac{\alpha}{1-\alpha}}\exp\left(-\frac{\beta}{1-\alpha}u\right),$$

where the right hand side is a function of u. It is easy to verify that the maximum of this function is reached when $u=\alpha/\beta$, and hence

$$t^{\alpha}\prod_{k=1}^{t}\left(1-\beta k^{-\alpha}\right)<e^{\frac{\beta}{1-\alpha}}\cdot\left(\frac{\alpha}{\beta}\right)^{\frac{\alpha}{1-\alpha}}\exp\left(-\frac{\alpha}{1-\alpha}\right).$$

The upper bound (36) follows naturally.

Proof of Equation (37). In order to derive an iterative relationship with respect to the left-hand-side of (37), we denote

$$X_t=\sum_{i=1}^{t}i^{-\nu}\prod_{k=i+1}^{t}\left(1-\beta k^{-\alpha}\right)$$

for any positive integer t. Then by definition, $X_1=1$ and the sequence $\{X_t\}_{t\geq 1}$ is featured by the iterative relation

$$X_{t+1}=\left(1-\beta(t+1)^{-\alpha}\right)X_t+(t+1)^{-\nu}. \tag{40}$$

For simplicity, we denote

$$C=\frac{1}{\nu-1}\left(\frac{2(\nu-\alpha)}{\beta}\right)^{\frac{\nu-\alpha}{1-\alpha}}.$$

then the lemma is equivalent to the claim that

$$t^{\nu-\alpha}X_t\leq C \tag{41}$$

holds for all positive integer t, and can hence be proved by an induction argument. On one hand, we observe that the iterative relation (40) implies

$$X_{t+1}<X_t+(t+1)^{-\nu},$$

hence X_t can be bounded by

$$X_t < \sum_{i=1}^{t} i^{-\nu} < \int_1^{\infty} x^{-\nu} \mathrm{d}x = \frac{1}{\nu - 1}.$$

Therefore, we can take

$$t^\star = \left(\frac{2(\nu - \alpha)}{\beta} \right)^{\frac{1}{1-\alpha}},$$

then for any $t \leq t^\star$, it can be guaranteed that

$$t^{\nu-\alpha} X_t < (t^\star)^{\nu-\alpha} \cdot \frac{1}{\nu - 1} = C.$$

On the other hand, assume that (41) holds for an integer $t = k \geq t^\star$, then for $t = k + 1$, we observe

$$\begin{aligned}
&t^{\nu-\alpha} X_t \\
&= (k+1)^{\nu-\alpha} \left[(1 - \beta(k+1)^{-\alpha}) X_k + (k+1)^{-\alpha} \right] \\
&\leq (k+1)^{\nu-\alpha} \cdot C k^{\alpha-\nu} - \beta C k^{\alpha-\nu} (k+1)^{\nu-2\alpha} + (k+1)^{-\alpha} \\
&= C + C \left[\left(\frac{k+1}{k} \right)^{\nu-\alpha} - 1 \right] - \beta C (k+1)^{-\alpha} \left(\frac{k+1}{k} \right)^{\nu-\alpha} + (k+1)^{-\alpha} \\
&= C + C \left(\frac{k+1}{k} \right)^{\nu-\alpha} \left[1 - \left(\frac{k}{k+1} \right)^{\nu-\alpha} \right] \\
&- \beta C (k+1)^{-\alpha} \left(\frac{k+1}{k} \right)^{\nu-\alpha} + (k+1)^{-\alpha} \\
&\overset{(i)}{<} C + C \left(\frac{k+1}{k} \right)^{\nu-\alpha} \cdot \frac{\nu - \alpha}{k+1} - \beta C (k+1)^{-\alpha} \left(\frac{k+1}{k} \right)^{\nu-\alpha} + (k+1)^{-\alpha} \\
&\overset{(ii)}{<} C - \frac{\beta}{2} C (k+1)^{-\alpha} \left(\frac{k+1}{k} \right)^{\nu-\alpha} + (k+1)^{-\alpha} \\
&\overset{(iii)}{<} C - (k+1)^{-\alpha} \left(\frac{k+1}{k} \right)^{\nu-\alpha} + (k+1)^{-\alpha} < C.
\end{aligned}$$

Throughout the process, we applied the following facts:

$$\begin{aligned}
&(i): \left(\frac{k}{k+1} \right)^{\nu-\alpha} = \left(1 - \frac{1}{k+1} \right)^{\nu-\alpha} \geq 1 - \frac{\nu - \alpha}{k+1}, \quad \text{since } \nu - \alpha \in (1 - \alpha, 1]; \\
&(ii): \frac{\nu - \alpha}{k+1} < \frac{\beta}{2} (k+1)^{-\alpha}, \quad \text{since } k > t^\star; \quad \text{and} \\
&(iii): \beta C = \frac{\beta}{\nu - 1} \left(\frac{2(\nu - \alpha)}{\beta} \right)^{\frac{\nu-\alpha}{1-\alpha}} > \frac{\beta}{\nu - 1} \frac{2(\nu - \alpha)}{\beta} > 2, \quad \text{since } \nu > 1 > \alpha.
\end{aligned}$$

An induction argument then validates (41) for every positive integer t.

Proof of Equation (38). In order to derive an iterative relationship with respect to the left-hand-side of (38), we denote

$$X_t = \sum_{i=1}^{t} i^{-2\alpha} \prod_{k=i+1}^{t} \left(1 - \beta k^{-\alpha} \right)$$

for any positive integer t. Then by definition, $X_1 = 1$ and the sequence $\{X_t\}_{t\geq 1}$ is featured by the iterative relation

$$X_t = \left(1 - \beta t^{-\alpha}\right) X_{t-1} + t^{-2\alpha}, \quad \text{for every } t > 1. \tag{42}$$

Direct calculation yields

$$\begin{aligned}
t^\alpha X_t &= \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) \left[(t-1)^\alpha X_{t-1}\right] + t^{-\alpha} \\
&= \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) \cdot \frac{1}{\beta} + \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) \left[(t-1)^\alpha X_{t-1} - \frac{1}{\beta}\right] + t^{-\alpha} \\
&= \frac{1}{\beta} + \left[\left(\frac{t}{t-1}\right)^\alpha - 1\right] \left(\frac{1}{\beta} + t^{-\alpha}\right) \\
&+ \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) \left[(t-1)^\alpha X_{t-1} - \frac{1}{\beta}\right].
\end{aligned}$$

Setting $Z_t = t^\alpha X_t - \frac{1}{\beta}$, we immediately obtain

$$Z_t = \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) Z_{t-1} + \left[\left(\frac{t}{t-1}\right)^\alpha - 1\right] \left(\frac{1}{\beta} + t^{-\alpha}\right).$$

Notice that for every $t > 1$ and $\alpha < 1$,

$$\left(\frac{t}{t-1}\right)^\alpha - 1 < \frac{t}{t-1} - 1 = \frac{1}{t-1} < \frac{2}{t}.$$

Hence, the sequence $\{Z_t\}_{t\geq 1}$ is upper bounded iteratively by

$$Z_t < \left(\frac{t}{t-1}\right)^\alpha \left(1 - \beta t^{-\alpha}\right) Z_{t-1} + \frac{2}{t} \left(\frac{1}{\beta} + t^{-\alpha}\right).$$

By induction, this directly implies

$$Z_t < \underbrace{t^\alpha \prod_{k=1}^{t} \left(1 - \beta k^{-\alpha}\right) Z_1}_{I_1} + \underbrace{2t^\alpha \sum_{i=1}^{t} \left(\frac{1}{\beta} i^{-\alpha-1} + i^{-2\alpha-1}\right) \prod_{k=i+1}^{t} (1 - \beta k^{-\alpha})}_{I_2},$$

where $Z_1 = 1^\alpha X_1 - \frac{1}{\beta} = 1 - \frac{1}{\beta}$. The proof now boils down to showing that both I_1 and I_2 converges to 0 as $t \to \infty$. On one hand, I_1 is bounded by

$$t^\alpha \exp\left(-\beta \sum_{k=1}^{t} k^{-\alpha}\right) < \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}} t^{\alpha-1}.$$

On the other hand, the convergence of I_2 is guaranteed by (37):

$$\begin{aligned}
I_2 &\leq \frac{4t^\alpha}{\beta} \sum_{i=1}^{t} i^{-\alpha-1} \prod_{k=i+1}^{t} (1 - \beta k^{-\alpha}) \\
&\leq \frac{2}{\alpha} \left(\frac{2}{\beta}\right)^{\frac{2-\alpha}{1-\alpha}} t^{\alpha-1}.
\end{aligned}$$

Combining these results, we obtain

$$t^\alpha X_t - \frac{1}{\beta} = Z_t < 5 \left(\frac{2}{\beta}\right)^{\frac{2-\alpha}{1-\alpha}} t^{\alpha-1},$$

which directly implies (38).

Proof of Equation (39). For every positive integer t, we begin by identifying the index i that maximizes

$$i^{-\alpha} \prod_{k=i+1}^{t} \left(1 - \beta k^{-\alpha}\right) := X_i^t.$$

Consider the ratio of two consecutive terms:

$$\begin{aligned} \frac{X_{i+1}}{X_i} &= \frac{(i+1)^{-\alpha}}{i^{-\alpha}(1 - \beta(i+1)^{-\alpha})} \\ &= \frac{i^{\alpha}}{(i+1)^{\alpha} - \beta}, \quad \text{for any } 1 \le i < t. \end{aligned}$$

Equivalently, we observe

$$\frac{X_i}{X_{i+1}} = 1 + i^{-\alpha}\left[(i+1)^{\alpha} - i^{\alpha} - \beta\right].$$

Since $(i+1)^{\alpha} - i^{\alpha} \searrow 0$ for any $\alpha \in (\frac{1}{2}, 1)$, we can conclude that when

$$2^{\alpha} - 1 - \beta < 0,$$

the sequence $X_{i\,1\le i\le t}^t$ is monotonically increasing; otherwise, there exists a constant $i^\star$ such that $\{X_i^t\}_{1\le i\le t}$ decreases when $i \le i^\star$ and increases when $i > i^\star$. Either way, the maximum over i is reached at either $i = 1$ or $i = t$. That is,

$$\max_{1\le i\le t} X_i^t = \max\left\{\prod_{k=2}^{t}\left(1 - \beta k^{-\alpha}\right), t^{-\alpha}\right\}$$

where the first term on the right hand side is bounded above by (36). The equation follows naturally. □

Lemma 6.7. *For every $\beta \in (0,1)$, $\alpha \in (\frac{1}{2}, 1)$ and sufficiently large T, it can be guaranteed that*

$$t^{-\alpha}\sum_{j=t}^{T}\prod_{k=t+1}^{j}(1 - \beta k^{-\alpha}) < 3\left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}}, \quad \textit{and} \tag{43}$$

$$t^{-\alpha}\sum_{j=t}^{\infty}\prod_{k=t+1}^{j}(1 - \beta k^{-\alpha}) - \frac{1}{\beta} \lesssim \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)t^{\alpha-1}. \tag{44}$$

Proof: We address the two inequalities in order.

Proof of (43). For simplicity, we denote

$$\begin{aligned} t^\star &= \left(\frac{2\alpha}{\beta}\right)^{\frac{1}{1-\alpha}}, \\ X_t &= \sum_{j=t}^{T}\prod_{k=t+1}^{j}\left(1 - \beta k^{-\alpha}\right), \quad \text{and} \\ Y_t &= t^{-\alpha}X_t. \end{aligned}$$

By definition, we observe that $X_T = 1$, $Y_T = T^{-\alpha}$ and that the sequences $\{X_t\}_{1\le t\le T}$ and $\{Y_t\}_{1\le t\le T}$ is featured by the iterative relation

$$X_t = 1 + \left[1 - \beta(t+1)^{-\alpha}\right]X_{t+1} \tag{45}$$

$$Y_t = t^{-\alpha} + \left(\frac{t+1}{t}\right)^{\alpha}\left[1 - \beta(t+1)^{-\alpha}\right]Y_{t+1}. \tag{46}$$

In what follows, we consider the bound on Y_t for both $t \ge t^\star$ and $1 \le t < t^\star$.

Bounding Y_t for $t \geq t^\star$. Without loss of generality, we assume that $T > t^\star$; the bound on Y_t for $t \geq t^\star$ is shown by an inverse induction from $t = T$ to $t = t^\star$. Namely, we observe that $Y_T = T^{-\alpha} < 1 < \frac{2}{\beta}$; if $Y_{t+1} < \frac{2}{\beta}$ holds for an integer $t \in (t^\star, T)$, then the iterative relation (46) implies

$$\begin{aligned} Y_t &< t^{-\alpha} + \left(\frac{t+1}{t}\right)^{\alpha} \left[1 - \beta(t+1)^{-\alpha}\right] \cdot \frac{2}{\beta} \\ &< \frac{2}{\beta} + t^{-\alpha} \left\{ \left[(t+1)^{\alpha} - t^{\alpha}\right] \frac{2}{\beta} - 1 \right\}. \end{aligned}$$

Hence, it boils down to showing

$$(t+1)^{\alpha} - t^{\alpha} \leq \frac{\beta}{2},$$

for which we observe

$$(t+1)^{\alpha} - t^{\alpha} < \alpha t^{\alpha-1} \leq \alpha \left(t^\star\right)^{\alpha-1} = \frac{2}{\beta}.$$

In this way, we prove that $Y_t \leq \frac{2}{\beta}$ for every $t \geq t^\star$.

Bounding Y_t for $1 \leq t < t^\star$. For every positive integer t satisfying $1 \leq t < t^\star$, the iterative relation (45)

$$\begin{aligned} Y_t = t^{-\alpha} X_t < X_t < X_{t^\star} + (t^\star - t) &< X_{t^\star} + t^\star \\ &= (t^\star)^{\alpha} Y_{t^\star} + t^\star \\ &< \left(\frac{2\alpha}{\beta}\right)^{\frac{\alpha}{1-\alpha}} \frac{2}{\beta} + \left(\frac{2\alpha}{\beta}\right)^{\frac{1}{1-\alpha}} \\ &= \left(\frac{2\alpha}{\beta}\right)^{\frac{1}{1-\alpha}} \left(\frac{1}{\alpha} + 1\right). \end{aligned}$$

Proof of (44). We firstly expand the left-hand-side of (44) as

$$\begin{aligned} & t^{-\alpha} \sum_{j=t}^{\infty} \prod_{k=t+1}^{j} (1 - \beta k^{-\alpha}) \\ &= \sum_{j=t}^{\infty} (j+1)^{-\alpha} \prod_{k=t+1}^{j} (1 - \beta k^{-\alpha}) + \sum_{j=t}^{\infty} (t^{-\alpha} - (j+1)^{-\alpha}) \prod_{k=t+1}^{j} (1 - \beta k^{-\alpha}). \end{aligned} \tag{47}$$

Here, the first term is bounded by applying the telescoping technique

$$\begin{aligned}
&\sum_{j=t}^{\infty}(j+1)^{-\alpha}\prod_{k=t+1}^{j}(1-\beta k^{-\alpha}) \\
&=\sum_{j=t}^{\infty}\frac{1}{\beta}\left[1-(1-\beta(j+1)^{-\alpha})\right]\prod_{k=t+1}^{j}(1-\beta k^{-\alpha}) \\
&=\frac{1}{\beta}\sum_{j=t}^{\infty}\left(\prod_{k=t+1}^{j}(1-\beta k^{-\alpha})-\prod_{k=t+1}^{j+1}(1-\beta k^{-\alpha})\right) \\
&=\frac{1}{\beta}\left(\sum_{j=t}^{\infty}\prod_{k=t+1}^{j}(1-\beta k^{-\alpha})-\sum_{j=t}^{\infty}\prod_{k=t+1}^{j+1}(1-\beta k^{-\alpha})\right) \\
&=\frac{1}{\beta}\left[\left(\prod_{k=t+1}^{t}(1-\beta k^{-\alpha})+\sum_{j=t+1}^{\infty}\prod_{k=t+1}^{j}(1-\beta k^{-\alpha})\right)-\sum_{j=t}^{\infty}\prod_{k=t+1}^{j+1}(1-\beta k^{-\alpha})\right] \\
&\overset{(i)}{=}\frac{1}{\beta}\left(1+\sum_{i=t}^{\infty}\prod_{k=t+1}^{i+1}(1-\beta k^{-\alpha})-\sum_{j=t}^{\infty}\prod_{k=t+1}^{j+1}(1-\beta k^{-\alpha})\right)=\frac{1}{\beta}. \qquad (48)
\end{aligned}$$

Notice that in (i) we invoked the convention

$$\prod_{k=t+1}^{t}(1-\beta k^{-\alpha})=1,$$

and a change of variable $i=j-1$.

Therefore, the proof boils down to showing that the second term is bounded by

$$\sum_{j=t}^{\infty}((j+1)^{-\alpha}-t^{-\alpha})\prod_{k=t+1}^{j}(1-\beta k^{-\alpha})=O(t^{\alpha-1}). \qquad (49)$$

Towards this end, we denote

$$m_t^j=\sum_{k=t}^{j}k^{-\alpha}. \qquad (50)$$

In terms of m_t^j, the left-hand-side of (49) is bounded by

$$\begin{aligned}
&\sum_{j=t}^{\infty}((j+1)^{-\alpha}-t^{-\alpha})\prod_{k=t+1}^{j}(1-\beta k^{-\alpha}) \\
&\asymp\sum_{j=t}^{\infty}((j+1)^{-\alpha}-t^{-\alpha})\exp\left(-\beta m_t^j\right).
\end{aligned}$$

Furthermore, we observe that on one hand, m_t^j is bounded by

$$m_t^j=\sum_{k=t}^{j}k^{-\alpha}>\int_t^{j+1}x^{-\alpha}\mathrm{d}x=\frac{(j+1)^{1-\alpha}-t^{1-\alpha}}{1-\alpha}; \qquad (51)$$

on the other hand, the term $(j+1)^{-\alpha}-t^{-\alpha}$ can be bounded by

$$\begin{aligned}
t^{-\alpha}-(j+1)^{-\alpha}&=(j+1)^{-\alpha}\left[\left(\frac{j+1}{t}\right)^{\alpha}-1\right] \\
&=(m_t^{j+1}-m_t^j)\cdot\left[\left(\frac{j+1}{t}\right)^{\alpha}-1\right].
\end{aligned}$$

In what follows, we aim to bound the last term in terms of m_t^j. Specifically, since

$$\left(\frac{j+1}{t}\right)^{1-\alpha} - 1 < \frac{(1-\alpha)m_t^j}{t^{1-\alpha}},$$

we obtain

$$\begin{aligned}\left(\frac{j+1}{t}\right)^{\alpha} - 1 &< \left[\frac{(1-\alpha)m_t^j}{t^{1-\alpha}} + 1\right]^{\frac{\alpha}{1-\alpha}} - 1 \\ &< t^{\alpha-1}\left((1-\alpha)m_t^j + 1\right)^{\frac{\alpha}{1-\alpha}}.\end{aligned}$$

Therefore, the left-hand-side of (49) is bounded by

$$\begin{aligned}&\sum_{j=t}^{\infty}((j+1)^{-\alpha} - t^{-\alpha}) \prod_{k=t+1}^{j}(1-\beta k^{-\alpha}) \\ &< \sum_{j=t}^{\infty}(m_t^{j+1} - m_t^j)t^{\alpha-1}\left((1-\alpha)m_t^j + 1\right)^{\frac{\alpha}{1-\alpha}} \exp\left(\beta t^{-\alpha} - \beta m_t^j\right) \\ &< t^{\alpha-1}\sum_{j=t}^{\infty}\left((1-\alpha)m_t^j + 1\right)^{\frac{\alpha}{1-\alpha}} \exp\left(-\beta m_t^j\right)(m_t^{j+1} - m_t^j) \\ &\asymp t^{\alpha-1}\int_0^{\infty}((1-\alpha)m+1)^{\frac{\alpha}{1-\alpha}}\exp(-\beta m)\mathrm{d}m = O(t^{\alpha-1});\end{aligned}$$

the integral on the last line is further bounded by

$$\begin{aligned}&\int_0^{\infty}((1-\alpha)m+1)^{\frac{\alpha}{1-\alpha}}\exp(-\beta m)\mathrm{d}m \\ &< \int_0^{\infty}\left(m^{\frac{\alpha}{1-\alpha}} + 2^{\frac{\alpha}{1-\alpha}}\right)\exp(-\beta m)\mathrm{d}m \\ &= 2^{\frac{\alpha}{1-\alpha}}\frac{1}{\beta} + \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right) \lesssim \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right).\end{aligned}$$

The bound (44) follows naturally. □

Next, we introduce additional quantities

$$\overline{\boldsymbol{\Lambda}}_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Q}_t\boldsymbol{\Gamma}\boldsymbol{Q}_t^{\top}, \tag{52}$$

where $\boldsymbol{\Gamma}$ is defined in (14) and, for each t, we set

$$\boldsymbol{Q}_t = \eta_t\sum_{j=t}^{T}\prod_{k=t+1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A}). \tag{53}$$

The matrix $\overline{\boldsymbol{\Lambda}}_T$ plays the role of a non-asymptotic covariance matrix for $\overline{\boldsymbol{\theta}}_T$. The following lemmas feature the properties of $\boldsymbol{Q}_t$ and $\overline{\boldsymbol{\Lambda}}_T$.

Lemma 6.8. *Let $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$ and $0 \le \eta_0 \le \frac{1}{2\lambda_\Sigma}$. Recall that $\boldsymbol{Q}_t$ is defined as in (53). Then for any positive integer $\ell < T$, the following relation hold:*

$$\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A}), \tag{54a}$$

Furthermore, there exist a sequence of matrix $\{\boldsymbol{S}_t\}_{1\le t\le T}$*, along with a constant* $\widetilde{C}$*, depending on* α, η_0 *and* γ*, such that the difference between* $\boldsymbol{Q}_t$ *and* $\boldsymbol{A}^{-1}$ *can be represented as*

$$\boldsymbol{Q}_t - \boldsymbol{A}^{-1} = \boldsymbol{S}_t - \boldsymbol{A}^{-1}\prod_{k=t}^{T}(\boldsymbol{I} - \eta_k \boldsymbol{A}), \tag{54b}$$

in which $\boldsymbol{S}_t$ *is independent of* T*, and its norm is bounded by*

$$\|\boldsymbol{S}_t\| \lesssim \eta_0\Gamma\left(\frac{1}{1-\alpha}\right)\left(\frac{2}{(1-\gamma)\lambda_0\eta_0}\right)^{\frac{1}{1-\alpha}} t^{\alpha-1}. \tag{54c}$$

Proof: We address these equations and bounds separately.

Proof of (54a) By definition, the sum of $\boldsymbol{Q}_t$ for $1 \le t \le T$ can be expressed as

$$\begin{aligned}\sum_{t=1}^{T}\boldsymbol{Q}_t &= \sum_{t=1}^{T}\eta_t\sum_{j=t}^{T}\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\ &= \sum_{j=1}^{T}\sum_{t=1}^{j}\eta_t\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}),\end{aligned}$$

where the second line follows by switching the order of summation. Here, the summand can be represented as

$$\begin{aligned}\eta_t\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) &= \boldsymbol{A}^{-1}[\boldsymbol{I}-(\boldsymbol{I}-\eta_t\boldsymbol{A})]\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\ &= \boldsymbol{A}^{-1}\left[\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) - \prod_{k=t}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right].\end{aligned} \tag{55}$$

Notice that summing the difference in the bracket from $t=1$ to $t=j$ would cancel all the terms in the middle and only the first and last terms would remain. Therefore,

$$\begin{aligned}\sum_{t=1}^{T}\boldsymbol{Q}_t &= \sum_{j=1}^{T}\sum_{t=1}^{j}\eta_t\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\ &= \sum_{j=1}^{T}\sum_{t=1}^{j}\boldsymbol{A}^{-1}\left[\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) - \prod_{k=t}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right]\\ &= \boldsymbol{A}^{-1}\sum_{j=1}^{T}\left[\boldsymbol{I} - \prod_{k=1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right],\end{aligned}$$

which directly implies (54a).

Construction of the decomposition (54b)**.** The definition of $\boldsymbol{Q}_t$ (53) directly implies that on one hand,

$$\boldsymbol{Q}_t = \eta_t\sum_{j=t}^{T}\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) = \eta_t\boldsymbol{I} + \eta_t\sum_{j=t+1}^{T}\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}),$$

and that

$$\begin{aligned}(\boldsymbol{I}-\eta_t\boldsymbol{A})\boldsymbol{Q}_t = \eta_t\sum_{j=t}^{T}\prod_{k=t}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) &= \eta_t\sum_{j=t}^{T-1}\prod_{k=t}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A}) + \eta_t\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\ &= \eta_t\sum_{j=t+1}^{T-1}\prod_{k=t}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A}) + \eta_t\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})-\prod_{k=t}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\
&=\left[\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right](\boldsymbol{I}-\eta_j\boldsymbol{A})-(\boldsymbol{I}-\eta_t\boldsymbol{A})\left[\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right]\\
&=\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})-\eta_j\left[\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right]\boldsymbol{A}-\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})+\eta_t\boldsymbol{A}\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\
&=\eta_t\boldsymbol{A}\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})-\eta_j\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{A}\\
&=(\eta_t-\eta_j)\boldsymbol{A}\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A}).
\end{aligned}$$

Therefore, by taking a difference between $\boldsymbol{Q}_t$ and $(\boldsymbol{I}-\eta_t\boldsymbol{A})\boldsymbol{Q}_t$, we obtain

$$\begin{aligned}
\eta_t\boldsymbol{A}\boldsymbol{Q}_t&=\boldsymbol{Q}_t-(\boldsymbol{I}-\eta_t\boldsymbol{A})\boldsymbol{Q}_t\\
&=\eta_t\boldsymbol{I}+\eta_t\sum_{j=t+1}^{T-1}\left[\prod_{k=t+1}^{j}(\boldsymbol{I}-\eta_k\boldsymbol{A})-\prod_{k=t}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right]-\eta_t\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\\
&=\eta_t\boldsymbol{I}+\eta_t\sum_{j=t+1}^{T-1}(\eta_t-\eta_j)\boldsymbol{A}\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})-\eta_t\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}).
\end{aligned}$$

Consequently, the difference between $\boldsymbol{Q}_t$ and $\boldsymbol{A}^{-1}$ can be expressed as

$$\boldsymbol{Q}_t-\boldsymbol{A}^{-1}=\underbrace{\sum_{j=t+1}^{T-1}(\eta_t-\eta_j)\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})}_{\boldsymbol{S}_t}-\boldsymbol{A}^{-1}\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}),\tag{56}$$

which takes the form of (54b).

Proof of (54c). Lemma 6.4 implies that

$$\begin{aligned}
\left\|\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right\|&\le\prod_{k=t+1}^{j-1}\left(1-\frac{1}{2}\eta_k(1-\gamma)\lambda_0\right)\\
&\le\prod_{k=t+1}^{j}\exp\left(-\frac{1}{2}\eta_k(1-\gamma)\lambda_0\right)\\
&=\exp\left(-\frac{1}{2}(1-\gamma)\lambda_0\sum_{k=t+1}^{j-1}\eta_k\right).
\end{aligned}$$

Recalling the definition of m_t^j (50),(51), we can bound the norm $\boldsymbol{S}_t$ by

$$\begin{aligned}\|\boldsymbol{S}_t\|_2 &\leq \sum_{j=t}^{\infty}(\eta_t-\eta_j)\left\|\prod_{k=t+1}^{j-1}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right\|_2\\ &\leq \sum_{j=t}^{\infty}(\eta_t-\eta_j)\exp\left(-\frac{1}{2}(1-\gamma)\lambda_0\sum_{k=t+1}^{j-1}\eta_k\right)\\ &= \sum_{j=t}^{\infty}\eta_j\left[\left(\frac{j}{t}\right)^{\alpha}-1\right]\exp\left(-\frac{1}{2}(1-\gamma)\lambda_0\eta_0 m_t^j\right).\end{aligned}$$

The right-hand-side can then be bounded by a similar argument to that of Lemma 6.7. □

The following Lemma features the compressive rate of the matrix $\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})$.

Lemma 6.9. *When $\eta_0 < \frac{1}{2\lambda_\Sigma}$, for every positive definite matrix $\boldsymbol{Y}\in\mathbb{R}^{d\times d}$, it can be guaranteed that*

$$T^{-\alpha}\sum_{t=1}^{T}\left(\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{Y}\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}^\top)\right)-\boldsymbol{X}\leq O(T^{\alpha-1})\boldsymbol{Y} \tag{57}$$

where $\boldsymbol{X}\succeq\boldsymbol{0}$ is the solution to the Lyapunov equation

$$\eta_0(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^\top)=\boldsymbol{Y}, \tag{58}$$

and the big O notation for matrix sequences is defined as in (25).

Proof: For simplicity, we define a sequence of matrices $\{\boldsymbol{X}_i\}_{i=0}^{T}$ as

$$\boldsymbol{X}_i=\begin{cases} i^{-\alpha}\sum_{t=1}^{i}\left(\prod_{k=t}^{i}(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{Y}\prod_{k=t}^{i}(\boldsymbol{I}-\eta_k\boldsymbol{A}^\top)\right), & \text{for}\quad 1\leq i\leq T;\\ \boldsymbol{0}, & \text{for}\quad i=0.\end{cases}$$

in this way, the left-hand-side of (57) can be represented by $\boldsymbol{X}_T-\boldsymbol{X}$. It can be easily verified that the sequence $\{\boldsymbol{X}_i\}_{i=1}^{T}$ is featured by the iterative relation

$$i^\alpha\boldsymbol{X}_i=(i-1)^\alpha(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{X}_{i-1}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)+(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{Y}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top),\quad \forall 1\leq i\leq T.$$

Direct calculation yields that for every $1\leq i\leq T$,

$$\begin{aligned}i^\alpha(\boldsymbol{X}_i-\boldsymbol{X}) &= (i-1)^\alpha(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{X}_{i-1}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)+(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{Y}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)-i^\alpha\boldsymbol{X}\\ &= (i-1)^\alpha(\boldsymbol{I}-\eta_i\boldsymbol{A})(\boldsymbol{X}_{i-1}-\boldsymbol{X})(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)\\ &\quad+(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{Y}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)-[i^\alpha-(i-1)^\alpha]\boldsymbol{X}-(i-1)^\alpha[\eta_i(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^\top)+\eta_i^2\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^\top].\end{aligned}$$

We now feature the terms on the last line. Firstly, we notice that (58) guarantees

$$\eta_i(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^\top)=\eta_0 i^{-\alpha}(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^\top)=i^{-\alpha}\boldsymbol{Y};$$

therefore, a simple reorganization yields

$$\begin{aligned}&(\boldsymbol{I}-\eta_i\boldsymbol{A})\boldsymbol{Y}(\boldsymbol{I}-\eta_i\boldsymbol{A}^\top)-[i^\alpha-(i-1)^\alpha]\boldsymbol{X}-(i-1)^\alpha[\eta_i(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^\top)+\eta_i^2\boldsymbol{A}\boldsymbol{X}\boldsymbol{A}^\top]\\ &=\left[1-\left(\frac{i-1}{i}\right)^{\alpha}\right]\boldsymbol{Y}-[i^\alpha-(i-1)^\alpha]\boldsymbol{X}-\eta_i(\boldsymbol{A}\boldsymbol{Y}+\boldsymbol{Y}\boldsymbol{A}^\top)+\eta_i^2\boldsymbol{A}(\boldsymbol{Y}-(i-1)^\alpha\boldsymbol{X})\boldsymbol{A}^\top.\end{aligned}$$

For these terms, we observe

$$\begin{aligned}&\left[1-\left(\frac{i-1}{i}\right)^{\alpha}\right]\boldsymbol{Y}\preceq\left[1-\left(\frac{i-1}{i}\right)\right]\boldsymbol{Y}=\frac{1}{i}\boldsymbol{Y};\\ &[i^\alpha-(i-1)^\alpha]\boldsymbol{X}=i^\alpha\left[1-\left(\frac{i-1}{i}\right)^{\alpha}\right]\boldsymbol{X}\preceq i^\alpha\left[1-\left(\frac{i-1}{i}\right)\right]\boldsymbol{X}=i^{\alpha-1}\boldsymbol{X}=O(i^{\alpha-1})\boldsymbol{Y};\\ &\eta_i(\boldsymbol{A}\boldsymbol{Y}+\boldsymbol{Y}\boldsymbol{A}^\top)=O(i^{-\alpha})\boldsymbol{Y};\\ &\eta_i^2\boldsymbol{A}(\boldsymbol{Y}-(i-1)^\alpha\boldsymbol{X})\boldsymbol{A}^\top=O(i^{-\alpha})\boldsymbol{Y}.\end{aligned}$$

Hence, the leading term decreases by the rate of $O(i^{\alpha-1})\boldsymbol{Y}$. As a result,

$$i^{\alpha}(\boldsymbol{X}_i - \boldsymbol{X}) = (\boldsymbol{I} - \eta_i \boldsymbol{A})[(i-1)^{\alpha}(\boldsymbol{X}_{i-1} - \boldsymbol{X})](\boldsymbol{I} - \eta_i \boldsymbol{A}^{\top}) + O(i^{\alpha-1})\boldsymbol{Y}$$

For simplicity, we denote $\boldsymbol{Z}_i = i^{\alpha}(\boldsymbol{X}_i - \boldsymbol{X})$. The relation above can then be represented as

$$\boldsymbol{Z}_i = (\boldsymbol{I} - \eta_i \boldsymbol{A})\boldsymbol{Z}_{i-1}(\boldsymbol{I} - \eta_i \boldsymbol{A}^{\top}) + O(i^{\alpha-1})\boldsymbol{Y}.$$

It is easy to verify by iteration that, there exists a quantity $\widetilde{C}$ independent of T, such that

$$\begin{aligned}
&\sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{Y} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}) \\
&\preceq \boldsymbol{Z}_T \preceq \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{Y} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top});
\end{aligned}$$

In other words,

$$\begin{aligned}
&T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{Y} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}) \\
&\preceq \boldsymbol{X}_T - \boldsymbol{X} \\
&\preceq T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{Y} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}).
\end{aligned}$$

By letting $\beta = \frac{1-\gamma}{2}\lambda_0\eta_0$, Lemma 6.4 implies

$$\begin{aligned}
&\left\| T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}) \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}) \right\| \\
&\leq T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} \|\boldsymbol{I} - \eta_k \boldsymbol{A}\| \prod_{k=i+1}^{T} \|\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}\| \\
&\leq T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (1 - \beta k^{-\alpha})^2 \\
&< T^{\alpha-1} \sum_{i=1}^{T} i^{-\alpha} \prod_{k=i+1}^{T} (1 - \beta k^{-\alpha}) \\
&= T^{\alpha-1} \cdot \frac{1}{\beta} \sum_{i=1}^{T} \left(\prod_{k=i+1}^{T} (1 - \beta k^{-\alpha}) - \prod_{k=i}^{T} (1 - \beta k^{-\alpha}) \right) \\
&= \frac{1}{\beta} T^{\alpha-1} \left(\prod_{k=T+1}^{T} (1 - \beta k^{-\alpha}) - \prod_{k=1}^{T} (1 - \beta k^{-\alpha}) \right) < \frac{1}{\beta} T^{\alpha-1};
\end{aligned}$$

notice that in the last two lines we invoked a similar approach to the telescoping method (48). Combining this with (28), we obtain

$$T^{-\alpha} \sum_{i=1}^{T} i^{\alpha-1} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{Y} \prod_{k=i+1}^{T} (\boldsymbol{I} - \eta_k \boldsymbol{A}^{\top}) = O(T^{\alpha-1})\boldsymbol{Y}.$$

This completes the proof. □

Lemma 6.10. *Let $\eta_t = \eta_0 t^{-\alpha}$ with $\alpha \in (\frac{1}{2}, 1)$ and $0 \leq \eta \leq \frac{1}{2\lambda_\Sigma}$. For $\boldsymbol{Q}_t$ defined as in (53), it can be guaranteed that*

$$\|\boldsymbol{Q}_t\| \leq 3\eta_0^{-\frac{\alpha}{1-\alpha}} \left(\frac{4\alpha}{\lambda_0(1-\gamma)}\right)^{\frac{1}{1-\alpha}}, \quad \forall t \in [T]; \tag{59}$$

$$\|\boldsymbol{A}\boldsymbol{Q}_t\| \leq 2 + \eta_0 \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}} \Gamma\left(\frac{1}{1-\alpha}\right) t^{\alpha-1}; \quad \textit{and} \tag{60}$$

$$\frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{Q}_t - \boldsymbol{A}^{-1}\|^2 \leq \frac{1}{\lambda_0(1-\gamma)} T^{\alpha-1} + \widetilde{C} T^{2\alpha-2} \tag{61}$$

where $\beta = \frac{1-\gamma}{2}\lambda_0\eta_0$ and $\widetilde{C}$ depends on $\alpha, \eta_0, \lambda_0, \gamma$.

Proof: According to Lemma 6.4, since $0 < \eta_t \leq \eta_0 \leq \frac{1}{2\lambda_\Sigma}$ for all t, the norm of $\boldsymbol{Q}_t$ is bounded by

$$\begin{aligned}
\|\boldsymbol{Q}_t\|_2 &\leq \eta_t \sum_{j=t}^{T} \prod_{k=t+1}^{j} \left(1 - \frac{1}{2}\lambda_0(1-\gamma)\eta_k\right) \\
&= \eta_0 t^{-\alpha} \cdot \sum_{j=t}^{T} \prod_{k=t+1}^{j} \left(1 - \frac{1}{2}\lambda_0(1-\gamma)\eta_0 k^{-\alpha}\right).
\end{aligned}$$

The bound (59) follows immediately by taking $\beta = \frac{1-\gamma}{2}\lambda_0\eta_0$ in (43). For the bound (60), observe due to (54b) that

$$\boldsymbol{A}\boldsymbol{Q}_t - \boldsymbol{I} = \boldsymbol{A}\boldsymbol{S}_t - \prod_{k=t}^{T}(\boldsymbol{I} - \eta_k \boldsymbol{A}).$$

Therefore, the triangle inequality and (54c) directly imply

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{Q}_t\| \leq \|\boldsymbol{I}\| + \|\boldsymbol{A}\boldsymbol{S}_t\| + \left\|\prod_{k=t}^{T}(\boldsymbol{I} - \eta_k\boldsymbol{A})\right\| &\leq 1 + \eta_0 \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}} \Gamma\left(\frac{1}{1-\alpha}\right) t^{\alpha-1} + 1 \\
&= 2 + \eta_0 \left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}} \Gamma\left(\frac{1}{1-\alpha}\right) t^{\alpha-1}.
\end{aligned}$$

Furthermore, it can be easily verified that

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{Q}_t - \boldsymbol{I}\|^2 &\leq \|\boldsymbol{A}\boldsymbol{S}_t\|^2 + \left\|\prod_{k=t}^{T}(\boldsymbol{I} - \eta_k\boldsymbol{A})\right\|^2 \\
&\leq \widetilde{C}' t^{2\alpha-2} + \prod_{k=t}^{T}(1 - \beta k^{-\alpha}) \\
&< \widetilde{C}' t^{2\alpha-2} + T^{\alpha}(t-1)^{-\alpha} \prod_{k=t}^{T}(1 - \beta k^{-\alpha}) \\
&= \widetilde{C}' t^{2\alpha-2} + T^{\alpha} \cdot \frac{1}{\beta}\left(\prod_{k=t}^{T}(1 - \beta k^{-\alpha}) - \prod_{k=t-1}^{T}(1 - \beta k^{-\alpha})\right);
\end{aligned}$$

The bound (61) follows by taking an average from $t = 1$ to $t = T$ and applying the telescoping method as in (48). □

The decomposition (54b), combined with the bounds (54c) Theorem 6.9, implies the following bound on the difference between $\overline{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$.

Theorem 6.11. *Let $\overline{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$ be defined as in (52) and (14) respectively. Define $\boldsymbol{X}(\boldsymbol{\Lambda}^\star)$ as the unique solution to the Lyapunov equation*

$$\eta_0(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}^\top) = \boldsymbol{\Lambda}^\star; \tag{62}$$

then as $T \to \infty$, the difference between $\overline{\boldsymbol{\Lambda}_T}$ and $\boldsymbol{\Lambda}^\star$ satisfies

$$\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star - T^{\alpha-1}\boldsymbol{X}(\boldsymbol{\Lambda}^\star) = O(T^{2\alpha-2})\boldsymbol{\Gamma}.$$

Here, the O notation for matrices is defined as in (25).

Proof: By definition, the difference between $\overline{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$ can be expressed as

$$\begin{aligned}
\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star &= \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t\boldsymbol{\Gamma}\boldsymbol{Q}_t^\top - \boldsymbol{A}^{-1}\boldsymbol{\Gamma}\boldsymbol{A}^{-\top}) \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})\boldsymbol{\Gamma}\boldsymbol{A}^{-\top} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}^{-1}\boldsymbol{\Gamma}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})^\top}_{I_1} \\
&\quad + \underbrace{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})\boldsymbol{\Gamma}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})^\top}_{I_2}.
\end{aligned} \tag{63}$$

We now proceed to bound the norms of I_1 and I_2 respectively.

Bounding the norm of I_1. Notice that the first term can be expressed by

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})\boldsymbol{\Gamma}\boldsymbol{A}^{-\top} &= -\frac{1}{T}\boldsymbol{A}^{-1}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\cdot\boldsymbol{\Gamma}\boldsymbol{A}^{-\top} \\
&= -\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\cdot\boldsymbol{A}^{-1}\boldsymbol{\Gamma}\boldsymbol{A}^{-\top} \\
&= -\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\cdot\boldsymbol{\Lambda}^\star.
\end{aligned} \tag{64}$$

where we applied Equation (54a) the interchangeability of products. Similarly, the second term can be expressed by

$$\boldsymbol{A}^{-1}\boldsymbol{\Gamma}\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})^\top = -\boldsymbol{\Lambda}^\star\cdot\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})^\top.$$

Lemma 6.4 directly implies that

$$\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A}) \le \frac{1}{T}\prod_{k=1}^{j}\left(1 - \frac{1-\gamma}{2}\lambda_0\eta_0 k^{-\alpha}\right).$$

For simplicity, we denote $\beta = \frac{1-\gamma}{2}\lambda_0\eta_0$; it can then be seen that the summand on the right-hand-side can be bounded by

$$\begin{aligned}
\sum_{j=1}^{T}\prod_{k=1}^{j}(1-\beta k^{-\alpha}) &< \sum_{j=1}^{T}\exp\left(\beta\sum_{k=1}^{j}k^{-\alpha}\right) \\
&= \sum_{j=1}^{T}\exp\left(-\beta m_1^j\right) \qquad (65)\\
&\asymp \sum_{j=1}^{T}\exp(-\beta m_1^j)(m_1^{j+1}-m_1^j)((1-\alpha)m_1^j)^{\frac{\alpha}{1-\alpha}} \\
&\lesssim (1-\alpha)^{\frac{\alpha}{1-\alpha}}\int_0^{\infty} m^{\frac{\alpha}{1-\alpha}}\exp(-\beta m)\mathrm{d}m \\
&= \beta^{-\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right), \qquad (66)
\end{aligned}$$

a constant independent of T. Therefore, I_1 is featured by

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{Q}_t-\boldsymbol{A}^{-1})\boldsymbol{\Gamma}\boldsymbol{A}^{-\top}+\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}^{-1}\boldsymbol{\Gamma}(\boldsymbol{Q}_t-\boldsymbol{A}^{-1})^{\top}=O(T^{-1})\boldsymbol{\Lambda}^{\star}.$$

Decomposing I_2. The term I_2 can be further expressed by invoking the decomposition (54b). Specifically, we observe

$$\begin{aligned}
I_2 = &\underbrace{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}_t\boldsymbol{\Gamma}\boldsymbol{S}_t^{\top}}_{I_{21}}+\underbrace{\frac{1}{T}\sum_{t=1}^{T}\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{\Lambda}^{\star}\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}^{\top})}_{I_{22}} \\
&\underbrace{-\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}_t\boldsymbol{\Gamma}\boldsymbol{A}^{-\top}\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}^{\top})+\frac{1}{T}\sum_{t=1}^{T}\sum_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{A}^{-1}\boldsymbol{\Gamma}\boldsymbol{S}_t^{\top}\right)}_{I_{23}}.
\end{aligned}$$

Since $\|\boldsymbol{S}_t\| = O(T^{\alpha-1})$, it is easy to verify that $I_{21} = O(T^{2\alpha-2})\boldsymbol{\Gamma}$; meanwhile, theorem 6.9 indicates that $I_{22} = T^{\alpha-1}\boldsymbol{X}(\boldsymbol{\Lambda}^{\star}) + O(T^{2\alpha-2})\boldsymbol{\Lambda}^{\star}$. As for the term I_{23}, we observe that

$$\begin{aligned}
\left\|\sum_{t=1}^{T}\boldsymbol{S}_t\prod_{k=t}^{T}(\boldsymbol{I}-\eta_k\boldsymbol{A}^{\top})\right\| &\lesssim \widetilde{C}\sum_{t=1}^{T}t^{\alpha-1}\prod_{k=t}^{T}(1-\beta k^{-\alpha}) \\
&< \widetilde{C}T^{2\alpha-1}\sum_{t=1}^{T}t^{\alpha-1}\prod_{k=t+1}^{T}(1-\beta k^{-\alpha}) \\
&= \widetilde{C}T^{2\alpha-1}\cdot\frac{1}{\beta}\sum_{t=1}^{T}\left(\prod_{k=t+1}^{T}-\prod_{k=t}^{T}\right)(1-\beta k^{-\alpha}) \\
&< \widetilde{C}\frac{T^{2\alpha-1}}{\beta},
\end{aligned}$$

where $\widetilde{C}$ is a problem-related quantity related to η_0, α and γ. Consequently, it can be guaranteed that $I_{23} = O(T^{2\alpha-2})\boldsymbol{\Gamma}$. In combination, we have shown that

$$\begin{aligned}
\overline{\boldsymbol{\Lambda}}_T-\boldsymbol{\Lambda}^{\star} &= I_1+I_{21}+I_{22}+I_{23} \\
&= O(T^{-1})\boldsymbol{\Lambda}^{\star}+O(T^{2\alpha-2})\boldsymbol{\Gamma}+(T^{\alpha-1}\boldsymbol{X}(\boldsymbol{\Lambda}^{\star})+O(T^{2\alpha-2})\boldsymbol{\Lambda}^{\star})+O(T^{2\alpha-2})\boldsymbol{\Gamma} \\
&= T^{\alpha-1}\boldsymbol{X}(\boldsymbol{\Lambda}^{\star})+O(T^{2\alpha-2})\boldsymbol{\Gamma}+O(T^{2\alpha-2})\boldsymbol{\Lambda}^{\star};
\end{aligned}$$

the theorem follows naturally by combining this with the fact that $\mathbf{\Lambda}^\star = O(1)\mathbf{\Gamma}$, due to (34). □

The following Lemma controls the minimum eigenvalue of $\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top$.

Lemma 6.12. *With $\boldsymbol{A}$ defined as (5) and $\overline{\boldsymbol{\Lambda}}_T$ defined as in (52), when*

$$T \geq 4\left(\frac{2}{(1-\gamma)\lambda_0\eta_0}\right)^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma(\frac{1}{1-\alpha})\mathsf{cond}(\Gamma), \tag{67}$$

it can be guaranteed that

$$\lambda_{\min}(\boldsymbol{A}\mathbf{\Lambda}_T\boldsymbol{A}^\top) \geq \frac{1}{2}\lambda_{\min}(\mathbf{\Gamma}) = \frac{1}{2\|\mathbf{\Gamma}^{-1}\|}.$$

Proof: We firstly decompose $\boldsymbol{A}\mathbf{\Lambda}_T\boldsymbol{A}^\top$ as

$$\begin{aligned}\boldsymbol{A}\mathbf{\Lambda}_T\boldsymbol{A}^\top &= \boldsymbol{A}\mathbf{\Lambda}^\star\boldsymbol{A}^\top + \boldsymbol{A}(\mathbf{\Lambda}_T - \mathbf{\Lambda}^\star)\boldsymbol{A}^\top \\ &= \mathbf{\Gamma} + \boldsymbol{A}(\mathbf{\Lambda}_T - \mathbf{\Lambda}^\star)\boldsymbol{A}^\top;\end{aligned}$$

here, (63) indicates that the difference $(\mathbf{\Lambda}_T - \mathbf{\Lambda}^\star)$ is bounded above by

$$\mathbf{\Lambda}_T - \mathbf{\Lambda}^\star \succeq \frac{1}{T}\sum_{t=1}^{T}\left[(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})\mathbf{\Gamma}\boldsymbol{A}^{-\top} + \boldsymbol{A}^{-1}\mathbf{\Gamma}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})^\top\right].$$

Furthermore, (54a) reveals that

$$\begin{aligned}&\boldsymbol{A}(\mathbf{\Lambda}_T - \mathbf{\Lambda}^\star)\boldsymbol{A}^\top \\ &= \frac{1}{T}\boldsymbol{A}\sum_{t=1}^{T}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})\mathbf{\Gamma}\boldsymbol{A}^{-\top}\boldsymbol{A}^\top + \frac{1}{T}\boldsymbol{A}\boldsymbol{A}^{-1}\sum_{t=1}^{T}\mathbf{\Gamma}(\boldsymbol{Q}_t - \boldsymbol{A}^{-1})^\top\boldsymbol{A}^\top \\ &= -\frac{1}{T}\boldsymbol{A}\boldsymbol{A}^{-1}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\mathbf{\Gamma} - \frac{1}{T}\mathbf{\Gamma}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})^\top\boldsymbol{A}^{-\top}\boldsymbol{A}^\top \\ &= -\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\mathbf{\Gamma} - \frac{1}{T}\mathbf{\Gamma}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})^\top.\end{aligned}$$

It has been revealed in (65) that

$$\left\|\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\right\| \leq \left(\frac{2}{(1-\gamma)\lambda_0\eta_0}\right)^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma(\frac{1}{1-\alpha}).$$

Consequently, the triangle inequality yields

$$\begin{aligned}\lambda_{\min}(\boldsymbol{A}\mathbf{\Lambda}_T\boldsymbol{A}^\top) &\geq \lambda_{\min}(\mathbf{\Gamma}) - \left\|\frac{1}{T}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})\mathbf{\Gamma} + \frac{1}{T}\mathbf{\Gamma}\sum_{j=1}^{T}\prod_{k=1}^{j}(\boldsymbol{I} - \eta_k\boldsymbol{A})^\top\right\| \\ &\geq \lambda_{\min}(\mathbf{\Gamma}) - \frac{2}{T}\left(\frac{2}{(1-\gamma)\lambda_0\eta_0}\right)^{\frac{1}{1-\alpha}}(1-\alpha)^{\frac{\alpha}{1-\alpha}}\Gamma(\frac{1}{1-\alpha})\|\mathbf{\Gamma}\|.\end{aligned}$$

The lemma follows immediately by applying the condition (67). □

6.2 Gaussian approximation

This section is concerned with Gaussian approximation of random variables. We will firstly introduce some Berry-Esseen bounds and then illustrate some properties of the multi-dimensional Gaussian distribution.

6.2.1 Berry-Esseen bounds

The following theorem, due to Shao & Zhang (2022), gives a Berry-Esseen bound for non-linear statistics which would be crucial for our analysis of the non-asymptotic distribution of the TD estimation error.

Theorem 6.13. *(Theorem 2.1 of Shao & Zhang (2022)) Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be independent random variables, and let $\boldsymbol{T}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ be a d-dimensional statistic that can be represented as*

$$\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{D},$$

where $\boldsymbol{W}$ is the sum of independent random variables

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{\xi}_i = \sum_{i=1}^{n} \boldsymbol{h}_i(\boldsymbol{X}_i)$$

for Borel measurable functions $\{\boldsymbol{h}_i\}_{1\le i\le n}$ that satisfies

$$\mathbb{E}[\boldsymbol{\xi}_i] = 0, \text{ for each } 1 \le i \le n \text{ and } \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top] = \boldsymbol{\Lambda}.$$

Further assume that $\lambda_{\min}(\boldsymbol{\Lambda}) \ge \sigma > 0$. Let $\boldsymbol{z} \sim N(0, \boldsymbol{I}_d)$. Then for any convex set $\mathcal{A} \subseteq \mathbb{R}^d$,

$$d_{\mathsf{C}}(\boldsymbol{T}, \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{z}) \le 259\sqrt{d}\sigma^{-\frac{3}{2}} \cdot \sum_{i=1}^{n} \mathbb{E}\|\boldsymbol{\xi}_i\|_2^3 + 2\sigma^{-1}\mathbb{E}\{\|\boldsymbol{W}\|_2 \Delta\} + 2\sigma^{-1} \sum_{i=1}^{n} \mathbb{E}\{\|\boldsymbol{\xi}_i\|_2 |\Delta - \Delta^{(i)}|\}, \tag{68}$$

for any random variables Δ and $\{\Delta^{(i)}\}_{1\le i\le n}$ such that $\Delta \ge \|D\|_2$ and $\Delta^{(i)}$ is independent of $\boldsymbol{X}_i$.

6.2.2 Properties of multivariate Gaussian distributions

The following theorem by Devroye et al. (2018) gives an upper bound for the total variation (TV) distance between two Gaussian random variables with same means and different covariance matrices.

Theorem 6.14 (Devroye et al. (2018))**.** *Let $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \in \mathbb{S}^{d\times d}$ be two positive definite matrices, and $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^d$. Then the TV distance between Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1)$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_2)$ is bounded by*

$$\begin{aligned}
&\min\left\{1, \frac{1}{100}\left\|\boldsymbol{\Lambda}_1^{-1/2}\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1^{-1/2} - \boldsymbol{I}_d\right\|_{\mathsf{F}}\right\} \\
&\le d_{\mathsf{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_2)) \\
&\le \frac{3}{2}\left\|\boldsymbol{\Lambda}_1^{-1/2}\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1^{-1/2} - \boldsymbol{I}_d\right\|_{\mathsf{F}}.
\end{aligned}$$

In the case of $\boldsymbol{\Lambda}_1 \succ \boldsymbol{\Lambda}_2$, the following proposition further specifies that the TV distance and the distance on convex sets are the same between $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_1)$ and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_2)$.

Proposition 6.15. *Let $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \in \mathbb{S}^{d\times d}$ be two positive definite matrices, and $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^d$. When $\boldsymbol{\Lambda}_1 \succ \boldsymbol{\Lambda}_2$, it can be guaranteed that*

$$d_{\mathsf{C}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_2)) = d_{\mathsf{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}_2)).$$

Proof: Without loss of generality, let $\boldsymbol{\mu} = \boldsymbol{0}$. Define $\mathcal{C}$ as

$$\mathcal{C} := \{\boldsymbol{x} \in \mathbb{R}^d \mid p_1(\boldsymbol{x}) < p_2(\boldsymbol{x})\},$$

where $p_1(\cdot)$ and $p_2(\cdot)$ denote the probability density functions of $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_1)$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_2)$ respectively; we aim to show that $\mathcal{C}$ is a convex set. Notice that the ratio between these two density functions is featured by

$$\begin{aligned}\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} &= \frac{\frac{1}{\sqrt{\det(\mathbf{\Lambda_1})}}\exp(-\boldsymbol{x}^\top \mathbf{\Lambda_1}^{-1}\boldsymbol{x})}{\frac{1}{\sqrt{\det(\mathbf{\Lambda_2})}}\exp(-\boldsymbol{x}^\top \mathbf{\Lambda_2}^{-1}\boldsymbol{x})} \\ &= \sqrt{\frac{\det(\mathbf{\Lambda_2})}{\det(\mathbf{\Lambda_1})}}\exp(\boldsymbol{x}^\top(\mathbf{\Lambda_2}^{-1} - \mathbf{\Lambda_1}^{-1})\boldsymbol{x}).\end{aligned}$$

Therefore by definition, the set $\mathcal{C}$ is characterized by

$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}^d \middle| \boldsymbol{x}^\top(\mathbf{\Lambda_2}^{-1} - \mathbf{\Lambda_1}^{-1})\boldsymbol{x} < \frac{1}{2}\log\frac{\det(\mathbf{\Lambda}_1)}{\det(\mathbf{\Lambda}_2)}\right\}.$$

Since $\mathbf{\Lambda_1} \succ \mathbf{\Lambda}_2$, the matrix $(\mathbf{\Lambda_2}^{-1} - \mathbf{\Lambda_1}^{-1})$ must be positive definite; consequently, $\mathcal{C}$ is a d-dimensional hyper-ellipsoid and is therefore convex. With $\mathcal{C}$ being convex, it is seen that

$$\begin{aligned}&\left|\mathbb{P}_{z\sim\mathcal{N}(\mathbf{0},\mathbf{\Lambda}_1)}(\boldsymbol{z}\in\mathcal{C}) - \mathbb{P}_{z\sim\mathcal{N}(\mathbf{0},\mathbf{\Lambda}_2)}(\boldsymbol{z}\in\mathcal{C})\right| \\ &\le \sup_{\mathcal{A}\in\mathscr{C}_d}\left|\mathbb{P}_{z\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_1)}(\boldsymbol{z}\in\mathcal{A}) - \mathbb{P}_{z\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_2)}(\boldsymbol{z}\in\mathcal{A})\right| \\ &= d_{\mathsf{C}}(\mathcal{N}(\mathbf{0},\mathbf{\Lambda}_1), \mathcal{N}(\mathbf{0},\mathbf{\Lambda}_2)),\end{aligned}$$

while in the meantime

$$\begin{aligned}\left|\mathbb{P}_{z\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_1)}(\boldsymbol{z}\in\mathcal{C}) - \mathbb{P}_{z\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_2)}(\boldsymbol{z}\in\mathcal{C})\right| &= d_{\mathsf{TV}}(\mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_1), \mathcal{N}(\boldsymbol{\mu},\mathbf{\Lambda}_2)) \\ &\ge d_{\mathsf{C}}(\mathcal{N}(\mathbf{0},\mathbf{\Lambda}_1), \mathcal{N}(\mathbf{0},\mathbf{\Lambda}_2)).\end{aligned}$$

Combining these two inequalities completes the proof. □

6.3 Concentration inequalities

The following Azuma's inequality for matrix martingales is shown in Tropp (2011).

Theorem 6.16. *Let $\{\boldsymbol{X}_k\}$ be a random sequence of matrices in $\mathbb{S}^{d\times d}$, and $\{\boldsymbol{A}_k\}$ be a sequence of fixed matrices in $\mathbb{S}^{d\times d}$. Assume that $\mathbb{E}_{k-1}[\boldsymbol{X}_k] = \mathbf{0}$ and $\boldsymbol{X}_k^2 \preceq \boldsymbol{A}_k^2$ almost surely. Define*

$$\sigma^2 := \left\|\sum_k \boldsymbol{A}_k^2\right\|,$$

then for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left\|\sum_k \boldsymbol{X}_k\right\| \ge \varepsilon\right) \le d\exp\left(-\frac{\varepsilon^2}{8\sigma^2}\right).$$

Specifically, for vector-valued martingales, the following proposition is shown by Kallenberg & Sztencel (1991).

Proposition 6.17. *Let $\{\boldsymbol{x}_i\}_{i\ge1}$ be a martingale in $\mathbb{R}^d$. Then there exists a martingale $\{\boldsymbol{y}_i\}_{i\ge1}$ in $\mathbb{R}^2$ such that for every $i > 1$, $\|\boldsymbol{y}_i\|_2$ has the same distribution as $\|\boldsymbol{x}_i\|_2$ and $\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|_2$ has the same distribution as $\|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2$.*

The following Azuma's inequality on vector-valued martingales comes as a direct corollary.

Corollary 6.18. *Let $\{\boldsymbol{x}_i\}_{i\geq 1}$ be a martingale in $\mathbb{R}^d$, and let $W_{\max}$ be a positive constant that satisfies*

$$W_{\max} \geq \sum_{i=1}^{t} \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2^2 \quad \textit{almost surely.}$$

Then it can be guaranteed with probability at least $1-\delta$ that

$$\|\boldsymbol{x}_t\|_2 \leq 2\sqrt{2W_{\max}\log\frac{3}{\delta}}.$$

The following Freedman's inequality on Hilbert spaces was shown by Peng et al. (2024) in their Theorem 2.1.

Theorem 6.19. *Let $\mathcal{X}$ be a Hilbert space, $\{\boldsymbol{X}_i\}_{i=1}^n$ be an $\mathcal{X}$-valued martingale difference sequence adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^n$, $\boldsymbol{Y}_i := \sum_{j=1}^i \boldsymbol{X}_j$ be the corresponding martingale, and $W_i = \sum_{j=1}^i \mathbb{E}\|\boldsymbol{X}_j\|^2 \mid \mathscr{F}_{j-1}$. Suppose that $\max_{i\in[n]}\|\boldsymbol{X}_i\| \leq b$ for some constant $b>0$. Then for any ε and $\sigma>0$, it can be guaranteed that*

$$\mathbb{P}\left(\exists k\in[n],\ \textit{s.t. } \|\boldsymbol{Y}_k\|\geq\varepsilon \textit{ and } W_k\leq\sigma^2\right) \leq 2\exp\left(-\frac{\varepsilon^2/2}{\sigma^2+b\varepsilon/3}\right).$$

7 Proof of main results

7.1 Proof of Theorem 3.2

In order to derive the high-probability convergence guarantee for $\overline{\boldsymbol{\Delta}}_T$, we firstly propose the following result on the time-uniform convergence of the original TD estimator $\boldsymbol{\Delta}_t$.

7.1.1 High-probability convergence guarantee for the original TD estimation error

Theorem 7.1. *Consider the TD algorithm (8) with independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$, where $\alpha\in(\frac{1}{2},1)$. For any $\delta\in(0,1)$, there exists $\eta_0>0$, such that with probability at least $1-\delta$, the norm of $\boldsymbol{\Delta}_t$ is bounded by*

$$\|\boldsymbol{\Delta}_t\|_2 \lesssim \frac{\eta_0}{\sqrt{2\alpha-1}}\left(\frac{(1-\gamma)\lambda_0\eta_0}{4\alpha}\right)^{-\frac{\alpha}{2(1-\alpha)}}\sqrt{\log\frac{9T}{\delta}}(2\|\boldsymbol{\theta}^\star\|_2+1)t^{-\frac{\alpha}{2}} \tag{69}$$

simultaneously for all $t\in[T]$.

Proof: Under the TD update rule (8), we represent $\boldsymbol{\Delta}_t$ as

$$\begin{aligned}\boldsymbol{\Delta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star &= (\boldsymbol{\theta}_{t-1} - \eta_t(\boldsymbol{A}_t\boldsymbol{\theta}_{t-1}-\boldsymbol{b}_t)) - \boldsymbol{\theta}^\star \\ &= \boldsymbol{\Delta}_{t-1} - \eta_t(\boldsymbol{A}\boldsymbol{\theta}_{t-1}-\boldsymbol{b}+\boldsymbol{\zeta}_t) \\ &= (\boldsymbol{I}-\eta_t\boldsymbol{A})\boldsymbol{\Delta}_{t-1} - \eta_t\boldsymbol{\zeta}_t,\end{aligned}$$

where $\boldsymbol{\zeta}_t$ is defined as

$$\boldsymbol{\zeta}_t = (\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}_{t-1} - (\boldsymbol{b}_t-\boldsymbol{b}). \tag{70}$$

Therefore by induction, $\boldsymbol{\Delta}_t$ can be expressed as a weighted sum of $\{\boldsymbol{\zeta}_i\}_{0\leq i<t}$, namely

$$\begin{aligned}\boldsymbol{\Delta}_t &= \prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\Delta_0 - \sum_{i=1}^{t}\boldsymbol{R}_i^t\boldsymbol{\zeta}_i \\ &= \prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\Delta_0 - \sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i) - \sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1}\end{aligned} \tag{71}$$

in which the matrices $\{\boldsymbol{R}_i^t\}_{1\le i\le t}$ are defined as

$$\boldsymbol{R}_i^t = \eta_i \prod_{k=i+1}^{t} (\boldsymbol{I} - \eta_k \boldsymbol{A}). \tag{72}$$

For simplicity, in this proof we denote

$$\beta = \frac{1-\gamma}{2}\lambda_0\eta_0, \quad R = \frac{5}{2}\|\boldsymbol{\theta}^\star\|_2 + \frac{1}{4}, \quad \text{and}$$
$$C = \frac{2\alpha-1}{512\eta_0^2}\left(\frac{\beta}{2\alpha}\right)^{\frac{\alpha}{1-\alpha}}.$$

For any given δ, we define $t^\star$ as a function of δ, specifically

$$t^\star(\delta) = \inf\left\{t \in \mathbb{N}^+ : \int_t^\infty \exp(-Cx^\alpha)\mathrm{d}x \le \frac{\delta}{9}\right\}, \tag{73}$$

and assume that

$$16\eta_0\sqrt{\frac{2}{2\alpha-1}}\sqrt{\log\frac{9t^\star}{\delta}} \le 1. \tag{74}$$

Furthermore, we denote, for simplicity,

$$\mathcal{H}_i = \{\max_{1\le j\le i}\|\boldsymbol{\Delta}_j\| \le R\}, \quad \text{and} \quad \tilde{\boldsymbol{\zeta}}_i = \boldsymbol{\zeta}_i \mathbb{1}(\mathcal{H}_{i-1}).$$

It can then be easily verified that

$$\|\tilde{\boldsymbol{\zeta}}_i\|_2 \le 4R + 2\|\boldsymbol{\theta}^\star\|_2 + 1.$$

In what follows, we prove the desired result in the following 4 steps:

1. Showing the existence of $t^\star(\delta)$ and featuring its relationship with δ;
2. Showing that $\mathbb{P}(\mathcal{H}_{t^\star}) \ge 1 - \frac{1}{3}\delta$;
3. Showing that $\mathbb{P}(\mathcal{H}_\infty) \ge 1 - \frac{2}{3}\delta$;
4. Showing the high-probability bound (69).

Step 1: the existence of $t^\star(\delta)$. By definition, $t^\star$ depends on the integral

$$\int_t^\infty \exp(-Cx^\alpha)\mathrm{d}x,$$

which we handle with a change of variable

$$u = x^\alpha.$$

It is easy to verify that

$$\mathrm{d}x = \frac{1}{\alpha}u^{\frac{1-\alpha}{\alpha}}\mathrm{d}u,$$

and hence

$$\int_t^\infty \exp(-Cx^\alpha)\mathrm{d}x = \frac{1}{\alpha}\int_{t^\alpha}^\infty u^{\frac{1-\alpha}{\alpha}}\exp(-Cu)\mathrm{d}u.$$

since $\alpha > 0.5$, it can be guaranteed that $u^{\frac{1-\alpha}{\alpha}} < u$, for any $u > t^\alpha \geq 1$. Hence, this integral is bounded by

$$\begin{aligned}\int_t^\infty \exp(-Cx^\alpha)\mathrm{d}x &< \frac{1}{\alpha}\int_{t^\alpha}^\infty u\exp(-Cu)\mathrm{d}u \\ &= \frac{1}{\alpha C^2}(Cu+1)\exp(-Cu).\end{aligned}$$

Consequently, in order that

$$\int_t^\infty \exp(-Cx^\alpha)\mathrm{d}x \leq \frac{\delta}{9},$$

it suffices to let

$$(Ct^\alpha+1)\exp(-Ct^\alpha) \leq \frac{\alpha C^2\delta}{9};$$

It is easy to verify that this condition is satisfied by taking

$$t = \left[\frac{2}{C}\left(\log\frac{9}{\alpha C^2\delta} \vee \frac{5}{2}\right)\right]^{\frac{1}{\alpha}}.$$

In other words, $t^\star(\delta)$ is a well-defined function of δ that scales as $O\left(\log^\alpha \frac{1}{\delta}\right)$ as $\delta \to 0$.

Step 2: Bounding $\mathbb{P}(\mathcal{H}_{t^\star})$. We aim to prove, through an induction argument, that

$$\mathbb{P}(\mathcal{H}_t) \geq 1 - \frac{t}{t^\star}\cdot\frac{\delta}{3}. \tag{75}$$

Since $\boldsymbol{\theta}_0 = 0$, the norm of $\boldsymbol{\Delta}_0$ is equal to that of $\boldsymbol{\theta}^\star$, and therefore $\mathbb{P}(\mathcal{H}_0) = 1$ is trivial. When (75) holds for $t-1$, we consider the probability of $\mathcal{H}_t$. We firstly observe that

$$\begin{aligned}\left\|\sum_{i=1}^t \boldsymbol{R}_i^t\tilde{\boldsymbol{\zeta}}_i\right\|_2^2 &\leq \sum_{i=1}^t \|\boldsymbol{R}_i^t\|^2\|\tilde{\boldsymbol{\zeta}}_i\|_2^2 \\ &< \eta_0^2(4R+2\|\boldsymbol{\theta}^\star\|_2+1)^2\underbrace{\sum_{i=1}^t i^{-2\alpha}\prod_{k=1}^t(1-\beta k^{-\alpha})}_{W_t}.\end{aligned}$$

The term W_t is bounded by $\sum_{i=1}^t i^{-2\alpha} < \frac{1}{2\alpha-1}$. Therefore, the vectors Azuma's inequality (Corollary 6.18) shows that, with probability at least $1-\frac{\delta}{3t^\star}$,

$$\begin{aligned}\left\|\sum_{i=1}^t \boldsymbol{R}_i^t\tilde{\boldsymbol{\zeta}}_i\right\|_2 &< 2\eta_0(4R+2\|\boldsymbol{\theta}^\star\|_2+1)\sqrt{\frac{2}{2\alpha-1}\log\frac{9t^\star}{\delta}} \\ &\leq \frac{1}{8}(4R+2\|\boldsymbol{\theta}^\star\|_2+1),\end{aligned}$$

where we invoked the condition (74) in the second inequality. The triangle inequality and a union bound argument yield that

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| &\leq \left\|\prod_{k=1}^t(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{\Delta}_0\right\|_2 + \left\|\sum_{i=1}^t \boldsymbol{R}_i^t\boldsymbol{\zeta}_i\right\|_2 \\ &= \left\|\prod_{k=1}^t(\boldsymbol{I}-\eta_k\boldsymbol{A})\boldsymbol{\Delta}_0\right\|_2 + \left\|\sum_{i=1}^t \boldsymbol{R}_i^t\tilde{\boldsymbol{\zeta}}_i\right\|_2 \\ &< \|\boldsymbol{\theta}^\star\|_2 + \frac{1}{8}(4R+2\|\boldsymbol{\theta}^\star\|_2+1) = R,\end{aligned}$$

with probability at least $\mathbb{P}(\mathcal{H}_{t-1}) - \frac{\delta}{3t^\star} > 1 - \frac{t}{t^\star}\cdot\frac{\delta}{3}$. The induction assumption then guarantees $\mathbb{P}(\mathcal{H}_{t^\star}) \geq 1-\frac{\delta}{3}$.

Step 3: Bounding $\mathbb{P}(\mathcal{H}_\infty)$. When $t > t^\star$, we invoke a tighter bound on W_t to bound the norm of $\boldsymbol{\Delta}_t$. Namely, Lemma 6.6 indicates that

$$W_t \le \frac{1}{2\alpha - 1}\left(\frac{2\alpha}{\beta}\right)^{\frac{\alpha}{1-\alpha}} t^{-\alpha}.$$

Furthermore, define

$$\delta_t = 3\exp\left(-\frac{2\alpha-1}{512\eta_0^2}\left(\frac{\beta}{2\alpha}\right)^{\frac{\alpha}{1-\alpha}} t^\alpha\right) = 3\exp(-Ct^\alpha);$$

it can then be seen that with probability at least $1-\delta_t$,

$$\begin{aligned}\left\|\sum_{i=1}^t \boldsymbol{R}_i^t \tilde{\boldsymbol{\zeta}}_i\right\|_2 &\le 2\eta_0(4R + 2\|\boldsymbol{\theta}^\star\|_2 + 1)\sqrt{2W_t \log\frac{3}{\delta_t}} \\ &\le 2\eta_0(4R + 2\|\boldsymbol{\theta}^\star\|_2 + 1)\cdot\frac{1}{16} = \frac{1}{2}R + \frac{1}{4}(2\|\boldsymbol{\theta}^\star\|_2 + 1).\end{aligned}$$

With a similar reasoning as the first step, the triangle inequality then indicates

$$\mathbb{P}(\mathcal{H}_t) \ge \mathbb{P}(\mathcal{H}_{t-1}) - \delta_t.$$

Hence, the probability of $\mathcal{H}_\infty$ is bounded by

$$\begin{aligned}\mathbb{P}(\mathcal{H}_\infty) &\ge \mathbb{P}(\mathcal{H}_{t^\star}) - \sum_{t=t^\star+1}^\infty \delta_t \\ &\ge \left(1 - \frac{\delta}{3}\right) - \sum_{t=t^\star+1}^\infty 3\exp(-Ct^\alpha) \\ &\ge \left(1 - \frac{\delta}{3}\right) - \int_{t^\star}^\infty 3\exp(-Cx^\alpha)\mathrm{d}x \\ &\ge \left(1 - \frac{\delta}{3}\right) - \frac{\delta}{3} = 1 - \frac{2}{3}\delta,\end{aligned}$$

where we applied the definition of $t^\star$ at the last inequality.

Step 4: Refining the bound on $\|\boldsymbol{\Delta}_t\|_2$. With $\mathbb{P}(\mathcal{H}_\infty)$ being bounded, we now proceed to refining the bound on $\|\boldsymbol{\Delta}_t\|_2$ for any $t \in \mathbb{N}$. Specifically, we observe, due to Lemma 6.6, that for every $t \in [T]$,

$$\left\|\prod_{k=1}^t (\boldsymbol{I} - \eta_k \boldsymbol{A})\boldsymbol{\Delta}_0\right\|_2 \le \exp\left(-\frac{\beta}{1-\alpha}t^{1-\alpha}\right)\|\boldsymbol{\Delta}_0\|_2 \le \left(\frac{\alpha}{2e\beta}\right)^{\frac{\alpha}{2(1-\alpha)}}\|\boldsymbol{\theta}^\star\|_2 t^{-\frac{\alpha}{2}}.$$

Meanwhile, the vector Azuma's inequality (Corollary 6.18) implies with probability at least $1 - \frac{\delta}{3T}$,

$$\left\|\sum_{i=1}^t \boldsymbol{R}_i^t \tilde{\boldsymbol{\zeta}}_i\right\|_2 \le 2\eta_0\left(\frac{2\alpha}{\beta}\right)^{\frac{\alpha}{2(1-\alpha)}}\sqrt{\frac{2}{2\alpha-1}\log\frac{9T}{\delta}}(4R + 2\|\boldsymbol{\theta}^\star\|_2 + 1)t^{-\frac{\alpha}{2}}.$$

The bound (69) then follows by invoking the triangle inequality and a union bound, and plugging in the definition of R and β. □

7.1.2 Completing the proof of Theorem 3.2

Similar to the proof of Theorem 7.1, we firstly expand $\overline{\boldsymbol{\Delta}}_T$ into three terms: a non-random term capturing the error introduced by the initialization $\boldsymbol{\theta}_0$, a martingale, and a sum of $i.i.d.$ vectors. Specifically, (71) implies

$$\begin{aligned}
\overline{\boldsymbol{\Delta}}_T &= \frac{1}{T}\sum_{t=1}^{T}\left(\prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\Delta_0 - \sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i) - \sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1}\right)\\
&= \frac{1}{T}\sum_{t=1}^{T}\prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\Delta_0 - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i) - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{t}\boldsymbol{R}_i^t(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1}\\
&= \frac{1}{T}\sum_{t=1}^{T}\prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\Delta_0 - \frac{1}{T}\sum_{i=1}^{T}\sum_{t=i}^{T}\boldsymbol{R}_i^t(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i) - \frac{1}{T}\sum_{i=1}^{T}\sum_{t=i}^{T}\boldsymbol{R}_i^t(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1},
\end{aligned}$$

where we change the order of summation in the last equation. Hence, the definition of $\boldsymbol{R}_i^t$ (72) and $\boldsymbol{Q}_t$ (53) indicates that

$$\begin{aligned}
&\sum_{t=1}^{T}\prod_{k=1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A}) = \frac{\boldsymbol{Q}_0}{\eta_0}, \quad \text{and}\\
&\sum_{t=i}^{T}\boldsymbol{R}_i^t = \sum_{t=i}^{T}\left(\eta_i\prod_{k=i+1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A})\right) = \eta_i\sum_{t=i}^{T}\prod_{k=i+1}^{t}(\boldsymbol{I}-\eta_k\boldsymbol{A}) = \boldsymbol{Q}_i.
\end{aligned}$$

Consequently, $\overline{\boldsymbol{\Delta}}_T$ can be expressed as

$$\overline{\boldsymbol{\Delta}}_T = \underbrace{\frac{1}{T\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0}_{I_1} - \underbrace{\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)}_{I_2} - \underbrace{\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1}}_{I_3}. \tag{76}$$

Here, I_1 is a non-random term that captures the error introduced by $\boldsymbol{\Delta}_0$, I_2 is the sum of $i.i.d.$ random vectors, and I_3 is a martingale. We address the bounds on these terms in the following paragraphs in order. Through this proof, we denote, for simplicity, β as

$$\beta := \frac{1-\gamma}{2}\lambda_0\eta_0$$

Bounding the norm of I_1. The norm of I_1 is bounded by the universal control over $\|\boldsymbol{Q}_t\|$ as indicated by Lemma 6.10. Specifically,

$$\left\|\frac{1}{T\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0\right\|_2 \le \frac{1}{T\eta_0}\|\boldsymbol{Q}_0\|\|\boldsymbol{\Delta}_0\| \le \frac{3}{T}\left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}}\|\boldsymbol{\Delta}_0\|_2. \tag{77}$$

Bounding the norm of I_2. The term I_2 is a sum of $i.i.d.$ random vectors, and can hence be bounded by invoking the Freedman's inequality on Hilbert spaces (Theorem 6.19). Specifically, define

$$\begin{aligned}
W_{\max} &:= \sum_{i=1}^{T}\mathbb{E}\left\|\frac{1}{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\right\|_2^2, \quad \text{and}\\
B_{\max} &:= \max_{1\le i\le T}\sup\left\|\frac{1}{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\right\|_2.
\end{aligned}$$

It can then be guaranteed with probability at least $1-\delta$ that

$$\left\|\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\right\|_2 \le 2\sqrt{W_{\max}\log\frac{2}{\delta}} + \frac{4}{3}B_{\max}\log\frac{2}{\delta}.$$

Furthermore, the terms $W_{\max}$ and $B_{\max}$ can be controlled by

$$\begin{aligned}
W_{\max} &= \frac{1}{T^2}\sum_{i=1}^{T}\mathbb{E}\|\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\|_2^2\\
&= \frac{1}{T^2}\sum_{i=1}^{T}\mathbb{E}\left[(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)^\top\boldsymbol{Q}_i^\top\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\right]\\
&= \frac{1}{T^2}\sum_{i=1}^{T}\mathbb{E}\left[\mathsf{Tr}\left(\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)^\top\boldsymbol{Q}_i^\top\right)\right]\\
&= \frac{1}{T^2}\sum_{i=1}^{T}\mathsf{Tr}\left(\boldsymbol{Q}_i\mathbb{E}\left[(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)^\top\right]\boldsymbol{Q}_i^\top\right)\\
&= \frac{1}{T}\mathsf{Tr}(\overline{\boldsymbol{\Lambda}}_T);
\end{aligned}$$

$$\text{and}\qquad B_{\max}\le\frac{1}{T}\sup_{1\le i\le T}\|\boldsymbol{Q}_i\|\sup\|\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i\|_2\le\frac{3\eta_0}{T}\left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}}(2\|\boldsymbol{\theta}^\star\|_2+1).$$

Hence, the norm of I_2 is bounded by

$$\left\|\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star-\boldsymbol{b}_i)\right\|_2\le 2\sqrt{\frac{\mathsf{Tr}(\overline{\boldsymbol{\Lambda}}_T)}{T}\log\frac{6}{\delta}}+\frac{4\eta_0}{T}\left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}}(2\|\boldsymbol{\theta}^\star\|_2+1)\log\frac{6}{\delta}. \tag{78}$$

with probability at least $1-\frac{\delta}{3}$.

Bounding the norm of I_3. The term I_3 is a martingale, to which the matrix Azuma's inequality is applicable. Towards this end, we need the universal bound on $\{\|\boldsymbol{\Delta}_{i-1}\|_2\}_{1\le i\le T}$, for which Theorem 7.1 comes in handy. Specifically, under the condition (74), it can be guaranteed with probability at least $1-\frac{\delta}{3}$ that

$$\|\boldsymbol{\Delta}_t\|_2\lesssim\eta_0\left(\frac{2\alpha}{\beta}\right)^{\frac{\alpha}{2(1-\alpha)}}\sqrt{\frac{2}{2\alpha-1}\log\frac{27T}{\delta}}(2\|\boldsymbol{\theta}^\star\|_2+1)t^{-\frac{\alpha}{2}},\quad\forall t\in[T].$$

For simplicity, we denote

$$\begin{aligned}
R &= \eta_0\left(\frac{2\alpha}{\beta}\right)^{\frac{\alpha}{2(1-\alpha)}}\sqrt{\frac{2}{2\alpha-1}\log\frac{27T}{\delta}}(2\|\boldsymbol{\theta}^\star\|_2+1),\\
\mathcal{H}_t &= \left\{\max_{1\le j\le t}j^{\frac{\alpha}{2}}\|\boldsymbol{\Delta}_j\|_2\le R\right\},\quad\text{and}\\
\tilde{\boldsymbol{\Delta}}_t &= \boldsymbol{\Delta}_t\mathbb{1}(\mathcal{H}_t).
\end{aligned}$$

Then by definition, the probability of $\mathcal{H}_t$ is bounded by

$$\mathbb{P}(\mathcal{H}_1)\ge\mathbb{P}(\mathcal{H}_2)\ge\ldots\ge\mathbb{P}(\mathcal{H}_T)\ge 1-\frac{\delta}{3}. \tag{79}$$

Meanwhile, the vector Azuma's inequality can be invoked to obtain that with probability at least $1-\frac{\delta}{3}$,

$$\left\|\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i-\boldsymbol{A})\tilde{\boldsymbol{\Delta}}_{i-1}\right\|_2\le 2\sqrt{2W_{\max}\log\frac{9}{\delta}},$$

where

$$W_{\max} = \sup \frac{1}{T^2} \sum_{i=1}^{T} \|\boldsymbol{Q}_i(\boldsymbol{A}_i - \boldsymbol{A})\tilde{\boldsymbol{\Delta}}_{i-1}\|_2^2 \lesssim \frac{1}{T^2} 3\eta_0^2 \left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}} \cdot 4R^2 \sum_{i=1}^{T} i^{-\alpha}$$
$$\asymp \frac{12\eta_0^2}{1-\alpha} \left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}} R^2 T^{-1-\alpha}.$$

In other words, it can be guaranteed with probability at least $1 - \frac{\delta}{3}$ that

$$\left\| \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{Q}_i(\boldsymbol{A}_i - \boldsymbol{A})\tilde{\boldsymbol{\Delta}}_{i-1} \right\|_2 \lesssim \eta_0 \left(\frac{2}{\beta}\right)^{\frac{1}{2(1-\alpha)}} R\sqrt{\frac{1}{1-\alpha} \log \frac{9}{\delta}}$$
$$\lesssim \eta_0^2 \left(\frac{2}{\beta}\right)^{\frac{2+\alpha}{2(1-\alpha)}} (6\|\boldsymbol{\theta}^\star\|_2 + 1)\sqrt{\frac{1}{(1-\alpha)(2\alpha-1)}} T^{-\frac{\alpha+1}{2}} \log \frac{27T}{\delta}. \tag{80}$$

Combining (76), (77), (78), (79) and (80) with the triangle inequality and a union bound argument, we obtain with probability $1 - \delta$ that

$$\|\overline{\boldsymbol{\Delta}}_T\|_2 \leq \frac{3}{T} \left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}} \|\boldsymbol{\Delta}_0\|_2$$
$$+ 2\sqrt{\frac{\mathsf{Tr}(\overline{\boldsymbol{\Lambda}}_T)}{T} \log \frac{6}{\delta}} + \frac{4\eta_0}{T} \left(\frac{2}{\beta}\right)^{\frac{1}{1-\alpha}} (2\|\boldsymbol{\theta}^\star\|_2 + 1) \log \frac{6}{\delta}$$
$$+ \eta_0^2 \left(\frac{2}{\beta}\right)^{\frac{2+\alpha}{2(1-\alpha)}} (6\|\boldsymbol{\theta}^\star\|_2 + 1)\sqrt{\frac{1}{(1-\alpha)(2\alpha-1)}} T^{-\frac{\alpha+1}{2}} \log \frac{27T}{\delta}.$$

In the mean time, since the difference between $\overline{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$ is bounded by

$$\mathsf{Tr}(\overline{\boldsymbol{\Lambda}}_T) - \mathsf{Tr}(\boldsymbol{\Lambda}^\star) \lesssim T^{\alpha-1} \mathsf{Tr}(\boldsymbol{X}(\boldsymbol{\Lambda}^\star)) \lesssim \frac{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)}{\beta} T^{\alpha-1},$$

Theorem 3.2 follows by taking

$$\tilde{C} = \eta_0^2 \left(\frac{2}{\beta}\right)^{\frac{2+\alpha}{2(1-\alpha)}} (6\|\boldsymbol{\theta}^\star\|_2 + 1)\sqrt{\frac{1}{(1-\alpha)(2\alpha-1)}}. \tag{81}$$

7.1.3 The necessity of the condition (74)

An important fact to note about Theorem 3.2 is that it relies on choosing an appropriate initial stepsize η_0 dependent of δ. The relationship is specified in Equation (74). This constraint, albeit only related to δ in logarithm terms, prevents δ to converge to 0 for any *fixed* η_0. One might naturally wonder if such a constraint could be relaxed and the convergence of $\overline{\boldsymbol{\Delta}}_T$ can be guaranteed with high probability for any fixed η_0 and arbitrarily small δ. The following proposition shows that the answer is *negative*, so some form of constraint on δ is necessary for a Chernoff-style convergence guarantee.

Proposition 7.2. *Consider the TD algorithm (8) with independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$, where η_0 is fixed and $\alpha \in (\frac{1}{2}, 1)$. There exists an MDP, such that for any bi-variate polynomial function* $\mathsf{poly}(\cdot, \cdot)$*, there exists $\delta \in (0, 1)$ and $T \in \mathbb{N}$, such that*

$$\mathbb{P}\left(\|\overline{\boldsymbol{\Delta}}_T\|_2 \geq \mathsf{poly}\left(T, \log \frac{1}{\delta}\right) \right) \geq \delta.$$

Proof: Consider the following MDP, where $|\mathcal{S}| = 3$, $\gamma = 0.9$, while the transition matrix $\boldsymbol{P}$ and the reward vector $\boldsymbol{r}$ are

$$\boldsymbol{P} = \begin{pmatrix} 0.1 & 0.1 & 0.8 \\ 0.1 & 0.1 & 0.8 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{r} = \begin{pmatrix} 0.1 \\ 0.1 \\ 1 \end{pmatrix}.$$

It is easy to verify that $\boldsymbol{\mu} = (0.1, 0.1, 0.8)^\top$. Furthermore, let $d = 1$ [3], $\boldsymbol{\phi}(1) = -0.5$,$\boldsymbol{\phi}(2) = -1$ and $\boldsymbol{\phi}(3) = 1$. Direct calculation then yields

$$\boldsymbol{A} = 0.54475, \quad \boldsymbol{b} = 0.785, \quad \text{and} \quad \boldsymbol{\theta}^\star = 1.441.$$

For any $T \in \mathbb{N}$, use $\mathcal{E}_T$ to denote the event that $s_t = 1$ and $s_t' = 2$ for all $t \in [T]$. In the case of $i.i.d.$ samples, clearly $\mathbb{P}(\mathcal{E}_T) = 0.01^T$. For any bi-variate polynomial function $\mathsf{poly}(\cdot, \cdot)$, let $\delta = 0.01^T$, then $\log \frac{1}{\delta} = T \log 100$. Therefore, $\mathsf{poly}(T, \log \frac{1}{\delta}) = \mathsf{poly}(T, T \log 100)$ reduces to a univariate polynomial function of T. It suffices to show that under the event $\mathcal{E}_T$, the norm of $\overline{\boldsymbol{\Delta}}_T$ grows faster than any polynomial function of T.

Under the event $\mathcal{E}_T$, direct calculation yields $\boldsymbol{A}_t = -0.2$ and $\boldsymbol{b}_t = -0.05$. Hence, the estimation error $\boldsymbol{\Delta}_t$ is featured by the iterative relation

$$\begin{aligned} \boldsymbol{\Delta}_t &= \boldsymbol{\Delta}_{t-1} - \eta_t(\boldsymbol{A}_t \boldsymbol{\Delta}_{t-1} + \boldsymbol{A}_t \boldsymbol{\theta}^\star - \boldsymbol{b}_t) \\ &= (1 + 0.2\eta_t)\boldsymbol{\Delta}_{t-1} + (0.2\boldsymbol{\theta}^\star - 0.05)\eta_t. \end{aligned}$$

It is easy to show by induction that, when $\boldsymbol{\theta}_0 = 0$ and hence $\boldsymbol{\Delta}_0 = -\boldsymbol{\theta}^\star$,

$$\begin{aligned} \boldsymbol{\Delta}_t &= (0.2\boldsymbol{\theta}^\star - 0.05) \sum_{i=1}^{t} \eta_i \prod_{k=i+1}^{t} (1 + 0.2\eta_k) + \prod_{k=1}^{t} (1 + 0.2\eta_k)\boldsymbol{\Delta}_0 \\ &= 5(0.2\boldsymbol{\theta}^\star - 0.05) \sum_{i=1}^{t} \left(\prod_{k=i}^{t} - \prod_{k=i+1}^{t} \right) (1 + 0.2\eta_k) + \prod_{k=1}^{t} (1 + 0.2\eta_k)(-\boldsymbol{\theta}^\star) \\ &= 5(0.2\boldsymbol{\theta}^\star - 0.05) \left(\prod_{k=1}^{t} (1 + 0.2\eta_k) - 1 \right) - \prod_{k=1}^{t} (1 + 0.2\eta_k)\boldsymbol{\theta}^\star \\ &= -0.25 \prod_{k=1}^{t} (1 + 0.2\eta_k) + (0.25 - \boldsymbol{\theta}^\star) \\ &= -0.25 \prod_{k=1}^{t} (1 + 0.2\eta_k) - 1.191. \end{aligned}$$

Consequently, the averaged TD estimation error is featured by

$$\overline{\boldsymbol{\Delta}}_T = -1.191 - 0.25 \frac{1}{T} \sum_{t=1}^{T} \prod_{k=1}^{t} (1 + 0.2\eta_k),$$

and therefore

$$\left\| \overline{\boldsymbol{\Delta}}_T \right\|_2 = 1.191 + 0.25 \frac{1}{T} \sum_{t=1}^{T} \prod_{k=1}^{t} (1 + 0.2\eta_k)$$

[3] In this case, many vectors and matrices reduce to scalars; however, we keep the boldface notation for simplicity.

We now consider the growth of $\left\|\overline{\boldsymbol{\Delta}}_T\right\|_2$ with polynomial-decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$, where $\alpha \in (\frac{1}{2}, 1)$. In this scenario, the norm of $\overline{\boldsymbol{\Delta}}_T$ is bounded by

$$\begin{aligned}
\left\|\overline{\boldsymbol{\Delta}}_T\right\|_2 &= 1.191 + 0.25\frac{1}{T}\sum_{t=1}^{T}\prod_{k=1}^{t}(1+0.2\eta_k) \\
&= 1.191 + \frac{0.25}{T}\cdot\sum_{t=1}^{T}\frac{1}{0.2\eta_t}\left(\prod_{k=1}^{t} - \prod_{k=1}^{t-1}\right)(1+0.2\eta_k) \\
&> 1.191 + \frac{0.25}{0.2\eta_T T}\sum_{t=1}^{T}\left(\prod_{k=1}^{t} - \prod_{k=1}^{t-1}\right)(1+0.2\eta_k) \\
&= 1.191 + \frac{5}{4\eta_0}T^{-\alpha-1}\left(\prod_{k=1}^{T}(1+0.2\eta_0 k^{-\alpha}) - 1\right).
\end{aligned}$$

Here, since $k^{-\alpha} \leq 1$, it can be guaranteed that

$$\begin{aligned}
\prod_{k=1}^{T}(1+0.2\eta_0 k^{-\alpha}) &> \prod_{k=1}^{T}(1+0.2\eta_0)^{k^{-\alpha}} \\
&= \exp\left(\log(1+0.2\eta_0)\sum_{k=1}^{T}k^{-\alpha}\right) \\
&> \exp\left(\log(1+0.2\eta_0)\cdot\frac{T^{1-\alpha}}{1-\alpha}\right).
\end{aligned}$$

Take $u = T^{1-\alpha}$, then the norm of $\overline{\boldsymbol{\Delta}}_T$ is bounded below by

$$\left\|\overline{\boldsymbol{\Delta}}_T\right\|_2 > 1.191 + \frac{5}{4\eta_0}u^{-\frac{\alpha+1}{1-\alpha}}\left[\exp\left(\frac{\log(1+0.2\eta_0)}{1-\alpha}u\right) - 1\right],$$

which grows exponentially with u and hence faster than any polynomial function of T. This completes the proof of the proposition. □

7.1.4 Comparing Theorem 3.2 with previous results

Table 1 provides a comparison of the guarantees of Theorem 3.2 withexisting high-probability bounds for TD estimator with linear function approximation.

A recent line of works by Chen et al. (2024) and Khodadadian & Zubeldia (2025) has established high-probability convergence guarantees for the *tabular* TD learning algorithm. Notably, their results do *not* require the stepsizes to depend on the probability tolerance level δ. For example, when translated to our notation, Theorem 6.1 of Khodadadian & Zubeldia (2025) stated that in the tabular case, with the choice of stepsizes $\eta_t = \eta_0(t+h)^{-\frac{1}{2}}$ where $\eta_0 > \sqrt{h}$, it can be guaranteed for all $\delta \in (0,1)$ that with probability $1-\delta$, the averaged TD(n) estimator $\overline{\boldsymbol{V}}_T$ satisfies

$$\|\overline{\boldsymbol{V}}_T - \boldsymbol{V}^\star\|_\infty^2 \lesssim \frac{1}{(1-\gamma)^2(1-\gamma^n)^2\lambda_0^2}\frac{\log(1/\delta)+|\mathcal{S}|}{T}.$$

We remark that this result is not applicable to our problem for the following reason.

1. When the discount factor $\gamma > 0$, this bound does not naturally apply to the TD(0) algorithm, since the $(1-\gamma^n)^2$ term in the denominator would be equal to 0 when $n = 0$.
2. Crucially, this bound depends on the fact that in the *tabular* case, as long as the reward is bounded, the (original) TD iterates are *uniformly bounded a.s.*; Please refer to Appendix B.8 of Khodadadian & Zubeldia (2025), where Assumption 5.2 is verified. This fact cannot be generalized to TD with linear function approximation, as we have illustrated in the proof of Proposition 7.2.

paper	stepsize	convergence guarantee	constraints on t, T, δ
Dalal et al. (2018)	$\eta_t = (t+1)^{-1}$	$\|\boldsymbol{\Delta}_t\|_2 \leq O\left(t^{-\min\{\frac{1}{2}, \frac{\lambda}{\lambda+1}\}}\right)$ [4]	$t \geq \tilde{C}(\log \frac{1}{\delta})$
Li et al. (2024b)	$\eta_t = \eta$	$\|\overline{\boldsymbol{\Delta}}_T\|_2 \leq O\left(T^{-\frac{1}{2}} \log^{\frac{1}{2}} \frac{T}{\delta}\right)$	$\eta\sqrt{\log \frac{T}{\delta}} \leq \tilde{C}$
Samsonov et al. (2024b)	$\eta_t = \eta$	$\|\overline{\boldsymbol{\Delta}}_T\|_2 \leq O\left(T^{-\frac{1}{2}} \log \frac{T}{\delta}\right)$	$\eta \log \frac{T}{\delta} \leq \tilde{C}$
Theorem 3.2	$\eta_t = \eta_0 t^{-\alpha}, \alpha \in \left(\frac{1}{2}, 1\right)$	$\|\overline{\boldsymbol{\Delta}}_T\|_2 \leq O\left(T^{-\frac{1}{2}}\sqrt{\log \frac{1}{\delta}}\right)$	$\eta_0\sqrt{\log \frac{1}{\delta\eta_0}} \leq \tilde{C}$

Table 1: Comparisons between our results on TD convergence rate with prior work. The quantity $\tilde{C}$, whose value may change at each occurrence, may depend on problem-related parameters other than T and δ. The conditions required in Theorem 3.2 are weaker than those in Samsonov et al. (2024b) and Li et al. (2024b) while delivering the same rates; see discussion in the paragraph below Theorem 3.2.

In fact, in the tabular case and under our setting of TD(0) with stepsizes $\eta_t = \eta_0 t^{-\alpha}, \alpha \in (\frac{1}{2}, 1)$, one can verify by induction that $\|\boldsymbol{V}_t\|_\infty \leq \frac{1}{1-\gamma}$ and therefore $\|\boldsymbol{V}_t\|_2 \leq \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}$ for all $t > 0$ almost surely. As a result, the guarantees in Theorem 7.1 and of Theorem 3.2 can be straightened, by removing the dependence of η_0 on δ. Specifically, Theorem 7.1 can be amended as:

For any $\eta_0 \in (0, \frac{1}{2\lambda_\Sigma})$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_2 \lesssim \frac{\eta_0}{\sqrt{2\alpha-1}} \left(\frac{(1-\gamma)\lambda_0\eta_0}{4\alpha}\right)^{-\frac{\alpha}{2(1-\alpha)}} \sqrt{\log\frac{9T}{\delta}} \frac{\sqrt{|\mathcal{S}|}}{1-\gamma} t^{-\frac{\alpha}{2}}$$

simultaneously for all $t \in [T]$.

In turn, Theorem 3.2 can also be modified as follows:

For any $\eta_0 \in (0, \frac{1}{2\lambda_\Sigma})$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_2 \lesssim \sqrt{\frac{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)}{T} \log\frac{1}{\delta}} + o(T^{-\frac{1}{2}} \log\frac{1}{\delta}).$$

7.2 Proof of Theorem 3.1

By triangle inequality, we firstly decompose the convex distance between $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ and its asymptotic distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)$ by

$$d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \leq d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \overline{\boldsymbol{\Lambda}}_T)) + d_{\mathsf{C}}(\mathcal{N}(\mathbf{0}, \overline{\boldsymbol{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)). \tag{82}$$

Here, recall that $\overline{\boldsymbol{\Lambda}}_T$ is defined as in (52). The following two theorems bound the two terms on the right-hand-side of (82) respectively; Theorem 3.1 follows directly by combining them with (82).

Theorem 7.3. *Consider the TD with Polyak-Ruppert averaging with independent samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$ for $\alpha \in (\frac{1}{2}, 1)$ and $\eta_0 < 1/2$. Then,*

$$d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \overline{\boldsymbol{\Lambda}}_T)) \lesssim \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}} \mathsf{Tr}(\boldsymbol{\Gamma}) \|\boldsymbol{\Gamma}^{-1}\| T^{-\frac{\alpha}{2}} + \tilde{C}\, T^{-\frac{1}{2}},$$

where $\tilde{C}$ is a problem related quantity independent of T.

Theorem 7.4. *Let $\boldsymbol{\Lambda}^\star$ and $\overline{\boldsymbol{\Lambda}}_T$ be defined as in (14) and (52), respectively, with the stepsizes $\eta_t = \eta_0 t^{-\alpha}$, for all $t > 0$, any $\alpha \in (\frac{1}{2}, 1)$ and $\eta_0 < 1/2$. Then,*

$$d_{\mathsf{C}}(\mathcal{N}(\mathbf{0}, \overline{\boldsymbol{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \le d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \overline{\boldsymbol{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim \frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{(1-\gamma)\lambda_0\eta_0} T^{\alpha-1} + \tilde{C}'\, T^{2\alpha-2},$$

where $\tilde{C}'$ is a problem related quantity independent of T.

7.2.1 Proof of Theorem 7.3

This proof is based on Theorem 6.13, the Berry-Esseen bound for multivariate non-linear statistics. Throughout the proof, we will firstly define corresponding statistics, and then bound the terms on the right-hand-side of (68).

Defining $\boldsymbol{W}$,$\boldsymbol{D}$ and other statistics. Recall from (76) that the averaged TD error can be represented as

$$\overline{\boldsymbol{\Delta}}_T = \frac{1}{T\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0 - \frac{1}{T}\sum_{i=1}^{T-1}\boldsymbol{Q}_i\boldsymbol{\zeta}_i,$$

where $\boldsymbol{Q}_t$ is defined as in (53). Therefore, the statistics $\boldsymbol{W}$ and $\boldsymbol{D}$ can be constructed by

$$\begin{aligned}\sqrt{T}\boldsymbol{A}\overline{\boldsymbol{\Delta}}_T &= \frac{\boldsymbol{A}}{\sqrt{T}\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0 - \frac{\boldsymbol{A}}{\sqrt{T}}\sum_{i=1}^{T}\boldsymbol{Q}_i\boldsymbol{\zeta}_i \\ &= \underbrace{\sum_{i=1}^{T}\boldsymbol{\xi}_i}_{\boldsymbol{W}} + \underbrace{\left(\frac{\boldsymbol{A}}{\sqrt{T}\eta_0}\boldsymbol{Q}_0\Delta_0 + \frac{\boldsymbol{A}}{\sqrt{T}}\sum_{i=1}^{T}\boldsymbol{Q}_i\boldsymbol{\omega}_i\right)}_{\boldsymbol{D}}, \end{aligned} \tag{83}$$

in which we define $\boldsymbol{\xi}_i$ and $\boldsymbol{\omega}_i$ as

$$\boldsymbol{\xi}_i = \frac{1}{\sqrt{T}}\boldsymbol{A}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i); \tag{84}$$

$$\boldsymbol{\omega}_i = (\boldsymbol{A}_i - \boldsymbol{A})\Delta_{i-1}. \tag{85}$$

Notice that since $\{(s_i, s_i')\}_{i=1}^T$ are independent, $\{(\boldsymbol{A}_i, \boldsymbol{b}_i)\}_{i=1}^T$ are also independent. By definition, it is easy to verify that $\mathbb{E}[\boldsymbol{\xi}_i] = 0$ and that

$$\begin{aligned}\sum_{i=1}^{T}\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] &= \frac{1}{T}\boldsymbol{A}\left\{\sum_{i=1}^{T}\boldsymbol{Q}_i\mathbb{E}[(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i)(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i)^\top]\boldsymbol{Q}_i^\top\right\}\boldsymbol{A}^\top \\ &= \boldsymbol{A}\left\{\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i\boldsymbol{\Gamma}\boldsymbol{Q}_i^\top\right\}\boldsymbol{A}^\top = \boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top.\end{aligned}$$

With the minimum eigenvalue bounded below in Lemma 6.12 by

$$\lambda_{\min}(\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top) \ge \frac{1}{2\|\boldsymbol{\Gamma}^{-1}\|},$$

Theorem 6.13 guarantees that

$$d_{\mathsf{C}}(\boldsymbol{A}\overline{\boldsymbol{\Delta}}_T, (\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top)^{\frac{1}{2}}\boldsymbol{z}) \lesssim \underbrace{\sqrt{d}\|\boldsymbol{\Gamma}^{-\frac{3}{2}}\| \cdot \sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{\xi}_i\|_2^3}_{I_1} + \underbrace{\|\boldsymbol{\Gamma}^{-1}\|\mathbb{E}\{\|\boldsymbol{W}\|_2\Delta\}}_{I_2} + \underbrace{\|\boldsymbol{\Gamma}^{-1}\|\sum_{i=1}^{n}\mathbb{E}\{\|\boldsymbol{\xi}_i\|_2|\Delta - \Delta^{(i)}|\}}_{I_3}, \tag{86}$$

where $\Delta = \|\boldsymbol{D}\|$ and $\Delta^{(i)}$ is any random variable independent of $(\boldsymbol{A}_i, \boldsymbol{b}_i)$. Now, the proof boils down to bounding the three terms on the right-hand-side of (86).

Bounding I_1. We firstly observe that

$$\begin{aligned}\|\boldsymbol{\xi}_i\|_2^3 &\leq T^{-3/2}\|\boldsymbol{A}\boldsymbol{Q}_i\|^3\|\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i\|_2^3 \\ &\lesssim T^{-3/2}\cdot\left(2+\eta_0\left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)i^{\alpha-1}\right)^3\cdot\|\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i\|_2^3,\end{aligned}$$

in which $\|\boldsymbol{A}\boldsymbol{Q}_i\|$ is bounded by Lemma 6.10. Furthermore, since $\max_{s\in\mathcal{S}}\|\boldsymbol{\phi}(s)\|_2 \leq 1$ and $\max_{s\in\mathcal{S}} r(s)\leq 1$, it is easy to verify that

$$\|\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i\|_2 < 2\|\boldsymbol{\theta}^\star\|_2 + 1.$$

Therefore, the term I_1 is bounded by

$$\sqrt{d}\sum_{i=1}^T\mathbb{E}\|\boldsymbol{\xi}_i\|_2^3 \lesssim \sqrt{d}\cdot\left(\|\boldsymbol{\Gamma}^{-1}\|\right)^{3/2}(2\|\boldsymbol{\theta}^\star\|_2+1)^3\left(\eta_0\left(\frac{1}{\beta}\right)^{\frac{1}{1-\alpha}}\Gamma\left(\frac{1}{1-\alpha}\right)\right)^3 T^{-\frac{1}{2}} = O(T^{-\frac{1}{2}}). \tag{87}$$

Bounding I_2. First of all, we define $\Delta = \|\boldsymbol{D}\|_2$. By Cauchy-Schwartz inequality, the second term of (86) is upper bounded by

$$\mathbb{E}\left\{\|\boldsymbol{W}\|_2\Delta\right\} \leq \sqrt{\mathbb{E}\|\boldsymbol{W}\|_2^2\mathbb{E}\Delta^2}.$$

Regarding $\mathbb{E}\|\boldsymbol{W}\|_2^2$, notice that

$$\mathbb{E}\|\boldsymbol{W}\|_2^2 = \mathbb{E}[\mathsf{Tr}(\boldsymbol{W}\boldsymbol{W}^\top)] = \mathsf{Tr}(\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top) = \mathsf{Tr}(\boldsymbol{\Gamma}) + O(T^{\alpha-1}),$$

where the last identity is a consequence of Theorem 6.11. In detail,

$$\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top - \boldsymbol{\Gamma} = \boldsymbol{A}(\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star)\boldsymbol{A}^\top = T^{\alpha-1}\boldsymbol{X}(\boldsymbol{\Lambda}^\star)\boldsymbol{A}^\top + \boldsymbol{O}(T^{2\alpha-2}) = \boldsymbol{O}(T^{\alpha-1}),$$

where $\boldsymbol{X}(\boldsymbol{\Lambda}^\star)$ is defined as in (62).

As for the expectation of $|\Delta|^2$, observe that

$$\begin{aligned}\mathbb{E}\Delta^2 &\leq 2\mathbb{E}\left\|\frac{1}{\sqrt{T}\eta_0}\boldsymbol{A}\boldsymbol{Q}_0\boldsymbol{\Delta}_0\right\|_2^2 + 2\mathbb{E}\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^T\boldsymbol{A}\boldsymbol{Q}_i\boldsymbol{\omega}_i\right\|_2^2 \\ &\leq \frac{2}{T\eta_0^2}\|\boldsymbol{A}\boldsymbol{Q}_0\|^2\|\boldsymbol{\Delta}_0\|_2^2 + \frac{2}{T}\sum_{i=1}^T\|\boldsymbol{A}\boldsymbol{Q}_i\|^2\mathbb{E}\|\boldsymbol{\omega}_i\|_2^2 \\ &\lesssim \frac{1}{T\eta_0^2}\cdot\left(\frac{2}{(1-\gamma)\lambda_0\eta_0}\right)^{\frac{2}{1-\alpha}}\|\boldsymbol{\Delta}_0\|_2^2 \\ &+ \frac{1}{T}\sum_{i=1}^T(2+O(i^{\alpha-1}))^2\mathbb{E}\|(\boldsymbol{A}_i-\boldsymbol{A})\|_2^2\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2 \\ &\lesssim \frac{1}{T}\sum_{i=1}^T\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2 + O\left(\frac{1}{T}\right).\end{aligned}$$

The first line is the inequality $\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 \leq 2\|\boldsymbol{x}\|_2^2 + 2\|\boldsymbol{y}\|_2^2$, for any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$; in the third line, we applied the upper bounds for $\|\boldsymbol{A}\boldsymbol{Q}_t\|$ established in Lemma 6.10. For the term $\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2$, the following lemma comes in handy.

Lemma 7.5. *Under the TD iteration (8) with i.i.d. samples and decaying stepsizes $\eta_t = \eta_0 t^{-\alpha}$ for $\alpha\in(\frac{1}{2},1)$ and $\eta_0\leq\frac{1}{4}$, it can be guaranteed that*

$$\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2 \leq \frac{2\eta_0\mathsf{Tr}(\boldsymbol{\Gamma})}{(1-\gamma)\lambda_0}t^{-\alpha} + O(t^{-1}).$$

holds for any $t>0$.

Proof: See Appendix 8.1. □

As an immediate result of Lemma 7.5, the last term is bounded by

$$\begin{aligned}\sum_{i=1}^{T}\mathbb{E}\|\boldsymbol{\Delta}_i\|_2^2 &\leq \sum_{i=1}^{T}\frac{2\eta_0}{(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{\Gamma})i^{-\alpha}+O(i^{-1})\\ &= \frac{2\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}(2\|\boldsymbol{\theta}^\star\|_2+1)^2T^{1-\alpha}+O(\log T).\end{aligned}$$

Therefore, $\mathbb{E}\Delta^2$ can be bounded by

$$\begin{aligned}\mathbb{E}\Delta^2 &\lesssim \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2+O\left(\frac{1}{T}\right)\\ &\leq \frac{1}{T}\left(\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{\Gamma})T^{1-\alpha}+O(\log T)\right)+O\left(\frac{1}{T}\right)\\ &= \frac{\eta_0\mathsf{Tr}(\boldsymbol{\Gamma})}{(1-\alpha)(1-\gamma)\lambda_0}T^{-\alpha}+O\left(\frac{\log T}{T}\right)\end{aligned}$$

So in conclusion, the second term of (86) can be bounded by

$$\mathbb{E}\{\|W\|\Delta\}\leq\sqrt{\mathbb{E}\|W\|_2^2\mathbb{E}\Delta^2}\lesssim\sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\mathsf{Tr}(\boldsymbol{\Gamma})T^{-\frac{\alpha}{2}}+O\left(\sqrt{\frac{\log T}{T}}\right). \tag{88}$$

Bounding I_3. The third term of (86) involves random variables $\{\Delta^{(i)}\}_{0\leq i\leq T}$ such that $\Delta^{(i)}$ is independent of (s_i,s_i'). For this purpose, we define $\{(\check{s}_i,\check{s}_i')\}_{i=1}^n$ as independent copies of $\{(s_i,s_i')\}$ and let $\check{\boldsymbol{A}}_i,\check{\boldsymbol{b}}_i$ be generated by $(\check{s}_i,\check{s}_i')$. Consider the following iterations for any i:

$$\boldsymbol{\theta}_j^{(i)}=\begin{cases}\boldsymbol{\theta}_j,\ \text{if } j<i;\\ \boldsymbol{\theta}_{j-1}-\eta_j(\check{\boldsymbol{A}}_j\boldsymbol{\theta}_{j-1}-\check{\boldsymbol{b}}_j),\ \text{if } j=i;\\ \boldsymbol{\theta}_{j-1}^{(i)}-\eta_j(\boldsymbol{A}_j\boldsymbol{\theta}_{j-1}^{(i)}-\boldsymbol{b}_j),\ \text{if } j>i.\end{cases} \tag{89}$$

It is easy to verify that for all $j\in[T]$, $\boldsymbol{\theta}_j^{(i)}$ is independent of (s_i,s_i'). Therefore, with the definitions of

$$\boldsymbol{\omega}_j^{(i)}=\begin{cases}(\check{\boldsymbol{A}}_j-\boldsymbol{A})(\boldsymbol{\theta}_{j-1}^{(i)}-\boldsymbol{\theta}^\star), & \text{if} \quad j=i;\\ (\boldsymbol{A}_j-\boldsymbol{A})(\boldsymbol{\theta}_{j-1}^{(i)}-\boldsymbol{\theta}^\star), & \text{otherwise.}\end{cases} \tag{90}$$

and

$$\Delta^{(i)}=\left\|\frac{1}{\sqrt{T}\eta_0}\boldsymbol{A}\boldsymbol{Q}_0\boldsymbol{\Delta}_0-\frac{1}{\sqrt{T}}\sum_{j=1}^{T}\boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{\omega}_j^{(i)}\right\|_2,$$

It can be guaranteed that $\Delta^{(i)}$ is independent of (s_i,s_i'). The proof now boils down to bounding the difference between Δ and $\Delta^{(i)}$, for which the triangle inequality implies

$$|\Delta-\Delta^{(i)}|\leq\left\|\frac{1}{\sqrt{T}}\sum_{j=1}^{T}\boldsymbol{A}\boldsymbol{Q}_j(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)})\right\|_2\leq\frac{1}{\sqrt{T}}\left\|\sum_{j=i}^{T}\boldsymbol{A}\boldsymbol{Q}_j(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)})\right\|_2.$$

By squaring and taking expectations on both sides, we observe that

$$\begin{aligned}\mathbb{E}|\Delta-\Delta^{(i)}|^2 &\le \frac{1}{T}\mathbb{E}\left\|\sum_{j=i}^{T}\boldsymbol{A}\boldsymbol{Q}_j(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)})\right\|_2^2 \\ &= \frac{1}{T}\mathbb{E}\sum_{j=i}^{T}\left\|\boldsymbol{A}\boldsymbol{Q}_j(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)})\right\|_2^2+\frac{2}{T}\mathbb{E}\left\{\sum_{i\le j<j'<T}\left[\boldsymbol{A}\boldsymbol{Q}_j\left(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right)\right]^\top\left[\boldsymbol{A}\boldsymbol{Q}_{j'}\left(\boldsymbol{\omega}_{j'}-\boldsymbol{\omega}_{j'}^{(i)}\right)\right]\right\}.\end{aligned} \tag{91}$$

For the last term on the right-hand-side of (91), recall from the definitions of $\boldsymbol{\omega}_j$ (85) and $\boldsymbol{\omega}_j^{(i)}$ (90), for any $j>i$,

$$\begin{aligned}\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)} &= [(\boldsymbol{A}_j-\boldsymbol{A})(\boldsymbol{\theta}_{j-1}-\boldsymbol{\theta}^\star)]-\left[(\boldsymbol{A}_j-\boldsymbol{A})\left(\boldsymbol{\theta}_{j-1}^{(i)}-\boldsymbol{\theta}^\star\right)\right] \\ &= (\boldsymbol{A}_j-\boldsymbol{A})\left(\boldsymbol{\theta}_{j-1}-\boldsymbol{\theta}_{j-1}^{(i)}\right)\end{aligned}$$

Hence, for $i\le j<j'<T$, the law of total expectation implies

$$\begin{aligned}&\mathbb{E}\left\{\left[\boldsymbol{A}\boldsymbol{Q}_j\left(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right)\right]^\top\left[\boldsymbol{A}\boldsymbol{Q}_{j'}\left(\boldsymbol{\omega}_{j'}-\boldsymbol{\omega}_{j'}^{(i)}\right)\right]\right\} \\ &= \mathbb{E}\mathbb{E}_{j'-1}\left\{\left[\boldsymbol{A}\boldsymbol{Q}_j\left(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right)\right]^\top\left[\boldsymbol{A}\boldsymbol{Q}_{j'}\left(\boldsymbol{\omega}_{j'}-\boldsymbol{\omega}_{j'}^{(i)}\right)\right]\right\} \\ &= \mathbb{E}\left\{\left[\boldsymbol{A}\boldsymbol{Q}_j\left(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right)\right]^\top\left[\boldsymbol{A}\boldsymbol{Q}_{j'}\mathbb{E}_{j'-1}\left(\boldsymbol{\omega}_{j'}-\boldsymbol{\omega}_{j'}^{(i)}\right)\right]\right\} \\ &= \mathbb{E}\left\{\left[\boldsymbol{A}\boldsymbol{Q}_j\left(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right)\right]^\top\left[\boldsymbol{A}\boldsymbol{Q}_{j'}\mathbb{E}_{j'-1}\left[(\boldsymbol{A}_{j'}-\boldsymbol{A})\left(\boldsymbol{\theta}_{j'-1}-\boldsymbol{\theta}_{j'-1}^{(i)}\right)\right]\right]\right\}=0,\end{aligned}$$

since $\mathbb{E}_{j'-1}[\boldsymbol{A}_{j'}-\boldsymbol{A}]=\boldsymbol{0}$. This effectively means that the last term on the right-hand-side of (91) is 0. Therefore, the difference between Δ and $\Delta^{(i)}$ is bounded by

$$\begin{aligned}\mathbb{E}|\Delta-\Delta^{(i)}|^2 &\le \frac{1}{T}\mathbb{E}\sum_{j=i}^{T}\left\|\boldsymbol{A}\boldsymbol{Q}_j(\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)})\right\|_2^2 \\ &\le \frac{1}{T}\sum_{j=i}^{T}\|\boldsymbol{A}\boldsymbol{Q}_j\|^2\mathbb{E}\left\|\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right\|_2^2 \\ &\lesssim \frac{1}{T}\cdot\sum_{j=i}^{T}(2+O(j^{\alpha-1}))\mathbb{E}\left\|\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}\right\|_2^2,\end{aligned} \tag{92}$$

where in the last line we applied Lemma 6.10 to bound the norm of $\boldsymbol{Q}_j$. The proof now boils down to bounding the expected difference between $\boldsymbol{\omega}_j$ and $\boldsymbol{\omega}_j^{(i)}$, for which we recall from definition that

$$\boldsymbol{\omega}_j-\boldsymbol{\omega}_j^{(i)}=\begin{cases}(\boldsymbol{A}_i-\check{\boldsymbol{A}}_i)(\boldsymbol{\theta}_{i-1}-\boldsymbol{\theta}^\star), & \text{for} \quad j=i;\\ (\boldsymbol{A}_j-\boldsymbol{A})\left(\boldsymbol{\theta}_{j-1}-\boldsymbol{\theta}_{j-1}^{(i)}\right), & \text{for} \quad j>i.\end{cases}$$

The norm of $(\boldsymbol{\theta}_{i-1}-\boldsymbol{\theta}^\star)$ can be bounded by Lemma 7.5; for the norm of $\left(\boldsymbol{\theta}_{j-1}-\boldsymbol{\theta}_{j-1}^{(i)}\right)$, the following proposition comes in handy.

Proposition 7.6. *With $\boldsymbol{\theta}_j^{(i)}$ defined in (89),it can be guaranteed that for every $i\in[T]$,*

$$\mathbb{E}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_i^{(i)}\|_2^2\le 2\eta_i^2\mathsf{Tr}(\boldsymbol{\Gamma})+2\eta_i^2\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2. \tag{93a}$$

Furthermore, for every $j > i$,

$$\mathbb{E}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^{(i)}\|_2^2 \leq \prod_{k=i+1}^{j} \left(1 - \frac{1}{2}(1-\gamma)\lambda_0\eta_k\right) \mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(i)}\|_2^2. \tag{93b}$$

Proof: See appendix 8.2. □

For $j = i$, direct calculation yields

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^{(i)}\|_2^2 &= \mathbb{E}\|(\boldsymbol{A}_i - \check{\boldsymbol{A}}_i)(\boldsymbol{\theta}_{i-1} - \boldsymbol{\theta}^\star)\|_2^2 \\ &\lesssim \mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2;\end{aligned} \tag{94}$$

and similarly, for $j > i$,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_j^{(i)}\|_2^2 &= \mathbb{E}\left\|(\boldsymbol{A}_j - \boldsymbol{A})\left(\boldsymbol{\theta}_{j-1} - \boldsymbol{\theta}_{j-1}^{(i)}\right)\right\|_2^2 \\ &\lesssim \mathbb{E}\left\|\boldsymbol{\theta}_{j-1} - \boldsymbol{\theta}_{j-1}^{(i)}\right\|_2^2.\end{aligned} \tag{95}$$

By combining (92), (94),(95) with (93b), we obtain

$$\begin{aligned}\mathbb{E}|\Delta - \Delta^{(i)}|^2 &\leq \frac{1}{T}\left((2 + O(i^{\alpha-1}))\mathbb{E}\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^{(i)}\|_2^2 + \sum_{j=i+1}^{T}(2 + O(j^{\alpha-1}))\mathbb{E}\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_j^{(i)}\|_2^2\right) \\ &\lesssim \frac{1}{T}\cdot(2 + O(i^{\alpha-1}))\underbrace{\mathbb{E}\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^{(i)}\|_2^2}_{I_1} \\ &\quad + \underbrace{\frac{1}{T}\sum_{j=i+1}^{T}(2 + O(j^{\alpha-1}))\prod_{k=i+1}^{j}\left(1 - \frac{1-\gamma}{2}\lambda_0\eta_k\right)\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(i)}\|_2^2}_{I_2}.\end{aligned}$$

With I_1 bounded by (94), the proof now boils down to bounding I_2. Notice that (93a) implies

$$\begin{aligned}&\sum_{j=i+1}^{T}\prod_{k=i+1}^{j}\left(1 - \frac{1-\gamma}{2}\lambda_0\eta_k\right)\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(i)}\|_2^2 \\ &\lesssim \left[\mathsf{Tr}(\boldsymbol{\Gamma}) + 2\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2\right]\eta_0^2 i^{-2\alpha}\sum_{j=i+1}^{T}\prod_{k=i+1}^{j}\left(1 - \frac{1-\gamma}{2}\lambda_0\eta_k\right),\end{aligned}$$

where we can invoke (44) in Lemma 6.7 to obtain

$$i^{-2\alpha}\sum_{j=i+1}^{T}\prod_{k=i+1}^{j}\left(1 - \frac{1-\gamma}{2}\lambda_0\eta_k\right) < \frac{2}{(1-\gamma)\lambda_0\eta_0}i^{-\alpha} + O(i^{-1}).$$

Combining these results with Lemma 7.5, we establish the following bound for $I_1 + I_2$:

$$I_1 + I_2 \lesssim \frac{\eta_0}{(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{\Gamma})i^{-\alpha} + O(i^{-1}).$$

Hence,

$$\mathbb{E}|\Delta - \Delta^{(i)}|^2 \lesssim \frac{1}{T}\frac{\eta_0}{(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{\Gamma})i^{-\alpha} + O(i^{-1}).$$

Therefore, by Cauchy-Schwartz inequality,

$$\left(\mathbb{E}\sum_{i=1}^{T}\|\boldsymbol{\xi}_i\||\Delta-\Delta^{(i)}|\right)^2 \le \left(\sum_{i=1}^{T}\mathbb{E}\|\boldsymbol{\xi}_i\|_2^2\right)\left(\sum_{i=1}^{T}\mathbb{E}|\Delta-\Delta^{(i)}|^2\right), \tag{96}$$

in which the two terms on the right-hand-side can be bounded respectively by

$$\sum_{i=1}^{T}\mathbb{E}\|\boldsymbol{\xi}_i\|_2^2 = \sum_{i=1}^{T}\mathbb{E}\mathsf{Tr}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top] = \mathsf{Tr}(\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top) = \mathsf{Tr}(\boldsymbol{\Gamma}) + O(T^{\alpha-1}) \tag{97}$$

and

$$\begin{aligned}\sum_{i=1}^{T}\mathbb{E}|\Delta-\Delta^{(i)}|^2 &\lesssim \frac{1}{T}\frac{\eta_0}{(1-\gamma)\lambda_0}\mathsf{Tr}(\boldsymbol{\Gamma})\sum_{i=1}^{T}\left(i^{-\alpha}+O(i^{-1})\right)\\ &= \frac{\eta_0\mathsf{Tr}(\boldsymbol{\Gamma})}{(1-\alpha)(1-\gamma)\lambda_0}T^{-\alpha}+O\left(\frac{\log T}{T}\right).\end{aligned} \tag{98}$$

A direct combination of (96), (97) and (98) yields the following bound for the third term of (86):

$$\mathbb{E}\sum_{i=1}^{T}\|\boldsymbol{\xi}_i\||\Delta-\Delta^{(i)}| \lesssim \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\mathsf{Tr}(\boldsymbol{\Gamma})T^{-\frac{\alpha}{2}}+O\left(\sqrt{\frac{\log T}{T}}\right). \tag{99}$$

Completing the proof. A direct combination of (86), (87), (88) and (99) yields

$$\left|\mathbb{P}(\sqrt{T}\boldsymbol{A}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A})-\mathbb{P}((\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top)^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right| \le \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\|\boldsymbol{\Gamma}^{-1}\|\mathsf{Tr}(\boldsymbol{\Gamma})T^{-\frac{\alpha}{2}}+O\left(\sqrt{\frac{\log T}{T}}\right).$$

Furthermore, for any convex set $\mathcal{A}\in\mathbb{R}^d$, the set

$$\mathcal{A}' = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x}\in\mathcal{A}\}$$

must also be convex; by definition, $\mathcal{A}$ and $\mathcal{A}'$ are related by

$$\begin{aligned}&\mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Lambda}}_T^{-1/2}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A}) = \mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A}'), \quad \text{and}\\ &\mathbb{P}((\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T\boldsymbol{A}^\top)^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A}) = \mathbb{P}(\overline{\boldsymbol{\Lambda}}_T^{1/2}\boldsymbol{z}\in\mathcal{A}').\end{aligned}$$

Therefore, for any convex set $\mathcal{A}\in\mathbb{R}^d$, it can be guaranteed that

$$\left|\mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A})-\mathbb{P}(\overline{\boldsymbol{\Lambda}}_T^{1/2}\boldsymbol{z}\in\mathcal{A})\right| \le \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\|\boldsymbol{\Gamma}^{-1}\|\mathsf{Tr}(\boldsymbol{\Gamma})T^{-\frac{\alpha}{2}}+O\left(\sqrt{\frac{\log T}{T}}\right).$$

Taking supremum over $\mathcal{A}$ completes the proof. □

7.2.2 Proof of Theorem 7.4

For any Borel set $\mathcal{A}\subset\mathbb{R}^d$, define $\mathcal{A}' = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x}\in\mathbb{R}^d\}$. By definition, when $\boldsymbol{z}$ is the d-dimensional standard Gaussian distribution,

$$\left|\mathbb{P}(\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})-\mathbb{P}(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right| = \left|\mathbb{P}(\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A}')-\mathbb{P}(\boldsymbol{A}\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}\in\mathcal{A}')\right|.$$

Since $\boldsymbol{A}$ is non-singular, the map $\mathcal{A}\mapsto\mathcal{A}'$ is a bijection. Hence by taking supremum over $\mathcal{A}$ and $\mathcal{A}'$ on both sides of the equation, we obtain

$$d_{\mathsf{TV}}(\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}, \boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}) = d_{\mathsf{TV}}(\boldsymbol{A}\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}, \boldsymbol{A}\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z});$$

In other words,

$$d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) = d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \boldsymbol{A}\overline{\mathbf{\Lambda}}_T\boldsymbol{A}^\top), \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}));$$

Therefore, the proof boils down to bounding the TV distance between $\mathcal{N}(\mathbf{0}, \boldsymbol{A}\overline{\mathbf{\Lambda}}_T\boldsymbol{A}^\top)$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$. Towards this end, Theorem 6.14 indicates that this TV distance converges at the same rate of

$$\left\|\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}\overline{\mathbf{\Lambda}}_T\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}} - \boldsymbol{I}_d\right\|_{\mathsf{F}} = \left\|\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}(\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star)\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}}\right\|_{\mathsf{F}}.$$

Here, $\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star$ featured by Theorem 6.11 as

$$\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star = T^{\alpha-1}\boldsymbol{X} + \boldsymbol{O}(T^{2\alpha-2}),$$

where $\boldsymbol{X}$ satisfies the Lyapunov equation

$$\eta_0(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}^\top) = \mathbf{\Lambda}^\star;$$

Therefore it is easy to verify that

$$\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}(\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star)\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}} = T^{\alpha-1}\boldsymbol{Y} + \boldsymbol{O}(T^{2\alpha-2}),$$

in which

$$\eta_0(\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}\mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{Y} + \boldsymbol{Y}\mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}}) = \boldsymbol{I}.$$

Let λ be any eigenvalue of $\boldsymbol{Y}$ and $\boldsymbol{x}$ be its corresponding eigenvector with $\|\boldsymbol{x}\|_2 = 1$. Then by definition,

$$\begin{aligned}
1 &= \eta_0\boldsymbol{x}^\top(\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}\mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{Y} + \boldsymbol{Y}\mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}})\boldsymbol{x} \\
&= \eta_0\lambda\boldsymbol{x}^\top(\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}\mathbf{\Gamma}^{\frac{1}{2}} + \mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}})\boldsymbol{x} \\
&\geq \eta_0\lambda \cdot \lambda_{\min}(\mathbf{\Gamma}^{-\frac{1}{2}}\boldsymbol{A}\mathbf{\Gamma}^{\frac{1}{2}} + \mathbf{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^\top\mathbf{\Gamma}^{-\frac{1}{2}}) \\
&\geq \frac{2(1-\gamma)\lambda_0\eta_0}{\mathsf{cond}(\mathbf{\Gamma}^{\frac{1}{2}})}\lambda.
\end{aligned}$$

Therefore, the Frobenius norm of $\boldsymbol{Y}$ is bounded by

$$\|\boldsymbol{Y}\|_{\mathsf{F}} \leq \sqrt{d}\lambda_{\max}(\boldsymbol{Y}) \leq \frac{\sqrt{d\mathsf{cond}(\mathbf{\Gamma}^{\frac{1}{2}})}}{2(1-\gamma)\lambda_0\eta_0}T^{\alpha-1}.$$

Hence the distance between $\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T)$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)$ is bounded by

$$d_{\mathsf{C}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) \leq \|\boldsymbol{Y}\|_{\mathsf{F}}T^{\alpha-1} + \|\boldsymbol{O}(T^{2\alpha-2})\|_{\mathsf{F}} = \frac{\sqrt{d\mathsf{cond}(\mathbf{\Gamma})}}{2(1-\gamma)\lambda_0\eta_0}T^{\alpha-1} + O(T^{2\alpha-2}).$$

□

7.2.3 Tightness of Theorem 7.4

As is indicated by Lemma 6.11, the difference between $\overline{\mathbf{\Lambda}}_T$ and $\mathbf{\Lambda}^\star$ is featured by

$$\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star \geq \boldsymbol{O}(T^{\alpha-1})$$

when T is sufficiently large. Therefore, since $\overline{\mathbf{\Lambda}}_T \succ \mathbf{\Lambda}^\star$, Proposition 6.15 guarantees that

$$d_{\mathsf{C}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) = d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)).$$

Meanwhile, Theorem 6.14 indicates the TV distance between $\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T)$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star))$ converges at the same rate, i.e.

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) &\approx \left\| \mathbf{\Lambda}^{\star -\frac{1}{2}} \overline{\mathbf{\Lambda}}_T \mathbf{\Lambda}^{\star -\frac{1}{2}} - \boldsymbol{I} \right\|_{\mathsf{F}} \\ &= \left\| \mathbf{\Lambda}^{\star -\frac{1}{2}} (\overline{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star) \mathbf{\Lambda}^{\star -\frac{1}{2}} \right\|_{\mathsf{F}} = O(T^{\alpha-1}). \end{aligned}$$

Consequently, the $O(T^{\alpha-1})$ rate by which $\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T)$ converges to $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)$ cannot be further improved. As a direct consequence, when $\alpha \in (\frac{2}{3}, 1)$, since $\alpha - 1 > -\frac{1}{3} > -\frac{\alpha}{2}$, it can be guaranteed through triangle inequality that

$$\begin{aligned} d_{\mathsf{C}}(\sqrt{T}\overline{\mathbf{\Delta}}_T, \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) &\geq d_{\mathsf{C}}(\mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) - d_{\mathsf{C}}(\sqrt{T}\overline{\mathbf{\Delta}}_T, \mathcal{N}(\mathbf{0}, \overline{\mathbf{\Lambda}}_T)) \\ &\geq \Theta(T^{\alpha-1}) - O(T^{-\frac{\alpha}{2}}) \geq \Theta(T^{-\frac{1}{3}}). \end{aligned}$$

7.3 Proof of Theorem 3.3

Applying the same reasoning as in the proof of Theorem 7.4,

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \widehat{\mathbf{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\star)) &= d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \boldsymbol{A}\widehat{\mathbf{\Lambda}}_T \boldsymbol{A}^\top), \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})) \\ &\lesssim \|\mathbf{\Gamma}^{-\frac{1}{2}} \boldsymbol{A} (\widehat{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star) \boldsymbol{A}^\top \mathbf{\Gamma}^{-\frac{1}{2}}\|_{\mathsf{F}} \\ &\leq \|\mathbf{\Gamma}^{-1}\| \|\boldsymbol{A} (\widehat{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star) \boldsymbol{A}^\top\|_{\mathsf{F}} \end{aligned}$$

In order to bound the right-hand-side of the inequality above, we firstly present the difference between $\widehat{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}^\star$ as

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_T - \mathbf{\Lambda}^\star &= \overline{\boldsymbol{A}}_T^{-1} \widehat{\mathbf{\Gamma}}_T \overline{\boldsymbol{A}}_T^{-\top} - \boldsymbol{A}^{-1} \mathbf{\Gamma} \boldsymbol{A}^{-\top} \\ &= \overline{\boldsymbol{A}}_T^{-1} (\widehat{\mathbf{\Gamma}}_T - \mathbf{\Gamma}) \overline{\boldsymbol{A}}_T^{-\top} + \left[\overline{\boldsymbol{A}}_T^{-1} \mathbf{\Gamma} \overline{\boldsymbol{A}}_T^{-\top} - \boldsymbol{A}^{-1} \mathbf{\Gamma} \boldsymbol{A}^{-\top} \right]. \end{aligned} \tag{100}$$

The proof now boils down to bounding the estimation error of $\overline{\boldsymbol{A}}^{-1}$ and $\widehat{\mathbf{\Gamma}}$, for which the following lemmas come in handy.

Lemma 7.7. *Let $\overline{\boldsymbol{A}}_T$ be defined as in (17) with $\{(s_t, s_t')\}_{0 \leq t \leq T}$ being i.i.d. samples. It can then be guaranteed with probability at least $1 - \delta$ that*

$$\|(\overline{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}^{-1})\boldsymbol{A}\|_{\mathsf{F}} \leq \frac{4\sqrt{2}}{\lambda_0(1-\gamma)\sqrt{T}} \sqrt{\log \frac{3}{\delta}},$$

with the proviso that

$$T \geq \frac{128}{\lambda_0(1-\gamma)} \log \frac{3}{\delta}. \tag{101}$$

Proof: See Appendix 8.3. □

Lemma 7.8. *Let $\mathbf{\Gamma}$ be defined as in (14) and $\widehat{\mathbf{\Gamma}}_T$ be defined as in (17). Then the difference between $\widehat{\mathbf{\Gamma}}_T$ and $\mathbf{\Gamma}$ can be bounded by*

$$\left\| \widehat{\mathbf{\Gamma}}_T - \mathbf{\Gamma} \right\|_{\mathsf{F}} \leq \frac{\sqrt{\mathsf{Tr}(\mathbf{\Gamma})}}{\lambda_0(1-\gamma)} (2\|\boldsymbol{\theta}^\star\|_2 + 1) \|\mathbf{\Gamma}^{-1}\| T^{-\frac{1}{3}} + o(T^{-\frac{1}{3}}),$$

with probability at least $1 - 2T^{-\frac{1}{3}} - \frac{\mathsf{Tr}(\mathbf{\Gamma})}{\lambda_0(1-\gamma)} T^{-\frac{1}{2}}$.

Proof: See Appendix 8.4. □

Based on Lemmas 7.7 and 7.8, we now consider the estimation error of $\widehat{\boldsymbol{\Lambda}}_T$. By triangle inequality, (100) implies that

$$\left\|\boldsymbol{A}(\widehat{\boldsymbol{\Lambda}}_T-\boldsymbol{\Lambda}^\star)\boldsymbol{A}^\top\right\|_{\mathsf{F}} \le \underbrace{\|\overline{\boldsymbol{A}}_T^{-1}\boldsymbol{A}\|^2\|\widehat{\boldsymbol{\Gamma}}_T-\boldsymbol{\Gamma}\|_{\mathsf{F}}}_{I_1} + \underbrace{\left\|\boldsymbol{A}\overline{\boldsymbol{A}}_T^{-1}\boldsymbol{\Gamma}\overline{\boldsymbol{A}}_T^{-\top}\boldsymbol{A}^\top-\boldsymbol{\Gamma}\right\|_{\mathsf{F}}}_{I_2}.$$

In what follows, we bound I_1 and I_2 in order.

Bounding I_1. The norm of $\widehat{\boldsymbol{\Gamma}}_T-\boldsymbol{\Gamma}$ is bounded by Lemma 7.8, while the norm of $\boldsymbol{A}\overline{\boldsymbol{A}}^{-1}$ is bounded by the fact that

$$\|\boldsymbol{A}\overline{\boldsymbol{A}}_T^{-1}\| \le \|\boldsymbol{I}\| + \|\boldsymbol{A}(\overline{\boldsymbol{A}}_T^{-1}-\boldsymbol{A}^{-1})\| \le 3$$

with probability at least $1-\delta$, according to Lemma 7.7 under condition (101). In combination, I_1 can be bounded by

$$I_1 \le \left(\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)}+(2\|\boldsymbol{\theta}^\star\|_2+1)\right)(2\|\boldsymbol{\theta}^\star\|_2+1)\sqrt{\frac{\log T}{T}}+o\left(\sqrt{\frac{\log T}{T}}\right) \tag{102}$$

with probability at least $1-2T^{-\frac{1}{2}}$.

Bounding I_2. The norm of I_2 is bounded by estimation error of $\overline{\boldsymbol{A}}^{-1}$. Namely, we further decompose I_2 by triangle inequality as

$$\begin{aligned} I_2 &\le \|(\boldsymbol{A}\overline{\boldsymbol{A}}_T^{-1}-\boldsymbol{I})\boldsymbol{\Gamma}\|_{\mathsf{F}} + \|\boldsymbol{\Gamma}(\overline{\boldsymbol{A}}_T^{-\top}\boldsymbol{A}^\top-\boldsymbol{I})\|_{\mathsf{F}} + \|(\overline{\boldsymbol{A}}_T^{-1}-\boldsymbol{A}^{-1})\boldsymbol{\Gamma}(\overline{\boldsymbol{A}}_T^{-\top}-\boldsymbol{A}^{-\top})\|_{\mathsf{F}} \\ &\le 2\|\boldsymbol{\Gamma}\|\|\boldsymbol{A}(\overline{\boldsymbol{A}}_T^{-1}-\boldsymbol{A}^{-1})\|_{\mathsf{F}} + \|\boldsymbol{\Gamma}\|\|\boldsymbol{A}(\overline{\boldsymbol{A}}_T^{-1}-\boldsymbol{A}^{-1})\|_{\mathsf{F}}^2, \end{aligned}$$

where we can apply Lemma 7.7 to conclude that

$$I_2 \lesssim \frac{\|\boldsymbol{\Gamma}\|}{\lambda_0(1-\gamma)}\sqrt{\frac{\log T}{T}}. \tag{103}$$

with probability at least $1-T^{-\frac{1}{2}}$. Theorem 3.3 follows by combining (102) and (103).

7.4 Proof of Theorem 3.4

We approach the theorem in two steps: first measuring the difference between the distribution of $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ and $\mathcal{N}(\boldsymbol{0},\widehat{\boldsymbol{\Lambda}}_T)$, and then bounding the probability that $\boldsymbol{\theta}$ is within the region $\mathcal{C}_\delta$.

Step 1: Measuring the difference between the distribution of $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ and $\mathcal{N}(\boldsymbol{0},\widehat{\boldsymbol{\Lambda}}_T)$. We firstly observe that by triangle inequality, for any convex set $\mathcal{A}\subset\mathbb{R}^d$ and a d-dimensional standard Gaussian random variable $\boldsymbol{z}$,

$$\begin{aligned} \left|\mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A})-\mathbb{P}(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right| &\le \underbrace{\left|\mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A})-\mathbb{P}(\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right|}_{I_1} \\ &\quad + \underbrace{\left|\mathbb{P}(\overline{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})-\mathbb{P}(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right|}_{I_2} + \underbrace{\left|\mathbb{P}(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})-\mathbb{P}(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right|}_{I_3}. \end{aligned}$$

Here, I_1 and I_2 measures the difference between the non-asymptotic and asymptotic distributions of $\overline{\boldsymbol{\Delta}}_T$, and are bounded by Theorems 7.3 and 7.4 respectively. Specifically,

$$\begin{aligned} &\left|\mathbb{P}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T\in\mathcal{A})-\mathbb{P}(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z}\in\mathcal{A})\right| \\ &\lesssim \sqrt{\frac{\eta_0}{(1-\alpha)(1-\gamma)\lambda_0}}\mathsf{Tr}(\boldsymbol{\Gamma})\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{\alpha}{2}} + \frac{\sqrt{d\,\mathsf{cond}(\boldsymbol{\Gamma})}}{2(1-\gamma)\lambda_0\eta_0}T^{\alpha-1} + O(T^{2\alpha-2}+T^{-\frac{1}{2}}). \end{aligned} \tag{104}$$

where $\boldsymbol{z}$ is a d-dimensional standard Gaussian random variable independent of the sample trajectory $\{(s_t, s_t')\}_{t\geq 0}$. Hence, the proof boils down to bounding I_3, which measures between the difference the centered Gaussian distribution with variance matrices $\widehat{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$. Towards this end, we use $\mathcal{E}$ to denote the event of

$$\begin{aligned}
&d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \\
&\lesssim \left(\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} + (2\|\boldsymbol{\theta}^\star\|_2 + 1)\right)(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|\sqrt{\frac{\log T}{T}} + o\left(\sqrt{\frac{\log T}{T}}\right).
\end{aligned}$$

it has been illustrated in Theorem 3.3 that

$$\mathbb{P}(\mathcal{E}) \geq 1 - 3T^{-\frac{1}{2}};$$

meanwhile by definition,

$$\begin{aligned}
&\left|\mathbb{P}\left(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right)\right| \\
&\lesssim \left(\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} + (2\|\boldsymbol{\theta}^\star\|_2 + 1)\right)(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|\sqrt{\frac{\log T}{T}} + o\left(\sqrt{\frac{\log T}{T}}\right).
\end{aligned}$$

Therefore, since $\boldsymbol{z}$ is independent of the sample trajectory, the law of total probability implies

$$\begin{aligned}
\left|\mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right)\right| &= \left|\mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right)(\mathbb{P}(\mathcal{E}) - 1) + \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}^{\mathsf{C}}\right)\mathbb{P}\left(\mathcal{E}^{\mathsf{C}}\right)\right| \\
&= \mathbb{P}(\mathcal{E}^{\mathsf{C}}) \cdot \left|\mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}^{\mathsf{C}}\right)\right| \\
&\leq \mathbb{P}(\mathcal{E}^{\mathsf{C}}) \leq 3T^{-\frac{1}{2}}.
\end{aligned}$$

Consequently, the triangle inequality directly implies that

$$\begin{aligned}
&\left|\mathbb{P}\left(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right)\right| \\
&\leq \left|\mathbb{P}\left(\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right)\right| + \left|\mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A}\right) - \mathbb{P}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z} \in \mathcal{A} \mid \mathcal{E}\right)\right| \\
&\leq \left(\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} + (2\|\boldsymbol{\theta}^\star\|_2 + 1)\right)(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|\sqrt{\frac{\log T}{T}} + o\left(\sqrt{\frac{\log T}{T}}\right) + 3T^{-\frac{1}{2}} \\
&= \left(\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} + (2\|\boldsymbol{\theta}^\star\|_2 + 1)\right)(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|\sqrt{\frac{\log T}{T}} + o\left(\sqrt{\frac{\log T}{T}}\right). \qquad (105)
\end{aligned}$$

Combining (104), (105) and taking the supremum of $\mathcal{A}$ over all convex sets, we obtain

$$d_{\mathsf{C}}\left(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Lambda}}_T)\right) \lesssim \frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{2(1-\gamma)\lambda_0\eta_0}T^{-\frac{1}{3}} + \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)}(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{1}{3}} + o(T^{-\frac{1}{3}}). \qquad (106)$$

Step 2: Bounding $\mathbb{P}(\boldsymbol{\theta}^\star \in \mathcal{C}_\delta)$. Define $\mathcal{C}_\delta'$ as

$$\mathcal{C}_\delta' := \left\{y = \sqrt{T}(\overline{\boldsymbol{\theta}}_T - \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{C}_\delta\right\}.$$

Then since $\mathcal{C}_\delta$ is convex, it is easy to verify that $\mathcal{C}_\delta'$ is also convex. Meanwhile on one hand,

$$\mathbb{P}(\boldsymbol{\theta}^\star \in \mathcal{C}_\delta) = \mathbb{P}\left(\sqrt{T}(\overline{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star) \in \mathcal{C}_\delta'\right);$$

On the other hand,

$$\mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\left(\overline{\boldsymbol{\theta}}_T-\frac{\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}}{\sqrt{T}}\in\mathcal{C}_\delta\middle|\overline{\boldsymbol{\theta}}_T,\widehat{\boldsymbol{\Lambda}}_T\right)=\mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{C}'_\delta\right).$$

As a direct consequence, (106) and the triangle inequality indicate that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\theta}^\star\in\mathcal{C}_\delta)&=\mathbb{P}\left(\sqrt{T}(\overline{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star)\in\mathcal{C}'_\delta\right)\\
&\geq\mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\left(\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\in\mathcal{C}'_\delta\right)-d_{\mathsf{C}}\left(\sqrt{T}\overline{\boldsymbol{\Delta}}_T,\mathcal{N}(\mathbf{0},\widehat{\boldsymbol{\Lambda}}_T)\right)\\
&=\mathbb{P}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}\left(\overline{\boldsymbol{\theta}}_T-\frac{\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}}{\sqrt{T}}\in\mathcal{C}_\delta\middle|\overline{\boldsymbol{\theta}}_T,\widehat{\boldsymbol{\Lambda}}_T\right)-d_{\mathsf{C}}\left(\sqrt{T}\overline{\boldsymbol{\Delta}}_T,\mathcal{N}(\mathbf{0},\widehat{\boldsymbol{\Lambda}}_T)\right)\\
&\geq 1-\delta-\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)}(2\|\boldsymbol{\theta}^\star\|_2+1)\|\boldsymbol{\Gamma}^{-1}\|T^{-\frac{1}{3}}-\frac{\sqrt{d\mathsf{cond}(\boldsymbol{\Gamma})}}{2(1-\gamma)\lambda_0\eta_0}T^{-\frac{1}{3}}-o(T^{-\frac{1}{3}}).
\end{aligned}$$

□

8 Proof of supportive lemmas and propositions

8.1 Proof of Lemma 7.5

The lemma is proved by developing an iterative relation for the series $\{\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2\}$. By definition,

$$\begin{aligned}
\boldsymbol{\Delta}_t&=(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^\star)-\eta_t(\boldsymbol{A}_t\boldsymbol{\theta}_{t-1}-\boldsymbol{b}_t)\\
&=(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^\star)-\eta_t\boldsymbol{A}_t(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^\star)-\eta_t(\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t)\\
&=(\boldsymbol{I}-\eta_t\boldsymbol{A}_t)\boldsymbol{\Delta}_{t-1}-\eta_t(\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t).
\end{aligned}$$

Therefore, under $i.i.d.$ samples, we can relate $\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2$ with $\mathbb{E}\|\boldsymbol{\Delta}_{t-1}\|_2^2$ by

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2&=\mathbb{E}\|(\boldsymbol{I}-\eta_t\boldsymbol{A}_t)\boldsymbol{\Delta}_{t-1}\|_2^2+2\eta_t\mathbb{E}\langle(\boldsymbol{I}-\eta_t\boldsymbol{A}_t)\boldsymbol{\Delta}_{t-1},\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\rangle+\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\\
&=\mathbb{E}\|(\boldsymbol{I}-\eta_t\boldsymbol{A}_t)\boldsymbol{\Delta}_{t-1}\|_2^2+2\eta_t^2\mathbb{E}\langle\boldsymbol{A}_t\boldsymbol{\Delta}_{t-1},\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\rangle+\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\\
&=\mathbb{E}\|(\boldsymbol{I}-\eta_t\boldsymbol{A}_t)\boldsymbol{\Delta}_{t-1}\|_2^2+\eta_t^2\left(\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\Delta}_{t-1}\|_2^2+\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\right)\\
&\quad+\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\\
&=\|\boldsymbol{\Delta}_{t-1}\|_2^2-\eta_t\boldsymbol{\Delta}_{t-1}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{\Delta}_{t-1}\\
&\quad+2\eta_t^2\boldsymbol{\Delta}_{t-1}^\top\mathbb{E}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]\boldsymbol{\Delta}_{t-1}+\frac{3}{2}\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\\
&\leq\|\boldsymbol{\Delta}_{t-1}\|_2^2-(\eta_t-2\eta_t^2)\boldsymbol{\Delta}_{t-1}^\top(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{\Delta}_{t-1}+2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2.
\end{aligned}$$

Notice that we applied Lemma 6.4 in the last inequality. Now, since the stepsize η_t satisfies $\eta_t\leq\eta_0\leq\frac{1}{4}$, Lemma 6.4 implies

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2&\leq\mathbb{E}\|\boldsymbol{\Delta}_{t-1}\|_2^2-\frac{1}{2}\eta_t\mathbb{E}[(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^\star)^\top(\boldsymbol{A}+\boldsymbol{A}^\top)(\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^\star)]+2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2\\
&<\left(1-(1-\gamma)\lambda_0\eta_t\right)\mathbb{E}\|\boldsymbol{\Delta}_{t-1}\|_2^2+2\eta_t^2\mathsf{Tr}(\boldsymbol{\Gamma}).
\end{aligned}$$

Therefore by induction, $\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2$ is bounded by

$$\mathbb{E}\|\boldsymbol{\Delta}_t\|_2^2<\prod_{i=1}^t\left(1-(1-\gamma)\lambda_0\eta_i\right)\|\boldsymbol{\Delta}_0\|_2^2+2\sum_{i=1}^n\eta_i^2\prod_{j=i+1}^n\left(1-(1-\gamma)\lambda_0\eta_j\right)\mathsf{Tr}(\boldsymbol{\Gamma}).\tag{107}$$

The proof now boils down to bounding the two terms on the right hand side of (107). For the first term, we observe

$$\prod_{i=1}^{t}(1-(1-\gamma)\lambda_0\eta_i) < \exp\left((1-\gamma)\lambda_0\eta_0\sum_{i=1}^{t}i^{-\alpha}\right) \lesssim \exp\left(-\frac{1-\gamma}{(1-\alpha)}\lambda_0\eta_0 t^{1-\alpha}\right) = o(t^{-1}); \tag{108}$$

for the second term, we can apply Lemma 6.6 with $\beta = (1-\gamma)\lambda_0\eta_0$ to show that

$$\sum_{i=1}^{n}\eta_i^2\prod_{j=i+1}^{n}(1-(1-\gamma)\lambda_0\eta_j) = \frac{\eta_0}{(1-\gamma)\lambda_0}t^{-\alpha}+O(t^{-1}). \tag{109}$$

The lemma follows immediately by plugging (108) and (109) into (107).

8.2 Proof of Proposition 7.6

We address the two properties in order.

Proof of Equation (93a). Recall from (8) and (89) that for any $i \in [T]$,

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \eta_i(\boldsymbol{A}_i\boldsymbol{\theta}_{i-1} - \boldsymbol{b}_i); \quad \text{and}$$
$$\boldsymbol{\theta}_i^{(i)} = \boldsymbol{\theta}_{i-1} - \eta_i(\check{\boldsymbol{A}}_i\boldsymbol{\theta}_{i-1} - \check{\boldsymbol{b}}_i).$$

Therefore, the difference between $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_i^{(i)}$ can be bounded by

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(i)}\|_2^2 &\le \eta_i^2\mathbb{E}\|(\boldsymbol{A}_i-\check{\boldsymbol{A}}_i)\boldsymbol{\theta}_{i-1} - (\boldsymbol{b}_i-\check{\boldsymbol{b}}_i)\|_2^2 \\ &= \eta_i^2\mathbb{E}\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}^\star + (\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1} - (\boldsymbol{b}_i-\boldsymbol{b})\|_2^2 \\ &\le \eta_i^2(2\mathbb{E}\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_{i-1}\|_2^2 + 2\mathbb{E}\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}^\star - (\boldsymbol{b}_i-\boldsymbol{b})\|_2^2) \\ &\le 2\eta_i^2\mathsf{Tr}(\boldsymbol{\Gamma}) + 2\eta_i^2\mathbb{E}\|\boldsymbol{\Delta}_{i-1}\|_2^2.\end{aligned}$$

Proof of Equation (93b). Recall from (8) and (89) that

$$\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1} - \eta_j(\boldsymbol{A}_j\boldsymbol{\theta}_{j-1} - \boldsymbol{b}_j);$$
$$\boldsymbol{\theta}_j^{(i)} = \boldsymbol{\theta}_{j-1}^{(i)} - \eta_j(\boldsymbol{A}_j\boldsymbol{\theta}_{j-1}^{(i)} - \boldsymbol{b}_j).$$

Therefore, the difference between $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_j^{(i)}$ can be expressed as

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^{(i)}\|_2^2 &= \mathbb{E}\|(\boldsymbol{I}-\eta_j\boldsymbol{A}_j)(\boldsymbol{\theta}_{j-1} - \boldsymbol{\theta}_{j-1}^{(i)})\|_2^2 \\ &\le \left(1-\frac{1}{2}(1-\gamma)\lambda_0\eta_j\right)\mathbb{E}\|\boldsymbol{\theta}_{j-1} - \boldsymbol{\theta}_{j-1}^{(i)}\|_2^2,\end{aligned}$$

where the inequality follows from Lemma 6.4. By induction, we obtain

$$\mathbb{E}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^{(i)}\|_2^2 \le \prod_{k=i+1}^{j}\left(1-\frac{1}{2}(1-\gamma)\lambda_0\eta_k\right)\mathbb{E}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(i)}\|_2^2.$$

8.3 Proof of Lemma 7.7

We firstly observe that since $\|\boldsymbol{\phi}(s)\|_2 \le 1$ for all $s \in \mathcal{S}$, it can be guaranteed that

$$\begin{aligned}\|\mathbf{vec}(\boldsymbol{A}_t)\|_2^2 = \|\boldsymbol{A}_t\|_{\mathsf{F}}^2 &= \mathsf{Tr}(\boldsymbol{A}_t^\top\boldsymbol{A}_t) \\ &= \mathsf{Tr}\left([\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')]\boldsymbol{\phi}(s_t)^\top\boldsymbol{\phi}(s_t)[\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')]^\top\right) \\ &\le \mathsf{Tr}\left([\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')][\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')]^\top\right) < 4\end{aligned}$$

holds uniformly for all $0 \le t \le T$. Hence the vector Azuma's inequality 6.18 directly implies that

$$\|\overline{\boldsymbol{A}}_T - \boldsymbol{A}\|_{\mathsf{F}} = \big\|\mathbf{vec}(\overline{\boldsymbol{A}}_T - \boldsymbol{A})\big\| \le \frac{4\sqrt{2}}{\sqrt{T}}\sqrt{\log\frac{3}{\delta}}$$

with probability at least $1-\delta$. In what follows, we aim to transform this error bound to the error bound for $\overline{\boldsymbol{A}}_T^{-1}$. Towards this end, we observe that condition (101) guarantees with probability at least $1-\delta$,

$$\begin{aligned}\big\|\boldsymbol{A}^{-1}(\overline{\boldsymbol{A}}_T - \boldsymbol{A})\big\| &\le \|\boldsymbol{A}^{-1}\|\|\overline{\boldsymbol{A}}_T - \boldsymbol{A}\| \\ &\le \lambda_0^{-1}(1-\gamma)^{-1}\frac{4\sqrt{2}}{\sqrt{T}}\sqrt{\log\frac{3}{\delta}} < \frac{1}{2}\end{aligned}$$

where the last inequality is guaranteed by the condition (101). Hence, the condition of Lemma 6.3 is satisfied with probability at least $1-\delta$ by taking $\boldsymbol{Y} = \boldsymbol{A}^{-1}\overline{\boldsymbol{A}}$ and $\boldsymbol{X} = \boldsymbol{I}$. Consequently, the difference between $\overline{\boldsymbol{A}}^{-1}\boldsymbol{A}$ and $\boldsymbol{I}$ is bounded by

$$\begin{aligned}\|(\overline{\boldsymbol{A}}_T^{-1} - \boldsymbol{A}^{-1})\boldsymbol{A}\|_{\mathsf{F}} \le 2\|(\overline{\boldsymbol{A}}_T - \boldsymbol{A})\boldsymbol{A}^{-1}\|_{\mathsf{F}} &\le \frac{4\sqrt{2}}{\sqrt{T}}\sqrt{\log\frac{3}{\delta}}\|\boldsymbol{A}^{-1}\| \\ &\le \frac{4\sqrt{2}}{\lambda_0(1-\gamma)\sqrt{T}}\sqrt{\log\frac{3}{\delta}}.\end{aligned}$$

□

8.4 Proof of Lemma 7.8

We firstly represent $\widehat{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma}$ as

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_T - \boldsymbol{\Gamma} = &\underbrace{\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)^\top - (\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right]}_{I_1} \\ &+ \underbrace{\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right] - \mathbb{E}[(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top]}_{I_2},\end{aligned}$$

where I_1 is mainly controlled by the estimation error of $\overline{\boldsymbol{\theta}}_t$, and I_2 only involves the randomness of $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}$. We now control the norms of these two terms in order.

Bounding the Frobenius norm of I_1. By definition, I_1 can be further decomposed as

$$\begin{aligned}I_1 &= \sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\theta}}_t - \boldsymbol{b}_t)^\top - (\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right] \\ &= \underbrace{\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right]}_{I_{11}} + \underbrace{\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top + (\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right]}_{I_{12}}.\end{aligned}$$

By triangle inequality, the Frobenius norm of I_{11} is bounded by

$$\begin{aligned}\left\|\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right]\right\|_{\mathsf{F}} &\le \sum_{t=1}^{T} w_{t,T}\left\|(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right\|_{\mathsf{F}} \\ &\le \sum_{t=1}^{T} w_{t,T}\sqrt{\mathsf{Tr}\left((\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right)} \\ &\le \sum_{t=1}^{T} w_{t,T}\cdot 4\|\overline{\boldsymbol{\Delta}}_t\|_2^2. \end{aligned} \tag{110}$$

Since

$$\mathbb{E}\|\overline{\boldsymbol{\Delta}}_t\|_2^2 = \frac{1}{t}\mathsf{Tr}(\overline{\boldsymbol{\Lambda}}_t) + \tilde{C}t^{-\alpha-1} = \frac{1}{t}\mathsf{Tr}(\boldsymbol{\Lambda}^\star) + o(t^{-1}) \le \frac{\mathsf{Tr}(\boldsymbol{\Gamma})}{\lambda_0^2(1-\gamma)^2 t} + o(t^{-1}),$$

the expectation of this bound can be bounded by

$$\begin{aligned}\mathbb{E}\left\|\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right]\right\|_{\mathsf{F}} &\le 4\sum_{t=1}^{T} w_{t,T}\cdot\mathbb{E}\|\overline{\boldsymbol{\Delta}}_t\|_2^2 \\ &\le 4\sum_{t=1}^{T}\frac{2t}{T(T+1)}\cdot\left(\frac{\mathsf{Tr}(\boldsymbol{\Gamma})}{\lambda_0^2(1-\gamma)^2 t} + o(t^{-1})\right) \\ &\asymp \frac{\mathsf{Tr}(\boldsymbol{\Gamma})}{\lambda_0^2(1-\gamma)^2 T} + o(T^{-1});\end{aligned}$$

Consequently, the Markov's inequality indicates

$$\|I_{11}\|_{\mathsf{F}} \le \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)\sqrt{T}} + o(T^{-\frac{1}{2}}) \tag{111}$$

with probability at least $1 - \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)\sqrt{T}}$.

Meanwhile, the norm of I_{12} is bounded by

$$\begin{aligned}&\left\|\sum_{t=1}^{T} w_{t,T}\left[(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top + (\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right]\right\|_{\mathsf{F}} \\ &\le \sum_{t=1}^{T} w_{t,T}\left\|(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top + (\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top\right\|_{\mathsf{F}} \\ &\le 2\sum_{t=1}^{T} w_{t,T}\left\|(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right\|_{\mathsf{F}} \\ &\le 2\sum_{t=1}^{T} w_{t,T}\sqrt{\mathsf{Tr}\left((\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)^\top(\boldsymbol{A}_t\overline{\boldsymbol{\Delta}}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right)} \\ &\le 4\sum_{t=1}^{T} w_{t,T}(2\|\boldsymbol{\theta}^\star\|_2 + 1)\|\overline{\boldsymbol{\Delta}}_t\|_2 \\ &= 4(2\|\boldsymbol{\theta}^\star\|_2 + 1)\sum_{t=1}^{T} w_{t,T}\|\overline{\boldsymbol{\Delta}}_t\|_2. \end{aligned} \tag{112}$$

Furthermore, since $\sum_{t=1}^{T} w_{t,T} = 1$, the Jensen's inequality guarantees

$$\begin{aligned}\mathbb{E}\left(\sum_{t=1}^{T} w_{t,T}\|\overline{\boldsymbol{\Delta}}_t\|_2\right)^2 &\leq \sum_{t=1}^{T} w_{t,T}\mathbb{E}\|\overline{\boldsymbol{\Delta}}_t\|_2^2 \\ &\leq \sum_{t=1}^{T} \frac{2t}{T(T+1)}\left(\frac{\mathsf{Tr}(\boldsymbol{\Gamma})}{\lambda_0^2(1-\gamma)^2 t} + o(t^{-1})\right) \\ &\asymp \frac{\mathsf{Tr}(\boldsymbol{\Gamma})}{\lambda_0^2(1-\gamma)^2 T} + o(T^{-1})\end{aligned}$$

Therefore, the Chebyshev's inequality indicates

$$\|I_{22}\|_{\mathsf{F}} \lesssim 4(2\|\boldsymbol{\theta}^\star\|_2 + 1)\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} T^{-\frac{1}{3}} + o(T^{-\frac{1}{3}}) \tag{113}$$

with probability at least $1 - T^{-\frac{1}{3}}$.

Hence by combining (111) and (113), we obtain

$$\|I_1\|_{\mathsf{F}} \lesssim 4(2\|\boldsymbol{\theta}^\star\|_2 + 1)\frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)} T^{-\frac{1}{3}} + o(T^{-\frac{1}{3}}) \tag{114}$$

with probability at least $1 - T^{-\frac{1}{3}} - \frac{\sqrt{\mathsf{Tr}(\boldsymbol{\Gamma})}}{\lambda_0(1-\gamma)\sqrt{T}}$.

Bounding the Frobenius norm of I_2. Notice that I_2 only involves the randomness of $i.i.d.$ samples $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}$, and

$$\begin{aligned}&\left\|\mathbf{vec}\left((\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right)\right\|_2 \\ &= \left\|(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top\right\|_{\mathsf{F}} \\ &= \sqrt{\mathsf{Tr}((\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)(\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t)^\top)} \\ &< (2\|\boldsymbol{\theta}^\star\|_2 + 1)^2\end{aligned}$$

holds uniformly for all $t \in [T]$. Consequently, the Frobenius norm of I_2 can be bounded by directly applying the vector Azuma's inequality: for every $\delta \in (0,1)$,

$$\|I_2\|_{\mathsf{F}} \leq 2\sqrt{2}\sqrt{\sum_{t=1}^{T} w_{t,T}^2}(2\|\boldsymbol{\theta}^\star\|_2 + 1)^2\sqrt{\log\frac{3}{\delta}}; \tag{115}$$

with probability at least $1 - \delta$. Since

$$\begin{aligned}\sum_{t=1}^{T} w_{t,T}^2 &= \frac{4}{T^2(T+1)^2}\sum_{t=1}^{T} t^2 \\ &= \frac{4}{T^2(T+1)^2}\cdot\frac{T(T+1)(2T+1)}{6} \asymp \frac{1}{T},\end{aligned}$$

Lemma 7.8 follows by combining (114) and (115) with $\delta = T^{-\frac{1}{3}}$. □

9 Experimental details

9.1 Online algorithm to construct the variance estimator

The code for our experiments can be found at https://github.com/Weichen-Wu-1996/TD_inference. Experiments were performed on a personal laptop (MacOS). No external GPU clusters or cloud resources were used.

Algorithm 1 Online algorithm for computing $\overline{\boldsymbol{\theta}}_T$ and $\widehat{\boldsymbol{\Lambda}}_T$

Require: MDP transition kernel $\mathcal{P}$, stationary distribution μ, discount factor γ, feature map $\boldsymbol{\phi}$, initial stepsize η_0, stepsize decay parameter $\alpha \in (0.5, 1)$, sample size T

1: Initialize $\boldsymbol{\theta} = \mathbf{0}$, $\overline{\boldsymbol{\theta}} = \mathbf{0}$, $\overline{\boldsymbol{A}} = \mathbf{0}$, $\widehat{\boldsymbol{\Gamma}} = \mathbf{0}$.
2: **for** $t = 1, \ldots, T$ **do**
  Sample $s \sim \mu$, $s' \sim \mathcal{P}(\cdot \mid s)$.
  Compute: $\boldsymbol{A} \leftarrow \boldsymbol{\phi}(s)[\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}(s')]$, $\boldsymbol{b} \leftarrow r(s)\,\boldsymbol{\phi}(s)$.
  Perform TD update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_0 t^{-\alpha}(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{b})$.
  Perform online updates:
  $\overline{\boldsymbol{\theta}} \leftarrow \overline{\boldsymbol{\theta}} + \frac{1}{t}\,(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}})$.
  $\overline{\boldsymbol{A}} \leftarrow \overline{\boldsymbol{A}} + \frac{1}{t}\,(\boldsymbol{A} - \overline{\boldsymbol{A}})$.
  $\widehat{\boldsymbol{\Gamma}} \leftarrow \widehat{\boldsymbol{\Gamma}} + \frac{2}{t+1}\big[(\boldsymbol{A}\overline{\boldsymbol{\theta}} - \boldsymbol{b})(\boldsymbol{A}\overline{\boldsymbol{\theta}} - \boldsymbol{b})^\top - \widehat{\boldsymbol{\Gamma}}\big]$.
3: **end for**
4: Construct the variance estimator:
5: $\widehat{\boldsymbol{\Lambda}} \leftarrow \overline{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{\Gamma}}\,\overline{\boldsymbol{A}}^{-\top}$.

Ensure: Averaged TD estimator $\overline{\boldsymbol{\theta}}$, covariance matrix estimator $\widehat{\boldsymbol{\Lambda}}$.

It is easy to verify that by definition (17), the estimators $\overline{\boldsymbol{A}}_T$ and $\widehat{\boldsymbol{\Gamma}}_T$ can be updated by

$$\overline{\boldsymbol{A}}_T = \overline{\boldsymbol{A}}_{T-1} + \frac{1}{T}(\boldsymbol{A}_T - \overline{\boldsymbol{A}}_{T-1}), \quad \text{and}$$
$$\widehat{\boldsymbol{\Gamma}}_T = \widehat{\boldsymbol{\Gamma}}_{T-1} + \frac{2}{T+1}\left((\boldsymbol{A}_T\overline{\boldsymbol{\theta}}_T - \boldsymbol{b}_T)(\boldsymbol{A}_T\overline{\boldsymbol{\theta}}_T - \boldsymbol{b}_T)^\top - \widehat{\boldsymbol{\Gamma}}_{T-1}\right).$$

Algorithm 1 reveals the complete process of generating both the averaged TD estimator $\overline{\boldsymbol{\theta}}_T$ and the variance estimator $\widehat{\boldsymbol{\Lambda}}_T$. Since Algorithm 1 iterates from $t = 0$ to $t = T - 1$ in a single loop and does not store the samples for each iteration, it has a space complexity of $O(Td^2 + d^3)$ and a space complexity of $O(d^2)$. Here, the d^3 term stems from calculating matrix inverse.

9.2 The Markov Decision Process and the feature map

Given a finite state space $\mathcal{S}$, define a sequence of MDP $\{\mathcal{M}_{\boldsymbol{q}}\}$ indexed by $\boldsymbol{q} \in \mathcal{Q} \subset \{q_+, q_-\}^{d-1}$ where for each $\boldsymbol{q}$, the transition kernel equals

$$P_{\boldsymbol{q}}(s' \mid s) = \begin{cases} q_s\mathbb{1}(s' = s) + \frac{1-q_s}{|\mathcal{S}|-d+1}\mathbb{1}(s' \geq d) & \text{for} \quad s < d; \\ \frac{\gamma}{|\mathcal{S}|-d+1}\mathbb{1}(s' \geq d) + \frac{1-q_{s'}}{d-1}\mathbb{1}(s' < d) & \text{for} \quad s \geq d. \end{cases} \tag{116}$$

and the reward function equals $r(s) = \mathbb{1}(s \geq d)$.

Here, for each $i \in [d-1]$, q_i is taken to be either q_+ or q_- where

$$q_+ := \gamma + (1-\gamma)^2\varepsilon, \qquad \text{and} \qquad q_- := \gamma - (1-\gamma)^2\varepsilon.$$

We further impose the constraint that the number of q_+'s and q_-'s in $\boldsymbol{q}$ are the same, namely,

$$\sum_{s=1}^{d-1}\mathbb{1}(q_s = q_+) = \sum_{s=1}^{d-1}\mathbb{1}(q_s = q_-) = (d-1)/2. \tag{117}$$

Here without loss of generality, assume d is an odd number. With these definitions in place, it can be easily verified that the stationary distribution for $\boldsymbol{P}$ obeys

$$\mu(s) = \begin{cases} \frac{1}{2(d-1)} & \text{for} \quad s < d; \\ \frac{1}{2(|\mathcal{S}|-d+1)} & \text{for} \quad s \geq d. \end{cases} \tag{118}$$

Moreover, suppose the feature map is taken to be

$$\boldsymbol{\phi}(s) = \boldsymbol{e}_{s\wedge d} \in \mathbb{R}^d,$$

then one can further verify that

$$\begin{aligned}\theta^\star(d) = V^\star(s) &= \frac{1}{1-\gamma^2-\sum_{i=1}^{d-1}\frac{\gamma^2(1-q_i)^2}{(d-1)(1-\gamma q_i)}},\\ \theta^\star(i) = V^\star(i) &= \frac{\gamma(1-q_i)}{1-\gamma q_i}V^\star(s), \text{ for } s \geq d \text{ and } i < d.\end{aligned} \tag{119}$$

From the expressions above, we remark that, the values of q and $V^\star(s)$ with $s \geq d$ are fixed for all $\boldsymbol{q} \in \mathcal{Q}$ which is ensured by the construction (117). It is shown in Theorem 2 of Li et al. (2024b) that the sequence $\{\mathcal{M}_{\boldsymbol{q}}\}_{q\in\mathcal{Q}}$ is hard to distinguish for a considerably large set $\mathcal{Q}$.

9.3 Parameter specification

In our experiments, we fix the size of the state space to $|\mathcal{S}| = 10$, the discount factor to be $\gamma = 0.2$, and let $\varepsilon = 0.01$. For the TD learning algorithm, we set the initial stepsize to be $\eta_0 = 5$, and total number of iterations to be $T = 2^{20}$. Data is recorded at every power of 2, so that the log plot has constant intervals on the x-axis. In order to show the role played by the stepsize decaying speed α, we fix $d = 3$ and let α iterate through the group $(\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1$; and in order to illustrate the influence of the dimension of the feature map, we fix $\alpha = \frac{2}{3}$ and let d iterate through the group of $(3, 5, 7, 9)$. For every given d, we set $\boldsymbol{q}$ to be

$$q_i = \begin{cases} q_+, & \text{for } 1 \leq i \leq \frac{d-1}{2};\\ q_-, & \text{for } \frac{d+1}{2} \leq i \leq d.\end{cases}$$

We verify our theoretical results through the following:

1. In order to illustrate the high-probability convergence of the TD estimation error as indicated by Theorem 3.2, we measure the 95% quantile of the L_2 norm of $\overline{\boldsymbol{\Delta}}_T$ in $10,000$ independent trials;
2. In order to illustrate the Berry-Esseen bound as indicated by Theorems 7.3 and 7.4, we measure the difference between the empirical and asymptotic distributions of $\overline{\boldsymbol{\Delta}}_T$ by

$$d(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^\star)) := \max_{1\leq j\leq d}\sup_{x\in\mathbb{R}}\left|\mathbb{P}(\sqrt{T}(\overline{\boldsymbol{\Delta}}_T)_j \leq x) - \mathbb{P}((\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z})_j \leq x)\right|,$$

   averaged among $10,000$ independent trials. Besides, as an intuitive demonstration, Figure 1 in the introduction section records the empirical and asymptotic distributions of $(\overline{\boldsymbol{\Delta}}_T)_1$ when $d = 3$ and $\alpha = \frac{2}{3}$.
3. In order to illustrate the efficiency of our variance estimator $\widehat{\boldsymbol{\Lambda}}_T$, we measure the averaged Frobenius norm of the difference between $\widehat{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$ among $10,000$ independent trials. As is indicated by 6.14, this is a direct proxy for the TV distance between the zero-mean Gaussian distributions with the two corresponding variance matrices.
4. In order to validate our statistical inference procedure, we run Algorithm 1 on the MDP with $d = 3$ using $\alpha = \frac{2}{3}$, construct both individual and simultaneous confidence intervals for $\boldsymbol{\theta}^\star$, and measure the coverage rates of these confidence intervals in $10,000$ independent trials. Details to follow in section 9.4.4.

9.4 Full experimental results

9.4.1 High-probability convergence guarantee for the TD estimator

Figure 3 shows the decay of the 95% quantile of $\|\overline{\boldsymbol{\Delta}}_T\|_2^2$; the slope of this log-log plot verifies our theoretical claim in Theorem 3.2 that $\|\overline{\boldsymbol{\Delta}}_T\|_2$ converges by the rate of $O(T^{-\frac{1}{2}})$ with high probability.

Plot (a) illustrates the behavior of the estimation error under different choices of α. As a reference, the black dotted line represents the 95% quantile of $\{\boldsymbol{Z}_T\}$ where $\boldsymbol{Z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star/T)$. Our theory predicts that when T grows large, this should be the asymptotic line for the empirical 95% quantiles when $\alpha \in (\frac{1}{2}, 1)$. This is exactly what we observe that $\alpha = \frac{2}{3}$ and $\alpha = \frac{3}{4}$. On the contrary, when $\alpha = 1$, the empirical quantiles do *not* converge to the asymptotic quantiles. This is also predicted by our theory that since $\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star = \boldsymbol{O}(T^{\alpha-1})$ (which does not converge when $\alpha = 1$), the variance of $\overline{\boldsymbol{\Delta}}_T$ will *not* converge to $\boldsymbol{\Lambda}^\star$ when $\alpha = 1$. This fact is also reflected in the $O(T^{\frac{\alpha}{2}-1})$ term in the bound of Theorem 3.2, which when $\alpha = 1$, converges by $O(T^{-\frac{1}{2}})$, the same rate as the leading term. As for the case of $\alpha = \frac{1}{2}$, we observe that when T is small, $\|\overline{\boldsymbol{\Delta}}_T\|_2$ is relatively large; it is not until a very large number of iterations (i.e., $T > 10^5$), that the empirical and asymptotic lines begin to overlap. This phenomenon is also predicted by the $O(T^{-\frac{\alpha+1}{2}})$ term in the bound of Theorem 3.2, which converges slower when $\alpha = \frac{1}{2}$. A more detailed explanation follows in Appendix 9.5.

Meanwhile, plot(b) shows that, fixing $\alpha = \frac{2}{3}$, the convergence is slower when d is larger. Both are in accordance with our conclusion in Theorem (1) that $\|\overline{\boldsymbol{\Delta}}_T\|_2$ converges by the rate of

$$O\left(\sqrt{\frac{\mathsf{Tr}(\boldsymbol{\Lambda}^\star)}{T}\log\frac{1}{\delta}}\right)$$

with probability at least $1-\delta$.

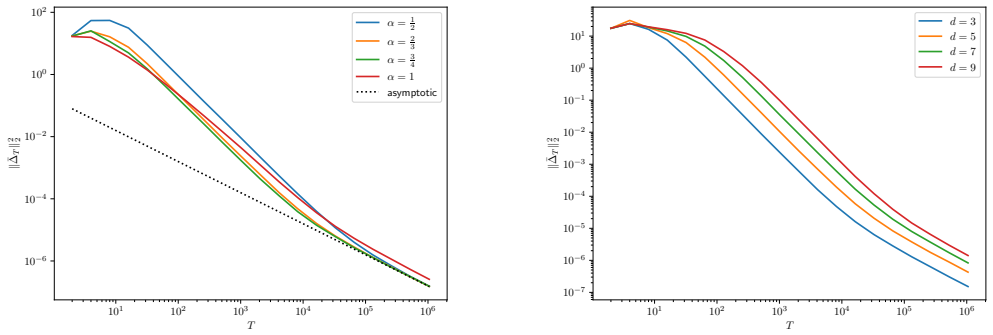

Figure 3: The decay of the 95% quantile of $\|\overline{\boldsymbol{\Delta}}_T\|_2^2$ when (a) $d = 3$, for different choices of α; and (b) $\alpha = \frac{2}{3}$, for different choices of d.

9.4.2 Berry-Esseen bound

Figure 4 illustrates how the finite-sample distribution of $\sqrt{T}\overline{\boldsymbol{\Delta}}_T$ converges to its asymptotic Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)$. For ease of computation, instead of the convex distance we evaluated the distance

$$\max_{1\le j\le d}\sup_{x\in\mathbb{R}}\left|\widehat{\mathbb{P}}(\sqrt{T}(\overline{\boldsymbol{\Delta}}_T)_j \le x) - \mathbb{P}_{\boldsymbol{z}}((\boldsymbol{\Lambda}^{\star\frac{1}{2}}\boldsymbol{z})_j \le x)\right|, \tag{120}$$

where $\widehat{\mathbb{P}}$ refers to the empirical distribution in our $10,000$ trials, and $\boldsymbol{z}$ is the d-dimensional standard Gaussian.

Notably, plot (a) indicates that the convergence is slower when $\alpha = \frac{1}{2}$, compared to the choices of $\alpha = \frac{2}{3}$ and $\alpha = \frac{3}{4}$. This directly contradicts the suggestion by Samsonov et al. (2024a) to choose $\alpha = \frac{1}{2}$ in order to yield a faster in-distribution convergence of the TD estimation error. On the other hand, when $\alpha = 1$, the empirical distribution does *not* converge to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)$; that verifies our claim in Appendix 7.2.3 that the $O(T^{\alpha-1})$ term in the Berry-Esseen bound is unavoidable.

Meanwhile, plot (b) shows the convergence generally is slower when d is larger.

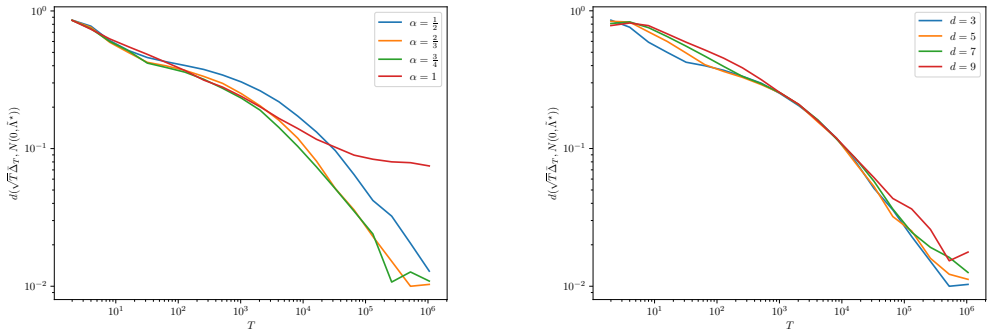

Figure 4: The decay of $d(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star))$, when (a) $d = 3$, for different choices of α; and (b) $\alpha = \frac{2}{3}$, for different choices of d.

9.4.3 Estimator of the covariance matrix

Figure 5 illustrates the rate by which our variance estimator $\widehat{\boldsymbol{\Lambda}}_T$ converges to the asymptotic variance $\boldsymbol{\Lambda}^\star$. Plot (a) shows that the convergence rate is similar for different choices of α when the number of iterations T is sufficiently large; plot(b) indicates that the convergence is slower when d is larger, the nearly parallel curves in the log-log plot demonstrate that the convergence rate with respect to T is similar among different choices of d.

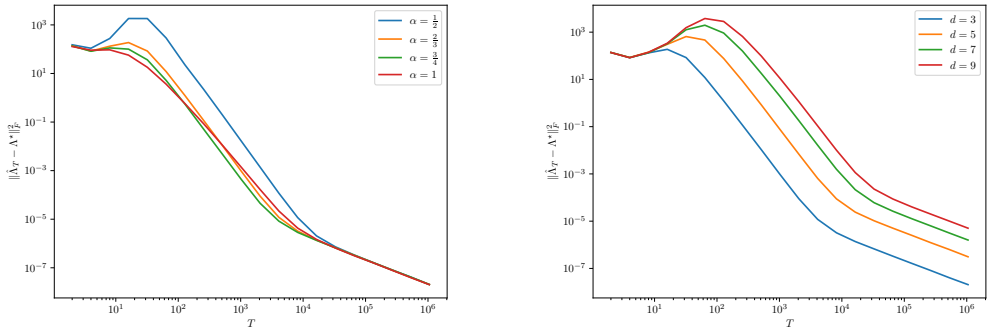

Figure 5: The decay of the difference between estimated and asymptotic covariance matrices, measured in squared Frobenius norm, when (a) $d = 3$, for different choices of α; and (b) $\alpha = \frac{2}{3}$, for different choices of d.

9.4.4 Confidence intervals

Finally, we validate our statistical inference framework on the MDP with $d = 3$ and TD learning with $\alpha = \frac{2}{3}$. We run 10,000 independent trials of Algorithm 1, and at each iteration T, we construct both the *individual* and the *simultaneous* confidence intervals. Specifically, the individual confidence intervals for the three entries of $\boldsymbol{\theta}^\star$ are constructed by

$$\mathcal{C}_j = \left[\overline{\boldsymbol{\theta}}_T - \Phi(1-\delta/2) \cdot \sqrt{\frac{(\widehat{\boldsymbol{\Lambda}}_T)_{jj}}{T}}, \overline{\boldsymbol{\theta}}_T + \Phi(1-\delta/2) \cdot \sqrt{\frac{(\widehat{\boldsymbol{\Lambda}}_T)_{jj}}{T}} \right], \quad \text{for} \quad j = 1, 2, 3.$$

Here, $\Phi(1-\delta/2)$ is the $1-\frac{\delta}{2}$ quantile of the standard Gaussian distribution. In our experiment, we take $\delta = 0.05$ and therefore $\Phi(1-\delta/2) = 1.96$. Figure 6 illustrates the emergence of $\overline{\boldsymbol{\theta}}_T$ and $\{\mathcal{C}_j\}_{j=1,2,3}$ in *one single trial*. As T grows larger, the three entries of the point estimate $\overline{\boldsymbol{\theta}}_T$ (blue

curves) converge to the corresponding entries of the optimal coefficients $\boldsymbol{\theta}^\star$ (black dotted lines), while the individual confidence intervals $\{\mathcal{C}_j\}_{j=1,2,3}$ (blue shades) also concentrate around $\boldsymbol{\theta}^\star$.

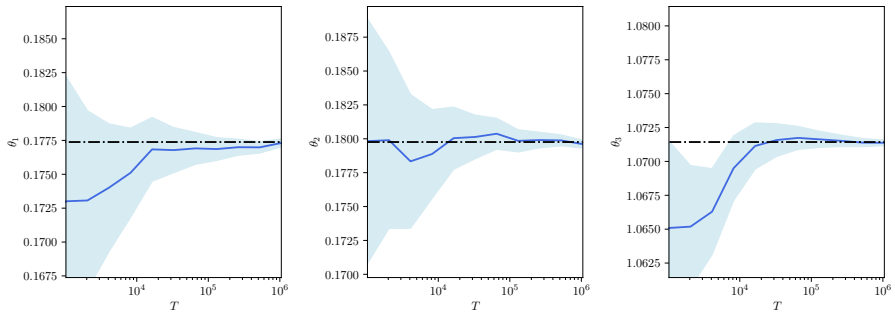

Figure 6: The emergence of the point estimate $\overline{\boldsymbol{\theta}}_T$ (blue curves) and individual confidence intervals (blue shades) on the three entries of the optimal coefficients $\boldsymbol{\theta}^\star$ (black dotted lines) for the MDP with $d = 3$, using TD with Polyak-Ruppert averaging, $\alpha = \frac{2}{3}$

Meanwhile, the simultaneous confidence intervals are constructed as

$$\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_T\|_\infty \le x^\star\}, \quad \text{where}$$
$$x^\star := \sup\left\{x : \mathbb{P}_{z\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_3)}\left(\left\|\widehat{\boldsymbol{\Lambda}}_T^{\frac{1}{2}}\boldsymbol{z}\right\|_\infty \le x\right) \le 1-\delta\right\}.$$

Essentially, $x^\star$ is the $1-\delta$ quantile of the ℓ_∞ norm of the $\mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Lambda}}_T)$ random variable. Calculating this quantity is a non-trivial task, so we use 1000 independent simulations from $\mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Lambda}}_T)$ and take the $1-\delta$ quantile of the ℓ_∞ norm of the samples as an estimate.

Figure 7 shows the coverage rates of the individual and simultaneous confidence intervals, defined as the percentage of times when the optimal coefficient $\boldsymbol{\theta}^\star$ (or its entries) are covered by the confidence intervals in the $10,000$ independent trials. As is indicated by the plot, the coverage rates all approach the target confidence level of $1-\delta = 95\%$ when T grows large.

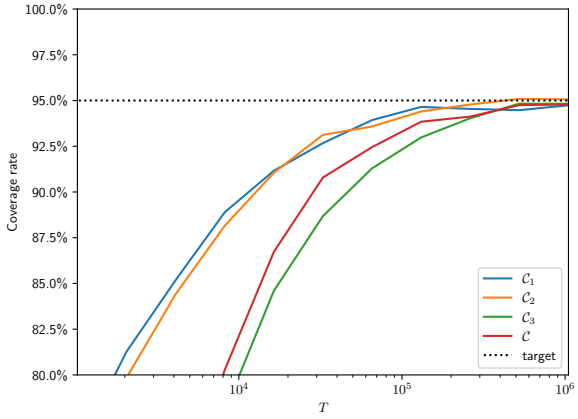

Figure 7: The emergence coverage rates of independent and simultaneous confidence intervals on the three entries of the optimal coefficients $\boldsymbol{\theta}^\star$ of the MDP with $d = 3$, using TD with Polyak-Ruppert averaging, $\alpha = \frac{2}{3}$. Black dotted line represents the target confidence level of 95%; blue, orange and green curves represent the coverage rates of the individual confidence intervals $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ correspondingly; the red curve represents the coverage rate of the confidence region (simultaneous confidence intervals) $\mathcal{C}$.

9.5 Decomposition of TD estimation error

In this section, we aim to further explore the emergence of the TD estimation error $\overline{\boldsymbol{\Delta}}_T$ in numerical simulations, in order to verify our key claims in the proofs of our theoretical results. Specifically, we aim to explore the role played by the choice of α through decomposing the TD estimation error. Recall that throughout our proofs, we repetitively used the decomposition

$$\overline{\boldsymbol{\Delta}}_T = \underbrace{\frac{1}{T\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0}_{I_1} - \underbrace{\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i)}_{I_2} - \underbrace{\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i - \boldsymbol{A})\boldsymbol{\Delta}_{i-1}}_{I_3},$$

where the matrices $\{\boldsymbol{Q}_i\}_{0\le i\le T}$ are defined by (53), and therefore decided by the matrix $\boldsymbol{A}$ as well as the choice of stepsizes $\{\eta_t\}_{1\le t\le T}$. Of the three terms on the right-hand-side:

1. The term I_1 is related only to the initial estimator $\boldsymbol{\theta}_0$ through $\boldsymbol{\Delta}_0$ and not to the samples $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}_{1\le t\le T}$. Therefore, we refer to this term as the *initialization error*. This term is deterministic, and according to our analysis, converges by the rate of

$$\left\|\frac{1}{T\eta_0}\boldsymbol{Q}_0\boldsymbol{\Delta}_0\right\|_2 \lesssim O(T^{-1}).$$

2. The term I_2 is related only to the samples $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}_{1\le t\le T}$ and not to any of the iterations $\{\boldsymbol{\theta}_t\}_{0\le t\le T}$. Therefore, we refer to this term as the *sampling error*. According to our analysis, this term corresponds to the asymptotic error:

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^\star)$$

   when $\alpha \in (\frac{1}{2}, 1)$; non-asymptotically speaking, the variance of this term is $\overline{\boldsymbol{\Lambda}}_T/T$ by definition, and $\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star = \boldsymbol{O}(T^{\alpha-1})$;

3. The term I_3 is related to both the samples $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}_{1\le t\le T}$ and the intermediate iterations $\{\boldsymbol{\theta}_t\}_{0\le t\le T}$. Hence, we refer to this term as the *iterative error*. According to our analysis, this term converges by the rate of

$$\left\|\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{Q}_i(\boldsymbol{A}_i - \boldsymbol{A})\boldsymbol{\Delta}_{i-1}\right\|_2 \lesssim O(T^{-\frac{\alpha}{2}}).$$

Comparing the magnitudes of the components. Figure 8 illustrates the convergence of these terms in our $10,000$ independent trials when $d = 3$ and $\alpha = \frac{2}{3}$. In the plot, the blue, orange and green lines represent the sample averages of $\|I_1\|_2^2$, $\|I_2\|_2^2$ and $\|I_3\|_2^2$ respectively; the red line represents the evolvement of $\|\overline{\boldsymbol{\Delta}}_T\|_2^2$, and the black dotted line represents the asymptotic variance $\mathsf{Tr}(\boldsymbol{\Lambda}^\star)/T$. As shown in the graph, the initialization error I_1 decreases the fastest; when $T < 10^3$, the iterative error I_3 is the dominant component in the estimation error, as the orange and red lines almost overlap. However, the slope of the orange line is steeper than that of the green line representing the sampling error I_2, an illustration that the iterative error converges faster as a function of T. In the mean time, the green line converges quickly to the black dotted line, verifying our claim that the sampling error corresponds to the asymptotic distribution. Finally, as T becomes sufficiently large ($> 10^5$ in this example), the green, red, and black dotted lines overlap with each other, an illustration that the sampling error dominates the estimation error, which converges to its asymptotic distribution.

Difference between $\overline{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$. Theorem 6.11 characterizes the difference between the "non-asymptotic variance" $\overline{\boldsymbol{\Lambda}}_T$ and the asymptotic variance $\boldsymbol{\Lambda}^\star$ as

$$\overline{\boldsymbol{\Lambda}}_T - \boldsymbol{\Lambda}^\star \asymp \boldsymbol{O}(T^{\alpha-1}),$$

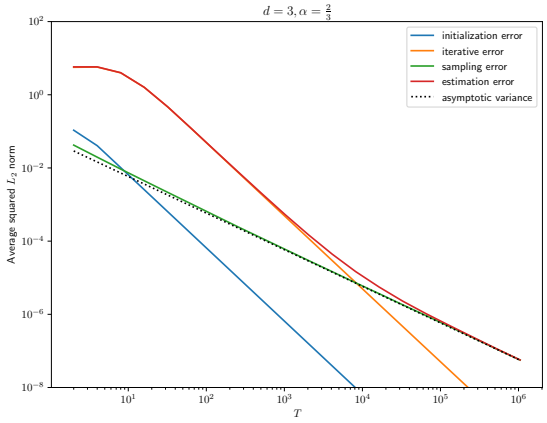

Figure 8: Decomposition of the TD estimation error through $10,000$ independent trials, using MDP with $d = 3$ and stepsizes $\eta_t = \eta_0 t^{-\frac{2}{3}}$

which results in our conclusion that the difference between the non-asymptotic and asymptotic distributions of $\overline{\boldsymbol{\Delta}}_T$ is lower bounded by $O(T^{\alpha-1})$. In order to illustrate this difference in variance through numerical simulations, we use $\check{\boldsymbol{\Lambda}}_T$ to denote the sample mean (averaged over our $10,000$ independent trials) of

$$\frac{1}{T}\left(\sum_{i=1}^{T} \boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i)\right)\left(\sum_{i=1}^{T} \boldsymbol{Q}_i(\boldsymbol{A}_i\boldsymbol{\theta}^\star - \boldsymbol{b}_i)\right)^\top.$$

By definition, $\check{\boldsymbol{\Lambda}}_T$ is an unbiased estimator of $\overline{\boldsymbol{\Lambda}}_T$. Figure 9 demonstrates the difference between $\check{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$, measured by the Frobenius norm of their difference, as T increases. It is clear from the plot that the difference converges slower when α is larger, and especially that it does not converge to 0 when $\alpha = 1$. Notice that the plot is truncated before $T = 10^4$ as when T becomes larger than the number of trials, the randomness within $\check{\boldsymbol{\Lambda}}_T$ dominates the difference, making the plot uninformative.

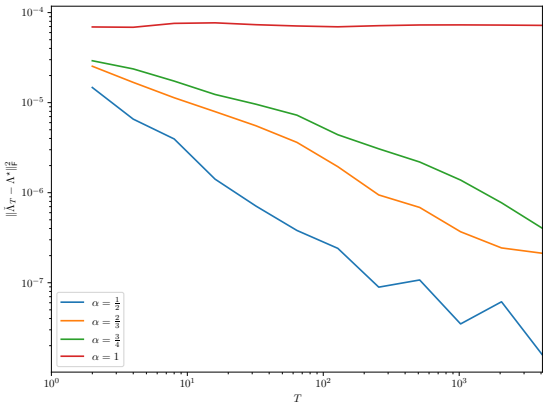

Figure 9: The difference between $\check{\boldsymbol{\Lambda}}_T$ and $\boldsymbol{\Lambda}^\star$, for $d = 3$ and different choices of α

10 Extension to general linear stochastic approximation

In this section, we explore on the conditions under which our theoretical analysis and statistical inference procedure can be extended to general Linear Stochastic Approximation (LSA) for the linear system

$$\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{b},$$

where a sample matrix $\boldsymbol{A}_t$ and a sample vector $\boldsymbol{b}_t$ are observed $i.i.d.$ at each time t, with expectations being $\boldsymbol{A}$ and $\boldsymbol{b}$ respectively. Similar to TD learning, we also run

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t(\boldsymbol{A}_t\boldsymbol{\theta}_{t-1} - \boldsymbol{b}_t),$$

with the same choice of stepsizes $\eta_t = \eta_0 t^{-\alpha}$, $\alpha \in (\frac{1}{2}, 1)$.

We first make the following assumptions on the matrix $\boldsymbol{A}$:

Assumption 10.1 (Stability)**.** There exists a $d \times d$ symmetric matrix $\boldsymbol{U} \succeq \boldsymbol{0}$ and a constant $\beta_0 > 0$, such that $\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top \succeq \beta_0 \boldsymbol{I}$;

Notice that in our setting of TD learning, this assumption is satisfied with $\boldsymbol{U} = \boldsymbol{I}$, and $\beta_0 = 2(1-\gamma)\lambda_0$; for general LSA, it is guaranteed as long as $-\boldsymbol{A}$ is *Hurwitz*, i.e., all the eigenvalues of $-\boldsymbol{A}$ has strictly negative real-parts. Under that condition, β_0 can be set to 1, and $\boldsymbol{U}$ is the unique solution to the Lyapunov equation $\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top = \boldsymbol{I}$. Meanwhile, we need the following constraints on the samples $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}$:

Assumption 10.2 (Sampling constraints)**.** With $\boldsymbol{U}$ specified as in Assumption 10.1,

1. There exists a constant $c_U > 0$, such that $\mathbb{E}[\boldsymbol{A}_t\boldsymbol{U}\boldsymbol{A}_t^\top] \preceq c_U(\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top)$;
2. There exists constants $c_A > 0, c_b > 0$, such that $\|\boldsymbol{A}_t\|_{\boldsymbol{U}} \le c_A$ and $\|\boldsymbol{b}_t\|_{\boldsymbol{U}} \le c_b$ almost surely.

Under our setting of TD learning, Assumption 10.2 is satisfied with $c_U = 1, c_A = 2$ and $c_b = 1$ by Lemma 6.4. For general LSA with $-\boldsymbol{A}$ being Hurwitz, as long as (2) is satisfied, (1) can also be guaranteed by setting $c_U = c_A\|\boldsymbol{U}\|$. As long as both Assumption 10.1 and Assumption 10.2 are satisfied, it is immediate that $\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{U}} \le c_A\|\boldsymbol{\theta}^\star\|_{\boldsymbol{U}} + c_b$ almost surely, and that for every $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\|(\boldsymbol{I} - \eta\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{U}}^2 &\le \mathbb{E}\|(\boldsymbol{I} - \eta\boldsymbol{A}_t)\boldsymbol{x}\|_{\boldsymbol{U}}^2 \\
&= \|\boldsymbol{x}\|_{\boldsymbol{U}}^2 - \eta\boldsymbol{x}^\top(\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top)\boldsymbol{x} + \eta^2\boldsymbol{x}^\top\mathbb{E}[\boldsymbol{A}_t\boldsymbol{U}\boldsymbol{A}_t^\top]\boldsymbol{x} \\
&\le \|\boldsymbol{x}\|_{\boldsymbol{U}}^2 - (\eta - c_U\eta^2)\boldsymbol{x}^\top(\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top)\boldsymbol{x};
\end{aligned}$$

as long as

$$\eta \le \frac{1}{2c_U} =: \eta_{\max},$$

this bound can be further elaborated as

$$\begin{aligned}
\|(\boldsymbol{I} - \eta\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{U}}^2 &\le \mathbb{E}\|(\boldsymbol{I} - \eta\boldsymbol{A}_t)\boldsymbol{x}\|_{\boldsymbol{U}}^2 \\
&\le \|\boldsymbol{x}\|_{\boldsymbol{U}}^2 - \frac{\eta}{2}\boldsymbol{x}^\top(\boldsymbol{A}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{A}^\top)\boldsymbol{x} \\
&\le \|\boldsymbol{x}\|_{\boldsymbol{U}}^2 - \frac{\beta_0\eta}{2}\|\boldsymbol{x}\|_2^2 \le \left(1 - \frac{\beta_0\eta}{2\|\boldsymbol{U}\|}\right)\|\boldsymbol{x}\|_{\boldsymbol{U}}^2;
\end{aligned}$$

Consequently, as long as $\eta \le \eta_{\max}$,

$$\|\boldsymbol{I} - \eta\boldsymbol{A}\|_{\boldsymbol{U}} \le \sqrt{1 - \frac{\beta_0\eta}{2\|\boldsymbol{U}\|}} \le 1 - \frac{\beta_0\eta}{4\|\boldsymbol{U}\|}.$$

This is a direct generalization of Lemma 6.4.

While the detailed derivations are well beyond the scope of this paper, our theoretical analysis in Theorems 3.2, 3.1, 3.3 can all be extended to general LSA with this generalized version of Lemma 6.4. Specifically:

1. For every $\delta \in (0,1)$, there exists a proper choice of η_0, such that with probability at least $1-\delta$,
$$\|\overline{\boldsymbol{\Delta}}_T\|_{\boldsymbol{U}} \lesssim \sqrt{\frac{\mathsf{Tr}(\boldsymbol{U}\boldsymbol{\Lambda}^\star)}{T}\log\frac{1}{\delta}};$$
2. As long as $\eta_0 \le \eta_{\max}$, it can be guaranteed that
$$d_{\mathsf{C}}(\sqrt{T}\overline{\boldsymbol{\Delta}}_T, \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim C_1 T^{-\frac{\alpha}{2}} + C_2 T^{\alpha-1},$$
with coefficients C_1 and C_2 related to $\boldsymbol{U}, \beta_0, c_U, c_A, c_b$. Therefore, choosing $\alpha = \frac{2}{3}$ would still yield an $O(T^{-1/3})$ convergence rate;
3. With variance estimator $\widehat{\boldsymbol{\Lambda}}_T$ defined in the same way as (18),
$$d_{\mathsf{TV}}(\mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Lambda}}_T), \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^\star)) \lesssim O(T^{-\frac{1}{3}})$$
with probability at least $1 - O(T^{-\frac{1}{3}})$;
4. Under the same statistical inference procedure as described in Section 3.4, the confidence region $\mathcal{C}_\delta$ still satisfies
$$\mathbb{P}(\boldsymbol{\theta}^\star \in \mathcal{C}_\delta) \ge 1 - \delta - O(T^{-\frac{1}{3}}).$$