Sharp high-probability sample complexities for policy evaluation with linear function approximation

Gen Li*† UPenn Statistics
Weichen Wu*‡ CMU Statistics
Yuejie Chi§ CMU ECE
Cong Ma¶ UChicago Statistics
Alessandro Rinaldo‡ CMU Statistics
Yuting Wei† UPenn Statistics

May 31, 2023

Abstract

This paper is concerned with the problem of policy evaluation with linear function approximation in discounted infinite horizon Markov decision processes. We investigate the sample complexities required to guarantee a predefined estimation error of the best linear coefficients for two widely-used policy evaluation algorithms: the temporal difference (TD) learning algorithm and the two-timescale linear TD with gradient correction (TDC) algorithm. In both the on-policy setting, where observations are generated from the target policy, and the off-policy setting, where samples are drawn from a behavior policy potentially different from the target policy, we establish the first sample complexity bound with high-probability convergence guarantee that attains the optimal dependence on the tolerance level. We also exhibit an explicit dependence on problem-related quantities, and show in the on-policy setting that our upper bound matches the minimax lower bound on crucial problem parameters, including the choice of the feature map and the problem dimension.

Keywords: policy evaluation, temporal difference learning, two-timescale stochastic approximation, minimax optimal, function approximation

Contents

*The first two authors contributed equally.
†Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
‡Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
§Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
¶Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

1 Introduction

Policy evaluation plays a critical role in many scientific and engineering applications in which practitioners aim to evaluate the performance of a target strategy based on either sequentially collected or a batch of offline data samples (Bojinov and Shephard, 2019; Dann et al., 2014; Murphy, 2003; Tang and Wiens, 2021). For example, in clinical trials (Tang and Wiens, 2021), real-time data acquisition might be expensive and risky; it is thus of essential value if historical data can be analyzed and information can be transferred to new tasks. While in other applications, such as mobile health (Bertsimas et al., 2022), it is practical to implement the desired policy and collect its feedback in a timely manner.

Mathematically, Markov decision processes (MDPs) provide a general framework to design policy evaluation methods in dynamic settings; reinforcement learning (RL) is often modeled using MDPs when the exact model configuration is not available (Bertsekas, 2017; Sutton and Barto, 2018). In this framework, a target policy is assessed through its corresponding value function. In practice, evaluating value functions often require an overwhelming number of samples due to the large dimensionality of the underlying state space. For this reason, RL methods are normally concerned with some form of function approximation. Dating back to the seminal work of Tsitsiklis and Van Roy (1997), there has been an extensive line of works that consider different types of function approximation, including linear function approximation (Bhandari et al., 2021; Fan et al., 2020), reproducing kernel Hilbert space (Duan et al., 2021; Farahmand et al., 2016), deep neural networks (Arulkumaran et al., 2017; Bertsekas and Tsitsiklis, 1995) or function approximation on the model itself (see, e.g. Jin et al. (2020); Li et al. (2021a); Wang et al. (2021a)), with a focus on improving the sample efficiency of RL algorithms.

Two settings: on-policy vs. off-policy. The main goal of this paper is to provide sharp statistical guarantees of policy evaluation algorithms with linear function approximation in two different settings. As the aforementioned examples already indicated, there are typically two different types of data-generating mechanisms to consider: the *on-policy* setting when we have access to the outcomes of the target policy and the *off-policy* setting, in which the only available data are generated from a behavior policy that is potentially different from the target policy.

In the on-policy setting, temporal difference (TD) learning is arguably the most popular algorithm (Sutton, 1988) for policy evaluation in RL practice, partly because it is easy to implement and lends itself well to function approximations. As a model-free algorithm, TD learning processes data in an online manner without explicitly modeling the environment and is, therefore, memory efficient. While the asymptotic convergence of TD with linear function approximation has been known since Tsitsiklis and Van Roy (1997), the finite-sample minimax optimality of TD has been established only recently for the tabular MDP (Li et al., 2023a). For TD learning with linear function approximation, several recent contributions have produced new non-asymptotic analyses and insights (e.g. Bhandari et al. (2021); Dalal et al. (2018a); Lakshminarayanan and Szepesvari (2018); Srikant and Ying (2019)), which partially unveil impacts of both the tolerance level and various problem-related parameters on its sample efficiency. However, minimax-optimal dependence on the tolerance level (i.e. target level of estimation accuracy) is only established in expectation instead of with high probability; furthermore, the optimal dependence on problem-related parameters, such as the size of the state space and the effective horizon, still remains unsettled, and it is unclear whether existing sample complexity bounds can be further improved. Failing to understand these questions, however, casts doubt on whether TD with linear function approximation is statistically efficient in practice, and brings difficulties to performing statistical inference based on TD estimators. In this paper, we seek to answer these questions by providing tighter characterizations of the performance of TD with linear function approximation.

In the off-policy setting, it is known that the error of TD learning with linear function approximation may diverge to infinity (Baird, 1995). In order to address this issue, Sutton et al. (2009) proposed a now popular alternative with two-timescale learning rates, called the linear TD with gradient correction (TDC) algorithm, which enjoys convergence guarantees in the off-policy case. In terms of finite-sample guarantees, although a number of recent efforts (see, e.g. Dalal et al. (2020, 2018b); Gupta et al. (2019); Kaledin et al. (2020); Wang et al. (2021b); Xu and Liang (2021)) tried to characterize the statistical performance of TDC for both *i.i.d.* and Markovian data, they remain inadequate in providing either a convergence guarantee with high-probability, an explicit dependence on salient problem parameters, or a sharp dependence on the sample size. The challenge lies in dealing with the statistical dependence between two separate iterate sequences at different timescales. To tackle this challenge, it calls for a new analysis framework for the TDC algorithm.

1.1 Our main contributions

This paper is concerned with evaluating the performance of a given target policy π in an infinite-horizon γ-discounted MDP with a finite but large number of states. The goal is to learn the best linear approximation of the value function in a pre-specified feature space given *i.i.d.* transition pairs drawn from the stationary distribution. In the on-policy setting, we focus on the TD learning algorithm; in the off-policy setting, we shift gear to the TDC learning algorithm. We summarize our main contributions as follows, with their exact statements and consequences postponed to later sections.

- Via a careful analysis of TD learning with Polyak-Ruppert averaging, we show that, in the on-policy setting, a number of samples of order

$$\widetilde{O}\left(\frac{\max_s\{\boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}(1+\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2)}{(1-\gamma)^2\varepsilon^2}\right)$$

  is sufficient to achieve an accuracy level (estimation error) of $\varepsilon > 0$, with high probability. Here, $\boldsymbol{\phi}(s) \in \mathbb{R}^d$ indicates the linear feature vector for the state s in the state space $\mathcal{S}$, $\boldsymbol{\theta}^\star$ is the best linear approximation coefficient of the value function, and $\boldsymbol{\Sigma}$ corresponds to the feature covariance matrix weighted by the stationary distribution. See Section 2 for the definitions of these parameters. Compared to prior work by Bhandari et al. (2021) and Srikant and Ying (2019), our sample complexity bound can be tighter by a factor of $\mathsf{cond}(\boldsymbol{\Sigma})$ which can be as large as $|\mathcal{S}|$ (the cardinality of the state

paper	algorithm	stepsize	sample complexity	error control
Bhandari et al. (2021)	TD	$\eta_t \asymp t^{-1}$	$O\left(\frac{\|\mathbf{\Sigma}^{-2}\|\|\mathbf{\Sigma}\|\|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2}{(1-\gamma)^2\varepsilon^2}\right)$	in expectation
Srikant and Ying (2019)	TD	$\eta_t \asymp T^{-1}$	$O\left(\frac{\|\mathbf{\Sigma}^{-2}\|\|\mathbf{\Sigma}\|\|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2}{(1-\gamma)^2\varepsilon^2}\right)$	in expectation
Dalal et al. (2018a)	TD	$\eta_t = t^{-1}$	$O\left(\frac{1}{\varepsilon^{\max\{2,1+\frac{1}{\lambda}\}}}\right)$	w. high-prob
This work	Averaged TD	$\eta_t = \eta$	$O\left(\frac{\|\mathbf{\Sigma}^{-1}\|\|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2}{(1-\gamma)^2\varepsilon^2}\right)$	w. high-prob

Table 1. Comparisons with prior results (up to logarithmic terms) in finding an ε-optimal solution using TD learning. Using the Polyak-Ruppert averaging, our result improves upon previous sample complexity bounds by a multiplicative factor of the condition number of $\mathbf{\Sigma}$, and is the first sample complexity with high-probability guarantee of convergence to match the minimax-optimal dependence on the tolerance level ε.

space). Our result is also the first to control ε-convergence with high probability that matches the minimax-optimal dependence on the tolerance level ε. To assess the tightness of this upper bound, we provide a minimax lower bound in Section 3.3, which certifies the optimal dependence of our bound on both the tolerance level ε and problem-related parameters $\mathbf{\Sigma}$ and $\boldsymbol{\theta}^\star$.

- In the off-policy setting, we establish a sample complexity bound for the TDC algorithm of order

$$\widetilde{O}\left(\frac{\rho_{\max}^7}{\lambda_1^4\lambda_2^3}\frac{\|\widetilde{\mathbf{\Sigma}}\|^2}{\varepsilon^2}(1+\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\mathbf{\Sigma}}}^2)\right),$$

where $\widetilde{\boldsymbol{\theta}}$ corresponds to the best linear approximation coefficient of the value function in the off-policy setting, $\widetilde{\mathbf{\Sigma}}$ is the feature covariance matrix under the behavior policy, $\rho_{\max}$ denotes the largest importance sampling ratio measuring the discrepancy between the target policy and the behavior policy, and lastly, λ_1 and λ_2 denote the smallest eigenvalues of some problem-dependent matrices. Details about these constants are deferred to Section 4. To the best of our knowledge, our bound is the first one to control ε-convergence with high probability that matches the minimax-optimal dependence on the tolerance level ε. At the same time, our sample complexity bound also provides an explicit dependence on the salient parameters.

Comparisons of our results to existing bounds and relevant commentary can be found in Table 1 and 2.

1.2 Other related works

In this section, we review several recent lines of works and provide a broader context of the current paper.

Finite-sample guarantees for policy evaluation. Classical analyses of policy evaluation algorithms have mainly focused on providing asymptotic guarantees given a fixed model (Szepesvári, 1998; Tsitsiklis and Van Roy, 1997). New tools developed in high-dimensional statistics and probability allow for a fine-grained understanding of these algorithms especially from a finite-sample and finite-time perspective. As argued in this paper, understanding how statistical errors depend on the effective horizon, dimension of the problem and the number of samples, is essential as it provides important insights on how these RL algorithms perform in practice. A highly incomplete list of prior art includes Bhandari et al. (2021); Boyan (1999); Dalal et al. (2018a); Jin et al. (2018); Khamaru et al. (2020); Lakshminarayanan and Szepesvari (2018); Srikant and Ying (2019) with a focus on the non-asymptotic analyses for model-free algorithms, and Agarwal et al. (2020); Li et al. (2023b); Pananjady and Wainwright (2021); Sidford et al. (2018) which derive non-asymptotic bounds for model-based algorithms.

paper	algorithm	stepsize	sample complexity	error control
Dalal et al. (2020)	**Projected** TDC	$\alpha_t = t^{-\alpha}$, $\beta_t = t^{-\beta}$	$O\left(\frac{1}{\varepsilon^{2\alpha}}\right), \alpha < 1$	w. high-prob
Kaledin et al. (2020)	TDC	$\alpha_t, \beta_t \asymp \frac{1}{T}$	$O\left(\frac{1}{\varepsilon^2}\right)$	in expectation
Xu and Liang (2021)	**Batched** TDC	$\alpha_t = \alpha$, $\beta_t = \beta$	$O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$	in expectation
This work	TDC	$\alpha_t, \beta_t \asymp \frac{1}{T}$	$O\left(\frac{1}{\varepsilon^2}\right)$	w. high-prob

Table 2. Comparisons with prior results (up to logarithmic terms) in finding an ε-optimal solution using TDC learning. We omit dependence on problem-related parameters in this table. Our sample complexity bound for TDC is the first to achieve high-probability convergence guarantee with non-varying stepsizes and without using projection steps or batched updates; in the mean time, we also provide explicit dependence on problem-related parameters.

Stochastic approximation. The idea of stochastic approximation (SA) (Lai, 2003; Robbins and Monro, 1951) lies at the core of the TD and TDC learning algorithms considered in this paper. With the intention of solving a deterministic fixed-point equation, SA methods perform stochastic updates based on approximations of the current residual. The asymptotic theory of SA methods are relatively well-developed, where SA iterates provably track the trajectory of a limiting ordinary differential equation (Borkar, 2009; Borkar and Meyn, 2000) and with properly decaying step sizes, the Polyak-Ruppert averaged iterates asymptotically follow the central limit theorem. Recently, non-asymptotic results have also been obtained for SA for different problems especially in the RL setting; see Lakshminarayanan and Szepesvari (2018); Mou et al. (2020); Moulines and Bach (2011); Nemirovski et al. (2009) and references therein. The TDC algorithm is a special case of two-timescale linear SA, whose convergence rates have also been investigated in Dalal et al. (2020); Gupta et al. (2019); Wu et al. (2020); Xu et al. (2019), among others.

Off-policy learning. Policy evaluation in the off-policy setting is closely related to offline or batch RL, which aims to learn purely based on historical data without actively exploring the environment. The main challenge here lies in the discrepancy between the behavior policy and the target or optimal policy. One natural approach is to use importance sampling (IS) in order to form an unbiased estimator of the target policy (Precup, 2000), and various different techniques have been applied to reduce the high variance of IS (see, e.g. Jiang and Li (2016); Kallus and Uehara (2020); Ma et al. (2022); Thomas and Brunskill (2016); Xie et al. (2019); Yang et al. (2020)). Non-asymptotic guarantees are also provided for off-policy evaluation using a fitted Q-iteration approach under linear function approximation in Duan et al. (2020). A recent line of works also considered finding the optimal policy using batch datasets (Jin et al., 2021; Li et al., 2022; Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021).

1.3 Notation

Throughout this paper, we denote by $\Delta(\mathcal{S})$ (resp. $\Delta(\mathcal{A})$) the probability simplex over the finite set $\mathcal{S}$ (resp. $\mathcal{A}$). For any positive integer n, we use $[n]$ to denote the set of positive integers that are no larger than n: $[n] = \{1, 2, ..., n\}$. When a function is applied to a vector, it should be understood as being applied in a component-wise fashion; for example, $\sqrt{\boldsymbol{z}} := [\sqrt{z_i}]_{1\le i\le n}$ and $|\boldsymbol{z}| := [|z_i|]_{1\le i\le n}$. For any vectors $\boldsymbol{z} = [a_i]_{1\le i\le n}$ and $\boldsymbol{w} = [w_i]_{1\le i\le n}$, the notation $\boldsymbol{z} \ge \boldsymbol{w}$ (resp. $\boldsymbol{z} \le \boldsymbol{w}$) stands for $z_i \ge w_i$ (resp. $z_i \le w_i$) for all $1 \le i \le n$. Additionally, we write $\mathbf{1}$ for the all-one vector, $\boldsymbol{I}$ for the identity matrix, and $\mathbb{1}\{\cdot\}$ for the indicator function.

For any matrix $\boldsymbol{P} = [P_{ij}]$, we denote $\|\boldsymbol{P}\|_1 := \max_i \sum_j |P_{ij}|$. Given a symmetric positive definite matrix $\boldsymbol{D}$, define the inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{D}}$ as $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{D}} = \boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{v}$ and the associated norm $\|\boldsymbol{v}\|_{\boldsymbol{D}} = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle_{\boldsymbol{D}}}$. For any matrix $\boldsymbol{M}$, we use $\|\boldsymbol{M}\|$ to denote its operator norm (i.e. the largest singular value), if not specified otherwise. Throughout this paper, we use $c, c_0, c_1, C, \cdots$ to denote universal constants that do not depend

either on the parameters of the MDP or the target levels (ε, δ); their exact values may change from line to line. Given two sequences, $\{f_t\}_{t\geq 0}$ and $\{g_t\}_{t\geq 0}$, we write $f_t \lesssim g_t$ (resp. $f_t \gtrsim g_t$) or $f_t = O(g_t)$ (resp. $g_t = O(f_t)$) if there exists some universal constant $c_1 > 0$, such that $f_t \leq c_1 g_t$ (resp. $f \geq c_1 g$). If both $f = O(g)$ and $g = O(f)$ hold simultaneously, we write $f_t \asymp g_t$ or $f_t = \Theta(g_t)$. We adopt the notation $f = \widetilde{O}(g)$ to indicate $f = O(g)$ up to logarithmic factors in g. For any symmetric matrix $\boldsymbol{X}$, we use $\lambda_{\min}(\boldsymbol{X})$ to denote its smallest eigenvalue.

2 Problem formulation

2.1 Model and settings

Markov decision process. Consider an infinite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ with discounted rewards, where $\mathcal{S}$ and $\mathcal{A}$ denote respectively the (finite) state space and action space, and $\gamma \in (0,1)$ indicates the discount factor (Bertsekas, 2017). The probability transition kernel of the MDP is given by $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, where for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\mathcal{P}(\cdot \mid s,a) \in \Delta(\mathcal{S})$ denotes the transition probability distribution from state s when action a is executed. The reward function is represented by the function $r : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$, where $r(s,a)$ denotes the immediate reward from state s when action a is taken; for simplicity, we assume throughout that all immediate rewards lie within $[0,1]$.

A policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ is an action selection rule that maps a state to a distribution over the set of actions; in particular, it is said to be stationary if it is time-invariant. The value function $V^\pi : \mathcal{S} \mapsto \mathbb{R}$ is used to measure the quality of a policy π, defined as

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right], \tag{1}$$

which is the expected discounted cumulative reward received by following the policy π under the MDP $\mathcal{M}$ when initialized at state $s_0 = s$. Here, $a_t \sim \pi(\cdot \mid s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$ for all $t \geq 0$. It can be easily verified that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$ for any π.

For a given policy π, we can define the reward function of every state $s \in \mathcal{S}$ as the expected reward for (s,a) when a is chosen according to π:

$$r(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]. \tag{2}$$

For simplicity, we introduce the vector notation for the reward function $\boldsymbol{r} := [r(s)]_{1\leq s\leq|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$, and the value function $\boldsymbol{V}^\pi = [V^\pi(s)]_{1\leq s\leq|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$. We can also define the transition matrix $\boldsymbol{P}^\pi$ for this given policy π, such that its (i,j) element represents the probability that state i is transited to state j under the policy π; formally,

$$P^\pi_{ij} = \sum_{a\in\mathcal{A}} \mathcal{P}(s_{t+1} = j \mid s_t = i, a_t = a)\pi(a_t = a \mid s_t = i). \tag{3}$$

We denote by μ the stationary distribution corresponding to the Markov chain when the transition follows $\boldsymbol{P}^\pi$, which we assume to be well-defined, and introduce the vector notation $\boldsymbol{\mu} := [\mu(s)]_{1\leq s\leq|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$.

Linear approximation for the value function. As discussed previously, it is often infeasible to collect a number of samples that scales with the ambient dimension $|\mathcal{S}|$. This motivates the search for lower dimension approximation of the value function, of which linear approximation emerges as a convenient option. Mathematically, for $\boldsymbol{\theta} \in \mathbb{R}^d$, define $V_{\boldsymbol{\theta}}(s)$ as

$$\forall s \in \mathcal{S}: \qquad V_{\boldsymbol{\theta}}(s) = \boldsymbol{\phi}(s)^\top \boldsymbol{\theta},$$

where $\boldsymbol{\phi}(s) \in \mathbb{R}^d$ is the feature vector associated with state $s \in \mathcal{S}$, with $d \leq |\mathcal{S}|$. The vector $\boldsymbol{\theta}$ of linear coefficients is shared across states.

Using matrix notation, we let

$$\boldsymbol{\Phi} := [\boldsymbol{\phi}(1), \boldsymbol{\phi}(2), \cdots, \boldsymbol{\phi}(|\mathcal{S}|)]^\top \in \mathbb{R}^{|\mathcal{S}|\times d}, \tag{4}$$

be the feature matrix that concatenates the feature vectors for all states and $\boldsymbol{V}_\theta = [V_{\boldsymbol{\theta}}(s)]_{s\in\mathcal{S}} \in \mathbb{R}^{|S|}$ be the linear approximation vector to the value function. It follows that

$$\boldsymbol{V}_\theta = \boldsymbol{\Phi}\boldsymbol{\theta}.$$

We impose the following mild assumption on the feature vectors.

Assumption 1. *The columns of $\boldsymbol{\Phi}$ are linearly independent with Euclidean norm uniformly bounded by one, i.e.* $\max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2 \leq 1$.

2.2 Policy evaluation with linear approximation

On-policy evaluation with linear approximation. The task of policy evaluation is to measure the value function $V^\pi(s)$ for every $s \in \mathcal{S}$ (see definition (1)) given a policy π of interest. In the **on-policy** setting, data samples are collected while the policy π is executed and a sequence of samples are obtained

$$\{(s_0, a_0, r_0), \ldots, (s_T, a_0, r_T)\}, \quad \text{where } a_t \sim \pi(\cdot \mid s_t),\ r_t = r(s_t, a_t).$$

In this setting, in order to find the best linear approximation to $\boldsymbol{V}^\pi$, we find it helpful to first introduce some shorthand notation. First, given the stationary distribution μ for $\boldsymbol{P}^\pi$, we let

$$\boldsymbol{D}_\mu = \mathsf{diag}\big(\mu(1), \mu(2), \cdots, \mu(|\mathcal{S}|)\big) \tag{5}$$

and denote with

$$\boldsymbol{\Sigma} := \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu \boldsymbol{\Phi} = \mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\right] \in \mathbb{R}^{d\times d} \tag{6}$$

the feature covariance matrix with respect to this stationary distribution.

The best linear approximation coefficients, $\boldsymbol{\theta}^\star$, is defined as the unique solution to the following projected Bellman equation (Tsitsiklis and Van Roy, 1997)

$$\boldsymbol{\Phi}\boldsymbol{\theta} = \Pi_{\boldsymbol{D}_\mu} \mathcal{T}^\pi \left(\boldsymbol{\Phi}\boldsymbol{\theta}\right). \tag{7}$$

Here, $\Pi_{\boldsymbol{D}_\mu}$ denotes the projection operator onto the column space of $\boldsymbol{\Phi}$ (namely, the subspace $\{\boldsymbol{\Phi}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^d\}$) w.r.t. the inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{D}_\mu}$, where for any vector $\boldsymbol{v} \in \mathbb{R}^{|S|}$ one has

$$\Pi_{\boldsymbol{D}_\mu}(\boldsymbol{v}) := \mathop{\arg\min}_{\boldsymbol{z}\,\in\,\{\boldsymbol{\Phi}\boldsymbol{x}\,\mid\,\boldsymbol{x}\in\mathbb{R}^d\}} \|\boldsymbol{z} - \boldsymbol{v}\|^2_{\boldsymbol{D}_\mu}.$$

The function $\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ is known as the *Bellman operator,* which is given by

$$\boldsymbol{v} \mapsto \mathcal{T}^\pi(\boldsymbol{v}) := \boldsymbol{r} + \gamma \boldsymbol{P}^\pi \boldsymbol{v}. \tag{8}$$

Off-policy evaluation with linear approximation. In contrast, in the **off-policy** setting, we observe a trajectory from a behavior policy π_{b} instead of the target policy π. The goal is then to learn the value function for the target policy π based on

$$\{(s_0, a_0, r_0), \ldots, (s_T, a_0, r_T)\}, \quad \text{where } a_t \sim \pi_{\mathsf{b}}(\cdot \mid s_t),\ r_t = r(s_t, a_t).$$

Let μ_{b} be the stationary distribution over $\mathcal{S}$ induced by the behavior π_{b}, and correspondingly let

$$\boldsymbol{D}_{\mu_b} := \mathsf{diag}\big(\mu_{\mathsf{b}}(1), \mu_{\mathsf{b}}(2), \cdots, \mu_{\mathsf{b}}(|\mathcal{S}|)\big).$$

We denote with $\Pi_{\boldsymbol{D}_{\mu_b}}$ the projection operator associated with $\boldsymbol{D}_{\mu_b}$, which is given explicitly as

$$\Pi_{\boldsymbol{D}_{\mu_b}} \boldsymbol{v} := \mathop{\arg\min}_{\boldsymbol{z}\,\in\,\{\boldsymbol{\Phi}\boldsymbol{x}\,\mid\,\boldsymbol{x}\in\mathbb{R}^d\}} \|\boldsymbol{z} - \boldsymbol{v}\|^2_{\boldsymbol{D}_{\mu_{\mathsf{b}}}}.$$

In the off-policy setting, instead of trying to solve the projected Bellman's equation (7), we aim at minimizing the Mean-Squared Projected Bellman Error (MSPBE):

$$\text{minimize}_{\boldsymbol{\theta}} \quad \mathsf{MSPBE}(\boldsymbol{\theta}) := \frac{1}{2}\|\boldsymbol{V}_{\boldsymbol{\theta}} - \Pi_{\boldsymbol{D}_{\mu_b}} \mathcal{T}^{\pi} \boldsymbol{V}_{\boldsymbol{\theta}}\|_{\boldsymbol{D}_{\mu_b}}^2. \tag{9}$$

Throughout, we shall denote the minimizer of the above problem (9) as $\widetilde{\boldsymbol{\theta}}^{\star}$. We remark here that the norm and the projection are both induced by $\boldsymbol{D}_{\mu_b}$, while the Bellman operator is again in terms of the target policy π. For this reason, solving (9) is different from solving the projected Bellman's equation (7); as a result, in general, $\boldsymbol{\theta}^{\star} \neq \widetilde{\boldsymbol{\theta}}^{\star}$.

3 On-policy evaluation with TD learning

In this section, we study the accuracy of the estimator of $\boldsymbol{\theta}^{\star}$ (cf. (7)) returned by the TD learning algorithm in the on-policy setting. Specifically, we seek to determine the tightest sample complexity for this algorithm that ensures an ε-close solution. To better highlight our analysis strategy, we only consider the stylized generative model[1] whereby, at each time stamp t, one acquires an independent sample pair

$$(s_t, s_t') \quad \text{where } s_t \overset{\text{i.i.d.}}{\sim} \mu,\ a_t \sim \pi(s_t), \ \text{ and } s_t' \sim \mathcal{P}(\cdot \mid s_t, a_t). \tag{10}$$

Here recall that μ is the stationary distribution corresponding to $\boldsymbol{P}^{\pi}$. Notice that in the on-policy setting, since we are focused on a fixed policy π and interested only in the state pairs $\{(s_t, s_t')\}_{t=0}^{T}$ and not the actions $\{a_t\}_{t=0}^{T}$, the Markov decision process reduces to a Markov reward process (MRP). Given a sequence of sample pairs $\{(s_t, s_t')\}_{t=0}^{T}$ and a given level of tolerance $\varepsilon > 0$, our goal is to derive a sharp lower bound on the number of samples T that is required for TD learning to produce an estimator $\widehat{\boldsymbol{\theta}}$ such that, with high probability,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\star}\|_{\boldsymbol{\Sigma}} \leq \varepsilon.$$

3.1 The TD learning algorithm

To motivate TD learning, it is helpful to first consider the properties of the best linear approximation coefficients $\boldsymbol{\theta}^{\star}$; see (7). For any sample transition (s_t, s_t') (see (10)), define the random quantities

$$\boldsymbol{A}_t := \boldsymbol{\phi}(s_t)\left(\boldsymbol{\phi}(s_t) - \gamma \boldsymbol{\phi}(s_t')\right)^{\top} \in \mathbb{R}^{d \times d}, \tag{11a}$$

$$\boldsymbol{b}_t := \boldsymbol{\phi}(s_t) r(s_t) \in \mathbb{R}^{d}, \tag{11b}$$

whose means are given respectively by

$$\boldsymbol{A} := \mathop{\mathbb{E}}_{s \sim \mu, s' \sim P^{\pi}(\cdot|s)} \left[\boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma \boldsymbol{\phi}(s')\right)^{\top}\right] = \boldsymbol{\Phi}^{\top} \boldsymbol{D}_{\mu} (\mathbf{I} - \gamma \boldsymbol{P}^{\pi}) \boldsymbol{\Phi} \in \mathbb{R}^{d \times d}, \tag{12a}$$

$$\boldsymbol{b} := \mathop{\mathbb{E}}_{s \sim \mu} [\boldsymbol{\phi}(s) r(s)] = \boldsymbol{\Phi}^{\top} \boldsymbol{D}_{\mu} \boldsymbol{r} \in \mathbb{R}^{d}. \tag{12b}$$

It turns out that the target vector $\boldsymbol{\theta}^{\star}$ satisfies the equation (Tsitsiklis and Van Roy, 1997)

$$\boldsymbol{\theta}^{\star} := \boldsymbol{A}^{-1} \boldsymbol{b}. \tag{13}$$

The TD learning algorithm leverages this representation by iteratively improving the linear approximation of the value function at each time stamp through the updates

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t (\boldsymbol{A}_t \boldsymbol{\theta}_t - \boldsymbol{b}_t), \quad t = 0, 1, 2, \ldots, \tag{14a}$$

[1] We believe that our framework can be potentially generalized to Markovian samples using similar techniques in Li et al. (2021b) which is beyond the scope of the current paper.

where, for each t, $\eta_t > 0$ denotes the learning rate or stepsize. After T iterations, the TD learning algorithm returns $\boldsymbol{\theta}_T$ as the estimator. In contrast, TD learning with Polyak-Ruppert averaging, or *averaged TD learning* in short, returns an average across all iterates

$$\overline{\boldsymbol{\theta}}_T = \frac{1}{T}\sum_{i=1}^{T}\boldsymbol{\theta}_i. \tag{14b}$$

While we are mainly concerned with the averaged estimator $\overline{\boldsymbol{\theta}}_T$, we also obtain some theoretical properties of $\boldsymbol{\theta}_T$ as a by-product of our analysis.

3.2 Sample complexity of TD learning

In this section, we present a finite-sample bound for the estimation error of $\overline{\boldsymbol{\theta}}_T$ assuming independent data, from which we derive a novel sample complexity guarantee for TD learning. Below, we denote by κ the condition number of $\boldsymbol{\Sigma}$ as follows

$$\kappa := \lambda_{\mathsf{max}}(\boldsymbol{\Sigma})/\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}) \geq 1. \tag{15}$$

Theorem 1. *There exist universal, positive constants $C_0, c_0 > 0$ and $c_1 > 0$, such that for any given $0 < \delta < 1$, the averaged TD learning estimator $\overline{\boldsymbol{\theta}}_T$ (14) after T iterations satisfies the bound*

$$\begin{aligned}\left\|\overline{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}} \leq C_0\Bigg\{&\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log(\frac{d}{\delta})}{T(1-\gamma)^2}}\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\right)\\ &+\frac{\|\boldsymbol{\Sigma}^{-1}\|\left[\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\right)\sqrt{\frac{\kappa\log(\frac{dT}{\delta})}{\eta(1-\gamma)^3}}+\frac{1}{\eta(1-\gamma)}\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}\right]}{T}\Bigg\}\end{aligned} \tag{16}$$

with probability at least $1-\delta$, provided that $\boldsymbol{\theta}_0 = \mathbf{0}$, $\eta_0 = \ldots = \eta_T = \eta < \frac{c_0(1-\gamma)}{\kappa\log(Td/\delta)}$ and

$$T \geq \frac{c_1\kappa(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1)^2\log^2\frac{\kappa dT(\|\boldsymbol{\theta}^\star\|_2+1)}{(1-\gamma)\delta}}{\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})}.$$

The proof of the theorem and the other results from this section can be found in Section 6. Theorem 1 directly implies the following corollary, which gives an upper bound for the sample complexity of TD learning with independent samples.

Corollary 1 (Sample complexity of TD learning)**.** *There exists a universal constant $c > 0$ such that, for any $\varepsilon \in (0, \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}})$ and $\delta \in (0,1)$, the averaged TD estimator (14b) achieves*

$$\left\|\boldsymbol{V}_{\bar{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\boldsymbol{\theta}^\star}\right\|_{\boldsymbol{D}_\mu} = \left\|\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}} \leq \varepsilon \tag{17}$$

with probability exceeding $1-\delta$, provided that

$$T \geq \frac{c\{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}\left(1+\left\|\boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}^2\right)\log\left(\frac{d}{\delta}\right)}{(1-\gamma)^2\varepsilon^2}. \tag{18}$$

Comparisons to prior literature. We remark that the best finite-sample results for TD learning obtained so far are given by (Bhandari et al., 2021, Theorem 2(c)) and (Srikant and Ying, 2019, Corollary 1), with decaying stepsizes $\eta_t \asymp t^{-1}$ and sample size-related stepsizes $\eta_t \asymp T^{-1}$ respectively. Translated into our notation, they both prove that in order for the *expected* estimation error to be controlled by ε, namely

$$\mathbb{E}\left\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}^2 \leq \varepsilon^2,$$

it suffices to take (up to some logarithmic factors)

$$T^{\mathsf{prior}} \asymp \frac{\kappa\|\boldsymbol{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{1+\boldsymbol{\Sigma}}^2\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}. \tag{19}$$

We refer readers to Appendix D.1 and D.2 for a detailed translation of their results. Comparing (18) and (19), our result improves upon previous works by a multiplicative factor of

$$\frac{T^{\text{prior}}}{T^{\text{ours}}} = \kappa,$$

the condition number of $\mathbf{\Sigma}$; κ can be as large as d, the dimension of the features, which can scale with $|\mathcal{S}|$.

As for sample complexity with high-probability convergence guarantees, the best result so far is given by Dalal et al. (2018a), who shows that in order for (17) to hold with probability at least $1-\delta$, it suffices to take

$$T \asymp \max\left\{\left(\frac{1}{\varepsilon}\right)^2\left(\log\frac{1}{\delta}\right)^3, \left(\frac{1}{\varepsilon}\right)^{1+1/\lambda_{\min}(\boldsymbol{A})}\left(\log\frac{1}{\delta}\right)^{1+1/\lambda_{\min}(\boldsymbol{A})}\right\}. \tag{20}$$

Comparing (18) and (20), we can see that our result improves on the dependence of both the error tolerance ε and the probability tolerance δ; in fact, our result is the first sample complexity for TD learning with high-probability convergence guarantee that matches the minimax-optimal dependence of ε and displays a clear dependence on the problem-related parameters, as would be shown in the following section.

3.3 Minimax lower bounds

To assess the tightness of our upper bounds in Corollary 1, in this section, we provide a minimax lower bound for the value function estimation problem with linear approximation. More specifically, the question we intend to answer is: for any target accuracy level ε, do there exist estimators that achieve an ε-approximation of $\boldsymbol{V}_{\boldsymbol{\theta}^\star}$ with fewer samples? As shown in the following result, the answer is, by and large, negative.

Theorem 2 (Minimax lower bound)**.** *Consider any $\frac{1}{2} < \gamma < 1$, $1 < d \le |\mathcal{S}|$, and $0 < \varepsilon < c_1 \max\{1, \|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}\}$ for some universal constant $c_1 > 0$. There exist universal constants $c_2, c_3 > 0$ such that for any estimator $\widehat{\boldsymbol{\theta}}$ based on T independent pairs $\{(s_t, s_t')\}_{t=1}^T$ as in (10), there exists a Markov reward process and a choice of the feature matrix $\mathbf{\Phi}$ such that*

$$\mathbb{P}\Big\{\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big\|_{\mathbf{\Sigma}} > c_2\varepsilon\Big\} \ge \frac{1}{4}, \tag{21}$$

provided that the number of samples T satisfies

$$T \le \frac{c_3\big\{\max_s \boldsymbol{\phi}(s)^\top \mathbf{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(1 + \big\|\boldsymbol{\theta}^\star\big\|_{\mathbf{\Sigma}}^2\big)}{(1-\gamma)\varepsilon^2}. \tag{22}$$

Remark 1. We remark that minimax lower bounds are also previously investigated in a general framework in (Duan et al., 2021) where the value function is approximated using a general reproducing kernel Hilbert space (RKHS). When it comes to linear function approximation, for completeness, we include in Section B a different but simpler construction tailored to the linear space. Compared to the results of Duan et al. (2021), our lower bound is stated in terms of different parameters, which allows us to evaluate the tightness of Corollary 1 directly. Instantiating both lower bounds, they do agree and equal to

$$O\left(\frac{d}{\varepsilon^2(1-\gamma)^3}\right), \tag{23}$$

as one plugs in the exact parameters from our construction.

As asserted by this theorem, no algorithm whatsoever can attain an ε-approximation of the best linear coefficient — in a minimax sense — unless the total sample size exceeds

$$O\left(\frac{\big\{\max_s \boldsymbol{\phi}(s)^\top \mathbf{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(1 + \big\|\boldsymbol{\theta}^\star\big\|_{\mathbf{\Sigma}}^2\big)}{(1-\gamma)\varepsilon^2}\right).$$

Consequently, the upper bounds developed in Corollary 1 are sharp in terms of the accuracy level ε, the dependence of the feature map $\mathbf{\Phi}$, the underlying coefficient $\boldsymbol{\theta}^\star$, and the covariance matrix $\mathbf{\Sigma}$. Therefore, it implies that the performances of the TD learning algorithms can not be further improved in the minimax sense other than a factor of $\frac{1}{1-\gamma}$—the effective horizon.

4 Off-policy evaluation with TDC learning

In this section, we aim to estimate the optimizer $\widetilde{\boldsymbol{\theta}}^\star$ of the optimization problem (9) in the off-policy setting by means of the TDC algorithm. We continue to focus on the case when samples are generated in the i.i.d. fashion by the behavior policy π_b. At each time stamp t, one obtains

$$(s_t, a_t, s_t') \quad \text{where } s_t \overset{\text{i.i.d.}}{\sim} \mu_b,\ a_t \sim \pi_{\mathsf{b}}(\cdot \mid s_t), \quad \text{and } s_t' \sim \mathcal{P}(\cdot \mid s_t, a_t). \tag{24}$$

Here, recall that μ_{b} is the stationary distribution corresponding to the behavior policy π_{b}. We first provide some intuition behind the TDC algorithm before describing novel bounds on its sample complexity for obtaining an ε-accurate solution.

4.1 The TDC algorithm

The TDC algorithm is designed to solve the optimization problem (9) using a two-timescale linear TD with gradient correction (Sutton et al., 2009). To provide some high-level ideas behind the design of this algorithm, it is helpful to rewrite the objective function in the following form by directly expanding the terms in expression (9).

Claim 1. *The quantity* $\mathsf{MSPBE}(\boldsymbol{\theta})$ *can be equivalently written as*

$$\mathsf{MSPBE}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right]^\top \left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^\top \mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right], \tag{25}$$

where $\delta_t \coloneqq r_t + \gamma\boldsymbol{\phi}(s_t')^\top\boldsymbol{\theta} - \boldsymbol{\phi}(s_t)^\top\boldsymbol{\theta}$ *is the temporal difference error.*

In light of the above expression, the gradient of $\mathsf{MSPBE}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ equals to

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathsf{MSPBE}(\boldsymbol{\theta}) &= \mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\left(\gamma\boldsymbol{\phi}(s_t') - \boldsymbol{\phi}_t\right)\boldsymbol{\phi}(s_t)^\top\right]\left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] \\
&= -\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] + \gamma\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top\right]\left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] \\
&= -\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t))\delta_t\right] + \gamma\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top\right]\boldsymbol{w}_t,
\end{aligned} \tag{26}$$

where in the last step we have defined

$$\boldsymbol{w}_t = \boldsymbol{w}(\boldsymbol{\theta}_t) = \left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t)\delta_t\right]. \tag{27}$$

and have used the importance weights

$$\rho_t \coloneqq \frac{\pi(a_t|s_t)}{\pi_{\mathsf{b}}(a_t|s_t)} \tag{28}$$

to replace the expectation w.r.t. π with the expectation w.r.t. π_{b}.

The high-level idea of TDC is to estimate the right hand side of (26) based on the sample trajectory (24), and then perform stochastic gradient updates for $\boldsymbol{\theta}_t$. However, the challenge is that the second term in the gradient of MSPBE (26) involves the product of two expectations. Simultaneously sampling and using the sample product is inappropriate due to their correlation. In order to address this issue, Sutton et al. (2008) and Sutton et al. (2009) introduced an auxiliary parameter $\boldsymbol{w}$ to estimate $\boldsymbol{w}(\boldsymbol{\theta}_t)$ by solving a linear stochastic approximation (SA) problem corresponding to the linear system

$$\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\boldsymbol{w} = \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t)\delta_t\right]. \tag{29}$$

Putting these ideas together, TDC amounts to the following two-timescale linear stochastic method

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}_{t+1} &= \widetilde{\boldsymbol{\theta}}_t - \alpha_t[\gamma\rho_t\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top\boldsymbol{w}_t - \rho_t\delta_t\boldsymbol{\phi}(s_t)]; \\
\boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \beta_t\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\boldsymbol{w}_t - \rho_t\delta_t\boldsymbol{\phi}(s_t)\right].
\end{aligned}$$

Here, the update of $\widetilde{\boldsymbol{\theta}}_t$ corresponds to a gradient step regarding (25), the update of $\boldsymbol{w}_t$ corresponds to linear SA for solving (29), and $\delta_t \coloneqq r_t + \gamma \boldsymbol{\phi}(s_t')^\top \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(s_t)^\top \widetilde{\boldsymbol{\theta}}_t$ is the temporal difference error. In addition, α_t, β_t are the corresponding stepsizes. For notational convenience, let us denote

$$\begin{aligned} \widetilde{\boldsymbol{A}}_t = \rho_t \boldsymbol{\phi}(s_t) \left(\boldsymbol{\phi}(s_t) - \gamma \boldsymbol{\phi}(s_t')\right)^\top, \quad \widetilde{\boldsymbol{b}}_t \coloneqq \rho_t \boldsymbol{\phi}(s_t) r_t, \\ \boldsymbol{\Pi}_t \coloneqq \rho_t \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t')^\top, \quad \widetilde{\boldsymbol{\Sigma}}_t \coloneqq \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top. \end{aligned} \tag{30}$$

With these definitions, the TDC iterates can be written compactly as

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \alpha_t (\widetilde{\boldsymbol{A}}_t \widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}}_t + \gamma \boldsymbol{\Pi}_t^\top \boldsymbol{w}_t); \tag{31a}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \beta_t (\widetilde{\boldsymbol{A}}_t \widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}}_t + \widetilde{\boldsymbol{\Sigma}}_t \boldsymbol{w}_t). \tag{31b}$$

4.2 Sample complexity of TDC

Our finite-sample characterization of TDC builds upon a careful analysis of the population dynamics of TDC, which we then show to be uniformly well approximated by the empirical dynamics of TDC via matrix concentration inequalities. Before stating our main result, we find it helpful to introduce some extra pieces of notation. Specifically, define the population parameters as

$$\widetilde{\boldsymbol{A}} \coloneqq \mathbb{E}_{\mu_\mathsf{b}, \pi_\mathsf{b}, \mathcal{P}}[\widetilde{\boldsymbol{A}}_t] = \mathbb{E}_{\mu_\mathsf{b}, \pi_\mathsf{b}, \mathcal{P}}[\rho_t \boldsymbol{\phi}(s_t) \left(\boldsymbol{\phi}(s_t) - \gamma \boldsymbol{\phi}(s_t')\right)^\top]; \tag{32a}$$

$$\widetilde{\boldsymbol{b}} \coloneqq \mathbb{E}_{\mu_\mathsf{b}}[\widetilde{\boldsymbol{b}}_t] = \mathbb{E}_{\mu_\mathsf{b}, \pi_\mathsf{b}}[\rho_t \boldsymbol{\phi}(s_t) r_t]; \tag{32b}$$

$$\boldsymbol{\Pi} \coloneqq \mathbb{E}_{\mu_\mathsf{b}, \pi_\mathsf{b}, \mathcal{P}}[\boldsymbol{\Pi}_t] = \mathbb{E}_{\mu_\mathsf{b}, \pi_\mathsf{b}, \mathcal{P}}[\rho_t \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t')^\top]; \tag{32c}$$

$$\widetilde{\boldsymbol{\Sigma}} \coloneqq \mathbb{E}_{\mu_\mathsf{b}}[\widetilde{\boldsymbol{\Sigma}}_t] = \mathbb{E}_{\mu_\mathsf{b}}[\boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top]. \tag{32d}$$

In addition, denote the parameters

$$\begin{aligned} \lambda_1 = \lambda_{\min}(\widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}}), \quad \lambda_2 = \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}), \quad \lambda_\Sigma = \|\widetilde{\boldsymbol{\Sigma}}^{-1}\| = 1/\lambda_2, \\ \widetilde{\kappa} = \lambda_\Sigma \cdot \|\widetilde{\boldsymbol{\Sigma}}\|, \quad \rho_{\max} = \max_{s,a} [\pi(a|s)/\pi_\mathsf{b}(a|s)]. \end{aligned} \tag{33}$$

With these notation in place, we are ready to state our main result for TDC learning, with its proof deferred to Section 7.

Theorem 3. *There exist universal constants $\widetilde{C}_0, \widetilde{c}_1 > 0$, such that for any given $0 \le \delta \le 1$, the output $\widetilde{\boldsymbol{\theta}}_T$ of the TDC learning iterate (31) at time T satisfies the bound*

$$\|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{C}_0 \frac{\rho_{\max}^2 \|\widetilde{\boldsymbol{\Sigma}}\|}{\lambda_1} \sqrt{\frac{\beta}{\lambda_2} \log \frac{2dT}{\delta}} (\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 2), \tag{34}$$

with probability at least $1 - \delta$, provided that

$$\begin{aligned} &\widetilde{\boldsymbol{\theta}}_0 = \boldsymbol{0}, \\ &\alpha_0 = \ldots = \alpha_T = \alpha, \qquad \beta_0 = \ldots = \beta_T = \beta, \\ &0 < \alpha < \frac{1}{\lambda_1 \lambda_\Sigma^2 \|\widetilde{\boldsymbol{\Sigma}}\| \log \frac{2dT}{\delta}}, \qquad \frac{\alpha}{\beta} = \frac{1}{128} \frac{\lambda_1 \lambda_2}{\rho_{\max}^2 (1 + \lambda_\Sigma \rho_{\max})}, \\ &T \ge \widetilde{c}_1 \frac{\log \|\widetilde{\boldsymbol{\theta}}^\star\|_2}{\alpha \lambda_1} \log \max \left\{ \sqrt{\widetilde{\kappa}}, \ \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} \sqrt{\frac{\alpha \lambda_1}{\log \frac{2dT}{\delta}}} \right\}. \end{aligned} \tag{35}$$

Remark 2. A similar result in terms of the ℓ_2 error (namely, $\|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_2$) can be derived in the same way as in (34). In particular, under the same conditions as in (35), it can be derived with probability at least $1 - \delta$ that

$$\|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_2 \lesssim \widetilde{C}_0 \frac{\rho_{\max}^2}{\lambda_1} \sqrt{\frac{\beta}{\lambda_2} \log \frac{2dT}{\delta}} (\|\widetilde{\boldsymbol{\theta}}^\star\|_2 + 2). \tag{36}$$

Since the proof follows in the similar fashion, we omit here for brevity.

Next, we state a direct consequence of Theorem 3 below, which gives an upper bound for the sample complexity of TDC.

Corollary 2. *There exists a universal constant $\widetilde{c}$ such that, for any $\delta \in (0,1)$ and $\varepsilon \in (0, \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}})$, the TDC estimator $\widetilde{\boldsymbol{\theta}}_T$ at iterate T satisfies the bound*

$$\|\boldsymbol{V}_{\widetilde{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\widetilde{\boldsymbol{\theta}}^\star}\|_{\boldsymbol{D}_{\mu_b}} = \|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} \leq \varepsilon \tag{37}$$

with probability exceeding $1-\delta$, provided that

$$T \geq \widetilde{c}\frac{\rho_{\max}^7}{\lambda_1^4\lambda_2^3}\frac{\|\widetilde{\boldsymbol{\Sigma}}\|^2}{\varepsilon^2}(1+\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2)\log\Big(\frac{d\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}}{\delta}\Big), \tag{38}$$

and the stepsize parameters α_t and β_t are chosen as

$$\alpha_t \asymp \frac{\log\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}}{T\lambda_1}, \qquad \beta_t = 128\frac{\rho_{\max}^2(1+\lambda_\Sigma\rho_{\max})}{\lambda_1\lambda_2}\alpha. \tag{39}$$

Comparisons to other sample complexity bounds for TDC. Let us compare our results in Theorem 3 and Corollary 2 with the state-of-the-art sample complexities for the TDC algorithm. The result that is most comparable to ours is obtained by Dalal et al. (2020), where a projected version of TDC is considered with decaying stepsizes $\alpha_t = O(t^{-\alpha})$ and $\beta_t = O(t^{-\beta})$ for $0 < \beta < \alpha < 1$. The sample complexity therein, with high-probability convergence guarantee at tolerance level ε, is of order $O\left(\left(\frac{1}{\varepsilon}\right)^{2\alpha}\right)$ without explicit dependence on the problem-related parameters. If one chooses $\alpha = 1-\delta$ with δ sufficiently small, their sample complexity bound can be improved, but it cannot achieve the rate $\Theta\left(\frac{1}{\varepsilon^2}\right)$. Regarding finite-sample in-expectation error control for TDC, the best result so far is developed by Kaledin et al. (2020), who shows that with the choice of $\alpha_t, \beta_t \asymp \frac{1}{T}$, the sample complexity for TDC with tolerance level ε can be upper bounded by $O\left(\frac{1}{\varepsilon^2}\right)$. Our result in Corollary 2 is the first sample complexity for the original TDC algorithm that guarantees high-probability convergence and achieves the minimax-optimal rate of $O\left(\frac{1}{\varepsilon^2}\right)$; it is also noteworthy that we display an explicit dependence on problem-related parameters. We also remark that Xu and Liang (2021) considers a variant of TDC where $\widetilde{\boldsymbol{\theta}}_t$ is updated *not* with every sample tuple (s_t, a_t, s_t'), but with every batch of samples, and obtains a sample complexity of order $O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))$.

5 Numerical experiments

In this section, we corroborate our theoretical results with illustrative numerical experiments. In what follows, we will consider the on-policy and off-policy settings respectively.

5.1 On-policy evaluation: averaged TD learning

In the on-policy setting, we will investigate the empirical performance of the averaged TD learning algorithm.

MDP setting. We consider a member of the family of MDPs constructed in proof of Theorem 2, which provides a minimax lower bound. This family of MDPs is designed to be difficult to distinguish between each other, and hence, is a natural instance for evaluating the performance of TD learning. For construction details of this MDP, we refer the reader to Appendix B. In these simulations, we set $|\mathcal{S}| = 10$, $\gamma = 0.2$, and choose the stepsize of TD as $\eta = 0.01$. We examine both the original and the averaged TD iterates when the feature dimension equals to $d = 3$ and $d = 9$. Under each setting, 100 independent trials for $T = 10^5$ iterations were conducted, and we report the mean value as well as the 95% confidence band for the estimation error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}$ for TD and $\|\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}$ for averaged TD.

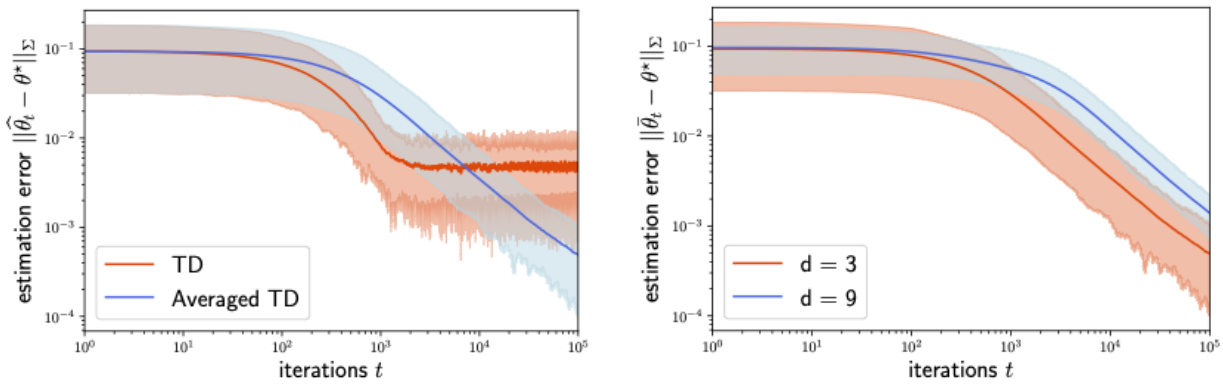

Figure 1. (a) Comparisons of the estimation error of TD and averaged TD when $d = 3$. (b) Comparisons of the estimation error for averaged TD with $d = 3$ and $d = 9$. Two curves in the middle represent their average errors, while the shaded areas represent the 95% confidence bands.

Experimental results. Figure 1(a) compares the performances of TD and averaged TD of an MDP with feature dimension $d = 3$. While the estimation error of TD levels off at around 5×10^{-3} after 10^3 iterations, the error of averaged TD keeps decreasing to below 5×10^{-4} when $T = 10^5$. In addition, Figure 1(b) demonstrates the estimation error of averaged TD for MDPs with feature dimension $d = 3$ and $d = 9$. The slopes of these curves on the right part of this log-log plot match our theoretical prediction: the estimation error decreases in the order of $O(t^{-1/2})$. Moreover, the difference between the two curves indicates that the lower-dimension problem enjoys a faster convergence rate.

5.2 Off-policy evaluation: TDC learning

In order to demonstrate the efficiency of TDC for off-policy evaluation, we compare its performance with that of the off-policy TD learning on *Baird's counterexample* (Baird, 1995).

Baird's counterexample. We start by introducing Baird's counterexample, which was constructed to illustrate the instability of TD learning in the off-policy regime. Consider an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, with the discount factor $\gamma = 0.9$, state space $\mathcal{S} = [7]$, action space $\mathcal{A} = \{0, 1\}$ and the reward function $r = 0$ for all states and actions. The action $a = 1$ transitions any initial state s to $s' = 7$, while the action $a = 0$ transitions any initial state s to $s' \in [6]$ with the same probability. The target policy π selects $a = 1$ at any given state s, while the behavior policy π_{b} takes $a = 0$ with probability $\frac{6}{7}$ and $a = 1$ with probability $\frac{1}{7}$. Formally, the MDP satisfies the equations (see also Figure 2 for an illustration)

$$\mathcal{P}(s'|s, 1) = \mathbb{1}\{s' = 7\}, \quad \forall s \in [7]; \qquad \mathcal{P}(s'|s, 0) = \frac{1}{6}\mathbb{1}\{1 \le s' \le 6\}, \quad \forall s \in [7];$$
$$\pi(1|s) = 1, \quad \forall s \in [7];$$
$$\pi_{\mathsf{b}}(0|s) = \frac{6}{7}, \quad \forall s \in [7]; \qquad \pi_{\mathsf{b}}(1|s) = \frac{1}{7}, \quad \forall s \in [7].$$

In this example, it is easy to check that the stationary distribution corresponding to the behavior policy π_{b} is the uniform distribution among all states, and that the value function is 0 for all states. We apply the following linear approximation of the value function: for $\boldsymbol{\theta} \in \mathbb{R}^8$,

$$\begin{cases} V(i) = 2\theta_i + \theta_8, & \text{for} \quad 1 \le i \le 6; \\ V(7) = \theta_7 + 2\theta_8. \end{cases} \tag{40}$$

We remark that with this approximation, the feature space has a higher dimension ($d = 8$) than the state space ($|\mathcal{S}| = 7$). Consequently, the optimal estimator $\widetilde{\boldsymbol{\theta}}^\star$ is not unique, and instead can be any $\boldsymbol{\theta} \in \mathbb{R}^8$ such

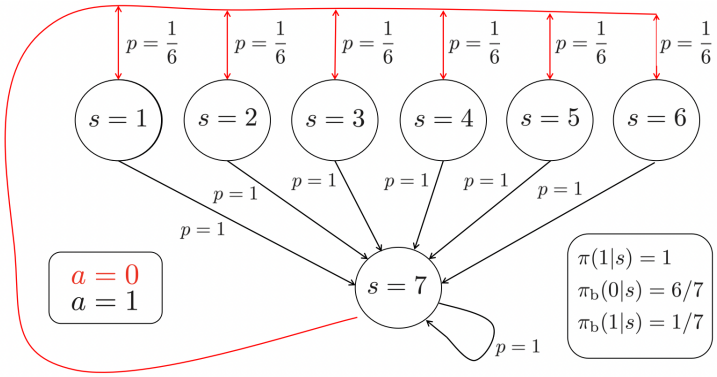

Figure 2. Baird's counterexample. Taking action $a = 1$ always leads to state $s = 7$, while taking $a = 0$ leads to one of the other six states with equal probability. The reward is set to be always zero.

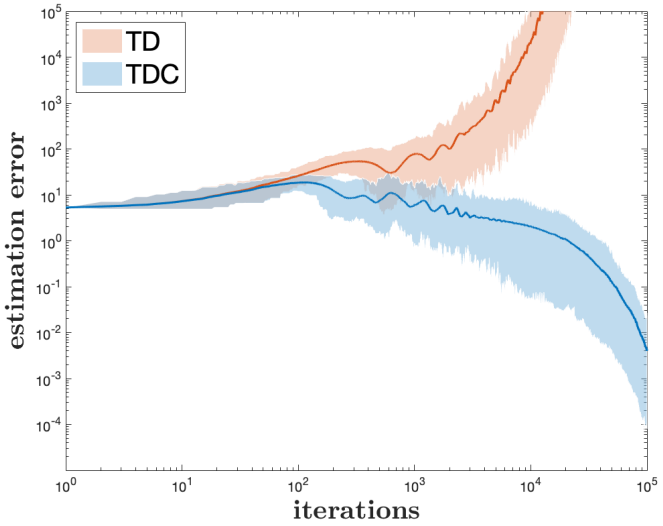

Figure 3. Performances of off-policy averaged TD (red, $\eta = 0.02$) and TDC (blue, $\alpha = 0.02$, $\beta = 0.002$). Two curves in the middle represent their average errors, while the shaded areas correspond to 95% confidence bands.

that the estimated value vector is $\boldsymbol{V}_{\boldsymbol{\theta}} = \mathbf{0}$. Technically, this issue can be circumvented by creating several identical states as state $s = 7$; we omit this detail here for simplicity, since we use $\|\widehat{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} = \|\boldsymbol{V}_{\widehat{\boldsymbol{\theta}}_t} - \boldsymbol{V}^\star\|_{\boldsymbol{D}_{\mu_b}}$ to evaluate the estimation error, and our experimental results would remain the same.

Experimental results. We perform 100 independent trials for both off-policy averaged TD learning (with stepsize $\eta = 0.02$) and TDC (with stepsizes $\alpha = 0.02, \beta = 0.002$), starting at $\widehat{\boldsymbol{\theta}}_0 = (1, 1, 1, 1, 1, 1, 10, 1)^\top$, as suggested by Baird (1995). In these experiments, we set $\alpha = \eta$ to ensure that the stepsize for $\boldsymbol{\theta}$-updates are the same between the two algorithms. Figure 3 demonstrates how the estimation error $\|\widehat{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}$ changes as two algorithms execute. As can be seen in this figure, TDC converges to an error of below 0.01 after $T = 10^5$ iterations while the off-policy averaged TD diverges to infinity.

6 Proof of Theorem 1 (TD learning)

For the sake of convenience, let us introduce the following notation

$$\boldsymbol{\Delta}_t \coloneqq \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star, \qquad \text{and} \qquad \overline{\boldsymbol{\Delta}}_t \coloneqq \overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star. \tag{41}$$

Step 1: a recursive relation. To understand the convergence behavior of $\overline{\boldsymbol{\Delta}}_t$, the idea is to first look at the following decomposition

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

where we define

$$\boldsymbol{\xi}_t \coloneqq (\boldsymbol{A}_t - \boldsymbol{A})\boldsymbol{\theta}_t - (\boldsymbol{b}_t - \boldsymbol{b}). \tag{42}$$

Here, the second line invokes the update rule (14a) and the identity $\boldsymbol{A}\boldsymbol{\theta}^\star = \boldsymbol{b}$, whereas the third line is obtained by properly rearranging terms. Applying the above relation recursively, one arrives at

$$\boldsymbol{\Delta}_t = (\mathbf{I} - \eta\boldsymbol{A})\boldsymbol{\Delta}_{t-1} - \eta\boldsymbol{\xi}_{t-1} = (\mathbf{I} - \eta\boldsymbol{A})^t\boldsymbol{\Delta}_0 - \eta\sum_{i=0}^{t-1}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i. \tag{43}$$

Step 2: a crude bound on $\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}$. We aim to establish, via an induction argument, that with probability at least $1-\delta$,

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(1 + \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}) + 2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} =: R_0 \tag{44}$$

simultaneously over all $0 \le t \le T$, as long as $0 < \eta_t \le \frac{c_3(1-\gamma)}{\kappa\log\frac{dT}{\delta}}$ for some sufficiently small constant $c_3 > 0$. As a side remark, this boundedness property saves us from enforcing additional projection steps as adopted in Bhandari et al. (2021).

To start with, note that the inequality (44) holds trivially for the base case with $t = 0$, given that $\kappa \ge 1$. Next, suppose that the hypothesis (44) holds for $\boldsymbol{\Delta}_0, \ldots, \boldsymbol{\Delta}_{t-1}$, and we intend to establish it for $\boldsymbol{\Delta}_t$ as well. Towards this end, invoking the decomposition (43) and the triangle inequality yields

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le \big\|(\mathbf{I} - \eta\boldsymbol{A})^t\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} + \eta\Big\|\sum_{i=0}^{t-1}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}. \tag{45}$$

As for the first term of (45), it is seen that

$$\begin{aligned}\big\|(\mathbf{I} - \eta\boldsymbol{A})^t\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} &= \big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \eta\boldsymbol{A})^t\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_0\big\|_2 \le \big\|\boldsymbol{\Sigma}^{1/2}\big\| \cdot \big\|\boldsymbol{\Sigma}^{-1/2}\big\| \cdot \big\|\mathbf{I} - \eta\boldsymbol{A}\big\|^t \cdot \big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_0\big\|_2 \\ &\le \sqrt{\kappa}\left(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})\right)^t\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} \le \sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}},\end{aligned} \tag{46}$$

where the last inequality arises from the definition of κ and the property (79g) (with the restriction that $\eta \le (1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$). When it comes to the second term of (45), the following lemma comes in handy.

Lemma 1. *Fix any quantity $R > 0$ and, for each $0 \le i \le T-1$, define the auxiliary random vector*

$$\widetilde{\boldsymbol{\xi}}_i \coloneqq \boldsymbol{\xi}_i\,\mathbb{1}\{\mathcal{H}_i\}, \qquad \text{where} \quad \mathcal{H}_i \coloneqq \Big\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R\Big\}. \tag{47}$$

Then, with probability at least $1-\delta/T$, simultaneously over the indices (l,u,t) such that $0 \le l \le u \le t-1 < T$ it holds that

$$\Big\|\sum_{i=l}^{u}(\boldsymbol{I} - \eta\boldsymbol{A})^{t-i-1}\widetilde{\boldsymbol{\xi}}_i\Big\|_{\boldsymbol{\Sigma}} \le 16\Big(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\Big)^{t-u-1}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R + 1)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}},$$

provided that $0 < \eta_t \le \frac{1-\gamma}{\kappa\log\frac{2dT}{\delta}}$.

Proof. See Section C.1. □

Under the induction hypothesis that $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_0$ for $0 \le i \le t-1$, we can invoke Lemma 1 (with $R = R_0, l = 0$ and $u = t-1$) to show that

$$\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} = \Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\mathbb{1}\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_0\}\Big\|_{\boldsymbol{\Sigma}} \le 16(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R_0 + 1)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \tag{48}$$

holds with probability at least $1-\delta/T$, provided that $0 < \eta \le \frac{1-\gamma}{\kappa\log\frac{dT}{\delta}}$. Combining (45), (46) and (48) together and recalling the definition (44) of R_0, we can easily verify that

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le \sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} + 16(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R_0 + 1)\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}} \le R_0, \tag{49}$$

with the proviso that $32\sqrt{\frac{\eta\kappa\log(2dT/\delta)}{1-\gamma}} \le 1$. The induction argument coupled with the union bound then establishes the claim (44).

Step 3: a refined bound on $\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}$. It turns out that the upper bound (44) is somewhat loose due to the complete ignorance of the contraction effect of $\mathbf{I}-\eta\boldsymbol{A}$; see (46). In what follows, we develop a strengthened bound. Define

$$t_{\mathsf{seg}} := \frac{c_1\log\max\{4\sqrt{\kappa}, \frac{16\kappa\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1}, \|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}\sqrt{\frac{1-\gamma}{\eta\kappa\log\frac{2dT}{\delta}}}\}}{\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})} \tag{50}$$

for some sufficiently large constant $c_1 > 0$. For any integer $k \ge 1$, we aim to establish that

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \frac{\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{k-1}} + \frac{3}{2}\Big) =: R_k \tag{51}$$

for any t obeying $kt_{\mathsf{seg}} \le t \le T$, provided that $0 < \eta \le \frac{c_3(1-\gamma)}{\kappa\log\frac{dT}{\delta}}$ for some small enough constant $c_3 > 0$.

Because of relation (45), we claim that it suffices to prove that

$$\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \le 32\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \frac{2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^k} + 1\Big)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}}, \qquad kt_{\mathsf{seg}} \le t \le T. \tag{52}$$

To see this: note that the first term on the right-hand side of (45) has already been bounded in (46), which combined with the definition (50) of t_{seg} indicates that

$$\big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\eta\boldsymbol{A})^t\boldsymbol{\Delta}_0\big\|_2 \le \sqrt{\kappa}\left(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})\right)^{t_{\mathsf{seg}}}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} \le \sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}} \tag{53}$$

for any $t \ge t_{\mathsf{seg}}$. Clearly, combining (52) with (45) and (53) shall immediately lead to the claim (51). The remainder of this step is thus devoted to demonstrating (52) inductively.

The base case (i.e. $k = 1$) follows immediately from our bounds (44) and (48) in Step 2, given that $\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}$ is sufficiently small. Suppose now that the claim (52) holds for a given integer $k \ge 1$ and any t obeying $kt_{\mathsf{seg}} \le t \le T$, and we intend to show that (52) continues to hold for $k+1$ and any t obeying $(k+1)t_{\mathsf{seg}} \le t \le T$. Towards this, we first single out the following straightforward decomposition

$$\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \le \Big\|\sum_{i=t-t_{\mathsf{seg}}+1}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} + \Big\|\sum_{i=0}^{t-t_{\mathsf{seg}}}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}},$$

which allows us to upper bound the two terms on the right-hand side above seperately.

- Under the induction hypothesis that $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_k$ for all i obeying $kt_{\mathsf{seg}} \le i \le T$, one can invoke Lemma 1 with $R = R_k, l = t - t_{\mathsf{seg}} + 1$ and $u = t-1$ to see that

$$\begin{aligned}\Big\| \sum_{i=t-t_{\mathsf{seg}}+1}^{t-1} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \Big\|_{\boldsymbol{\Sigma}} &= \Big\| \sum_{i=t-t_{\mathsf{seg}}+1}^{t-1} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \mathbb{1}\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_k\} \Big\|_{\boldsymbol{\Sigma}} \\ &\le 16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R_k + 1\big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \\ &\le 24\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \frac{2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{k+1}} + 1\Big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}},\end{aligned}$$

  where the last line uses the definition (51) of R_k and holds as long as $\frac{\eta\kappa \log\frac{2dT}{\delta}}{1-\gamma}$ is sufficiently small.

- In addition, we make the observation that: for any $t \ge t_{\mathsf{seg}}$,

$$\begin{aligned}\Big\| \sum_{i=0}^{t-t_{\mathsf{seg}}} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \Big\|_{\boldsymbol{\Sigma}} &= \Big\| \sum_{i=0}^{t-t_{\mathsf{seg}}} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \mathbb{1}\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_0\} \Big\|_{\boldsymbol{\Sigma}} \\ &\le 16(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}))^{t_{\mathsf{seg}}-1}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R_0 + 1)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \\ &\le 8\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + 1\big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}}.\end{aligned}$$

  Here, the first equality uses the crude bound $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_0$ for all i (see (44)), the second to last inequality utilizes Lemma 1 with $R = R_0, l = 0$ and $u = t - t_{\mathsf{seg}}$, whereas the last inequality relies on the definition (44) of R_0 and invokes the fact that $\sqrt{\kappa}\left(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})\right)^{t_{\mathsf{seg}}-1} \le \min\left\{\frac{1}{4}, \frac{1}{4\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}\right\}$ with our choice (50) of t_{seg}.

Combine the previous two bounds to reach

$$\begin{aligned}\Big\| \sum_{i=0}^{t-1} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \Big\|_{\boldsymbol{\Sigma}} &\le 24\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \frac{2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{k+1}} + 1\Big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}} + 8\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + 1\big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \\ &\le 32\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \frac{2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{k}} + 1\Big)\sqrt{\frac{\kappa \log\frac{2dT}{\delta}}{\eta(1-\gamma)}}.\end{aligned}$$

This finishes the induction step and in turn establishes (52) (and hence (51)).

As a straightforward consequence, the bounds (44) and (51) imply that

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le \begin{cases} R_0, & 0 \le t < t'_{\mathsf{seg}}, \\ 32\sqrt{\frac{\eta\kappa \log\frac{2dT}{\delta}}{1-\gamma}}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + 2), & t'_{\mathsf{seg}} \le t < T, \end{cases} \tag{54}$$

where

$$t'_{\mathsf{seg}} := c_2 t_{\mathsf{seg}} \log\left(\kappa(\|\boldsymbol{\Delta}_0\|_2 + 1)\right) \tag{55}$$

for some large enough constant $c_2 > 0$. To see this, note that for any $t \ge t'_{\mathsf{seg}}$, it is guaranteed that the second term on the right-hand side of (51) obeys $\frac{4\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{\lfloor t/t_{\mathsf{seg}} \rfloor}} \le 2$, thus confirming the second case in (54).

Step 4: controlling $\|\overline{\boldsymbol{\Delta}}_T\|_{\boldsymbol{\Sigma}}$. Now we are positioned to control $\overline{\boldsymbol{\Delta}}_T$. The key is to write $\overline{\boldsymbol{\Delta}}_T$ as a linear combination of $\{\boldsymbol{\xi}_i\}_{0 \le i \le T-1}$ as follows, which is a direct consequence of the relation (43):

$$\overline{\boldsymbol{\Delta}}_T = \frac{1}{T}\sum_{j=1}^{T} \boldsymbol{\Delta}_j = \frac{1}{T}\sum_{j=1}^{T} (\mathbf{I} - \eta\boldsymbol{A})^j \boldsymbol{\Delta}_0 - \frac{1}{T}\sum_{j=1}^{T} \eta \sum_{i=0}^{j-1} (\mathbf{I} - \eta\boldsymbol{A})^{j-i-1} \boldsymbol{\xi}_i$$

$$= \frac{1}{T}\sum_{j=1}^{T}(\mathbf{I}-\eta\boldsymbol{A})^j\boldsymbol{\Delta}_0 - \frac{1}{T}\sum_{i=0}^{T-1}\eta\sum_{j=i+1}^{T}(\mathbf{I}-\eta\boldsymbol{A})^{j-i-1}\boldsymbol{\xi}_i$$
$$= \frac{1}{T\eta}\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0 - \frac{1}{T}\boldsymbol{\Delta}_0 - \frac{1}{T}\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i, \tag{56}$$

where the middle line follows from swapping the summation over i and j, and in the last line we define

$$\boldsymbol{A}_i^{(t)} \coloneqq \eta\sum_{j=i+1}^{t}(\mathbf{I}-\eta\boldsymbol{A})^{j-i-1} = \boldsymbol{A}^{-1}\big(\mathbf{I}-(\mathbf{I}-\eta\boldsymbol{A})^{t-i}\big). \tag{57}$$

We claim that the following two inequalities hold, the first deterministically and the second with probability of at least $1-\delta$ (with their proofs deferred to Section C.2)

$$\big\|\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} \le \frac{2\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}; \tag{58a}$$
$$\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \lesssim \left\{\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T(1-\gamma)^2}} + \frac{\|\boldsymbol{\Sigma}^{-1}\|}{T}\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}\right\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\big). \tag{58b}$$

Putting the above two inequalities together with (56), we arrive at

$$\begin{aligned}
\|\overline{\boldsymbol{\Delta}}_T\|_{\boldsymbol{\Sigma}} &\le \Big\|\frac{1}{T\eta}\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0\Big\|_{\boldsymbol{\Sigma}} + \Big\|\frac{1}{T}\boldsymbol{\Delta}_0\Big\|_{\boldsymbol{\Sigma}} + \Big\|\frac{1}{T}\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \\
&\lesssim \frac{1}{\eta T}\frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} + \frac{1}{T}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} + \left\{\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T(1-\gamma)^2}} + \frac{\|\boldsymbol{\Sigma}^{-1}\|}{T}\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}\right\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\big) \\
&\asymp \frac{1}{\eta T}\frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} + \left\{\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T(1-\gamma)^2}} + \frac{\|\boldsymbol{\Sigma}^{-1}\|}{T}\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}\right\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\big),
\end{aligned}$$

where the last line follows since $\|\boldsymbol{\Sigma}^{-1}\| \ge 1$ (see (79h)) and $\eta < 1$. This finishes the proof of Theorem 1.

7 Proof of Theorem 3 (TDC learning)

Firstly, let us analyze the population dynamics of TDC. It turns out that the convergence of this dynamics can be described via a contractive linear mapping. Given this nice property of population TDC, we shall decompose the empirical TDC into two parts: the first part can be controlled via the aforementioned population dynamics, and the rest is treated as a stochastic component, which is controlled via matrix martingale concentration.

7.1 Population analysis

First recall that the population parameters are defined as

$$\begin{aligned}
\widetilde{\boldsymbol{A}} &\coloneqq \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}},\mathcal{P}}[\widetilde{\boldsymbol{A}}_t] = \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}},\mathcal{P}}[\rho_t\boldsymbol{\phi}(s_t)\left(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')\right)^\top]; \\
\widetilde{\boldsymbol{b}} &\coloneqq \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}}}[\widetilde{\boldsymbol{b}}_t] = \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}}}[\rho_t\boldsymbol{\phi}(s_t)r_t]; \\
\boldsymbol{\Pi} &\coloneqq \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}},\mathcal{P}}[\boldsymbol{\Pi}_t] = \mathbb{E}_{\mu_{\mathrm{b}},\pi_{\mathrm{b}},\mathcal{P}}[\rho_t\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top]; \\
\widetilde{\boldsymbol{\Sigma}} &\coloneqq \mathbb{E}_{\mu_{\mathrm{b}}}[\boldsymbol{\Sigma}_t] = \mathbb{E}_{\mu_{\mathrm{b}}}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top].
\end{aligned}$$

Corresponding to the empirical version of TDC as given in (31), we can define its population analogue of TDC as

$$\breve{\boldsymbol{\theta}}_{t+1} = \breve{\boldsymbol{\theta}}_t - \alpha(\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}} + \gamma\boldsymbol{\Pi}^\top\breve{\boldsymbol{w}}_t),$$

$$\breve{\boldsymbol{w}}_{t+1} = \breve{\boldsymbol{w}}_t - \beta(\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}} + \widetilde{\boldsymbol{\Sigma}}\breve{\boldsymbol{w}}_t), \tag{59}$$

where sampled parameters are replaced by their corresponding expectations. In this section, we analyze the population dynamics of TDC as given above; in order to control the finite-sample dynamics, we bound the difference of these two in the section to follow.

Since $\boldsymbol{\phi}(s_t)$ is independent of the transition, the expectation of $\widetilde{\boldsymbol{\Sigma}}_t$ is independent of which policy is being adopted. Hence, $\widetilde{\boldsymbol{\Sigma}}$ can also be presented as

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}} &= \sum_{s_t \in \mathcal{S}} \mu_{\mathsf{b}}(s_t) \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top \\
&= \sum_{s_t \in \mathcal{S}} \mu_{\mathsf{b}}(s_t) \left(\sum_{a_t \in \mathcal{A}} \pi(a_t|s_t) \right) \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top \\
&= \sum_{s_t \in \mathcal{S}} \sum_{a_t \in \mathcal{A}} \mu_{\mathsf{b}}(s_t) \pi_{\mathsf{b}}(a_t|s_t) \left(\frac{\pi(a_t|s_t)}{\pi_{\mathsf{b}}(a_t|s_t)} \right) \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top = \mathbb{E}_{\mu_{\mathsf{b}}, \pi_{\mathsf{b}}}[\rho_t \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top].
\end{aligned} \tag{60}$$

In view of this relation, $\widetilde{\boldsymbol{A}}$ admits another characterization, namely

$$\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{\Sigma}} - \gamma \boldsymbol{\Pi}. \tag{61}$$

Consequently, the fixed point $(\breve{\boldsymbol{\theta}}^\star, \boldsymbol{w}^\star)$ of the population dynamics obeys

$$\begin{cases} \widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}^\star - \widetilde{\boldsymbol{b}} + \gamma \boldsymbol{\Pi}^\top \boldsymbol{w}^\star = \boldsymbol{0}, \\ \widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}^\star - \widetilde{\boldsymbol{b}} + \widetilde{\boldsymbol{\Sigma}} \boldsymbol{w}^\star = \boldsymbol{0}. \end{cases}$$

As long as $\widetilde{\boldsymbol{A}}$ is invertible, this set of conditions is equivalent to

$$\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}^\star = \widetilde{\boldsymbol{b}}, \qquad \text{and} \qquad \boldsymbol{w}^\star = \boldsymbol{0}.$$

In order to study the population dynamics, it is useful to consider two auxiliary parameters

$$\begin{aligned}
\breve{\boldsymbol{\Delta}}_t &\coloneqq \breve{\boldsymbol{\theta}}_t - \breve{\boldsymbol{\theta}}^\star, \\
\breve{\boldsymbol{z}}_t &\coloneqq \breve{\boldsymbol{w}}_t + \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}} \breve{\boldsymbol{\Delta}}_t;
\end{aligned}$$

here $\breve{\boldsymbol{\Delta}}_t$ tracks the convergence of $\breve{\boldsymbol{\theta}}_t$ to $\breve{\boldsymbol{\theta}}^\star$, and $\breve{\boldsymbol{z}}_t$ tracks the size of the residual $\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}} + \widetilde{\boldsymbol{\Sigma}}\breve{\boldsymbol{w}}_t$. With these two parameters in place, the population dynamics satisfy

$$\begin{bmatrix} \breve{\boldsymbol{\Delta}}_t \\ \breve{\boldsymbol{z}}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} - \alpha \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}} & -\alpha\gamma \boldsymbol{\Pi}^\top \\ -\alpha(\boldsymbol{I} - \gamma \boldsymbol{\Sigma}^{-1} \boldsymbol{\Pi}) \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}} & \boldsymbol{I} - \beta \widetilde{\boldsymbol{\Sigma}} - \alpha\gamma(\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Pi}) \boldsymbol{\Pi}^\top \end{bmatrix} \begin{bmatrix} \breve{\boldsymbol{\Delta}}_{t-1} \\ \breve{\boldsymbol{z}}_{t-1} \end{bmatrix}. \tag{62}$$

To analyze this optimization dynamics, for every positive constant $\varkappa \in (0, 1)$, consider

$$\breve{\boldsymbol{x}}_t \coloneqq \begin{bmatrix} \breve{\boldsymbol{\Delta}}_t \\ \varkappa \breve{\boldsymbol{z}}_t \end{bmatrix}$$

then $\breve{\boldsymbol{x}}_t$ yields

$$\breve{\boldsymbol{x}}_t = \underbrace{\begin{bmatrix} \boldsymbol{I} - \alpha \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}} & -\frac{1}{\varkappa}\alpha\gamma \boldsymbol{\Pi}^\top \\ -\varkappa\alpha(\boldsymbol{I} - \gamma \boldsymbol{\Sigma}^{-1} \boldsymbol{\Pi}) \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}} & \boldsymbol{I} - \beta \widetilde{\boldsymbol{\Sigma}} - \alpha\gamma(\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Pi}) \boldsymbol{\Pi}^\top \end{bmatrix}}_{=: \boldsymbol{\Psi}} \breve{\boldsymbol{x}}_{t-1}. \tag{63}$$

It is known that how fast $\breve{\boldsymbol{x}}_t$ converges to $\boldsymbol{0}$ is determined by the spectral norm of $\boldsymbol{\Psi}$, which is characterized in the lemma below.

Lemma 2. *Suppose that*

$$\lambda_1 = \lambda_{\min}(\widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}}), \quad \lambda_2 = \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}), \quad \lambda_\Sigma = \|\widetilde{\boldsymbol{\Sigma}}^{-1}\| = 1/\lambda_2.$$

Then as long as the following conditions hold:

$$\beta \gtrsim \lambda_\Sigma \rho_{\max} \alpha, \tag{64a}$$

$$\varkappa\beta \gtrsim \alpha, \tag{64b}$$

$$\alpha\gamma(\rho_{\max} + \gamma\lambda_\Sigma \rho_{\max}^2) \ll \beta\lambda_w, \tag{64c}$$

$$\frac{\alpha\gamma\rho_{\max}}{\varkappa} + \varkappa\alpha(1 + \gamma\lambda_\Sigma\rho_{\max})\lambda_\Sigma(2\rho_{\max})^2 \ll \sqrt{\alpha\lambda_1\beta\lambda_w} \tag{64d}$$

it holds true that

$$\|\boldsymbol{\Psi}\| \leq 1 - \frac{1}{2}\alpha\lambda_1.$$

Therefore, the mapping $\boldsymbol{\Psi}$ is contractive, thus ensuring the linear convergence of $\boldsymbol{x}_t$, with the proviso that $\alpha\lambda_1 < 2$.

7.2 Finite-sample analysis

Armed with the population analysis, the proof for Theorem 3 is completed if we can make a connection of the finite-sample performances to that of the population ones.

Step 1: a recursive relation. Firstly, we define two noise variables

$$\begin{aligned}
\boldsymbol{\nu}_t &\coloneqq (\widetilde{\boldsymbol{A}}_t - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_t - (\widetilde{\boldsymbol{b}}_t - \widetilde{\boldsymbol{b}}) + \gamma(\boldsymbol{\Pi}_t - \boldsymbol{\Pi})^\top \boldsymbol{w}_t, \\
\boldsymbol{\eta}_t &\coloneqq (\widetilde{\boldsymbol{A}}_t - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_t - (\widetilde{\boldsymbol{b}}_t - \widetilde{\boldsymbol{b}}) + (\widetilde{\boldsymbol{\Sigma}}_t - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{w}_t.
\end{aligned}$$

As a result, TDC can be rewritten as

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}_{t+1} &= \widetilde{\boldsymbol{\theta}}_t - \alpha(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}} + \gamma\boldsymbol{\Pi}^\top \boldsymbol{w}_t) - \alpha\boldsymbol{\nu}_t; \\
\boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \beta(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}} + \widetilde{\boldsymbol{\Sigma}}\boldsymbol{w}_t) - \beta\boldsymbol{\eta}_t.
\end{aligned}$$

Using the same notations as in Section 7.1, we observe that the following iteration holds true for finite-sample TDC:

$$\boldsymbol{x}_{t+1} = \boldsymbol{\Psi}\boldsymbol{x}_t - \boldsymbol{\zeta}_t,$$

in which

$$\boldsymbol{\zeta}_t = \begin{bmatrix} \alpha\boldsymbol{\nu}_t \\ \varkappa(\alpha(1 - \gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\nu}_t + \beta\boldsymbol{\eta}_t) \end{bmatrix}. \tag{65}$$

Hence,

$$\boldsymbol{x}_t = \boldsymbol{\Psi}^t \boldsymbol{x}_0 - \sum_{i=0}^{t-1} \boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i, \tag{66}$$

where $\boldsymbol{x}_0 = [\boldsymbol{\Delta}_0^\top, \varkappa\boldsymbol{z}_0^\top]^\top$. Since the norm of $\boldsymbol{\Psi}$ has been bounded by Lemma 2, bounding the norm of $\boldsymbol{x}_t$ boils down to bounding the second term of (66). In the following, with a slight abuse of notation, for any $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{2d}$ with $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, we will define $\|\boldsymbol{x}\|_{\widetilde{\boldsymbol{\Sigma}}}^2$ as

$$\|\boldsymbol{x}\|_{\widetilde{\boldsymbol{\Sigma}}}^2 = \|\boldsymbol{x}_1\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\boldsymbol{x}_2\|_{\widetilde{\boldsymbol{\Sigma}}}^2.$$

with this definition, it is easy to see that

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}^2 = \|\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \varkappa^2 \|\boldsymbol{w}_t + \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}}^2.$$

Hence, the norms of $\widetilde{\boldsymbol{\Delta}}_t$, $\boldsymbol{w}_t$ and $\boldsymbol{x}_t$ can be related by the inequalities

$$\begin{cases} \|\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \le \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}; \\ \|\boldsymbol{w}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \frac{1}{\varkappa}\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}; \\ \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \|\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}} + |\boldsymbol{w}_t\|_{\widetilde{\boldsymbol{\Sigma}}}. \end{cases} \tag{67}$$

Step 2: crude bound for $\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}$. We first aim to establish, via an induction argument, that with probability at least $1-\delta$,

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \le 2\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}} + 80\varkappa\beta\rho_{\max}\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+1) =: \widetilde{R}_0 \tag{68}$$

holds simulatanesouly for all $0 \le t \le T$. To start with, note that the inequality (68) holds trivially for the base case with $t=0$. Next, suppose that the hypothesis (68) holds for $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$, and we intend to establish it for $\boldsymbol{x}_t$ as well. Towards this end, involking the decomposition (66) and the triangle inequality yields

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} = \left\|\boldsymbol{\Psi}^t\boldsymbol{x}_0\right\|_{\widetilde{\boldsymbol{\Sigma}}} + \left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{69}$$

As for the first term of (69), it is seen that

$$\left\|\boldsymbol{\Psi}^t\boldsymbol{x}_0\right\|_{\widetilde{\boldsymbol{\Sigma}}} \le \|\boldsymbol{x}_0\|_{\widetilde{\boldsymbol{\Sigma}}} = \|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{70}$$

When it comes to the second term of (69), the following lemma comes in handy.

Lemma 3. *Fix any quantity $\widetilde{R} > 0$ and, for each $0 \le i \le T-1$, define the random vector*

$$\widetilde{\boldsymbol{\zeta}}_i := \boldsymbol{\zeta}_i \mathbb{1}\{\widetilde{\mathcal{H}}_i\}, \quad \text{where} \quad \widetilde{\mathcal{H}}_i := \left\{\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{R}\right\}. \tag{71}$$

Then, with probability at least $1-\delta/T$,

$$\left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{1}{\varkappa}\widetilde{R} + 1), \tag{72}$$

provided that the stepsizes α, β satisfy the conditions (64) and that $0 < \alpha < \frac{1}{\lambda_1\lambda_\Sigma^2\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}$.

Proof. See Section C.4. □

Putting relations (69) and (72) together, we find

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} = \|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}} + C\sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{1}{\varkappa}\widetilde{R}_0 + 1) \le \widetilde{R}_0$$

by definition of $\widetilde{R}_0$ in (68), provided that $\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\beta\rho_{\max} \le c$ for some constant $c > 0$ small enough. Therefore, by induction assumption, one has

$$\begin{aligned} \mathbb{P}\left\{\max_{0\le i\le t}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} &\le \mathbb{P}\left\{\max_{0\le i<t-1}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} + \mathbb{P}\left\{\max_{0\le i<t-1}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{R}_0, \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} \\ &\le \frac{(t-1)\delta}{T} + \mathbb{P}\left\{\left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}} \gtrsim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \widetilde{R}_0 + 1)\right\} \\ &\le \frac{(t-1)\delta}{T} + \frac{\delta}{T} = \frac{t\delta}{T}. \end{aligned} \tag{73}$$

This completes our claim at this step.

Step 3: refined bound for $\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}$. It turns out that the upper bound (68) can be tightened by taking into account the contraction effect of $\boldsymbol{\Psi}$. In what follows, we develop a strengthened bound. Define

$$\widetilde{t}_{\mathsf{seg}} := \frac{\widetilde{c}_1 \log \max\left\{\sqrt{\widetilde{\kappa}}, \frac{\sqrt{\widetilde{\kappa}}\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}}{\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+1}, \|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}\sqrt{\frac{\alpha\lambda_1}{\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}}\frac{1}{\varkappa\beta\rho_{\max}}\right\}}{\alpha\lambda_1} \tag{74}$$

for some sufficiently large constant $\widetilde{c}_1 > 0$, where $\widetilde{\kappa}$ is the condition number of $\widetilde{\boldsymbol{\Sigma}}$. For any integer $k > 1$, we claim that with probability at least $1-\delta$,

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \varkappa\beta\rho_{\max}\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}}{2^{k-1}} + \frac{3}{2}\right) =: \widetilde{R}_k \tag{75}$$

for any t obeying $k\widetilde{t}_{\mathsf{seg}} \le t \le T$, provided that $\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max} \le c$ for some constant c small enough. The proof of this claim is essentially the same as that of Step 3 for proving Theorem 1, and we will omit it here. Therefore, by defining

$$\widetilde{t}'_{\mathsf{seg}} := \left(2 + \frac{1}{\log 2}\log\|\widetilde{\boldsymbol{\theta}^\star}\|_{\widetilde{\boldsymbol{\Sigma}}}\right)\widetilde{t}_{\mathsf{seg}}, \tag{76}$$

we can conclude that with probability at least $1-\delta$, for all $t \ge \widetilde{t}'_{\mathsf{seg}}$,

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \varkappa\beta\rho_{\max}\sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 2\right). \tag{77}$$

Recall that this bound holds for any $\varkappa \in (0,1)$ satisfying the conditions (64). Hence, Theorem 3 follows by taking $\varkappa = 8\rho_{\max}\sqrt{\frac{\alpha}{\lambda_1\beta\lambda_2}}$ and

$$\frac{\alpha}{\beta} = \frac{1}{128}\frac{\lambda_1\lambda_2}{\rho_{\max}^2(1+\lambda_\Sigma\rho_{\max})}.$$

8 Discussion

Our primary contribution in this paper is obtaining high-probability sample complexity bounds for both the TD and TDC algorithms for policy evaluation in the γ-discounted infinite-horizon MDPs. For TD learning with Polyak-Ruppert averaging, we improve upon existing results in terms of both the accuracy level ε and other problem-related parameters like the effective horizon $\frac{1}{1-\gamma}$, the weighted feature covariance $\boldsymbol{\Sigma}$ and the optimal linear estimator $\boldsymbol{\theta}^\star$. We have also established a minimax lower bound and showed that our upper bound is near-minimax optimal by a factor of $\frac{1}{1-\gamma}$. For TDC with linear function approximation, we provide the first sample complexity bound that achieves the optimal dependence on the error tolerance ε, and characterize the exact dependence on problem-related constants at the same time.

Our analysis leaves open several directions for future investigation; we close by sampling a few of them. Regarding TD learning, a natural direction of future work is to close the $\frac{1}{1-\gamma}$ gap between our upper bound and the minimax lower bound. Notably, this gap also appears in the bounds of Duan et al. (2021) for least-square TD in general when no restriction of the variance for the temporal difference residual is imposed. In terms of TDC, while our result provides a tight control of the same size T, the dependence on problem-related constants can be potentially improved. Moreover, it is noteworthy that the analysis in this work is based on the assumption of *i.i.d.* transition pairs drawn from the stationary distribution; it is of natural interest to generalize these results to other scenarios such as Markovian trajectories. Moving beyond linear function approximation, understanding the sample complexities for policy evaluation with other function classes is also an interesting direction.

Acknowledgements

W. Wu and A. Rinaldo are supported in part by the NIH Grant R01 NS121913. Y. Chi is supported in part by the grants ONR N00014-19-1-2404 and NSF CCF-2106778. Y. Wei is supported in part by the NSF grants CCF-2106778, DMS-2147546/2015447 and NSF CAREER award DMS-2143215 and Google Research Scholar Award.

A Preliminary facts

The following two lemmas consider the basic properties of important matrices and vectors that would be useful in the proof of the main theorems in the paper.

Lemma 4. *Recall the definitions of* $\boldsymbol{\Phi}$, $\boldsymbol{D}_\mu$ *and* $\boldsymbol{\Sigma}$ *in* (4), (5) *and* (6), *respectively. Then one has*

$$\left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| = 1, \qquad \text{and} \qquad \left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}^{\pi}\boldsymbol{D}_\mu^{-\frac{1}{2}}\right\| = 1. \tag{78}$$

Proof. For notational convenience, let $\widetilde{\boldsymbol{\Phi}} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\boldsymbol{P}_{\boldsymbol{D}_\mu} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}^{\pi}\boldsymbol{D}_\mu^{-\frac{1}{2}}$. First of all, it is seen that

$$\left\|\widetilde{\boldsymbol{\Phi}}\right\| = \sqrt{\left\|\widetilde{\boldsymbol{\Phi}}^\top\widetilde{\boldsymbol{\Phi}}\right\|} = \sqrt{\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\|} = \sqrt{\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\|} = 1.$$

When it comes to $\left\|\boldsymbol{P}_{\boldsymbol{D}_\mu}\right\|$, we make the observation that

$$\left\|\boldsymbol{P}_{\boldsymbol{D}_\mu}\right\| = \sqrt{\left\|\boldsymbol{P}_{\boldsymbol{D}_\mu}\boldsymbol{P}_{\boldsymbol{D}_\mu}^\top\right\|} = \sqrt{\left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu^{\frac{1}{2}}\right\|} = \sqrt{\left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\left(\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu\right)\boldsymbol{D}_\mu^{-\frac{1}{2}}\right\|} = 1.$$

To see why the last identity holds, observe that $\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu$ is a stochastic matrix, that is $\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu$ contains nonnegative elements, and

$$\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu\boldsymbol{1} = \boldsymbol{1}.$$

In addition, $\boldsymbol{D}_\mu^{\frac{1}{2}}\left(\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu\right)\boldsymbol{D}_\mu^{-\frac{1}{2}}$ is similar to $\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu$. As a result, by the Perron-Frobenious theorem,

$$\begin{aligned}\left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\left(\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu\right)\boldsymbol{D}_\mu^{-\frac{1}{2}}\right\| &= \max_i\left|\lambda_i(\boldsymbol{D}_\mu^{\frac{1}{2}}\left(\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu\right)\boldsymbol{D}_\mu^{-\frac{1}{2}})\right| \\ &= \max_i\left|\lambda_i(\boldsymbol{P}\boldsymbol{D}_\mu^{-1}\boldsymbol{P}^\top\boldsymbol{D}_\mu)\right| = 1,\end{aligned}$$

where $\lambda_i(\boldsymbol{B})$ denotes the i-th eigenvalue of the matrix $\boldsymbol{B}$. □

Lemma 5. *Suppose that* $\|\boldsymbol{r}\|_\infty \le 1$. *For any* $0 \le \gamma < 1$, *the matrix* $\boldsymbol{\Sigma}$ *defined in* (6) *and the vector* $\boldsymbol{b}$ *defined in* (12b) *obey*

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}} \succeq (1-\gamma)^2\boldsymbol{I}, \tag{79a}$$

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}} \succeq (1-\gamma)^2\boldsymbol{I}, \tag{79b}$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-2}, \tag{79c}$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-2}, \tag{79d}$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-1}, \tag{79e}$$

$$\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Phi}^\top\boldsymbol{D}_{\boldsymbol{\mu}}\right\| \le \max_{s\in\mathcal{S}}\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s), \tag{79f}$$

$$\|\boldsymbol{I}-\eta\boldsymbol{A}\| \le 1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}), \quad \forall 0<\eta<\frac{1-\gamma}{4\|\boldsymbol{\Sigma}\|}, \tag{79g}$$

$$\|\boldsymbol{\Sigma}\| \le 1, \qquad \|\boldsymbol{\Sigma}^{-1}\| \ge 1, \tag{79h}$$

$$\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}\right\|_2 \le 1. \tag{79i}$$

Proof. We shall establish each of these claims separately as follows.

Proof of Eqn. (79a) and (79b). We start with the lower bound on $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. To begin with, observe that

$$\begin{aligned}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}} &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\mathbf{I}-\gamma\boldsymbol{P})\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} - \gamma\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu^{\frac{1}{2}}\left(\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_\mu^{-\frac{1}{2}}\right)\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} \\ &= \mathbf{I}-\gamma\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}},\end{aligned}$$

where

$$\widetilde{\boldsymbol{\Phi}} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} \qquad \text{and} \qquad \boldsymbol{P}_{\boldsymbol{D}_\mu} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_\mu^{-\frac{1}{2}}. \tag{80}$$

Therefore, any unit vector $\boldsymbol{x}$ (i.e. $\|\boldsymbol{x}\|_2=1$) necessarily satisfies

$$\begin{aligned}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} &= \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\right\|_2^2 \geq \left(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\right)^2 \\ &= \left(1-\gamma\boldsymbol{x}^\top\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}}\boldsymbol{x}\right)^2.\end{aligned}$$

Further, Lemma 4 tells us that

$$\left|\boldsymbol{x}^\top\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}}\boldsymbol{x}\right| \leq \|\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}}\| \leq \|\widetilde{\boldsymbol{\Phi}}\|^2\|\boldsymbol{P}_{\boldsymbol{D}_\mu}\| = 1. \tag{81}$$

Putting the preceding two bounds together, we demonstrate that

$$\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x} \geq \left(1-\gamma\right)^2$$

for any unit vector $\boldsymbol{x}$, thus concluding the proof of (79a). The proof for (79b) follows from an identical argument and is omitted for brevity.

Proof of Eqn. (79c), (79d) and (79e). With the above bounds in place, we can further obtain

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| = \left\|\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^{-1}\right\| \leq \frac{1}{\lambda_{\min}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)} \leq \frac{1}{(1-\gamma)^2},$$

where $\lambda_{\min}(\boldsymbol{B})$ denotes the smallest eigenvalue of $\boldsymbol{B}$, and the last inequality comes from (79b). This establishes (79c). The inequality (79d) follows from a similar argument. This also implies that

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| = \sqrt{\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|} \leq \frac{1}{1-\gamma},$$

as claimed in (79e).

Proof of Eqn. (79g). Recalling that $\boldsymbol{\Sigma}=\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu\boldsymbol{\Phi}$, we can arrange terms to derive

$$\begin{aligned}\boldsymbol{A}+\boldsymbol{A}^\top &= \boldsymbol{\Phi}^\top\boldsymbol{D}_\mu(\mathbf{I}-\gamma\boldsymbol{P})\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top(\mathbf{I}-\gamma\boldsymbol{P}^\top)\boldsymbol{D}_\mu\boldsymbol{\Phi} \\ &= 2\boldsymbol{\Sigma} - \gamma\boldsymbol{\Sigma}^{\frac{1}{2}}\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu\boldsymbol{P}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{P}^\top\boldsymbol{D}_\mu\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}\left\{2\mathbf{I} - \gamma\left(\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}} + \widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}^\top\widetilde{\boldsymbol{\Phi}}\right)\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &\succeq \boldsymbol{\Sigma}^{\frac{1}{2}}\left\{2\mathbf{I} - 2\gamma\|\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}}\|\mathbf{I}\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &\succeq 2(1-\gamma)\boldsymbol{\Sigma},\end{aligned}$$

where $\widetilde{\boldsymbol{\Phi}}$ and $\boldsymbol{P}_{\boldsymbol{D}_\mu}$ are defined in (80). Here, the last line follows since $\left\|\widetilde{\boldsymbol{\Phi}}^\top\boldsymbol{P}_{\boldsymbol{D}_\mu}\widetilde{\boldsymbol{\Phi}}\right\| \leq 1$ — a fact that has already been shown in (81). In addition, the following identity

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{\Phi}}^\top\left(\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}\right)\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Phi}}^\top\left(\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}^\top\right)\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Sigma}^{\frac{1}{2}}$$

allows us to bound

$$\begin{aligned}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\| &= \big\|\widetilde{\boldsymbol{\Phi}}^\top\big(\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}\big)\widetilde{\boldsymbol{\Phi}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Phi}}^\top\big(\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}^\top\big)\widetilde{\boldsymbol{\Phi}}\big\| \\ &\le \big\|\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}\big\|^2\big\|\widetilde{\boldsymbol{\Phi}}\big\|^4\|\boldsymbol{\Sigma}\| = \|\mathbf{I}-\gamma\boldsymbol{P}_{\boldsymbol{D}_\mu}\|^2\|\boldsymbol{\Sigma}\| \\ &\le \big(1+\gamma\big\|\boldsymbol{P}_{\boldsymbol{D}_\mu}\big\|\big)^2\|\boldsymbol{\Sigma}\| \le 4\|\boldsymbol{\Sigma}\|,\end{aligned}$$

where the last line makes use of Lemma 4. This essentially tells us that

$$\begin{aligned}&\mathbf{0} \preceq \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}} \preceq 4\|\boldsymbol{\Sigma}\|\mathbf{I} \\ \Longrightarrow\qquad &\boldsymbol{A}\boldsymbol{A}^\top \preceq 4\|\boldsymbol{\Sigma}\|\boldsymbol{\Sigma}.\end{aligned}$$

Putting the preceding bounds together implies that: for any $0<\eta<\frac{1-\gamma}{4\|\boldsymbol{\Sigma}\|}$ one has

$$\begin{aligned}\mathbf{0} \preceq \big(\mathbf{I}-\eta\boldsymbol{A}\big)\big(\mathbf{I}-\eta\boldsymbol{A}^\top\big) &= \mathbf{I}-\eta(\boldsymbol{A}+\boldsymbol{A}^\top)+\eta^2\boldsymbol{A}\boldsymbol{A}^\top \\ &\preceq \mathbf{I}-2\eta(1-\gamma)\boldsymbol{\Sigma}+4\eta^2\|\boldsymbol{\Sigma}\|\boldsymbol{\Sigma} \\ &= \mathbf{I}-\big\{2\eta(1-\gamma)-4\eta^2\|\boldsymbol{\Sigma}\|\big\}\boldsymbol{\Sigma} \\ &\preceq \mathbf{I}-\eta(1-\gamma)\boldsymbol{\Sigma} \\ &\preceq \big(1-\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\big)\mathbf{I},\end{aligned}$$

thus indicating that

$$\big\|\mathbf{I}-\eta\boldsymbol{A}\big\| \le \sqrt{\|\big(\mathbf{I}-\eta\boldsymbol{A}\big)\big(\mathbf{I}-\eta\boldsymbol{A}^\top\big)\|} \le \sqrt{1-\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})} \le 1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}).$$

Proof of Eqn. (79h). For any unit vector $\boldsymbol{u}$, the assumption $\max_s\|\boldsymbol{\phi}(s)\|_2\le 1$ guarantees that

$$\|\boldsymbol{\Phi}\boldsymbol{u}\|_\infty \le \max_s|\boldsymbol{\phi}(s)^\top\boldsymbol{u}| \le \max_s\|\boldsymbol{\phi}(s)\|_2\|\boldsymbol{u}\|_2 \le 1,$$

where in the last inequality we have used Cauchy-Schwartz inequality. Consequently, for any unit vector $\boldsymbol{u}$, by Hölder's inequality,

$$\boldsymbol{u}^\top\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu\boldsymbol{\Phi}\boldsymbol{u} \le \|\boldsymbol{\Phi}\boldsymbol{u}\|_\infty\cdot\mathbf{1}^\top\boldsymbol{D}_\mu\mathbf{1} \le 1,$$

thus proving that $\|\boldsymbol{\Sigma}\|\le 1$. This immediately implies that $\|\boldsymbol{\Sigma}^{-1}\|\ge 1/\|\boldsymbol{\Sigma}\|\ge 1$.

Proof of Eqn. (79i). Finally, we observe that

$$\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}\big\|_2 = \big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{r}\big\|_2 \le \big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Phi}^\top\boldsymbol{D}_\mu^{\frac{1}{2}}\big\|\cdot\big\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{r}\big\|_2 \overset{\text{(i)}}{\le} \big\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{r}\big\|_2 \le 1$$

as claimed. Here, (i) follows from Lemma 4 and (ii) holds true since $\big\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{r}\big\|_2 = \sqrt{\sum_s\mu(s)\big(r(s)\big)^2} \le \sqrt{\sum_s\mu(s)} = 1$. □

The following lemmas, about the concentration of $\widehat{\boldsymbol{A}}$, will be useful in our analysis.

Lemma 6. *Consider any $0<\delta<1$, and suppose that $T\gtrsim\log\big(\frac{d}{\delta}\big)$. Then the vector $\boldsymbol{b}$ defined in (12b) obeys that, with probability exceeding $1-\delta$,*

$$\big\|\boldsymbol{A}^{-1}\big(\widehat{\boldsymbol{b}}-\boldsymbol{b}\big)\big\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{\frac{\max_{s\in\mathcal{S}}\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\Big(\frac{d}{\delta}\Big)}.$$

Proof. See Section C.5. □

Lemma 7. *For any $0<\delta<1$, it follows that $\widehat{\boldsymbol{A}}$ is invertible and that*

$$\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\widehat{\boldsymbol{A}}\big)\boldsymbol{\Sigma}^{-1/2}\big\| \lesssim \sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\Big(\frac{d}{\delta}\Big)}$$

with probability at least $1-\delta$, as long as $T\ge c_2\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log(\frac{d}{\delta})$ for some sufficiently large constant $c_2>0$.

Proof. See Section C.5. □

B Proof of Theorem 2 (minimax lower bounds)

This theorem is proved by constructing a set of MDP instances that are hard to distinguish among each other. Based on this construction, the estimation error can be lower bounded via Fano's inequality, which reduces to control the KL-divergence between marginal likelihood functions. We start by constructing a sequence of hard MDP instances.

Construction of MDP instances and their properties. Given the 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 to

$$P_{\boldsymbol{q}}(s' \,|\, s) = \begin{cases} q_s \mathbb{1}(s' = s) + \frac{1-q_s}{|\mathcal{S}|-d+1}\mathbb{1}(s' \geq d) & \text{for} \quad s < d; \\ \frac{\gamma}{|\mathcal{S}|-d+1}\mathbb{1}(s' \geq d) + \frac{1-q_{s'}}{d-1}\mathbb{1}(s' < d) & \text{for} \quad s \geq d. \end{cases} \tag{82}$$

and the reward function equals to $r(s) = \mathbb{1}(s \geq d)$.

Here, for each $i \in [d-1]$, q_i is taken to be either q_+ or q_- where

$$q_+ \coloneqq \gamma + (1-\gamma)^2\varepsilon, \qquad \text{and} \qquad q_- \coloneqq \gamma - (1-\gamma)^2\varepsilon.$$

We further impose the constraint that the number of q_+'s and q_-'s in $\boldsymbol{q}$ are the same, namely,

$$\sum_{s=1}^{d-1} \mathbb{1}(q_s = q_+) = \sum_{s=1}^{d-1} \mathbb{1}(q_s = q_-) = (d-1)/2. \tag{83}$$

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

$$\mu(s) = \begin{cases} \frac{1}{2(d-1)} & \text{for} \quad s < d; \\ \frac{1}{2(|\mathcal{S}|-d+1)} & \text{for} \quad s \geq d. \end{cases} \tag{84}$$

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

$$\theta^\star(d) = V^\star(s) = \frac{1}{1 - \gamma^2 - \sum_{i=1}^{d-1} \frac{\gamma^2(1-q_i)^2}{(d-1)(1-\gamma q_i)}}, \tag{85}$$

$$\theta^\star(i) = V^\star(i) = \frac{\gamma(1-q_i)}{1-\gamma q_i} V^\star(s), \text{ for } s \geq d \text{ and } i < d. \tag{86}$$

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 (83).

Calculations of several key quantities. Based on the above constructions, let us compute several key quantities. To begin with, some direct algebra leads to

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}^\top \boldsymbol{D}_{\boldsymbol{\mu}} \boldsymbol{\Phi} = \sum_{s=1}^{d-1} \frac{1}{2(d-1)} \boldsymbol{e}_s \boldsymbol{e}_s^\top + \frac{1}{2} \boldsymbol{e}_d \boldsymbol{e}_d^\top,$$

as well as

$$\boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(s) = \begin{cases} 2(d-1) & \text{for} \quad s < d; \\ 2 & \text{for} \quad s \geq d. \end{cases}$$

As a consequence, one has

$$\max_s \{\boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(s)\} \asymp d. \tag{87}$$

Next, we move on to compute $\left\|\boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}$. First notice that for $\varepsilon \le \frac{c_1\gamma}{1-\gamma}$ with constant c_1 small enough, $(1-\gamma)^2\varepsilon \le c_1\gamma(1-\gamma)$ and hence, $1-\gamma q_+, 1-\gamma q_- \asymp 1-\gamma$, which guarantees that $V^\star(s) \asymp \frac{1}{1-\gamma}$. In view of these calculations, it satisfies that

$$\begin{aligned}
\left\|\boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}^2 = \sum_{i=1}^{d-1} \frac{1}{2(d-1)} \theta^{\star 2}(i) + \frac{1}{2}\theta^{\star 2}(d) &= \sum_{i=1}^{d-1} \frac{1}{2(d-1)} \left[\frac{\gamma(1-q_i)}{1-\gamma q_i} V^\star(s)\right]^2 + \frac{1}{2}[V^\star(s)]^2 \\
&\asymp \sum_{i=1}^{d-1} \frac{1}{2(d-1)} \left[\frac{\gamma(1-\gamma)}{1-\gamma}\frac{1}{1-\gamma}\right]^2 + \frac{1}{2}\left[\frac{1}{1-\gamma}\right]^2 \\
&\asymp \frac{1}{(1-\gamma)^2}.
\end{aligned} \tag{88}$$

Application of Fano's inequality. Armed with the properties derived above, we are ready to establish the desired lower bound. First notice that for $\boldsymbol{q}, \boldsymbol{q}' \in \mathcal{Q}$, if at some $i \in [d-1]$, $q_i \neq q_i'$, then

$$\begin{aligned}
|\theta^\star(i) - \theta'^\star(i)| = \gamma V^\star(s) \left|\frac{1-q_i}{1-\gamma q_i} - \frac{1-q_i'}{1-\gamma q_i'}\right| &= \gamma V^\star(s) \frac{2\varepsilon(1-\gamma)^3}{(1-\gamma q_i)(1-\gamma q_i')} \\
&\gtrsim (2\gamma)\frac{1}{1-\gamma}\frac{\varepsilon(1-\gamma)^3}{(1-\gamma)^2} \gtrsim \varepsilon,
\end{aligned}$$

where the penultimate inequality follows from $V^\star(s) \asymp \frac{1}{1-\gamma}$. Consequently, we can bound $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}'^\star\|_{\boldsymbol{\Sigma}}^2$ as

$$\left\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}'^\star\right\|_{\boldsymbol{\Sigma}}^2 \ge \sum_{s=1}^{d-1} |\theta^\star(s) - \theta'^\star(s)|^2 \frac{1}{2(d-1)} \gtrsim \varepsilon^2 \frac{1}{d-1} \sum_{s=1}^{d-1} \mathbb{1}(q_s \neq q_s').$$

This relation guarantees that if $\sum_{s=1}^{d-1} \mathbb{1}(q_s \neq q_s') \ge (d-1)/16$, one has

$$\left\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}'^\star\right\|_{\boldsymbol{\Sigma}} \gtrsim \varepsilon. \tag{89}$$

In other words, if we want each $\boldsymbol{\theta}^\star$ to be ε apart from each other, it is sufficient to construct a set $\mathcal{Q}$ where every $\boldsymbol{q}$ and $\boldsymbol{q}'$ are $(d-1)/16$ apart in Hamming distance. By virtue of the Gilbert-Varshamov lemma (Gilbert, 1952), there exists a set $\mathcal{Q}$ such that

$$M := |\mathcal{Q}| \ge e^{d/16} \qquad \text{and} \qquad \sum_{s=1}^{d-1} \mathbb{1}(q_s \neq q_s') \ge \frac{d}{16} \quad \text{for any } q, q' \in \mathcal{Q} \text{ obeying } q \neq q'. \tag{90}$$

The Fano method transforms the problem of estimating $\boldsymbol{\theta}^\star$ into an M-ary testing problem among the above MDPs $\{\mathbb{P}_{\boldsymbol{q}^1}, \mathbb{P}_{\boldsymbol{q}^2} \ldots, \mathbb{P}_{\boldsymbol{q}^M}\}$. More specifically, in view of Fano's inequality (Tsybakov (2009)), the probability of interest thus satisfies

$$\mathbb{P}\Big(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}} \gtrsim \varepsilon\Big) \ge 1 - \frac{1}{\log M}\Big(\frac{1}{M^2}\sum_{j,k=1}^{M} \mathsf{KL}(\mathbb{P}_{\boldsymbol{q}^j}^T \parallel \mathbb{P}_{\boldsymbol{q}^k}^T) + \log 2\Big), \tag{91}$$

given T independent sample pairs $\{(s_t, s_t')\}_{t=1}^T$. To control the right hand side, we proceed by computing the KL-divergence between every $\mathbb{P}_{\boldsymbol{q}}$ and $\mathbb{P}_{\boldsymbol{q}'}$. Here $\mathbb{P}_{\boldsymbol{q}}$ denotes the joint distribution of (s, s') when the transition is made according to $P_{\boldsymbol{q}}(s' \,|\, s)$ (cf. (82)). More specifically, given $s \sim \mu_q$ and $s'|s \sim P_{\boldsymbol{q}}(s'|s)$, one has

$$\mathbb{P}_{\boldsymbol{q}}(s, s') = \mu(s)P(s'|s) = \begin{cases} \frac{1}{2(d-1)} q_s \mathbb{1}(s' = s), & \text{for} \quad s < d, s' < d; \\ \frac{1-q_s}{2(d-1)(S-d+1)}, & \text{for} \quad s < d, s' > d; \\ \frac{1-q_{s'}}{2(d-1)(S-d+1)}, & \text{for} \quad s > d, s' < d; \\ \frac{\gamma}{2(S-d+1)^2}, & \text{for} \quad s > d, s' > d. \end{cases}$$

Recognizing the relation between the KL divergence and the χ^2 divergence, $\mathsf{KL}(\mathbb{P}_{\boldsymbol{q}} \,\|\, \mathbb{P}_{\boldsymbol{q}'})$ satisfies

$$\begin{aligned}
\mathsf{KL}(\mathbb{P}_{\boldsymbol{q}} \,\|\, \mathbb{P}_{\boldsymbol{q}'}) &\le \chi^2(\mathbb{P}_{\boldsymbol{q}'} \,\|\, \mathbb{P}_{\boldsymbol{q}}) \\
&= \sum_{s,s'} \frac{(\mathbb{P}_{\boldsymbol{q}}(s,s') - \mathbb{P}_{\boldsymbol{q}'}(s,s'))^2}{\mathbb{P}_{\boldsymbol{q}}(s,s')} \\
&= \left(\sum_{s<d,s'<d} + \sum_{s<d,s'\ge d} + \sum_{s\ge d,s'<d} + \sum_{s\ge d,s'\ge d} \right) \frac{(\mathbb{P}_{\boldsymbol{q}}(s,s') - \mathbb{P}_{\boldsymbol{q}'}(s,s'))^2}{\mathbb{P}_{\boldsymbol{q}}(s,s')} \\
&= \sum_{s=1}^{d-1} \frac{1}{2(d-1)} \frac{(q_s - q'_s)^2}{q_s} + \sum_{s<d,s'\ge d} \frac{1}{2(d-1)(S-d+1)} \frac{[(1-q_s) - (1-q'_s)^2]}{1-q_s} \\
&+ \sum_{s\ge d,s'<d} \frac{1}{2(d-1)(S-d+1)} \frac{[(1-q_{s'}) - (1-q'_{s'})^2]}{1-q_{s'}} + 0 \\
&\lesssim \sum_{s=1}^{d-1} \frac{1}{2(d-1)} \frac{[2\varepsilon(1-\gamma)^2]^2}{1-\gamma} + \sum_{s<d} \frac{1}{2(d-1)} \frac{[2\varepsilon(1-\gamma)^2]^2}{1-\gamma} + \sum_{s'<d} \frac{1}{2(d-1)} \frac{[2\varepsilon(1-\gamma)^2]^2}{1-\gamma} \\
&\asymp \varepsilon^2(1-\gamma)^3.
\end{aligned}$$

As a result, we have

$$\mathsf{KL}\big(\mathbb{P}_{\boldsymbol{q}}^T \,\|\, \mathbb{P}_{\boldsymbol{q}'}^T\big) \lesssim \varepsilon^2(1-\gamma)^3 T. \tag{92}$$

Substituting the above relation into (91) gives

$$\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big\|_{\boldsymbol{\Sigma}} \gtrsim \varepsilon\Big) \ge 1 - \frac{1}{d/16}\Big(c\varepsilon^2(1-\gamma)^3 T + \log 2\Big).$$

To prove Theorem 2, it is enough to take the above together with relations (87) and (88).

C Proofs of auxiliary lemmas and claims

C.1 Proof of Lemma 1

Here and throughout, we denote by $\mathbb{E}_i[\cdot]$ the expectation conditioned on the probability space generated by the samples $\{(s_j, s'_j)\}_{j\le i}$ (more formally, $\mathbb{E}_i[\cdot]$ represents the expectation conditioned on the filtration $\mathcal{F}_i$ — the σ-algebra generated by $\{(s_j, s'_j)\}_{j\le i}$). It is then easy to check that $\{(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\widetilde{\boldsymbol{\xi}}_i\}$ forms a martingale difference sequence, which motivates us to apply matrix Freedman's inequality.

To this end, one needs to upper bound the following two quantities

$$W \coloneqq \sum_{i=l}^{u} \mathbb{E}_{i-1}\Big[\big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\big\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\Big], \quad \text{and} \quad B \coloneqq \max_{i:l\le i\le u} \big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i \mathbb{1}\{\mathcal{H}_i\}\big\|_2, \tag{93}$$

which we accomplish in the sequel. For notational convenience, we set

$$\alpha \coloneqq \Big(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\Big)^{t-u-1}. \tag{94}$$

Control of W. Direct calculations yield

$$\begin{aligned}
W &= \sum_{i=l}^{u} \mathbb{E}_{i-1}\Big[\boldsymbol{\xi}_i^\top(\mathbf{I}-\eta\boldsymbol{A}^\top)^{t-i-1}\boldsymbol{\Sigma}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i \mathbb{1}\{\mathcal{H}_i\}\Big] \\
&\le \sum_{i=l}^{u} \big\|(\mathbf{I}-\eta\boldsymbol{A}^\top)^{t-i-1}\boldsymbol{\Sigma}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\big\| \cdot \mathbb{E}_{i-1}\big[\|\boldsymbol{\xi}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big]
\end{aligned}$$

$$\overset{\text{(i)}}{\leq} \sum_{i=l}^{u} \|\boldsymbol{\Sigma}\| \left(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})\right)^{2t-2i-2} 2 \max_{i:l\leq i\leq u} \Big\{\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] + \mathbb{E}_{i-1}\big[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2\big]\Big\}$$
$$\overset{\text{(ii)}}{\leq} \frac{4\|\boldsymbol{\Sigma}\|\alpha^2}{\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})} \max_{i:l\leq i\leq u} \Big\{\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] + \mathbb{E}_{i-1}\big[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2\big]\Big\}, \tag{95}$$

where (i) follows from the property (79g) (together with the assumption $\eta < (1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$) and the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$, and (ii) uses the elementary upper bound for the sum of geometric series as well as the definition (94) of α.

We then turn attention to $\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big]$ and $\mathbb{E}_{i-1}\big[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2\big]$. First, given that $\mathbb{E}_{i-1}[\boldsymbol{A}_i\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}] = \boldsymbol{A}\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}$, one can derive

$$\begin{aligned}
&\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] \leq \mathbb{E}_{i-1}\big[\|\boldsymbol{A}_i\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] \\
&\quad = \mathbb{E}_{i-1}\Big[\boldsymbol{\theta}_i^\top\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)\boldsymbol{\phi}(s_i)^\top\boldsymbol{\phi}(s_i)\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\Big] \\
&\quad \leq \max_s \|\boldsymbol{\phi}(s)\|_2^2 \cdot \mathbb{E}_{i-1}\Big[\boldsymbol{\theta}_i^\top\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\Big] \\
&\quad \overset{\text{(i)}}{\leq} 2\max_s \|\boldsymbol{\phi}(s)\|_2^2\Big(\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\phi}(s_i)\boldsymbol{\phi}(s_i)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big] + \gamma^2\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\phi}(s_i')\boldsymbol{\phi}(s_i')^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\Big) \\
&\quad \overset{\text{(ii)}}{=} 2\max_s \|\boldsymbol{\phi}(s)\|_2^2\Big(\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big] + \gamma^2\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\Big) \\
&\quad \overset{\text{(iii)}}{\leq} 4\|\boldsymbol{\theta}_i\|_{\boldsymbol{\Sigma}}^2\mathbb{1}\{\mathcal{H}_i\} \leq 4\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big)^2\mathbb{1}\{\mathcal{H}_i\} \\
&\quad \leq 4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R)^2,
\end{aligned} \tag{96}$$

where (i) relies on the elementary inequality $(\boldsymbol{a}+\boldsymbol{b})(\boldsymbol{a}+\boldsymbol{b})^\top \preceq 2\boldsymbol{a}\boldsymbol{a}^\top + 2\boldsymbol{b}\boldsymbol{b}^\top$, (ii) follows from the definition (6) of $\boldsymbol{\Sigma}$ and the fact that $s_i, s_i' \sim \mu$ in this case, (iii) holds due to the assumption $\max_s \|\boldsymbol{\phi}(s)\|_2 \leq 1$, and the last inequality results from the definition (47) of the event $\mathcal{H}_i$. Similarly, one can derive

$$\mathbb{E}_{i-1}[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2] \leq \mathbb{E}_{i-1}[\|\boldsymbol{b}_i\|_2^2] = \mathbb{E}_{i-1}\big[\|\boldsymbol{\phi}(s_i)r(s_i)\|_2^2\big] \leq 1, \tag{97}$$

where the last inequality holds since $\max_s \|\boldsymbol{\phi}(s)\|_2 \leq 1$ and $\max_s |r(s)| \leq 1$. Substitution into (95) yields

$$W \leq \frac{4\kappa}{\eta(1-\gamma)}\alpha^2\Big\{4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + R)^2 + 1\Big\} =: W_{\max}. \tag{98}$$

Control of B. By definition of B, one can write

$$\begin{aligned}
B &= \max_{i:l\leq i\leq u} \big\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\} = \max_{i:l\leq i\leq u} \big\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\} \\
&\leq \|\boldsymbol{\Sigma}\| \max_{i:l\leq i\leq u} \big\|\mathbf{I}-\eta\boldsymbol{A}\big\|^{t-i-1} \cdot \max_{i:l\leq i\leq u} \big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\} \\
&\leq \alpha\|\boldsymbol{\Sigma}\| \max_{i:l\leq i\leq u} \Big\{\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\} + \big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{b}_i-\boldsymbol{b})\big\|_2\Big\},
\end{aligned} \tag{99}$$

where the last step results from (79g) (with the restriction that $\eta < (1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$) and the definition (94) of α. It then suffices to control the two terms on the right-hand side of (99). To begin with, we have

$$\begin{aligned}
\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2 &\leq \big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|\|\boldsymbol{\theta}_i\|_{\boldsymbol{\Sigma}} \\
&\leq \Big(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_i\boldsymbol{\Sigma}^{-\frac{1}{2}}\| + \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|\Big)\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big).
\end{aligned}$$

Recall from (135) that $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_i\boldsymbol{\Sigma}^{-\frac{1}{2}}\| \leq 2\max_s \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(s)\|_2^2$, and similarly $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\| \leq 2\max_s \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(s)\|_2^2$. We then have

$$\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2 \leq 4\max_s \big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big). \tag{100}$$

Regarding the second term of (99), direct calculations give

$$\begin{aligned}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{b}_i-\boldsymbol{b})\|_2^2 &\le 2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}_i\|_2^2+2\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}\|_2^2 = 2\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s_i)r(s_i)\big\|_2^2+2\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}_{s\sim\mu}\big[\boldsymbol{\phi}(s)r(s)\big]\big\|_2^2\\ &\le 4\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\max_s|r(s)|^2\le 4\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}. \end{aligned}\tag{101}$$

Substituting the preceding two bounds into (99), we arrive at

$$\begin{aligned}B &\le 4\alpha\|\boldsymbol{\Sigma}\|\Big(\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\max_{i:i<t}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big)\mathbb{1}\{\mathcal{H}_i\}+\sqrt{\max_s\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}}\Big)\\ &\le 4\alpha\|\boldsymbol{\Sigma}\|\Big(\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)+\sqrt{\max_s\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}}\Big)\\ &\le 4\alpha\|\boldsymbol{\Sigma}\|\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)\\ &\le 4\alpha\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)=4\kappa\alpha\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)=:B_{\max}. \end{aligned}\tag{102}$$

Here, the last line follows from the assumption $\max\|\boldsymbol{\phi}(s)\|_2\le 1$, while the second to last inequality holds since $\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\ge 1$ (cf. (137)).

Invoking matrix Freedman's inequality. Equipped with the above bounds (98) and (102), we are ready to apply Freedman's inequality (Tropp, 2011, Corollary 1.3) (or a version in (Li et al., 2023a, Section A)), which asserts that

$$\begin{aligned}\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\widetilde{\boldsymbol{\xi}}_i\Big\|_{\boldsymbol{\Sigma}} &\le 2\sqrt{W_{\max}\log\frac{2dT}{\delta}}+\frac{4}{3}B_{\max}\log\frac{2dT}{\delta}\\ &= \alpha\cdot\left\{2\sqrt{\frac{4\kappa}{\eta(1-\gamma)}\Big\{4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R)^2+1\Big\}\log\frac{2dT}{\delta}}+\frac{16\kappa}{3}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)\log\frac{2dT}{\delta}\right\}\\ &\le 16(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}))^{t-u-1}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \end{aligned}\tag{103}$$

holds with probability at least $1-\delta/T$, provided that $0<\eta\le\frac{1}{\kappa(1-\gamma)\log\frac{2dT}{\delta}}$. Here in the last line, we identify α with $(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}))^{t-u-1}$. The proof is completed by observing that any $0<\eta\le\frac{1-\gamma}{\kappa\log\frac{2dT}{\delta}}$ satisfies the two requirements $0<\eta\le\frac{1}{\kappa(1-\gamma)\log\frac{2dT}{\delta}}$ and $\eta<(1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$ (given that $\|\boldsymbol{\Sigma}\|\le 1$ according to (79h)).

C.2 Proof of the inequalities (58a) and (58b)

Proof of the inequality (58a). Combining the triangle inequality with the definition (57) ensures that

$$\begin{aligned}\big\|\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} &\le \|\boldsymbol{A}^{-1}\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T+1}\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}\\ &= \big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Delta}_0\big\|_2+\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{I}-\eta\boldsymbol{A})^{T+1}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Delta}_0\|_2\\ &\le \big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-1}\big\|\cdot\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}+\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|^2\cdot\|\mathbf{I}-\eta\boldsymbol{A}\|^{T+1}\cdot\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\ &\le \frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\left\{1+\left(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\mathsf{min}}(\boldsymbol{\Sigma})\right)^{T+1}\right\}\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\ &\le \frac{2\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} \end{aligned}\tag{104}$$

as claimed. Here, the second to last step follows from (79e) and (79g), provided that $\eta\le(1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$.

Proof of the inequality (58b). Again, the triangle inequality together with the definition (57) yields

$$\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\le\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}+\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}$$

$$\leq \Big\|\boldsymbol{A}^{-1}\sum_{i=0}^{T-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_{\boldsymbol{\Sigma}}+\Big\|\boldsymbol{A}^{-1}\sum_{i=0}^{T-1}(\boldsymbol{b}_i-\boldsymbol{b})\Big\|_{\boldsymbol{\Sigma}}+\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}, \tag{105}$$

leaving us with three terms to handle. Here in the second line, we substitute in the definition of $\boldsymbol{\xi}_i$ (42).

- The second term of (105) can be bounded by Lemma 6, which asserts that

$$\frac{1}{T}\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}\big(\boldsymbol{b}_i-\boldsymbol{b}\big)\Big\|_{\boldsymbol{\Sigma}}\lesssim\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\Big(\frac{d}{\delta}\Big)} \tag{106}$$

  holds with probability at least $1-\delta$, as long as $T\gtrsim\log\frac{d}{\delta}$.

- For the third term of (105), invoking the property (79e) again yields

$$\begin{aligned}\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}&=\Big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Sigma}^{-1}\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_2\\&\leq\big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-1}\big\|\cdot\Big\|\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_2\\&\leq\frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\Big\|\sum_{i=0}^{T-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}.\end{aligned} \tag{107}$$

  Repeating the same analysis as in Step 3 to see that

$$\Big\|\sum_{i=0}^{T-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\leq 16(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+3)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \tag{108}$$

  with probability at least $1-\delta$. Substitution into (107) leads to

$$\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\leq 16\big(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+3\big)\big\|\boldsymbol{\Sigma}^{-1}\big\|\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}. \tag{109}$$

- It then boils down to bounding the first term of (105). In light of (54), we decompose it as follows

$$\Big\|\frac1T\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2\leq\Big\|\frac1T\sum_{i=0}^{\widetilde{t}_{\mathsf{seg}}-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2+\Big\|\frac1T\sum_{i=\widetilde{t}_{\mathsf{seg}}}^{T-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2. \tag{110}$$

  Bounding these terms requires the following lemma, whose proof is deferred to Section C.6.

  Lemma 8. *Fix any $R>0$ and define a collection of auxiliary random vectors for $0\leq i\leq T-1$*

$$\boldsymbol{\theta}_i'\coloneqq\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\},\qquad\mathcal{H}_i\coloneqq\big\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\leq R\big\}, \tag{111}$$

  Then for any indices (l,u,t) obeying $0\leq l\leq u\leq T-1$, one has with probability at least $1-\delta$ that

$$\Big\|\frac{1}{u-l+1}\sum_{i=l}^{u}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\Big\|_2\leq\frac{16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{u-l+1}} \tag{112}$$

  provided that

$$u-l+1\geq\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}.$$

Apply Lemma 8 with $R = R_0, l = 0$ and $u = t'_{\mathsf{seg}} - 1$ to obtain with probability of at least $1-\delta$ that

$$\Big\|\frac{1}{t'_{\mathsf{seg}}}\sum_{i=0}^{t'_{\mathsf{seg}}-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2 = \Big\|\frac{1}{t'_{\mathsf{seg}}}\sum_{i=0}^{t'_{\mathsf{seg}}-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\mathbb{1}\{\|\boldsymbol{\Delta}_i\|\le R_0\}\Big\|_2$$
$$\le \frac{16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R_0\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{t'_{\mathsf{seg}}}},$$

as long as $t'_{\mathsf{seg}} \ge \frac{4\|\boldsymbol{\Sigma}^{-1}\|\log\frac{2d}{\delta}}{9} \ge \frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}$. Here, the identity holds since $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \le R_0$ for $i \le t'_{\mathsf{seg}}-1$ with probability of at least $1-\delta$. Similarly, invoke Lemma 8 with $R = 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2), l = t'_{\mathsf{seg}}$ and $u = T-1$ to obtain with probability of at least $1-\delta$ that

$$\Big\|\frac{1}{T-t'_{\mathsf{seg}}}\sum_{i=t'_{\mathsf{seg}}}^{T-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2 \le \frac{16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2)\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T-t'_{\mathsf{seg}}}}$$
$$\le \frac{16\big(1.5\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T-t'_{\mathsf{seg}}}}$$

provided that $T - t'_{\mathsf{seg}} \ge \frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}$. Here, the last inequality arises from the relation

$$32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2) \le 0.5\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2,$$

which is an immediate consequence of the assumption that $\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}$ is sufficiently small. Therefore,

$$\Big\|\frac{1}{T}\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2 \le \frac{32\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\sqrt{\frac{t'_{\mathsf{seg}}}{T}}R_0+1\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T}}. \tag{113}$$

Combining the preceding bounds (106), (109) and (113) with (105), we reach the conclusion that with probability of at least $1-\delta$,

$$\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(t)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \asymp \left\{\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T(1-\gamma)^2}}+\frac{\|\boldsymbol{\Sigma}^{-1}\|}{T}\sqrt{\frac{\kappa\log\frac{dT}{\delta}}{\eta(1-\gamma)^3}}\right\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\big),$$

as long as $T \ge t'_{\mathsf{seg}}\kappa\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2$, where we use the definition (44) of R_0. It thus establishes the inequality (58b).

C.3 Proof of Lemma 2

We first decompose $\boldsymbol{\Psi}$ into

$$\boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{I}-\alpha\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}-\beta\widetilde{\boldsymbol{\Sigma}}\end{bmatrix}+\begin{bmatrix}\boldsymbol{0} & -\frac{1}{\kappa}\alpha\gamma\boldsymbol{\Pi}^\top\\ -\kappa\alpha(1-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & -\alpha\gamma(\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\Pi}^\top\end{bmatrix}.$$

Then the triangle inequality together with the properties of the operator norm tells us that

$$\|\boldsymbol{\Psi}\| \le \max\{\|\boldsymbol{I}-\alpha\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}\|,\|\boldsymbol{I}-\beta\widetilde{\boldsymbol{\Sigma}}\|\}+\|\frac{1}{\kappa}\alpha\gamma\boldsymbol{\Pi}^\top\|$$
$$+\|\kappa\alpha(\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}\|+\|\alpha\gamma(\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\Pi}^\top\|.$$

Note that by definition of λ_w and λ_w, we find

$$\|\boldsymbol{I}-\alpha\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}\| \le 1-\alpha\lambda_\theta,$$

$$\|\boldsymbol{I} - \beta\widetilde{\boldsymbol{\Sigma}}\| \leq 1 - \beta\lambda_w.$$

In addition, some direct algebra suggests

$$\begin{aligned}
\left\|\frac{1}{\kappa}\alpha\gamma\boldsymbol{\Pi}^\top\right\| &\leq \frac{\alpha\gamma\rho_{\max}}{\kappa}, \\
\|\kappa\alpha(\boldsymbol{I} - \gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}\| &\leq \kappa\alpha(1+\gamma\lambda_\Sigma\rho_{\max})\lambda_\Sigma(2\rho_{\max})^2, \\
\|\alpha\gamma(\boldsymbol{I} - \gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\Pi}^\top\| &\leq \alpha\gamma(\rho_{\max} + \gamma\lambda_\Sigma\rho_{\max}^2).
\end{aligned}$$

In summary, as long as

$$\begin{aligned}
\alpha\gamma(\rho_{\max} + \gamma\lambda_\Sigma\rho_{\max}^2) &\ll \beta\lambda_w, \\
\frac{\alpha\gamma\rho_{\max}}{\kappa} + \kappa\alpha(1+\gamma\lambda_\Sigma\rho_{\max})\lambda_\Sigma(2\rho_{\max})^2 &\ll \sqrt{\alpha\lambda_\theta\beta\lambda_w},
\end{aligned}$$

one has

$$\|\boldsymbol{\Psi}\| \leq 1 - \frac{1}{2}\alpha\lambda_\theta.$$

C.4 Proof of Lemma 3

Using the same notation of $\mathbb{E}_{i-1}$ as in Section C.1, we observe that $\{\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\}$ forms a martingale difference sequence. Furthermore, define

$$\widetilde{W} := \sum_{i=0}^{t-1}\mathbb{E}_{i-1}\left[\left\|\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \mathbb{1}\left\{\widetilde{\mathcal{H}}_i\right\}\right], \quad \text{and} \quad \widetilde{B} := \max_{i:0\leq i\leq t-1}\left\|\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i\mathbb{1}\left\{\widetilde{\mathcal{H}}_i\right\}\right\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{114}$$

In order to bound $\widetilde{W}$ and $\widetilde{B}$, we will firstly need to bound the norm of $\widetilde{\boldsymbol{\zeta}}_i$, as is shown in the following paragraph.

Controlling the norm of $\boldsymbol{\zeta}_i$. We firstly observe that since $\|\boldsymbol{\phi}(s)\|_2 \leq 1$ and $r(s) \leq 1$ for all $s \in \mathcal{S}$, with similar logic as (96), (97), (100) and (101), the following bounds hold true:

- For any $\mathcal{F}_{i-1}$-measurable $\widetilde{\boldsymbol{\theta}}_i \in \mathbb{R}^d$, the norm of $(\widetilde{\boldsymbol{A}}_i - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_i$ is bounded by

$$\mathbb{E}_{i-1}\left\|(\widetilde{\boldsymbol{A}}_i - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_i\right\|_2^2 \leq 4\rho_{\max}^2\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right), \quad \text{and} \tag{115}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}(\widetilde{\boldsymbol{A}}_i - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_i\right\|_2 \leq 4\rho_{\max}\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right); \tag{116}$$

- For any $\mathcal{F}_{i-1}$-measurable $\boldsymbol{z}_i \in \mathbb{R}^d$, the norm of $(\boldsymbol{\Pi}_i - \boldsymbol{\Pi})^\top\boldsymbol{z}_i$ is bounded by

$$\mathbb{E}_{i-1}\left\|(\boldsymbol{\Pi}_i - \boldsymbol{\Pi})^\top\boldsymbol{z}_i\right\|_2^2 \leq \rho_{\max}^2\|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \quad \text{and} \tag{117}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{\Pi}_i - \boldsymbol{\Pi})^\top\boldsymbol{z}_i\right\|_2 \leq 2\rho_{\max}\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}; \tag{118}$$

- For any $\mathcal{F}_{i-1}$-measurable $\boldsymbol{z}_i \in \mathbb{R}^d$, the norm of $(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{z}_i$ is bounded by

$$\mathbb{E}_{i-1}\left\|\left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{\Sigma}}\right)\boldsymbol{z}_i\right\|_2^2 \leq \|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \quad \text{and} \tag{119}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{\Sigma}}\right)\boldsymbol{z}_i\right\|_2 \leq 2\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}; \tag{120}$$

- The norm of $\widetilde{\boldsymbol{b}}_i - \widetilde{\boldsymbol{b}}$ is bounded by

$$\mathbb{E}_{i-1}\left\|\widetilde{\boldsymbol{b}}_i - \widetilde{\boldsymbol{b}}\right\|_2^2 \leq \rho_{\max}^2, \quad \text{and} \tag{121}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\left(\widetilde{\boldsymbol{b}}_i - \widetilde{\boldsymbol{b}}\right)\right\|_2^2 \leq 4\rho_{\max}^2\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}. \tag{122}$$

Therefore, by triangle inequality, the norm of $\boldsymbol{\nu_i}$ can be bounded by

$$\mathbb{E}_{i-1}\left\|\boldsymbol{\nu}_i\right\|_2^2 \lesssim \rho_{\max}^2\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right)+1+\gamma^2\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right], \quad \text{and} \tag{123}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\nu}_i\right\|_2 \lesssim \rho_{\max}\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right)+\gamma\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}+1\right]; \tag{124}$$

similarly, the norm of $\boldsymbol{\eta}_i$ can be bounded by

$$\mathbb{E}_{i-1}\left\|\boldsymbol{\eta}_i\right\|_2^2 \lesssim \rho_{\max}^2\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right)+1\right]+\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \quad \text{and} \tag{125}$$

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\eta_i}\right\|_2 \lesssim \max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\left\{\rho_{\max}\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right)+1\right]+\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right\}. \tag{126}$$

By combining (123) and (125) with the definition of $\boldsymbol{\zeta}_i$ (65), we obtain the following bound:

$$\begin{aligned}
\mathbb{E}_{i-1}\|\boldsymbol{\zeta}_i\|_2^2 &\lesssim \alpha^2\mathbb{E}_{i-1}\|\boldsymbol{\nu}_i\|_2^2+\varkappa^2\alpha^2\|\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi}\|^2\mathbb{E}_{i-1}\|\boldsymbol{\nu}_i\|_2^2+\varkappa^2\beta^2\mathbb{E}_{i-1}\|\boldsymbol{\eta}_i\|_2^2 \\
&\lesssim \alpha^2\left(1+\varkappa^2(1+\gamma\lambda_\Sigma\rho_{\max})^2\right)\cdot\rho_{\max}^2\left[4\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right)+1+\gamma^2\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right] \\
&\quad+\varkappa^2\beta^2\cdot\left\{\rho_{\max}^2\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right)+1\right]+\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\right\} \\
&\lesssim \varkappa^2\beta^2\rho_{\max}^2\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2+\frac{1}{\varkappa^2}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2+1\right),
\end{aligned} \tag{127}$$

and

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\zeta}_i\right\|_2 &\lesssim \alpha\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\nu}_i\|_2+\alpha\varkappa\|\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi}\|\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\nu}_i\|_2+\varkappa\beta\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\eta_i\right\|_2 \\
&\lesssim \alpha\left(1+\varkappa(1+\gamma\lambda_\Sigma\rho_{\max})\right)\cdot\rho_{\max}\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\cdot \\
&\quad\left[2\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right)+\gamma\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}+1\right]+ \\
&\quad\varkappa\beta\cdot\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\left\{\rho_{\max}\left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right)+1\right]+\|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\right\} \\
&\lesssim \varkappa\beta\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\frac{2}{\varkappa}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}}+1\right)
\end{aligned} \tag{128}$$

Control of $\widetilde{W}$ and $\widetilde{B}$. With the norm of $\widetilde{\boldsymbol{\zeta}}_i$ bounded, we can apply similar techniques as in equations (95), (98), (99) and (102) of Section C.1 to construct the following bound for $\widetilde{W}$:

$$\begin{aligned}
\widetilde{W} &\le \|\widetilde{\boldsymbol{\Sigma}}\|\sum_{i=0}^{t-1}\|\boldsymbol{\Psi}^{t-i-1}\|^2\cdot\mathbb{E}_{i-1}\left[\|\boldsymbol{\zeta}_i\|_2^2\mathbb{1}\left\{\widetilde{\mathcal{H}}_i\right\}\right] \\
&\lesssim \|\widetilde{\boldsymbol{\Sigma}}\|\sum_{i=0}^{t-1}(1-\frac{1}{2}\alpha\lambda_\theta)^{2t-2i-2}\varkappa^2\beta^2\rho_{\max}^2(2\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+2\widetilde{R}+1)^2 \\
&\lesssim \frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_\theta}\varkappa^2\beta^2\rho_{\max}^2(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\frac{1}{\varkappa}\widetilde{R}+1)^2,
\end{aligned} \tag{129}$$

and the following bound for $\widetilde{B}$:

$$\begin{aligned}
\widetilde{B} &\le \|\widetilde{\boldsymbol{\Sigma}}\|\max_{i:0\le i\le t-1}\left\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\zeta}_i\mathbb{1}\left\{\mathcal{H}_i\right\}\right\|_2 \\
&\lesssim \|\widetilde{\boldsymbol{\Sigma}}\|\varkappa\beta\rho_{\max}\max_s\left\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\right\}(\|\widetilde{\boldsymbol{\theta}}^\star\|_2+\widetilde{R}+1)=:\widetilde{B}_{\max}.
\end{aligned} \tag{130}$$

Invoking the matrix Freedman's inequality. With $\widetilde{W}$ and $\widetilde{B}$ bounded, we again invoke the matrix Freedman's inequality (Tropp, 2011, Corollary 1.3) to assert that

$$\left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\right\|_2 \le 2\sqrt{\widetilde{W}_{\max}\log\frac{2dT}{\delta}}+\frac{4}{3}\widetilde{B}_{\max}\log\frac{2dT}{\delta}$$

$$\lesssim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_\theta}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+\frac{1}{\varkappa}\widetilde{R}+1) \tag{131}$$

holds with probability at least $1-\delta/T$, proveded that $0<\alpha<\frac{1}{\lambda_\theta\lambda_\Sigma^2\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}$.

C.5 Proof of Lemma 6 and Lemma 7

Proof of Lemma 7: controlling $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\|$. We intend to invoke the matrix Bernstein inequality to establish the advertised bound Tropp (2015). Note that

$$\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}=\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\boldsymbol{A}_t)\boldsymbol{\Sigma}^{-\frac{1}{2}}}_{=:\boldsymbol{Z}_t}. \tag{132}$$

In order to control it, we need to first control the following two quantities:

$$v:=\max_t\left\{\max\left\{\left\|\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]\right\|,\left\|\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]\right\|\right\}\right\}\qquad\text{and}\qquad B:=\max_t\|\boldsymbol{Z}_t\|.$$

Step 1: controlling $\left\|\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]\right\|$. Towards this, we first make the observation that

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]&=\mathbb{E}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\boldsymbol{A}_t)\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}-\boldsymbol{A}_t)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&\preceq\mathbb{E}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_t^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&=\mathop{\mathbb{E}}_{s\sim\mu,\,s'\sim P(\cdot|s)}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)^\top\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&\preceq\max_{s,s'}\left\{\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)^\top\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)\right\}\mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&\preceq\max_{s,s'}\left\{2\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)+2\gamma^2\boldsymbol{\phi}(s')^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s')\right\}\cdot\left\{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\}\\
&\preceq\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\mathbf{I},
\end{aligned}\tag{133}$$

where the second line holds since $\mathbb{E}\big[(\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}])(\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}])^\top\big]\preceq\mathbb{E}[\boldsymbol{M}\boldsymbol{M}^\top]$ for any random matrix $\boldsymbol{M}$, the second to last inequality holds since $(\boldsymbol{a}-\boldsymbol{b})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\boldsymbol{b})\le 2\boldsymbol{a}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{a}+2\boldsymbol{b}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{b}$, and the last inequality comes from the assumption $\gamma<1$ and Lemma 5.

Step 2: controlling $\left\|\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]\right\|$. Similarly, one can obtain

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]&=\mathbb{E}\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{A}-\boldsymbol{A}_t)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}(\boldsymbol{A}-\boldsymbol{A}_t)\boldsymbol{\Sigma}^{-\frac{1}{2}}\right]\\
&\preceq\mathbb{E}\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\right]\\
&=\mathbb{E}\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}\big(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')\big)\boldsymbol{\phi}(s_t)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s_t)\big(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')\big)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right]\\
&\preceq\max_s\left\{\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\right\}\mathbb{E}\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}\big(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')\big)\big(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_t')\big)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right]\\
&\preceq\max_s\left\{\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\right\}\cdot 2\mathbb{E}\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}\big(\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top+\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t')^\top\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}\right]\\
&\preceq 4\max_s\left\{\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\right\}\mathbf{I}.
\end{aligned}$$

Here, the second to last bound follows from the elementary inequality $(\boldsymbol{a}-\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})^\top\preceq 2\boldsymbol{a}\boldsymbol{a}^\top+2\boldsymbol{b}\boldsymbol{b}^\top$ and the assumption $\gamma<1$, whereas the last line makes use of the facts $s_t\sim\mu$, $s_t'\sim\mu$ and the definition (6) of $\boldsymbol{\Sigma}$. It then boils down to upper bounding $\max_s\left\{\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\right\}$, which can be accomplished as follows

$$\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)=\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\left\{\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)$$

$$\leq \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2^2 \cdot \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|$$
$$\leq \frac{\max_s \boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}.$$

Here, the last line arises from Lemma 5. Putting the above bounds together yields

$$\mathbb{E}\left[\boldsymbol{Z}_t^\top \boldsymbol{Z}_t\right] \preceq \frac{4\max_s \boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\mathbf{I}. \tag{134}$$

Step 3: controlling $\|\boldsymbol{Z}_t\|$. Our starting point is the following triangle inequality

$$\begin{aligned}\|\boldsymbol{Z}_t\| &= \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\boldsymbol{A}_t)\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| \leq \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| + \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| \\ &\leq \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \cdot \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| + 1 \\ &\leq \frac{1}{1-\gamma}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| + 1,\end{aligned}$$

where the last inequality follows from Lemma 5. In addition, we see that

$$\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| \leq \max_s \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| + \gamma \max_{s,s'}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s')\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| \leq 2\max_s \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2^2. \tag{135}$$

This combined with the preceding bounds yields

$$\|\boldsymbol{Z}_t\| \leq \frac{2\max_s \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2^2}{1-\gamma} + 1 \leq \frac{4\max_s \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2^2}{1-\gamma} = \frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{1-\gamma}. \tag{136}$$

Here, the inequality follows since

$$\max_s \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2^2 \geq \mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\right] = \mathop{\mathbb{E}}_{s\sim\mu}\left[\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\big)\right] = \mathrm{tr}(\mathbf{I}_d) = d \geq 1. \tag{137}$$

Step 4: invoking the matrix Bernstein inequality. With the above bounds in mind, we are ready to apply the matrix Bernstein inequality (Tropp, 2015) to obtain that: with probability at least $1-\delta$ one has

$$\begin{aligned}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| &\lesssim \sqrt{\frac{1}{T^2}\sum_{t=0}^{T-1}\max\left\{\left\|\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]\right\|, \left\|\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]\right\|\right\}\log\left(\frac{d}{\delta}\right)} + \frac{\max_t\|\boldsymbol{Z}_t\|\log(\frac{d}{\delta})}{T}. \\ &\overset{\text{(i)}}{\lesssim} \sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\left(\frac{d}{\delta}\right)} + \frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log(\frac{d}{\delta})}{T(1-\gamma)} \\ &\overset{\text{(ii)}}{\asymp} \sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\left(\frac{d}{\delta}\right)}.\end{aligned} \tag{138}$$

Here, (i) results from the bounds (133), (134) and (136), while (ii) holds as long as $T \gtrsim \max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\left(\frac{d}{\delta}\right)$. In addition, if $T \geq \frac{c_2\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\left(\frac{d}{\delta}\right)}{(1-\gamma)^2}$ for some constant c_2 large enough, then one has $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| < 1$. Suppose that $\widehat{\boldsymbol{A}}$ is not invertible. Given that $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ are both invertible, this means that one can find a unit vectors $\boldsymbol{u}$ obeying $\boldsymbol{A}^{-1}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{u} = \mathbf{0}$, which in turn implies

$$\begin{aligned}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{u} &= \boldsymbol{u}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{u} - \boldsymbol{u}^\top\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{u} \\ &= 1-0=1\end{aligned}$$

and hence contradicts the condition $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| < 1$. As a result, we conclude that $\widehat{\boldsymbol{A}}$ is invertible as long as $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}-\widehat{\boldsymbol{A}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| < 1$.

Proof of Lemma 6: controlling $\|\boldsymbol{A}^{-1}(\widehat{\boldsymbol{b}}-\boldsymbol{b})\|_{\boldsymbol{\Sigma}}$. First of all, it is seen that

$$\|\boldsymbol{A}^{-1}(\widehat{\boldsymbol{b}}-\boldsymbol{b})\|_{\boldsymbol{\Sigma}} = \Big\|\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})\Big\|_2 = \Big\|\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{z}_t\Big\|_2,$$

where we define the vector $\boldsymbol{z}_t := \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})$. Therefore, we need to look at the properties of $\boldsymbol{z}_t$. Towards this end, we observe that

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{z}_t^\top\boldsymbol{z}_t\big] &= \mathbb{E}\Big[(\boldsymbol{b}_t-\boldsymbol{b})^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})\Big] \preceq \mathbb{E}\Big[\boldsymbol{b}_t^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{b}_t\Big] \\
&\overset{\text{(i)}}{\leq} \Big\{\max_{s\in\mathcal{S}}|r(s)|^2\Big\}\mathbb{E}\Big[\boldsymbol{\phi}(s_t)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s_t)\Big] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}\Big[\boldsymbol{\phi}(s_t)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s_t)\Big] \\
&= \mathbb{E}\Big[\boldsymbol{\phi}(s_t)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s_t)\Big] \\
&\leq \Big\{\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2\Big\}\cdot\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\big\| \\
&\overset{\text{(iii)}}{\leq} \frac{1}{(1-\gamma)^2}\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2,
\end{aligned}$$

where (i) holds since $\boldsymbol{b}_t = \boldsymbol{\phi}(s_t)r(s_t)$, (ii) follows from the assumption $\max_s|r(s)|\leq 1$, and (iii) arises from Lemma 5. Additionally,

$$\begin{aligned}
\max_t\big\|\boldsymbol{z}_t\big\|_2 &\leq \max_t\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{b}_t\big\|_2 + \big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{b}\big\|_2 \\
&\overset{\text{(iv)}}{\leq} 2\max_s\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)r(s)\big\|_2 \\
&\overset{\text{(v)}}{\leq} 2\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\big\|_2 \\
&\leq 2\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\big\|\cdot\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2 \\
&\leq \frac{2}{1-\gamma}\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2,
\end{aligned}$$

where (iv) holds since $\boldsymbol{b}_t = \boldsymbol{\phi}(s_t)r(s_t)$ and $\boldsymbol{b} = \mathbb{E}_{s\sim\mu}\big[\boldsymbol{\phi}(s)r(s)\big]$, (v) comes from the assumption $\max_s|r(s)|\leq 1$, and the last line is due to Lemma 5. Consequently, the matrix Bernstein inequality Tropp (2015) yields

$$\begin{aligned}
\|\boldsymbol{A}^{-1}(\widehat{\boldsymbol{b}}-\boldsymbol{b})\|_{\boldsymbol{\Sigma}} = \Big\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_t\Big\|_2 &\lesssim \sqrt{\frac{1}{T^2}\sum_{t=0}^{T-1}\mathbb{E}\big[\boldsymbol{z}_t^\top\boldsymbol{z}_t\big]\log\Big(\frac{d}{\delta}\Big)} + \frac{1}{T}\max_t\big\|\boldsymbol{z}_t\big\|_2\log\Big(\frac{d}{\delta}\Big) \\
&\lesssim \frac{\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2}{1-\gamma}\sqrt{\frac{1}{T}\log\Big(\frac{d}{\delta}\Big)} + \frac{\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2}{1-\gamma}\cdot\frac{1}{T}\log\Big(\frac{d}{\delta}\Big) \\
&\asymp \frac{\max_{s\in\mathcal{S}}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2}{1-\gamma}\sqrt{\frac{1}{T}\log\Big(\frac{d}{\delta}\Big)}
\end{aligned} \tag{139}$$

with probability at least $1-\delta$, as long as $T\gtrsim\log\big(\frac{d}{\delta}\big)$.

C.6 Proof of Lemma 8

Recall from the proof of Lemma 1 that $\mathbb{E}_i[\cdot]$ represents the expectation conditioned on the probability space generated by the samples $\{(s_j,s_j')\}_{j\leq i}$. It is easy to check that $\{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\}$ forms a martingale difference sequence, and we seek to bound $\Big\|\frac{1}{u-l+1}\sum_{i=l}^{u}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\Big\|_2$ via matrix Freedman's inequality.

The key is to control the following quantities (here, we abuse notation whenever it is clear from context):

$$W \coloneqq \sum_{i=l}^{u} \mathbb{E}_{i-1}\Big[\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\big\|_2^2\Big] \quad \text{and} \quad B \coloneqq \max_{i:l\le i\le u}\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\big\|_2. \tag{140}$$

Control of B. To begin with, observe that

$$\begin{aligned}
B &= \max_{i:l\le i\le u}\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Sigma}^{-1/2}\big\|\cdot\big\|\boldsymbol{\theta}_i'\big\|_{\boldsymbol{\Sigma}} \\
&\le \frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{1-\gamma}\max_{i:l\le i\le u}\big\{\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big\}\mathbb{1}\{\mathcal{H}_i\} \\
&\le \frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{1-\gamma}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big) =: B_{\max},
\end{aligned}$$

where the second to last inequality comes from (136) and the triangle inequality, and the last line is due to the definition of $\mathcal{H}_i$.

Control of W. Moreover, one can derive

$$\begin{aligned}
W &\coloneqq \sum_{i=l}^{u}\mathbb{E}_{i-1}\Big[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}_i'\|_2^2\Big] \\
&= \sum_{i=l}^{u}\boldsymbol{\theta}_i'^\top\boldsymbol{\Sigma}^{1/2}\mathbb{E}_{i-1}\Big[\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}_i-\boldsymbol{A})^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Sigma}^{-1/2}\Big]\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}_i' \\
&\le \sum_{i=l}^{u}\frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}_i'\|_2^2 \\
&\le \frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\sum_{i=l}^{u}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big)^2\mathbb{1}\{\mathcal{H}_i\} \\
&\le \frac{4(u-l+1)\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)^2 =: W_{\max},
\end{aligned}$$

where the first inequality arises from (134), and the last inequality makes use of the definition of $\mathcal{H}_i$.

With the above bounds in place, we can apply Freedman's inequality (Tropp, 2011, Corollary 1.3) for matrix martingales to demonstrate that

$$\begin{aligned}
\Big\|\frac{1}{u-l+1}\sum_{i=l}^{u}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\Big\|_2 &\le \frac{2}{u-l+1}\sqrt{W_{\max}\log\frac{2d}{\delta}}+\frac{4}{3u-l+1}B_{\max}\log\frac{2d}{\delta} \\
&\le \frac{8\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)}{1-\gamma}\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{u-l+1}}+\frac{16\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{3(1-\gamma)(u-l+1)}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big) \\
&\le \frac{16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)}{1-\gamma}\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{u-l+1}}
\end{aligned}$$

with probability at least $1-\delta$, as long as $u-l+1\ge\frac{4\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}$.

D Comparisons with previous works

D.1 Comparisons with Srikant and Ying (2019)

Srikant and Ying (2019) bounded the expectation of TD estimation error $\mathbb{E}\|\boldsymbol{\theta}_T-\boldsymbol{\theta}^\star\|_2^2$ with Markov samples by an iterative relation. For fair comparisons, we apply their ideas to bounding the error in $\boldsymbol{\Sigma}$-norm with independent samples.

Iterative relation on $\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2$. Recall from the TD update rule (14) that

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

Therefore, the $\boldsymbol{\Sigma}$-norm of $\boldsymbol{\Delta}_{t+1}$ can be expressed as

$$\begin{aligned}\|\boldsymbol{\Delta}_{t+1}\|_{\boldsymbol{\Sigma}}^2 = \|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 &- 2\eta_t\langle\boldsymbol{\Delta}_t, \boldsymbol{A}_t\boldsymbol{\Delta}_t\rangle_{\boldsymbol{\Sigma}} + \eta_t^2\|\boldsymbol{A}_t\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2\\ &- 2\eta_t\langle\boldsymbol{\Delta}_t, \boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\rangle_{\boldsymbol{\Sigma}} + 2\eta_t^2\langle\boldsymbol{A}_t\boldsymbol{\Delta}_t, \boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\rangle_{\boldsymbol{\Sigma}} + \eta_t^2\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

Notice that by definition,

$$\mathbb{E}_t\langle\boldsymbol{\Delta}_t, \boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\rangle_{\boldsymbol{\Sigma}} = \langle\boldsymbol{\Delta}_t, \boldsymbol{A}\boldsymbol{\theta}^\star - \boldsymbol{b}\rangle = 0,$$

and that a basic property of inner product yields

$$2\langle\boldsymbol{A}_t\boldsymbol{\Delta}_t, \boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\rangle_{\boldsymbol{\Sigma}} \le \|\boldsymbol{A}_t\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 + \|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2.$$

Therefore, we can apply the law of total expectations to obtain the following iterative relation:

$$\mathbb{E}\|\boldsymbol{\Delta}_{t+1}\|_{\boldsymbol{\Sigma}}^2 = \mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 - \underbrace{2\eta_t\mathbb{E}[\boldsymbol{\Delta}_t^\top(\boldsymbol{A}^\top\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A})\boldsymbol{\Delta}_t]}_{I_1} + \underbrace{2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2}_{I_2} + \underbrace{2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2}_{I_3}. \tag{141}$$

We now turn to bounding I_1, I_2 and I_3 in order.

Bounding I_1. In order to lower bound I_1 as a function of $\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2$, we firstly express it as

$$\begin{aligned}\boldsymbol{\Delta}_t^\top(\boldsymbol{A}^\top\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A})\boldsymbol{\Delta}_t &= \boldsymbol{\Delta}_t^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}^\top\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{A})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_t\\ &\ge \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_t\|_2^2\lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}\right)\\ &= \|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2\lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}\right).\end{aligned}$$

Recall from (79e) that

$$\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\| \le (1-\gamma)^{-1},$$

so the minimal eigenvalue of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}$ is lower bounded by

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}\right) &\ge \lambda_{\min}(\boldsymbol{\Sigma})\cdot\left[\gamma_{\min}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\right) + \gamma_{\min}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)\right]\\ &\ge \frac{2\lambda_{\min}(\boldsymbol{\Sigma})}{\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\|} \ge 2\lambda_{\min}(\boldsymbol{\Sigma})(1-\gamma).\end{aligned}$$

This directly implies that I_1 is lower bounded by

$$I_1 \ge 2\eta_t(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2. \tag{142}$$

Bounding I_2. We aim to upper bound I_2 as a function of η_t^2 and $\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2$, so that when η_t is sufficiently small, I_2 is negligible compared to I_1. Specifically, for any $\boldsymbol{A}_t$ generated by (11a) and any $\boldsymbol{\Delta}_t \in \mathbb{R}^d$, we observe

$$\begin{aligned}\|\boldsymbol{A}_t\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{\Delta}_t^\top\boldsymbol{A}_t\boldsymbol{\Sigma}\boldsymbol{A}_t\boldsymbol{\Delta}_t &\le \|\boldsymbol{\Delta}_t\|_2^2\|\boldsymbol{A}\|^2\|\boldsymbol{\Sigma}\| \le 4\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Delta}_t\|_2^2\\ &\le 4\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Delta}_t\|_2^2 = 4\kappa\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2,\end{aligned}$$

where we recall κ as the condition number of $\boldsymbol{\Sigma}$. Therefore, as long as

$$\eta_t \le \frac{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})}{4\kappa},$$

it can be guaranteed that $I_2 \le \frac{1}{2}I_1$.

Bounding I_3. In order to compare with our result (Theorem 1 and Corollary 1), we aim to bound I_3 as a function of $\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}$. Towards this end, we firstly notice that

$$\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t = \boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\boldsymbol{\theta}^\star - \gamma\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top\boldsymbol{\theta}^\star - r(s_t)\boldsymbol{\phi}(s_t).$$

Therefore, we can upper bound $\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2$ by

$$\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2 \le 3\mathop{\mathbb{E}}_{s\sim\mu}\|\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 3\mathop{\mathbb{E}}_{s\sim\mu,s'\sim\mathcal{P}(\cdot|s)}\|\boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 3\mathop{\mathbb{E}}_{s\sim\mu}\|r(s)\boldsymbol{\phi}(s)\|_{\boldsymbol{\Sigma}}^2,$$

where the three terms on the right-hand-side can be bounded respectively by

$$\begin{aligned}\mathop{\mathbb{E}}_{s\sim\mu}\|\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 = \mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\theta}^{\star\top}\boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}\boldsymbol{\phi}(s)\right)\boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^\star\right] &\le \mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\theta}^{\star\top}\boldsymbol{\phi}(s)\|\boldsymbol{\Sigma}\|\boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^\star\right]\\ &= \|\boldsymbol{\Sigma}\|\boldsymbol{\theta}^{\star\top}\mathop{\mathbb{E}}_{s\sim\mu}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top]\boldsymbol{\theta}^\star\\ &= \|\boldsymbol{\Sigma}\|\boldsymbol{\theta}^{\star\top}\boldsymbol{\Sigma}\boldsymbol{\theta}^\star = \|\boldsymbol{\Sigma}\|\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2;\end{aligned}$$

$$\begin{aligned}\mathop{\mathbb{E}}_{s\sim\mu,s'\sim\mathcal{P}(\cdot|s)}\|\boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 &= \mathop{\mathbb{E}}_{s\sim\mu,s'\sim\mathcal{P}(\cdot|s)}\left[\boldsymbol{\theta}^{\star\top}\boldsymbol{\phi}(s')\left(\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}\boldsymbol{\phi}(s)\right)\boldsymbol{\phi}(s')^\top\boldsymbol{\theta}^\star\right]\\ &\le \mathop{\mathbb{E}}_{s\sim\mu,s'\sim\mathcal{P}(\cdot|s)}\left[\boldsymbol{\theta}^{\star\top}\boldsymbol{\phi}(s')\|\boldsymbol{\Sigma}\|\boldsymbol{\phi}(s')^\top\boldsymbol{\theta}^\star\right]\\ &= \|\boldsymbol{\Sigma}\|\boldsymbol{\theta}^{\star\top}\mathop{\mathbb{E}}_{s'\sim\mu}[\boldsymbol{\phi}(s')\boldsymbol{\phi}(s')^\top]\boldsymbol{\theta}^\star\\ &= \|\boldsymbol{\Sigma}\|\boldsymbol{\theta}^{\star\top}\boldsymbol{\Sigma}\boldsymbol{\theta}^\star = \|\boldsymbol{\Sigma}\|\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2,\end{aligned}$$

$$\text{and}\quad \mathop{\mathbb{E}}_{s\sim\mu}\|r(s)\boldsymbol{\phi}(s)\|_{\boldsymbol{\Sigma}}^2 \le \max_{s\in\mathcal{S}} r^2(s)\|\boldsymbol{\phi}(s)\|_2^2\|\boldsymbol{\Sigma}\| \le \|\boldsymbol{\Sigma}\|.$$

Consequently, I_3 can be upper bounded by

$$I_3 \le 6\eta_t^2\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 1\right). \tag{143}$$

Bounding $\mathbb{E}\|\boldsymbol{\Delta}_T\|_{\boldsymbol{\Sigma}}^2$. By combining (141), (142) and (143) and recalling that $I_2 \le \frac{1}{2}I_1$ when η_t is sufficiently small, we obtain

$$\mathbb{E}\|\boldsymbol{\Delta}_{t+1}\|_{\boldsymbol{\Sigma}}^2 \le (1-(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta_t)\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 + 6\eta_t^2\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 1\right). \tag{144}$$

Therefore, for constant stepsizes $\eta_0 = \eta_1 = \ldots = \eta_T = \eta$, it is easy to verify by induction that

$$\mathbb{E}\|\boldsymbol{\Delta}_T\|_{\boldsymbol{\Sigma}}^2 \le (1-(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta)^T\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2 + \frac{6\eta\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})}.$$

Hence, in order to guarantee $\mathbb{E}\|\boldsymbol{\Delta}_T\|_{\boldsymbol{\Sigma}}^2 \le \varepsilon^2$, it suffices to take

$$\frac{\eta\|\boldsymbol{\Sigma}\|\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})} \lesssim \varepsilon^2; \quad \text{and} \quad \exp\left(-(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta T\right)\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2 \lesssim \varepsilon^2.$$

This implies the following upper bound for the sample complexity:

$$T \asymp \frac{\kappa\|\boldsymbol{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}, \tag{145}$$

with the proviso that we take the stepsize $\eta \asymp \frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\frac{1}{T}$ and that $T \gtrsim \|\boldsymbol{\Sigma}^{-2}\|(1-\gamma)^{-2}$.

D.2 Comparisons with Bhandari et al. (2021)

Theorem 2(c) in (Bhandari et al., 2021) shows that with decaying stepsizes $\eta_t = \frac{\beta}{\lambda+t}$ where

$$\beta = \frac{2\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)}, \quad \lambda = \frac{16\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^2}, \tag{146}$$

the expected ℓ_2 norm of TD estimation error is bounded by

$$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_2^2 \leq \frac{\nu}{\lambda+T}, \tag{147}$$

where

$$\nu = \max\left\{\frac{8\sigma^2\|\mathbf{\Sigma}^{-2}\|}{(1-\gamma)^2}, \frac{16\|\boldsymbol{\theta}^\star\|_2^2\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^2}\right\}. \tag{148}$$

- Suppose the maximum is attained at the second term for ν and T is sufficiently large, (147) is simplified as

  $$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_2^2 \lesssim \frac{16\|\boldsymbol{\theta}^\star\|_2^2\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^2 T}.$$

  In order for $\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2 \leq \varepsilon^2$, it suffices to take

  $$\frac{\varepsilon^2}{\|\mathbf{\Sigma}\|} \geq \frac{16\|\boldsymbol{\theta}^\star\|_2^2\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^2 T} \geq \mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_2^2,$$

  which implies the following sample complexity:

  $$T \asymp \frac{\|\mathbf{\Sigma}^{-1}\|\|\mathbf{\Sigma}\|\|\boldsymbol{\theta}^\star\|_2^2}{(1-\gamma)^2\varepsilon^2}$$

- Suppose that the first term on the right hand side of expression (148) is larger, (147) can be simplified as

  $$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_2^2 \lesssim \frac{\sigma^2\|\mathbf{\Sigma}^{-2}\|}{(1-\gamma)^2 T}.$$

  Then similarly, the sample complexity is

  $$T \asymp \frac{\|\mathbf{\Sigma}^{-2}\|\|\mathbf{\Sigma}\|\sigma^2}{(1-\gamma)^2\varepsilon^2},$$

  where $\sigma^2 = \mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star - \boldsymbol{b}_t\|_2^2$.

In the worst-case scenario, it satisfies $\sigma^2 \asymp \|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2 + 1$. Therefore, the sample complexity implied by Theorem 2(c) of Bhandari et al. (2021) scales as

$$T \asymp \frac{\kappa\|\mathbf{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{\mathbf{\Sigma}}^2 + 1\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}. \tag{149}$$

References

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier.

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

Bertsekas, D. P. and Tsitsiklis, J. N. (1995). Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pages 560–564. IEEE.

Bertsimas, D., Klasnja, P., Murphy, S., and Na, L. (2022). Data-driven interpretable policy construction for personalized mobile health. In *2022 IEEE International Conference on Digital Health (ICDH)*, pages 13–22. IEEE.

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

Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

Boyan, J. A. (1999). Least-squares temporal difference learning. In *ICML*, pages 49–56.

Dalal, G., Szorenyi, B., and Thoppe, G. (2020). A tale of two-timescale reinforcement learning with the tightest finite-time bound. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3701–3708.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018a). Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dalal, G., Thoppe, G., Szörényi, B., and Mannor, S. (2018b). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233.

Dann, C., Neumann, G., Peters, J., et al. (2014). Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883.

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

Duan, Y., Wang, M., and Wainwright, M. J. (2021). Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874.

Gilbert, E. N. (1952). A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522.

Gupta, H., Srikant, R., and Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Kaledin, M., Moulines, E., Naumov, A., Tadic, V., and Wai, H.-T. (2020). Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research*, 21(1):6742–6804.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2020). Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*.

Lai, T. L. (2003). Stochastic approximation. *The Annals of Statistics*, 31(2):391–406.

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

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023a). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*.

Li, G., Chen, Y., Chi, Y., Gu, Y., and Wei, Y. (2021a). Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting. *Advances in Neural Information Processing Systems*, 34:16671–16685.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023b). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021b). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473.

Ma, C., Zhu, B., Jiao, J., and Wainwright, M. J. (2022). Minimax off-policy evaluation for multi-armed bandits. *IEEE Transactions on Information Theory*, 68(8):5314–5339.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. (2020). On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *arXiv preprint arXiv:2004.04719*.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Pananjady, A. and Wainwright, M. J. (2021). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196.

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pages 993–1000.

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008). A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616.

Szepesvári, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070.

Tang, S. and Wiens, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.

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

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Wang, B., Yan, Y., and Fan, J. (2021a). Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in neural information processing systems*, 34:23009–23022.

Wang, Y., Zou, S., and Zhou, Y. (2021b). Non-asymptotic analysis for two time-scale tdc with general smooth function approximation. *Advances in Neural Information Processing Systems*, 34:9747–9758.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32.

Xu, T. and Liang, Y. (2021). Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR.

Xu, T., Zou, S., and Liang, Y. (2019). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561.