High-probability sample complexities for policy evaluation with linear function approximation

Gen Li, *Member, IEEE,* Weichen Wu, Yuejie Chi, *Fellow, IEEE,* Cong Ma, *Member, IEEE,* Alessandro Rinaldo, Yuting Wei, *Member, IEEE*

***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.**

***Index Terms*—policy evaluation, temporal difference learning, two-timescale stochastic approximation, minimax optimal, function approximation**

I. 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 [1], [2], [3], [4]. For example, in clinical trials [3], 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 [5], 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 [6], [7]. 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 [8], there has been an extensive line of works that consider different types of function approximation, including linear function approximation [9], [10], reproducing kernel Hilbert space [11], [12], deep neural networks [13], [14] or function approximation on the model itself (see, e.g. [15], [16], [17]), with a focus on improving the sample efficiency of RL algorithms.

a) 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 [18] 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 [8], the finite-sample minimax optimality of TD has been established only recently for the tabular MDP [19]. For TD learning with linear function approximation, several recent contributions have produced new non-asymptotic analyses and insights (e.g. [9], [20], [21], [22]), 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

The first two authors contributed equally.

The work of Gen Li was supported in part by the CUHK Direct Grant for Research. The work of Weichen Wu and Alessandro Rinaldo were supported in part by the NIH Grant R01 NS121913. The work of Yuejie Chi was supported in part by the grants ONR N00014-19-1-2404 and NSF CCF-2106778. The work of Yuting Wei was supported in part by the NSF grants CCF-2106778, DMS-2147546/2015447 and NSF CAREER award DMS-2143215 and Google Research Scholar Award. (Correspondent author: Yuting Wei.)

Gen Li is with the Department of Statistics, The Chinese University of Hong Kong, Hong Kong.

Weichen Wu is with the Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Yuejie Chi is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Cong Ma is with the Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

Akessandro Rinaldo is with the Department of Statistics and Data Science, University of Texas at Austin, Austin, UT 78712, USA.

Yuting Wei is with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA; email: ytwei@wharton.upenn.edu

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 [23]. In order to address this issue, [24] 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. [25], [26], [27], [28], [29], [30]) 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.

A. 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 II for the definitions of these parameters. Compared to prior work by [9] and [21], 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 space). Our result also controls ε-convergence with high probability that matches the minimax-optimal dependence on the tolerance level ε with lowest burn-in cost. To assess the tightness of this upper bound, we provide a minimax lower bound in Section III-C, which certifies the optimal dependence of our bound on both the tolerance level ε and problem-related parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}^\star$.
- In the off-policy setting, we establish a sample complexity bound for the TDC algorithm of order
$$\widetilde{O}\left(\frac{\rho_{\max}^7}{\lambda_1^4\lambda_2^3}\frac{\|\widetilde{\boldsymbol{\Sigma}}\|^2}{\varepsilon^2}(1+\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2)\right),$$
where $\widetilde{\boldsymbol{\theta}}$ corresponds to the best linear approximation coefficient of the value function in the off-policy setting, $\widetilde{\boldsymbol{\Sigma}}$ 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 IV. 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 I and II.

B. Other related works

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

a) Finite-sample guarantees for policy evaluation.: Classical analyses of policy evaluation algorithms have mainly focused on providing asymptotic guarantees given a fixed model [8], [32]. 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 [33], [34], [21], [9], [20], [22], [35] with a focus on the non-asymptotic analyses for model-free algorithms, and [36], [37], [38], [39] which derive non-asymptotic bounds for model-based algorithms.

b) Stochastic approximation.: The idea of stochastic approximation (SA) [40], [41] 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 [42], [43] and with properly decaying step sizes, the Polyak-Ruppert averaged iterates asymptotically follow the central limit theorem. Recently, non-asymptotic results have also been

paper	algorithm	stepsize	sample complexity	error control
[9]	TD	$\eta_t \asymp t^{-1}$	$O\left(\frac{\|\mathbf{\Sigma}^{-2}\|\|\mathbf{\Sigma}\|}{(1-\gamma)^4\varepsilon^2}\right)$	in expectation
[21]	TD	$\eta_t \asymp T^{-1}$	$O\left(\frac{\|\mathbf{\Sigma}^{-2}\|\|\mathbf{\Sigma}\|}{(1-\gamma)^4\varepsilon^2}\right)$	in expectation
[20]	TD	$\eta_t = t^{-1}$	$O\left(\frac{1}{\varepsilon^{\max\{2,1+\frac{1}{\lambda}\}}}\right), \lambda \in (0, \lambda_{\min}(\boldsymbol{A}))$	w. high-prob
[31]	Averaged TD	$\eta_t \asymp T^{-1/2}$	$O\left(\frac{\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^4\varepsilon^2} \vee \frac{\|\mathbf{\Sigma}^{2}\|\|\mathbf{\Sigma}^{-4}\|}{(1-\gamma)^6}\right)$	w. high-prob
This work	Averaged TD	$\eta_t = \eta$	$O\left(\frac{\|\mathbf{\Sigma}^{-1}\|}{(1-\gamma)^4\varepsilon^2} \vee \frac{\|\mathbf{\Sigma}^{2}\|\|\mathbf{\Sigma}^{-3}\|}{(1-\gamma)^4}\right)$	w. high-prob

TABLE I. Comparisons with prior results (up to logarithmic terms) in finding an ε-optimal solution using TD learning. Using the Polyak-Ruppert averaging, our high-probability sample complexity bound improves upon previous works in the dependence on the tolerance level ε and problem-related parameters.

obtained for SA for different problems especially in the RL setting; see [44], [45], [22], [46] and references therein. The TDC algorithm is a special case of two-timescale linear SA, whose convergence rates have also been investigated in [28], [47], [48], [30], among others.

c) 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 [49], and various different techniques have been applied to reduce the high variance of IS (see, e.g. [50], [51], [52], [53], [54], [55]). Non-asymptotic guarantees are also provided for off-policy evaluation using a fitted Q-iteration approach under linear function approximation in [56]. A recent line of works also considered finding the optimal policy using batch datasets [57], [58], [59], [60], [61].

C. 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, $\{f_t\}_{t\ge 0}$ and $\{g_t\}_{t\ge 0}$, 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 $f_t \le c_1 g_t$ (resp. $f \ge c_1 g$). 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.

II. Problem Formulation

A. Model and settings

a) 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 [7]. 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

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

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 $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$

paper	algorithm	stepsize	sample complexity	error control
[28]	**Projected** TDC	$\alpha_t = t^{-\alpha},\ \beta_t = t^{-\beta}$	$O\left(\frac{1}{\varepsilon^{2\alpha}}\right), \alpha < 1$	w. high-prob
[29]	TDC	$\alpha_t, \beta_t \asymp \frac{1}{T}$	$O\left(\frac{1}{\varepsilon^2}\right)$	in expectation
[26]	**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 II. 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.

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 $\boldsymbol{\mu} := [\mu(s)]_{1\leq s\leq|\mathcal{S}|} \in \mathbb{R}^{|\mathcal{S}|}$.

b) 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.* $\max_{s\in\mathcal{S}} \|\boldsymbol{\phi}(s)\|_2 \leq 1$.

B. Policy evaluation with linear approximation

a) 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

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

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 [8]

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

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

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

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

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

b) 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

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

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

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

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):

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

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, $\boldsymbol{\theta}^\star \neq \widetilde{\boldsymbol{\theta}}^\star$.

III. 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

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

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,

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

[1]We believe that our framework can be potentially generalized to Markovian samples using similar techniques in [62]. We will briefly discuss the techniques and difficulties in the following sections, but the full details are beyond the scope of the current paper.

A. 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

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

It turns out that the target vector $\boldsymbol{\theta}^\star$ satisfies the equation [8]

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

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

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

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.

B. 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*

$$\left\|\overline{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}} \tag{16}$$

$$\leq C_0\left\{\sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log(\frac{d}{\delta})}{T(1-\gamma)^2}}\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\right)\right.$$
$$\left.+\frac{\|\boldsymbol{\Sigma}^{-1}\|\left[\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\right)\sqrt{\frac{\kappa\log(\frac{dT}{\delta})}{\eta(1-\gamma)^3}}+\frac{1}{\eta(1-\gamma)}\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}\right]}{T}\right\} \tag{17}$$

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*

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

a) Proof sketch and technical novelty: An essential step in our proof of Theorem 1 is to guarantee with high probability that the estimation error of the original TD estimator, $\boldsymbol{\Delta}_t$, is bounded by a time-invariant value. Towards this end, we combine the matrix Freedman's inequality with an induction argument. For the technical details, we refer the readers to Steps 2 and 3 in Section 6. Controlling the norm of $\boldsymbol{\Delta}_t$ with high probability in a time-invariant manner, paves way for bounding the norm of $\overline{\boldsymbol{\Delta}}_T$ with high probability without the need of performing another projection step to restrict the norm of $\boldsymbol{\Delta}_t$ during the TD learning iterations.

b) Generalize to Markov samples: We give a brief introduction of how our results can be extended to Markov samples. The main difficulty towards this end lies in bounding the sequence of temporal difference errors

$$\{\boldsymbol{A}_t\boldsymbol{\theta}_t - \boldsymbol{b}_t\}_{t\geq 0}$$

which is no longer a Martingale difference process, so the Freedman's inequality is not directly applicable anymore. In order to tackle this problem, the popular strategy is to assume that the Markov chain mixes at a geometric rate. Specifically, for arbitrarily small $\varepsilon > 0$, there exists a positive integer $t_{\mathsf{mix}}(\varepsilon) \asymp \log\left(\frac{1}{\varepsilon}\right)$, such that for any $t \geq t_{\mathsf{mix}}$,

$$\|\mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{A}_t] - \boldsymbol{A}\| < \varepsilon, \quad \text{and} \quad \|\mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{b}_t] - \boldsymbol{b}\|_2 < \varepsilon. \tag{18}$$

Here, $\mathbb{E}_i[\cdot]$ represents the expectation conditioned on the filtration $\mathcal{F}_i$ — the σ-algebra generated by $\{(s_j, s_j')\}_{j\leq i}$. Under this assumption, the temporal difference error can be decomposed as

$$\begin{aligned}\boldsymbol{A}_t\boldsymbol{\theta}_t - \boldsymbol{b}_t &= \boldsymbol{A}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-t_{\mathsf{mix}}})\\ &+ \left[\left(\boldsymbol{A}_t - \mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{A}_t]\right)\boldsymbol{\theta}_{t-t_{\mathsf{mix}}} - \left(\boldsymbol{b}_t - \mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{b}_t]\right)\right]\\ &+ \left[\left(\mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{A}_t] - \boldsymbol{A}\right)\boldsymbol{\theta}_{t-t_{\mathsf{mix}}} - \left(\mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{b}_t] - \boldsymbol{b}\right)\right].\end{aligned}$$

On the right hand side, the last term can be bounded by the mixing property (18); the first term can be further expanded as

$$\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-t_{\mathsf{mix}}} = \sum_{j=t-t_{\mathsf{mix}}}^{t-1}(\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j)$$
$$= \sum_{j=t-t_{\mathsf{mix}}}^{t-1}\eta(\boldsymbol{A}_t\boldsymbol{\theta}_j - \boldsymbol{b}_j),$$

and bounded in terms of the step size η and the mixing time t_{mix}; the second term can be controlled by separating the sequence

$$\left\{\left(\boldsymbol{A}_t - \mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{A}_t]\right)\boldsymbol{\theta}_{t-t_{\mathsf{mix}}} - \left(\boldsymbol{b}_t - \mathbb{E}_{t-t_{\mathsf{mix}}}[\boldsymbol{b}_t]\right)\right\}_{t\geq 0}$$

into t_{mix} disjoint Martingale difference processes and invoking the matrix Freedman's inequality. We leave the details to our future work.

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*

$$\left\|\boldsymbol{V}_{\overline{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\boldsymbol{\theta}^\star}\right\|_{\boldsymbol{D}_\mu} = \left\|\overline{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}} \leq \varepsilon \tag{19}$$

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

$$T \geq \frac{c\left\{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\right\}\left(1+\left\|\boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}^2\right)\log\left(\frac{d}{\delta}\right)}{(1-\gamma)^2\varepsilon^2}. \tag{20}$$

c) Comparisons to prior literature: We remark that the best finite-sample results for TD learning obtained so far are given by [9, Theorem 2(c)] and [21, 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)

$$T^{\mathsf{prior}} \asymp \frac{\kappa\|\boldsymbol{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}. \tag{21}$$

We refer readers to Appendix D-A and D-B for a detailed translation of their results. Comparing (20) and (21), our result improves upon previous works by a multiplicative factor of

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

the condition number of $\boldsymbol{\Sigma}$; 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, another recent work [20] shows that in order for (19) to hold with probability at least $1-\delta$, it suffices to take

$$T \asymp \max\left\{\left(\frac{1}{\varepsilon}\right)^2\left(\log\frac{1}{\delta}\right)^3, \tag{22}$$

$$\left.\left(\frac{1}{\varepsilon}\right)^{1+1/\lambda_{\min}(\boldsymbol{A})}\left(\log\frac{1}{\delta}\right)^{1+1/\lambda_{\min}(\boldsymbol{A})}\right\}. \tag{23}$$

Comparing (20) and (22), 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.

After the initial post of the current paper, we are pointed to the work [31], which provides a general treatment of linear stochastic approximation with Polyak-Ruppert averaging. Their results lead to the same sample complexity as Corollary 1 with a slightly higher burn-in cost. We include the detailed comparisons of their result in Section D-D. We also point out the works of [63] and [64], which derived similar results regarding the error bound for averaged TD learning in expectation. Our result, as shown in Theorem 1, improves upon theirs in the sense that we provide high probability guarantees and offer explicit dependencies on problem-related parameters. Detailed comparisons can be found in Section D.3.

C. 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* $\frac{1}{2} < \gamma < 1$, $1 < d \leq |\mathcal{S}|$, *and* $0 < \varepsilon < c_1 \max\{1, \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}\}$ *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* $\boldsymbol{\Phi}$ *such that*

$$\mathbb{P}\Big\{ \big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big\|_{\boldsymbol{\Sigma}} > c_2\varepsilon \Big\} \geq \frac{1}{4}, \tag{24}$$

provided that the number of samples T *satisfies*

$$T \leq \frac{c_3 \big\{ \max_s \boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(s) \big\} \big(1 + \big\|\boldsymbol{\theta}^\star\big\|_{\boldsymbol{\Sigma}}^2\big)}{(1-\gamma)\varepsilon^2}. \tag{25}$$

Remark 1. We remark that minimax lower bounds are also previously investigated in a general framework in [12] 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 [12], 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

$$O\left(\frac{d}{\varepsilon^2(1-\gamma)^3}\right), \tag{26}$$

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

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

Consequently, the upper bounds developed in Corollary 1 are sharp in terms of the accuracy level ε, the dependence of the feature map $\boldsymbol{\Phi}$, the underlying coefficient $\boldsymbol{\theta}^\star$, and the covariance matrix $\boldsymbol{\Sigma}$. 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.

We believe that the gap in terms of $\frac{1}{1-\gamma}$ mainly comes from the function approximation paradigm. In fact, as far as we know, with linear function approximation, there has been no minimax optimal results established for this problem either for the model based method (e.g. LSTD) or for the variance-reduced approach, both of which are known to be minimax optimal in the tabular setting; the latter is also proved to be instance-optimal from [33] and [64]. We conjecture the minimax optimal dependency of $\frac{1}{1-\gamma}$ to be the same as that of the tabular setting and TD with LFA to be minimax optimal. Establishing this result, however, requires developing completely new analysis tools, particularly in dealing with the structure of variance across different steps, which we leave as an interesting open direction.

IV. 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

$$\begin{aligned} &(s_t, a_t, s_t'), \qquad \text{where} \\ &s_t \overset{\text{i.i.d.}}{\sim} \mu_b,\ a_t \sim \pi_b(\cdot \mid s_t), \quad \text{and } s_t' \sim \mathcal{P}(\cdot \mid s_t, a_t). \end{aligned} \tag{27}$$

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.

A. The TDC algorithm

The TDC algorithm is designed to solve the optimization problem (9) using a two-timescale linear TD with gradient correction [24]. 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*

$$\begin{aligned} \mathsf{MSPBE}(\boldsymbol{\theta}) = \frac{1}{2} &\mathbb{E}_{\mu_b,\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right]^\top \\ &\left\{\mathbb{E}_{\mu_b}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1} \mathbb{E}_{\mu_b,\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right], \end{aligned} \tag{28}$$

where $\delta_t := r_t + \gamma \boldsymbol{\phi}(s_t')^\top \boldsymbol{\theta} - \boldsymbol{\phi}(s_t)^\top \boldsymbol{\theta}$ *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

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}}\mathsf{MSPBE}(\boldsymbol{\theta}) \\
&= \mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\left(\gamma\boldsymbol{\phi}(s_t') - \boldsymbol{\phi}(s_t)\right)\boldsymbol{\phi}(s_t)^\top\right] \\
&\left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] \\
&= -\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] + \gamma\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top\right] \\
&\left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi,\mathcal{P}}\left[\boldsymbol{\phi}(s_t)\delta_t\right] \\
&= -\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t))\delta_t\right] + \gamma\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t')\boldsymbol{\phi}(s_t)^\top\right]\boldsymbol{w}_t,
\end{aligned} \tag{29}$$

where in the last step we have defined

$$\begin{aligned}
\boldsymbol{w}_t &= \boldsymbol{w}(\boldsymbol{\theta}_t) \\
&= \left\{\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\right\}^{-1}\mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t)\delta_t\right].
\end{aligned} \tag{30}$$

and have used the importance weights

$$\rho_t \coloneqq \frac{\pi(a_t|s_t)}{\pi_{\mathsf{b}}(a_t|s_t)} \tag{31}$$

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 (29) based on the sample trajectory (27), and then perform stochastic gradient updates for $\boldsymbol{\theta}_t$. However, the challenge is that the second term in the gradient of MSPBE (29) 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, [65] and [24] 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

$$\mathbb{E}_{\mu_{\mathsf{b}}}\left[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top\right]\boldsymbol{w} = \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}\left[\rho_t\boldsymbol{\phi}(s_t)\delta_t\right]. \tag{32}$$

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 (28), the update of $\boldsymbol{w}_t$ corresponds to linear SA for solving (32), 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

$$\begin{aligned}
\widetilde{\boldsymbol{A}}_t &= \rho_t\boldsymbol{\phi}(s_t)\left(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t')\right)^\top, \\
\widetilde{\boldsymbol{b}}_t &\coloneqq \rho_t\boldsymbol{\phi}(s_t)r_t, \\
\boldsymbol{\Pi}_t &\coloneqq \rho_t\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top, \\
\widetilde{\boldsymbol{\Sigma}}_t &\coloneqq \boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top.
\end{aligned} \tag{33}$$

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

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \alpha_t(\widetilde{\boldsymbol{A}}_t\widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}}_t + \gamma\boldsymbol{\Pi}_t^\top\boldsymbol{w}_t); \tag{34a}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \beta_t(\widetilde{\boldsymbol{A}}_t\widetilde{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{b}}_t + \widetilde{\boldsymbol{\Sigma}}_t\boldsymbol{w}_t). \tag{34b}$$

B. 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

$$\widetilde{\boldsymbol{A}} \coloneqq \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}[\widetilde{\boldsymbol{A}}_t] = \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}[\rho_t\boldsymbol{\phi}(s_t)\left(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_t')\right)^\top]; \tag{35a}$$

$$\widetilde{\boldsymbol{b}} \coloneqq \mathbb{E}_{\mu_{\mathsf{b}}}[\widetilde{\boldsymbol{b}}_t] = \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}}}[\rho_t\boldsymbol{\phi}(s_t)r_t]; \tag{35b}$$

$$\boldsymbol{\Pi} \coloneqq \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}[\boldsymbol{\Pi}_t] = \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}},\mathcal{P}}[\rho_t\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t')^\top]; \tag{35c}$$

$$\widetilde{\boldsymbol{\Sigma}} \coloneqq \mathbb{E}_{\mu_{\mathsf{b}}}[\widetilde{\boldsymbol{\Sigma}}_t] = \mathbb{E}_{\mu_{\mathsf{b}}}[\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top]. \tag{35d}$$

In addition, denote the parameters

$$\begin{aligned}
\lambda_1 &= \lambda_{\min}(\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}}), \\
\lambda_2 &= \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}), \\
\lambda_\Sigma &= \|\widetilde{\boldsymbol{\Sigma}}^{-1}\| = 1/\lambda_2, \\
\widetilde{\kappa} &= \lambda_\Sigma\cdot\|\widetilde{\boldsymbol{\Sigma}}\|, \\
\rho_{\max} &= \max_{s,a}[\pi(a|s)/\pi_{\mathsf{b}}(a|s)].
\end{aligned} \tag{36}$$

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

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 (34) at time T satisfies the bound*

$$\|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{C}_0\frac{\rho_{\max}^2\|\widetilde{\boldsymbol{\Sigma}}\|}{\lambda_1}\sqrt{\frac{\beta}{\lambda_2}\log\frac{2dT}{\delta}}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 2), \tag{37}$$

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

$$\begin{aligned}
&\widetilde{\boldsymbol{\theta}}_0 = \boldsymbol{0}, \\
&\alpha_0 = \ldots = \alpha_T = \alpha, \\
&\beta_0 = \ldots = \beta_T = \beta, \\
&0 < \alpha < \frac{1}{\lambda_1\lambda_\Sigma^2\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}, \\
&\frac{\alpha}{\beta} = \frac{1}{128}\frac{\lambda_1\lambda_2}{\rho_{\max}^2(1+\lambda_\Sigma\rho_{\max})}, \\
&T \ge \widetilde{c}_1\frac{\log\|\widetilde{\boldsymbol{\theta}}^\star\|_2}{\alpha\lambda_1}\log\max\left\{\sqrt{\widetilde{\kappa}},\ \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}\sqrt{\frac{\alpha\lambda_1}{\log\frac{2dT}{\delta}}}\right\}.
\end{aligned} \tag{38}$$

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 (37). In particular, under the same conditions as in (38), it can be derived with probability at least $1-\delta$ that

$$\|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_2 \lesssim \widetilde{C}_0\frac{\rho_{\max}^2}{\lambda_1}\sqrt{\frac{\beta}{\lambda_2}\log\frac{2dT}{\delta}}(\|\widetilde{\boldsymbol{\theta}}^\star\|_2 + 2). \tag{39}$$

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

a) Proof sketch and technical novelty: Our proof of Theorem 3 considers the convergence of the vector

$$\boldsymbol{x}_t := \begin{bmatrix} \tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}^\star \\ \varkappa \left[\tilde{\boldsymbol{w}}_t + \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{A}} (\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}^\star) \right] \end{bmatrix},$$

where $\varkappa \in (0,1)$ is a constant to be specified. Firstly, we identify the conditions on α, β and $\varkappa$ that guarantee the exponential convergence of $\boldsymbol{x}_t$ in the noise-free scenario; after this, we again combine the matrix Freedman's inequality and an induction argument to bound the norm of $\boldsymbol{x}_t$ for $i.i.d.$ samples with high probability by a time-invariant value in terms of the step size. And finally, with a careful choice of α, β and $\varkappa$, we establish the finite-sample guarantee as is shown in Theorem 3. The main technical novelty of this proof lies in the construction of the vector $\boldsymbol{x}_t$ and the choice of the paramter $\varkappa$.

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*

$$\|\boldsymbol{V}_{\widetilde{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\widetilde{\boldsymbol{\theta}}^\star}\|_{\boldsymbol{D}_{\mu_b}} = \|\widetilde{\boldsymbol{\theta}}_T - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} \leq \varepsilon \tag{40}$$

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

$$T \geq \widetilde{c} \frac{\rho_{\max}^7}{\lambda_1^4 \lambda_2^3} \frac{\|\widetilde{\boldsymbol{\Sigma}}\|^2}{\varepsilon^2} (1 + \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2) \log\Big(\frac{d \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}}{\delta}\Big), \tag{41}$$

and the stepsize parameters α_t and β_t are chosen as

$$\alpha_t \asymp \frac{\log \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}}{T \lambda_1}, \qquad \beta_t = 128 \frac{\rho_{\max}^2 (1 + \lambda_\Sigma \rho_{\max})}{\lambda_1 \lambda_2} \alpha. \tag{42}$$

b) 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 [28], 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 [29], 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. The key to achieving this again lies in our combination of the matrix Freedman's inequality with an induction argument; the details of the proof is postponed to Section VII. We also remark that [26] 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 $O(\frac{1}{\varepsilon^2}\log(\frac{1}{\varepsilon}))$.

V. 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.

A. On-policy evaluation: averaged TD learning

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

a) 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.

b) 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.

B. 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* [23].

a) 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)

$$\mathcal{P}(s'|s, 1) = \mathbb{1}\{s' = 7\}, \quad \forall s \in [7];$$

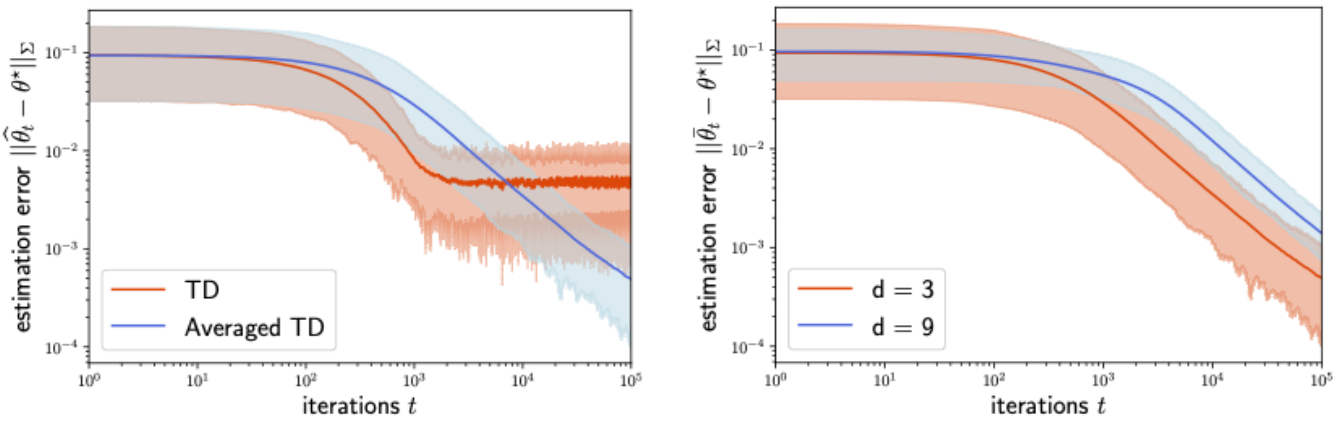

Fig. 1. (a) Comparisons of the estimation error of TD and averaged TD when $d = 3$. (b) Comparisons of the estimation error for averaged TD with $d = 3$ and $d = 9$. Two curves in the middle represent their average errors, while the shaded areas represent the 95% confidence bands.

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

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 $\boldsymbol{\theta} \in \mathbb{R}^8$,

$$\begin{cases} V(i) = 2\theta_i + \theta_8, & \text{for} \quad 1 \le i \le 6; \\ V(7) = \theta_7 + 2\theta_8. \end{cases} \tag{43}$$

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 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 $\|\widehat{\boldsymbol{\theta}}_t - \widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} = \|\boldsymbol{V}_{\widehat{\boldsymbol{\theta}}_t} - \boldsymbol{V}^\star\|_{\boldsymbol{D}_{\mu_{\mathsf{b}}}}$ to evaluate the estimation error, and our experimental results would remain the same.

b) 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 [23]. 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.

VI. Proof of Theorem 1 (TD learning)

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

$$\boldsymbol{\Delta}_t \coloneqq \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star, \qquad \text{and} \qquad \overline{\boldsymbol{\Delta}}_t \coloneqq \overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star. \tag{44}$$

a) 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

$$\boldsymbol{\xi}_t \coloneqq (\boldsymbol{A}_t - \boldsymbol{A})\boldsymbol{\theta}_t - (\boldsymbol{b}_t - \boldsymbol{b}). \tag{45}$$

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

$$\begin{aligned}
\boldsymbol{\Delta}_t &= (\mathbf{I} - \eta\boldsymbol{A})\boldsymbol{\Delta}_{t-1} - \eta\boldsymbol{\xi}_{t-1}\\
&= (\mathbf{I} - \eta\boldsymbol{A})^t\boldsymbol{\Delta}_0 - \eta\sum_{i=0}^{t-1}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i.
\end{aligned} \tag{46}$$

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

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(1 + \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}) + 2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} =: R_0 \tag{47}$$

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 [9].

To start with, note that the inequality (47) holds trivially for

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3394685

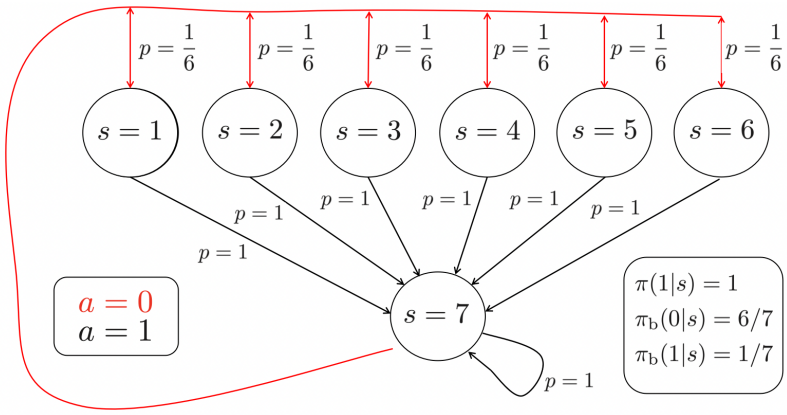

Fig. 2. Baird's counterexample. Taking action $a = 1$ always leads to state $s = 7$, while taking $a = 0$ leads to one of the other six states with equal probability. The reward is set to be always zero.

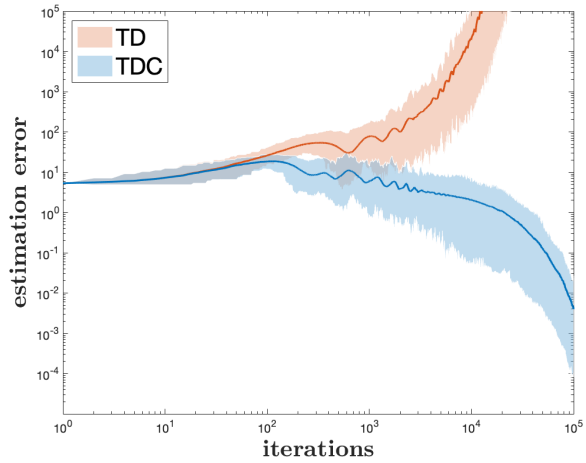

Fig. 3. Performances of off-policy averaged TD (red, $\eta = 0.02$) and TDC (blue, $\alpha = 0.02$, $\beta = 0.002$). Two curves in the middle represent their average errors, while the shaded areas correspond to 95% confidence bands.

the base case with $t = 0$, given that $\kappa \geq 1$. Next, suppose that the hypothesis (47) 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 (46) and the triangle inequality yields

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \leq \big\|(\mathbf{I} - \eta \boldsymbol{A})^t \boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} + \eta \Big\| \sum_{i=0}^{t-1} (\mathbf{I} - \eta \boldsymbol{A})^{t-i-1} \boldsymbol{\xi}_i \Big\|_{\boldsymbol{\Sigma}}. \tag{48}$$

As for the first term of (48), it is seen that

$$\begin{aligned} &\big\|(\mathbf{I} - \eta \boldsymbol{A})^t \boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} \\ &= \big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \eta \boldsymbol{A})^t \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_0\big\|_2 \\ &\leq \big\|\boldsymbol{\Sigma}^{1/2}\big\| \cdot \big\|\boldsymbol{\Sigma}^{-1/2}\big\| \cdot \big\|\mathbf{I} - \eta \boldsymbol{A}\big\|^t \cdot \big\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Delta}_0\big\|_2 \\ &\leq \sqrt{\kappa} \left(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\right)^t \|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} \leq \sqrt{\kappa} \|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}, \end{aligned} \tag{49}$$

where the last inequality arises from the definition of κ and the property (88g) (with the restriction that $\eta \leq (1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$). When it comes to the second term of (48), the following lemma comes in handy.

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

$$\widetilde{\boldsymbol{\xi}}_i := \boldsymbol{\xi}_i \mathbb{1}\{\mathcal{H}_i\}, \qquad \text{where} \quad \mathcal{H}_i := \Big\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}} \leq R\Big\}. \tag{50}$$

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

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

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

Proof. See Section C-A. □

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

$u = t-1$) to show that

$$\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} = \Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\mathbb{1}\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\le R_0\}\Big\|_{\boldsymbol{\Sigma}} \le 16(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R_0+1)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}} \tag{51}$$

holds with probability at least $1-\delta/T$, provided that $0<\eta\le\frac{1-\gamma}{\kappa\log\frac{dT}{\delta}}$. Combining (48), (49) and (51) together and recalling the definition (47) of R_0, we can easily verify that

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le \sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} + 16(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R_0+1)\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}} \le R_0 \tag{52}$$

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 (47).

c) Step 3: a refined bound on $\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}$: It turns out that the upper bound (47) is somewhat loose due to the complete ignorance of the contraction effect of $\mathbf{I}-\eta\boldsymbol{A}$; see (49). In what follows, we develop a strengthened bound. Define

$$t_{\mathsf{seg}} := \frac{c_1\log\max\{4\sqrt{\kappa},\ \frac{16\kappa\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1},\ \|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}\sqrt{\frac{1-\gamma}{\eta\kappa\log\frac{2dT}{\delta}}}\}}{\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})} \tag{53}$$

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

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \le 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\frac{\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{k-1}}+\frac{3}{2}\Big) =: R_k \tag{54}$$

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 (48), we claim that it suffices to prove that

$$\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \le 32\Big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\frac{2\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^k}+1\Big)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}}, \tag{55}$$

$$\text{when } kt_{\mathsf{seg}}\le t\le T. \tag{56}$$

To see this: note that the first term on the right-hand side of (48) has already been bounded in (49), which combined with the definition (53) of t_{seg} indicates that

$$\big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\eta\boldsymbol{A})^t\boldsymbol{\Delta}_0\big\|_2 \tag{57}$$

$$\le\sqrt{\kappa}\left(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\right)^{t_{\mathsf{seg}}}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}} \tag{58}$$

$$\le\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}} \tag{59}$$

for any $t\ge t_{\mathsf{seg}}$. Clearly, combining (56) with (48) and (59) shall immediately lead to the claim (54). The remainder of this step is thus devoted to demonstrating (56) inductively.

The base case (i.e. $k=1$) follows immediately from our bounds (47) and (51) in Step 2, given that $\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}$ is sufficiently small. Suppose now that the claim (56) 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 (56) 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 (54) 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 $t\ge t_{\mathsf{seg}}$,

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

  Here, the first equality uses the crude bound $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\le R_0$ for all i (see (47)), 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 (47) of R_0 and invokes

the fact that $\sqrt{\kappa}\left(1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\right)^{t_{\mathsf{seg}}-1} \leq \min\left\{\frac{1}{4}, \frac{1}{4\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}\right\}$ with our choice (53) of t_{seg}.

Combine the previous two bounds to reach

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

This finishes the induction step and in turn establishes (56) (and hence (54)).

As a straightforward consequence, the bounds (47) and (54) imply that

$$\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}} \leq \begin{cases} R_0, & 0\leq t<t'_{\mathsf{seg}}, \\ 32\sqrt{\frac{\eta\kappa\log\frac{2dT}{\delta}}{1-\gamma}}(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+2), & t'_{\mathsf{seg}}\leq t<T, \end{cases} \tag{60}$$

where

$$t'_{\mathsf{seg}} \coloneqq c_2 t_{\mathsf{seg}}\log\big(\kappa(\|\boldsymbol{\Delta}_0\|_2+1)\big) \tag{61}$$

for some large enough constant $c_2>0$. To see this, note that for any $t\geq t'_{\mathsf{seg}}$, it is guaranteed that the second term on the right-hand side of (54) obeys $\frac{4\sqrt{\kappa}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}}{2^{\lfloor t/t_{\mathsf{seg}}\rfloor}}\leq 2$, thus confirming the second case in (60).

d) 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 $\{\boldsymbol{\xi}_i\}_{0\leq i\leq T-1}$ as follows, which is a direct consequence of the relation (46):

$$\begin{aligned}
\overline{\boldsymbol{\Delta}}_T &= \frac{1}{T}\sum_{j=1}^{T}\boldsymbol{\Delta}_j \\
&= \frac{1}{T}\sum_{j=1}^{T}(\mathbf{I}-\eta\boldsymbol{A})^j\boldsymbol{\Delta}_0-\frac{1}{T}\sum_{j=1}^{T}\eta\sum_{i=0}^{j-1}(\mathbf{I}-\eta\boldsymbol{A})^{j-i-1}\boldsymbol{\xi}_i \\
&= \frac{1}{T}\sum_{j=1}^{T}(\mathbf{I}-\eta\boldsymbol{A})^j\boldsymbol{\Delta}_0-\frac{1}{T}\sum_{i=0}^{T-1}\eta\sum_{j=i+1}^{T}(\mathbf{I}-\eta\boldsymbol{A})^{j-i-1}\boldsymbol{\xi}_i \\
&= \frac{1}{T\eta}\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0-\frac{1}{T}\boldsymbol{\Delta}_0-\frac{1}{T}\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i,
\end{aligned} \tag{62}$$

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

$$\boldsymbol{A}_i^{(t)} \coloneqq \eta\sum_{j=i+1}^{t}(\mathbf{I}-\eta\boldsymbol{A})^{j-i-1} = \boldsymbol{A}^{-1}\big(\mathbf{I}-(\mathbf{I}-\eta\boldsymbol{A})^{t-i}\big). \tag{63}$$

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-B)

$$\big\|\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}} \leq \frac{2\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}; \tag{64a}$$

$$\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}} \lesssim \left\{\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{T(1-\gamma)^2}}\right. \tag{64b}$$

$$\left.+\frac{\|\boldsymbol{\Sigma}^{-1}\|}{T}\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}\right\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+1\big). \tag{64c}$$

Putting the above two inequalities together with (62), we arrive at

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

where the last line follows since $\|\boldsymbol{\Sigma}^{-1}\|\geq 1$ (see (88h)) and $\eta<1$. This finishes the proof of Theorem 1.

VII. 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.

A. Population analysis

First recall that the population parameters are defined as

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

Corresponding to the empirical version of TDC as given in (34), we can define its population analogue of TDC as

$$\begin{aligned}
\breve{\boldsymbol{\theta}}_{t+1} &= \breve{\boldsymbol{\theta}}_t-\alpha(\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}_t-\widetilde{\boldsymbol{b}}+\gamma\boldsymbol{\Pi}^\top\breve{\boldsymbol{w}}_t), \\
\breve{\boldsymbol{w}}_{t+1} &= \breve{\boldsymbol{w}}_t-\beta(\widetilde{\boldsymbol{A}}\breve{\boldsymbol{\theta}}_t-\widetilde{\boldsymbol{b}}+\widetilde{\boldsymbol{\Sigma}}\breve{\boldsymbol{w}}_t),
\end{aligned} \tag{65}$$

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

$$\begin{aligned}\widetilde{\boldsymbol{\Sigma}} &= \sum_{s_t\in\mathcal{S}} \mu_{\mathsf{b}}(s_t)\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top \\ &= \sum_{s_t\in\mathcal{S}} \mu_{\mathsf{b}}(s_t)\left(\sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)\right)\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top \\ &= \sum_{s_t\in\mathcal{S}}\sum_{a_t\in\mathcal{A}} \mu_{\mathsf{b}}(s_t)\pi_{\mathsf{b}}(a_t|s_t)\left(\frac{\pi(a_t|s_t)}{\pi_{\mathsf{b}}(a_t|s_t)}\right)\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top \qquad (66)\\ &= \mathbb{E}_{\mu_{\mathsf{b}},\pi_{\mathsf{b}}}[\rho_t\boldsymbol{\phi}(s_t)\boldsymbol{\phi}(s_t)^\top]. \qquad (67)\end{aligned}$$

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

$$\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{\Sigma}} - \gamma\boldsymbol{\Pi}. \qquad (68)$$

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

$$\begin{bmatrix}\breve{\boldsymbol{\Delta}}_t\\ \breve{\boldsymbol{z}}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \alpha\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & -\alpha\gamma\boldsymbol{\Pi}^\top\\ -\alpha(\boldsymbol{I}-\gamma\boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi})\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & \boldsymbol{I} - \beta\widetilde{\boldsymbol{\Sigma}} - \alpha\gamma(\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\Pi}^\top\end{bmatrix} \cdot \begin{bmatrix}\breve{\boldsymbol{\Delta}}_{t-1}\\ \breve{\boldsymbol{z}}_{t-1}\end{bmatrix}. \qquad (69)$$

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

$$\breve{\boldsymbol{x}}_t = \boldsymbol{\Psi}\breve{\boldsymbol{x}}_{t-1}, \qquad (70)$$

where $\boldsymbol{\Psi}$ represents the matrix

$$\begin{bmatrix}\boldsymbol{I} - \alpha\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & -\frac{1}{\varkappa}\alpha\gamma\boldsymbol{\Pi}^\top\\ -\varkappa\alpha(\boldsymbol{I}-\gamma\boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi})\widetilde{\boldsymbol{A}}^\top\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{A}} & \boldsymbol{I} - \beta\widetilde{\boldsymbol{\Sigma}} - \alpha\gamma(\boldsymbol{I}-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\Pi}^\top\end{bmatrix}. \qquad (71)$$

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:

$$\beta \gtrsim \lambda_\Sigma\rho_{\max}\alpha, \qquad (72a)$$

$$\varkappa\beta \gtrsim \alpha, \qquad (72b)$$

$$\alpha\gamma(\rho_{\max} + \gamma\lambda_\Sigma\rho_{\max}^2) \ll \beta\lambda_w, \qquad (72c)$$

$$\frac{\alpha\gamma\rho_{\max}}{\varkappa} + \varkappa\alpha(1+\gamma\lambda_\Sigma\rho_{\max})\lambda_\Sigma(2\rho_{\max})^2 \ll \sqrt{\alpha\lambda_1\beta\lambda_w} \qquad (72d)$$

it holds true that

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

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

B. 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.

a) 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 VII-A, 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

$$\boldsymbol{\zeta}_t = \begin{bmatrix}\alpha\boldsymbol{\nu}_t\\ \varkappa(\alpha(1-\gamma\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi})\boldsymbol{\nu}_t + \beta\boldsymbol{\eta}_t)\end{bmatrix}. \qquad (73)$$

Hence,

$$\boldsymbol{x}_t = \boldsymbol{\Psi}^t\boldsymbol{x}_0 - \sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i, \qquad (74)$$

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 (74). 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

$$\begin{cases} \|\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \le \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}; \\ \|\boldsymbol{w}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \frac{1}{\varkappa}\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}; \\ \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \|\widetilde{\boldsymbol{\Delta}}_t\|_{\widetilde{\boldsymbol{\Sigma}}} + |\boldsymbol{w}_t\|_{\widetilde{\boldsymbol{\Sigma}}}. \end{cases} \tag{75}$$

b) 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 $1-\delta$,

$$\begin{aligned} \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} &\le 2\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}} + 80\varkappa\beta\rho_{\max}\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 1) \\ &=: \widetilde{R}_0 \end{aligned} \tag{76}$$

holds simulatanesouly for all $0 \le t \le T$. To start with, note that the inequality (76) holds trivially for the base case with $t = 0$. Next, suppose that the hypothesis (76) 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 (74) and the triangle inequality yields

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \le \left\|\boldsymbol{\Psi}^t\boldsymbol{x}_0\right\|_{\widetilde{\boldsymbol{\Sigma}}} + \left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{77}$$

As for the first term of (77), it is seen that

$$\left\|\boldsymbol{\Psi}^t\boldsymbol{x}_0\right\|_{\widetilde{\boldsymbol{\Sigma}}} \le \|\boldsymbol{x}_0\|_{\widetilde{\boldsymbol{\Sigma}}} = \|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{78}$$

When it comes to the second term of (77), 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*

$$\widetilde{\boldsymbol{\zeta}}_i := \boldsymbol{\zeta}_i \mathbb{1}\{\widetilde{\mathcal{H}}_i\}, \quad \text{where} \quad \widetilde{\mathcal{H}}_i := \left\{\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{R}\right\}. \tag{79}$$

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

$$\left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{1}{\varkappa}\widetilde{R} + 1), \tag{80}$$

provided that the stepsizes α, β satisfy the conditions (72) and that $0 < \alpha < \frac{1}{\lambda_1\lambda_\Sigma^2\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}$.

Proof. See Section C-D. □

Putting relations (77) and (80) together, we find

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

by definition of $\widetilde{R}_0$ in (76), 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

$$\begin{aligned} &\mathbb{P}\left\{\max_{0\le i\le t}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} \\ &\le \mathbb{P}\left\{\max_{0\le i<t-1}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} \\ &+ \mathbb{P}\left\{\max_{0\le i<t-1}\|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \le \widetilde{R}_0, \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} > \widetilde{R}_0\right\} \\ &\le \frac{(t-1)\delta}{T} + \mathbb{P}\left\{\left\|\sum_{i=0}^{t-1}\boldsymbol{\Psi}^{t-i-1}\widetilde{\boldsymbol{\zeta}}_i\right\|_{\widetilde{\boldsymbol{\Sigma}}}\right. \\ &\left.\gtrsim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\varkappa\beta\rho_{\max}(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \widetilde{R}_0 + 1)\right\} \\ &\le \frac{(t-1)\delta}{T} + \frac{\delta}{T} = \frac{t\delta}{T}. \end{aligned} \tag{81, 82}$$

This completes our claim at this step.

c) Step 3: refined bound for $\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}}$: It turns out that the upper bound (76) can be tightened by taking into account the contraction effect of $\boldsymbol{\Psi}$. In what follows, we develop a strengthened bound. Define

$$\widetilde{t}_{\mathsf{seg}} := \frac{\widetilde{c}_1\log\max\left\{\sqrt{\widetilde{\kappa}}, \frac{\sqrt{\widetilde{\kappa}}\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}}{\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}+1}, \|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}\sqrt{\frac{\alpha\lambda_1}{\|\widetilde{\boldsymbol{\Sigma}}\|\log\frac{2dT}{\delta}}}\frac{1}{\varkappa\beta\rho_{\max}}\right\}}{\alpha\lambda_1} \tag{83}$$

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 $1-\delta$,

$$\begin{aligned} \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} &\lesssim \varkappa\beta\rho_{\max}\sqrt{\frac{1}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{\|\widetilde{\boldsymbol{\Delta}}_0\|_{\widetilde{\boldsymbol{\Sigma}}}}{2^{k-1}} + \frac{3}{2}\right) \\ &=: \widetilde{R}_k \end{aligned} \tag{84}$$

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

$$\begin{aligned} &\widetilde{t}'_{\mathsf{seg}} \\ &:= \left(2 + \frac{1}{\log 2}\log\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}\right)\widetilde{t}_{\mathsf{seg}}, \end{aligned} \tag{85}$$

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

$$\|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{\Sigma}}} \lesssim \varkappa\beta\rho_{\max}\sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha\lambda_1}\log\frac{2dT}{\delta}}\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 2\right). \tag{86}$$

Recall that this bound holds for any $\varkappa \in (0,1)$ satisfying the conditions (72). Hence, Theorem 3 follows by taking $\varkappa =$

$8\rho_{\max}\sqrt{\frac{\alpha}{\lambda_1\beta\lambda_2}}$ and

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

VIII. 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 [12] 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.

Appendix 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*

$$\left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| = 1, \qquad \text{and} \qquad \left\|\boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}^\pi\boldsymbol{D}_\mu^{-\frac{1}{2}}\right\| = 1. \quad (87)$$

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

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

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

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

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

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

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 (12c) obey*

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac{1}{2}} \succeq (1-\gamma)^2\boldsymbol{I}, \quad (88a)$$

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-\frac{1}{2}} \succeq (1-\gamma)^2\boldsymbol{I}, \quad (88b)$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-2}, \quad (88c)$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-2}, \quad (88d)$$

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \le (1-\gamma)^{-1}, \quad (88e)$$

$$\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Phi}^\top\boldsymbol{D}_{\boldsymbol{\mu}}\right\| \le \max_{s\in\mathcal{S}}\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s), \quad (88f)$$

$$\|\boldsymbol{I}-\eta\boldsymbol{A}\| \le 1-\frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}), \quad \forall 0<\eta<\frac{1-\gamma}{4\|\boldsymbol{\Sigma}\|}, \quad (88g)$$

$$\|\boldsymbol{\Sigma}\| \le 1, \qquad \|\boldsymbol{\Sigma}^{-1}\| \ge 1, \quad (88h)$$

$$\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}\right\|_2 \le 1. \quad (88i)$$

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

a) Proof of Eqn. (88a) and (88b): 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

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

where

$$\widetilde{\boldsymbol{\Phi}} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-\frac{1}{2}} \qquad \text{and} \qquad \boldsymbol{P}_{\boldsymbol{D}_\mu} := \boldsymbol{D}_\mu^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{D}_\mu^{-\frac{1}{2}}. \quad (89)$$

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

$$\left| \boldsymbol{x}^\top \widetilde{\boldsymbol{\Phi}}^\top \boldsymbol{P}_{\boldsymbol{D}_\mu} \widetilde{\boldsymbol{\Phi}} \boldsymbol{x} \right| \leq \|\widetilde{\boldsymbol{\Phi}}^\top \boldsymbol{P}_{\boldsymbol{D}_\mu} \widetilde{\boldsymbol{\Phi}}\| \leq \|\widetilde{\boldsymbol{\Phi}}\|^2 \|\boldsymbol{P}_{\boldsymbol{D}_\mu}\| = 1. \quad (90)$$

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 (88a). The proof for (88b) follows from an identical argument and is omitted for brevity.

b) Proof of Eqn. (88c), (88d) *and* (88e)*:* With the above bounds in place, we can further obtain

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

where $\lambda_{\min}(\boldsymbol{B})$ denotes the smallest eigenvalue of $\boldsymbol{B}$, and the last inequality comes from (88b). This establishes (88c). The inequality (88d) 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 (88e).

c) Proof of Eqn. (88g)*:* Recalling that $\boldsymbol{\Sigma} = \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu \boldsymbol{\Phi}$, we can arrange terms to derive

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

where $\widetilde{\boldsymbol{\Phi}}$ and $\boldsymbol{P}_{\boldsymbol{D}_\mu}$ are defined in (89). 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 (90). 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

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

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

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

thus indicating that

$$\begin{aligned}\left\| \mathbf{I} - \eta \boldsymbol{A} \right\| &\leq \sqrt{\left\| \left(\mathbf{I} - \eta \boldsymbol{A}\right) \left(\mathbf{I} - \eta \boldsymbol{A}^\top\right) \right\|} \\ &\leq \sqrt{1 - \eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq 1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}).\end{aligned}$$

d) Proof of Eqn. (88h)*:* For any unit vector $\boldsymbol{u}$, the assumption $\max_s \|\boldsymbol{\phi}(s)\|_2 \leq 1$ guarantees that

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

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

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

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

e) Proof of Eqn. (88i)*:* Finally, we observe that

$$\begin{aligned}\left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{b} \right\|_2 &= \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu^{\frac{1}{2}} \boldsymbol{D}_\mu^{\frac{1}{2}} \boldsymbol{r} \right\|_2 \\ &\leq \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Phi}^\top \boldsymbol{D}_\mu^{\frac{1}{2}} \right\| \cdot \left\| \boldsymbol{D}_\mu^{\frac{1}{2}} \boldsymbol{r} \right\|_2 \\ &\overset{\text{(i)}}{\leq} \left\| \boldsymbol{D}_\mu^{\frac{1}{2}} \boldsymbol{r} \right\|_2 \leq 1\end{aligned}$$

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

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 $T \gtrsim \log\left(\frac{d}{\delta}\right)$. Then the vector $\boldsymbol{b}$ defined in (12c) obeys that, with probability exceeding $1 - \delta$,*

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

Proof. See Section C-E. □

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

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

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

Proof. See Section C-E. □

Appendix 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.

a) 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

$$P_{\boldsymbol{q}}(s' \,|\, s) = \begin{cases} q_s\mathbb{1}(s'=s) + \frac{1-q_s}{|\mathcal{S}|-d+1}\mathbb{1}(s'\geq d) & \text{for} \quad s<d; \\ \frac{\gamma}{|\mathcal{S}|-d+1}\mathbb{1}(s'\geq d) + \frac{1-q_{s'}}{d-1}\mathbb{1}(s'<d) & \text{for} \quad s\geq d. \end{cases} \tag{91}$$

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

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

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

$$\sum_{s=1}^{d-1}\mathbb{1}(q_s = q_+) = \sum_{s=1}^{d-1}\mathbb{1}(q_s = q_-) = (d-1)/2. \tag{92}$$

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

$$\mu(s) = \begin{cases} \frac{1}{2(d-1)} & \text{for} \quad s<d; \\ \frac{1}{2(|\mathcal{S}|-d+1)} & \text{for} \quad s\geq d. \end{cases} \tag{93}$$

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

$$\theta^\star(d) = V^\star(s) = \frac{1}{1-\gamma^2-\sum_{i=1}^{d-1}\frac{\gamma^2(1-q_i)^2}{(d-1)(1-\gamma q_i)}}, \tag{94}$$

$$\theta^\star(i) = V^\star(i) = \frac{\gamma(1-q_i)}{1-\gamma q_i}V^\star(s), \text{ for } s\geq d \text{ and } i<d. \tag{95}$$

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 (92).

b) 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

$$\max_s\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\} \asymp d. \tag{96}$$

Next, we move on to compute $\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}$. First notice that for $\varepsilon \leq \frac{c_1\gamma}{1-\gamma}$ with constant c_1 small enough, $(1-\gamma)^2\varepsilon \leq c_1\gamma(1-\gamma)$ 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

$$\begin{aligned} \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 &= \sum_{i=1}^{d-1}\frac{1}{2(d-1)}\theta^{\star 2}(i) + \frac{1}{2}\theta^{\star 2}(d) \qquad (97) \\ &= \sum_{i=1}^{d-1}\frac{1}{2(d-1)}\left[\frac{\gamma(1-q_i)}{1-\gamma q_i}V^\star(s)\right]^2 + \frac{1}{2}[V^\star(s)]^2 \\ &\asymp \sum_{i=1}^{d-1}\frac{1}{2(d-1)}\left[\frac{\gamma(1-\gamma)}{1-\gamma}\frac{1}{1-\gamma}\right]^2 + \frac{1}{2}\left[\frac{1}{1-\gamma}\right]^2 \\ &\asymp \frac{1}{(1-\gamma)^2}. \qquad (98) \end{aligned}$$

c) 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

$$\begin{aligned} \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}'^\star\|_{\boldsymbol{\Sigma}}^2 &\geq \sum_{s=1}^{d-1}|\theta^\star(s) - \theta'^\star(s)|^2\frac{1}{2(d-1)} \\ &\gtrsim \varepsilon^2\frac{1}{d-1}\sum_{s=1}^{d-1}\mathbb{1}(q_s \neq q_s'). \end{aligned}$$

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

$$\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}'^\star\|_{\boldsymbol{\Sigma}} \gtrsim \varepsilon. \tag{99}$$

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 [66], there exists a set $\mathcal{Q}$ such that

$$M := |\mathcal{Q}| \geq e^{d/16} \qquad \text{and} \qquad \sum_{s=1}^{d-1} \mathbb{1}(q_s \neq q_s') \geq \frac{d}{16} \tag{100}$$

for any $q, q' \in \mathcal{Q}$ obeying $q \neq q'$.

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 ([67]), the probability of interest thus satisfies

$$\begin{aligned}&\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big\|_{\boldsymbol{\Sigma}} \gtrsim \varepsilon\Big) \\ &\geq 1 - \frac{1}{\log M}\Big(\frac{1}{M^2}\sum_{j,k=1}^{M} \mathsf{KL}(\mathbb{P}_{\boldsymbol{q}^j}^T \,\|\, \mathbb{P}_{\boldsymbol{q}^k}^T) + \log 2\Big),\end{aligned} \tag{101}$$

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. (91)). More specifically, given $s \sim \mu_q$ and $s'|s \sim P_{\boldsymbol{q}}(s'|s)$, one has

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

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

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

As a result, we have

$$\mathsf{KL}\big(\mathbb{P}_{\boldsymbol{q}}^T \,\|\, \mathbb{P}_{\boldsymbol{q}'}^T\big) \lesssim \varepsilon^2(1-\gamma)^3 T. \tag{102}$$

Substituting the above relation into (101) gives

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

To prove Theorem 2, it is enough to take the above together with relations (96) and (98).

Appendix C
Proofs of auxiliary lemmas and claims

A. 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 $\{(s_j, s_j')\}_{j\leq i}$ (more formally, $\mathbb{E}_i[\cdot]$ represents the expectation conditioned on the filtration $\mathcal{F}_i$ — the σ-algebra generated by $\{(s_j, s_j')\}_{j\leq i}$). 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

$$\begin{aligned}W &:= \sum_{i=l}^{u} \mathbb{E}_{i-1}\Big[\big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\big\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\Big], \quad \text{and} \\ B &:= \max_{i:l\leq i\leq u} \big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i \mathbb{1}\{\mathcal{H}_i\}\big\|_2,\end{aligned} \tag{103}$$

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

$$\alpha := \Big(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\Big)^{t-u-1}. \tag{104}$$

a) Control of W: Direct calculations yield

$$\begin{aligned}W &= \sum_{i=l}^{u} \mathbb{E}_{i-1}\Big[\boldsymbol{\xi}_i^\top (\mathbf{I} - \eta\boldsymbol{A}^\top)^{t-i-1}\boldsymbol{\Sigma}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i \mathbb{1}\{\mathcal{H}_i\}\Big] \\ &\leq \sum_{i=l}^{u} \big\|(\mathbf{I} - \eta\boldsymbol{A}^\top)^{t-i-1}\boldsymbol{\Sigma}(\mathbf{I} - \eta\boldsymbol{A})^{t-i-1}\big\| \\ &\quad \cdot \mathbb{E}_{i-1}\big[\|\boldsymbol{\xi}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] \\ &\overset{\text{(i)}}{\leq} \sum_{i=l}^{u} \|\boldsymbol{\Sigma}\| \left(1 - \frac{1}{2}\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\right)^{2t-2i-2} \\ &\quad \cdot 2 \max_{i:l\leq i\leq u} \Big\{\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i - \boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big] \\ &\quad + \mathbb{E}_{i-1}\big[\|\boldsymbol{b}_i - \boldsymbol{b}\|_2^2\big]\Big\} \\ &\overset{\text{(ii)}}{\leq} \frac{4\|\boldsymbol{\Sigma}\|\alpha^2}{\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})} \cdot \max_{i:l\leq i\leq u} \Big\{\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i - \boldsymbol{A})\boldsymbol{\theta}_i\|_2^2 \mathbb{1}\{\mathcal{H}_i\}\big]\end{aligned}$$

$$+\mathbb{E}_{i-1}\big[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2\big]\Big\}, \tag{105}$$

where (i) follows from the property (88g) (together with the assumption $\eta<(1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$) and the elementary inequality $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2\le 2\|\boldsymbol{a}\|_2^2+2\|\boldsymbol{b}\|_2^2$, and (ii) uses the elementary upper bound for the sum of geometric series as well as the definition (104) 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

$$\begin{aligned}
&\mathbb{E}_{i-1}\big[\|(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\|_2^2\mathbb{1}\{\mathcal{H}_i\}\big]\\
&\le \mathbb{E}_{i-1}\big[\|\boldsymbol{A}_i\boldsymbol{\theta}_i\|_2^2\mathbb{1}\{\mathcal{H}_i\}\big]\\
&=\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)\boldsymbol{\phi}(s_i)^\top\\
&\qquad\boldsymbol{\phi}(s_i)\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\\
&\le \max_s\|\boldsymbol{\phi}(s)\|_2^2\cdot\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)\\
&\qquad\big(\boldsymbol{\phi}(s_i)-\gamma\boldsymbol{\phi}(s_i')\big)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\\
&\overset{\text{(i)}}{\le} 2\max_s\|\boldsymbol{\phi}(s)\|_2^2\Big(\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\phi}(s_i)\boldsymbol{\phi}(s_i)^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\\
&\qquad+\gamma^2\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\phi}(s_i')\boldsymbol{\phi}(s_i')^\top\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\Big)\\
&\overset{\text{(ii)}}{=} 2\max_s\|\boldsymbol{\phi}(s)\|_2^2\Big(\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\\
&\qquad+\gamma^2\mathbb{E}_{i-1}\big[\boldsymbol{\theta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\}\big]\Big)\\
&\overset{\text{(iii)}}{\le} 4\|\boldsymbol{\theta}_i\|_{\boldsymbol{\Sigma}}^2\mathbb{1}\{\mathcal{H}_i\}\le 4\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big)^2\mathbb{1}\{\mathcal{H}_i\}\\
&\le 4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R)^2,
\end{aligned} \tag{106}$$

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 $\max_s\|\boldsymbol{\phi}(s)\|_2\le 1$, and the last inequality results from the definition (50) of the event $\mathcal{H}_i$. Similarly, one can derive

$$\mathbb{E}_{i-1}[\|\boldsymbol{b}_i-\boldsymbol{b}\|_2^2]\le\mathbb{E}_{i-1}[\|\boldsymbol{b}_i\|_2^2]=\mathbb{E}_{i-1}\big[\|\boldsymbol{\phi}(s_i)r(s_i)\|_2^2\big]\le 1, \tag{107}$$

where the last inequality holds since $\max_s\|\boldsymbol{\phi}(s)\|_2\le 1$ and $\max_s|r(s)|\le 1$. Substitution into (105) yields

$$W\le\frac{4\kappa}{\eta(1-\gamma)}\alpha^2\Big\{4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R)^2+1\Big\}=:W_{\max}. \tag{108}$$

b) Control of B: By definition of B, one can write

$$\begin{aligned}
B&=\max_{i:l\le i\le u}\big\|\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\}\\
&=\max_{i:l\le i\le u}\big\|\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\}\\
&\le\|\boldsymbol{\Sigma}\|\max_{i:l\le i\le u}\big\|\mathbf{I}-\eta\boldsymbol{A}\big\|^{t-i-1}\cdot\max_{i:l\le i\le u}\big\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{\xi}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\}\\
&\le\alpha\|\boldsymbol{\Sigma}\|\max_{i:l\le i\le u}\Big\{\big\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2\mathbb{1}\{\mathcal{H}_i\}\\
&\qquad+\big\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{b}_i-\boldsymbol{b})\big\|_2\Big\},
\end{aligned} \tag{109}$$

where the last step results from (88g) (with the restriction that $\eta<(1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$) and the definition (104) of α. It then suffices to control the two terms on the right-hand side of (109). To begin with, we have

$$\begin{aligned}
&\big\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2\\
&\le\big\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Sigma}^{-\frac12}\big\|\|\boldsymbol{\theta}_i\|_{\boldsymbol{\Sigma}}\\
&\le\Big(\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{A}_i\boldsymbol{\Sigma}^{-\frac12}\|+\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac12}\|\Big)\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big).
\end{aligned}$$

Recall from (146) that $\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{A}_i\boldsymbol{\Sigma}^{-\frac12}\|\le 2\max_s\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(s)\|_2^2$, and similarly $\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{A}\boldsymbol{\Sigma}^{-\frac12}\|\le 2\max_s\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\phi}(s)\|_2^2$. We then have

$$\begin{aligned}
&\big\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\big\|_2\\
&\le 4\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big).
\end{aligned} \tag{110}$$

Regarding the second term of (109), direct calculations give

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-\frac12}(\boldsymbol{b}_i-\boldsymbol{b})\|_2^2\\
&\le 2\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{b}_i\|_2^2+2\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{b}\|_2^2\\
&=2\big\|\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{\phi}(s_i)r(s_i)\big\|_2^2+2\big\|\boldsymbol{\Sigma}^{-\frac12}\mathbb{E}_{s\sim\mu}\big[\boldsymbol{\phi}(s)r(s)\big]\big\|_2^2\\
&\le 4\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\max_s|r(s)|^2\\
&\le 4\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}.
\end{aligned} \tag{111}$$

Substituting the preceding two bounds into (109), we arrive at

$$\begin{aligned}
B&\le 4\alpha\|\boldsymbol{\Sigma}\|\Big(\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\max_{i:i<t}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\big)\\
&\qquad\mathbb{1}\{\mathcal{H}_i\}+\sqrt{\max_s\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}}\Big)\\
&\le 4\alpha\|\boldsymbol{\Sigma}\|\Big(\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)\\
&\qquad+\sqrt{\max_s\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\}}\Big)\\
&\le 4\alpha\|\boldsymbol{\Sigma}\|\max_s\big\{\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\big\}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)\\
&\le 4\alpha\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)\\
&=4\kappa\alpha\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)=:B_{\max}.
\end{aligned} \tag{112}$$

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. (148)).

c) Invoking matrix Freedman's inequality: Equipped with the above bounds (108) and (112), we are ready to apply Freedman's inequality [68, Corollary 1.3] (or a version in [19, Section A]), which asserts that

$$\begin{aligned}
&\Big\|\sum_{i=0}^{t-1}(\mathbf{I}-\eta\boldsymbol{A})^{t-i-1}\widetilde{\boldsymbol{\xi}}_i\Big\|_{\boldsymbol{\Sigma}}\\
&\le 2\sqrt{W_{\max}\log\frac{2dT}{\delta}}+\frac43 B_{\max}\log\frac{2dT}{\delta}\\
&=\alpha\cdot\Bigg\{2\sqrt{\frac{4\kappa}{\eta(1-\gamma)}\Big\{4(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R)^2+1\Big\}\log\frac{2dT}{\delta}}\\
&\quad+\frac{16\kappa}{3}\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1\big)\log\frac{2dT}{\delta}\Bigg\}\\
&\le 16(1-\frac12\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}))^{t-u-1}\\
&\quad(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R+1)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}}
\end{aligned} \tag{113}$$

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 (88h)).

B. Proof of the inequalities (64a) and (64c)

a) Proof of the inequality (64a): Combining the triangle inequality with the definition (63) ensures that

$$\begin{aligned}
&\big\|\boldsymbol{A}_0^{(T+1)}\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\
&\le\|\boldsymbol{A}^{-1}\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}+\|\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T+1}\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}\\
&=\big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Delta}_0\big\|_2\\
&+\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Sigma}^{-\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{T+1}\boldsymbol{\Sigma}^{-\frac12}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Delta}_0\|_2\\
&\le\big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-1}\big\|\cdot\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\
&+\big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-\frac12}\big\|^2\cdot\|\mathbf{I}-\eta\boldsymbol{A}\|^{T+1}\cdot\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\
&\le\frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\left\{1+\left(1-\frac12\eta(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\right)^{T+1}\right\}\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}\\
&\le\frac{2\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\big\|\boldsymbol{\Delta}_0\big\|_{\boldsymbol{\Sigma}}
\end{aligned}\tag{114}$$

as claimed. Here, the second to last step follows from (88e) and (88g), provided that $\eta\le(1-\gamma)/(4\|\boldsymbol{\Sigma}\|)$.

b) Proof of the inequality (64c): Again, the triangle inequality together with the definition (63) yields

$$\begin{aligned}
&\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}_i^{(T)}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\\
&\le\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}+\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\\
&\le\Big\|\boldsymbol{A}^{-1}\sum_{i=0}^{T-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_{\boldsymbol{\Sigma}}+\Big\|\boldsymbol{A}^{-1}\sum_{i=0}^{T-1}(\boldsymbol{b}_i-\boldsymbol{b})\Big\|_{\boldsymbol{\Sigma}}\\
&+\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}},
\end{aligned}\tag{115}$$

leaving us with three terms to handle. Here in the second line, we substitute in the definition of $\boldsymbol{\xi}_i$ (45).

- The second term of (115) can be bounded by Lemma 6, which asserts that
$$\frac{1}{T}\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}\big(\boldsymbol{b}_i-\boldsymbol{b}\big)\Big\|_{\boldsymbol{\Sigma}}\lesssim\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\Big(\frac{d}{\delta}\Big)}\tag{116}$$
holds with probability at least $1-\delta$, as long as $T\gtrsim\log\frac{d}{\delta}$.
- For the third term of (115), invoking the property (88e) again yields
$$\begin{aligned}
&\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\\
&=\Big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\Sigma}^{-1}\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_2\\
&\le\big\|\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac12}\big\|\cdot\big\|\boldsymbol{\Sigma}^{-1}\big\|\cdot\Big\|\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_2\\
&\le\frac{\|\boldsymbol{\Sigma}^{-1}\|}{1-\gamma}\Big\|\sum_{i=0}^{T-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}.
\end{aligned}\tag{117}$$
Repeating the same analysis as in Step 3 to see that
$$\Big\|\sum_{i=0}^{T-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\le16(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+3)\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)}}\tag{118}$$
with probability at least $1-\delta$. Substitution into (117) leads to
$$\begin{aligned}
&\Big\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\mathbf{I}-\eta\boldsymbol{A})^{T-i}\boldsymbol{\xi}_i\Big\|_{\boldsymbol{\Sigma}}\\
&\le16\big(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+3\big)\|\boldsymbol{\Sigma}^{-1}\|\sqrt{\frac{\kappa\log\frac{2dT}{\delta}}{\eta(1-\gamma)^3}}.
\end{aligned}\tag{119}$$
- It then boils down to bounding the first term of (115). In light of (60), we decompose it as follows
$$\begin{aligned}
&\Big\|\frac1T\sum_{i=0}^{T-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2\\
&\le\Big\|\frac1T\sum_{i=0}^{\widetilde{t}_{\mathsf{seg}}-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2\\
&+\Big\|\frac1T\sum_{i=\widetilde{t}_{\mathsf{seg}}}^{T-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2.
\end{aligned}\tag{120}$$
Bounding these terms requires the following lemma, whose proof is deferred to Section C-F.

Lemma 8. *Fix any $R>0$ and define a collection of auxiliary random vectors for $0\le i\le T-1$*
$$\boldsymbol{\theta}_i'\coloneqq\boldsymbol{\theta}_i\mathbb{1}\{\mathcal{H}_i\},\qquad\mathcal{H}_i\coloneqq\big\{\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}\le R\big\},\tag{121}$$
Then for any indices (l,u,t) obeying $0\le l\le u\le T-1$, one has with probability at least $1-\delta$ that
$$\begin{aligned}
&\Big\|\frac{1}{u-l+1}\sum_{i=l}^{u}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\Big\|_2\\
&\le\frac{16\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}+R\big)}{1-\gamma}\sqrt{\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{u-l+1}}
\end{aligned}\tag{122}$$
provided that
$$u-l+1\ge\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}.$$
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
$$\Big\|\frac{1}{t'_{\mathsf{seg}}}\sum_{i=0}^{t'_{\mathsf{seg}}-1}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\Big\|_2$$

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

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,

$$\Big\| \frac{1}{T} \sum_{i=0}^{T-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A}^{-1} (\boldsymbol{A}_i - \boldsymbol{A}) \boldsymbol{\theta}_i \Big\|_2$$
$$\le \frac{32\big(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} + \sqrt{\frac{t'_{\mathsf{seg}}}{T}} R_0 + 1\big)}{1-\gamma} \sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\phi}(s) \log \frac{2d}{\delta}}{T}}. \tag{123}$$

Combining the preceding bounds (116), (119) and (123) with (115), we reach the conclusion that with probability of at least $1-\delta$,

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

as long as $T \ge t'_{\mathsf{seg}} \kappa \|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2$, where we use the definition (47) of R_0. It thus establishes the inequality (64c).

C. 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

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

In addition, some direct algebra suggests

$$\Big\| \frac{1}{\kappa}\alpha\gamma \boldsymbol{\Pi}^\top \Big\| \le \frac{\alpha\gamma\rho_{\max}}{\kappa},$$
$$\|\kappa\alpha(\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi}) \widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{A}}\| \le \kappa\alpha(1 + \gamma\lambda_\Sigma\rho_{\max})\lambda_\Sigma(2\rho_{\max})^2, \quad \text{and}$$
$$\|\alpha\gamma(\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Pi}) \boldsymbol{\Pi}^\top\| \le \alpha\gamma(\rho_{\max} + \gamma\lambda_\Sigma\rho_{\max}^2).$$

In summary, as long as

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

one has

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

D. Proof of Lemma 3

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

$$\widetilde{W} := \sum_{i=0}^{t-1} \mathbb{E}_{i-1}\Big[\big\|\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i\big\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \mathbb{1}\Big\{\widetilde{\mathcal{H}}_i\Big\}\Big], \quad \text{and}$$
$$\widetilde{B} := \max_{i:0\le i\le t-1} \Big\|\boldsymbol{\Psi}^{t-i-1}\boldsymbol{\zeta}_i \mathbb{1}\Big\{\widetilde{\mathcal{H}}_i\Big\}\Big\|_{\widetilde{\boldsymbol{\Sigma}}}. \tag{124}$$

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.

a) Controlling the norm of $\boldsymbol{\zeta}_i$: We firstly observe that since $\|\boldsymbol{\phi}(s)\|_2 \le 1$ and $r(s) \le 1$ for all $s \in \mathcal{S}$, with similar logic as (106), (107), (110) and (111), 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

$$\mathbb{E}_{i-1}\Big\|(\widetilde{\boldsymbol{A}}_i - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_i\Big\|_2^2 \le 4\rho_{\max}^2\Big(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2\Big), \quad \text{and} \tag{125}$$

$$\Big\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}(\widetilde{\boldsymbol{A}}_i - \widetilde{\boldsymbol{A}})\widetilde{\boldsymbol{\theta}}_i\Big\|_2 \le 4\rho_{\max}\max_s\Big\{\boldsymbol{\phi}(s)\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\phi}(s)\Big\}\Big(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}\Big); \tag{126}$$

- 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

$$\mathbb{E}_{i-1} \left\| (\boldsymbol{\Pi}_i - \boldsymbol{\Pi})^\top \boldsymbol{z}_i \right\|_2^2 \le \rho_{\max}^2 \|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \quad \text{and} \tag{127}$$

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} (\boldsymbol{\Pi}_i - \boldsymbol{\Pi})^\top \boldsymbol{z}_i \right\|_2 \le 2\rho_{\max} \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}; \tag{128}$$

- 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

$$\mathbb{E}_{i-1} \left\| \left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{\Sigma}} \right) \boldsymbol{z}_i \right\|_2^2 \le \|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \quad \text{and} \tag{129}$$

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \left(\widetilde{\boldsymbol{\Sigma}}_i - \widetilde{\boldsymbol{\Sigma}} \right) \boldsymbol{z}_i \right\|_2 \le 2 \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \|\boldsymbol{z}_i\|_{\widetilde{\boldsymbol{\Sigma}}}; \tag{130}$$

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

$$\mathbb{E}_{i-1} \left\| \widetilde{\boldsymbol{b}}_i - \widetilde{\boldsymbol{b}} \right\|_2^2 \le \rho_{\max}^2, \quad \text{and} \tag{131}$$

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \left(\widetilde{\boldsymbol{b}}_i - \widetilde{\boldsymbol{b}} \right) \right\|_2^2 \le 4\rho_{\max}^2 \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\}. \tag{132}$$

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

$$\mathbb{E}_{i-1} \|\boldsymbol{\nu}_i\|_2^2 \lesssim \rho_{\max}^2 \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right) + 1 + \gamma^2 \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right], \tag{133}$$

and

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\nu}_i \right\|_2 \lesssim \rho_{\max} \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right) + \gamma \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}} + 1 \right]; \tag{134}$$

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

$$\mathbb{E}_{i-1} \|\boldsymbol{\eta}_i\|_2^2 \lesssim \rho_{\max}^2 \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right) + 1 \right] + \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2, \tag{135}$$

and

$$\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\eta_i} \right\|_2 \lesssim \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \left\{ \rho_{\max} \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right) + 1 \right] + \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right\}. \tag{136}$$

By combining (133) and (135) with the definition of $\boldsymbol{\zeta}_i$ (73), we obtain the following bound:

$$\begin{aligned}
&\mathbb{E}_{i-1} \|\boldsymbol{\zeta}_i\|_2^2 \\
&\lesssim \alpha^2 \mathbb{E}_{i-1} \|\boldsymbol{\nu}_i\|_2^2 + \varkappa^2 \alpha^2 \|\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Pi}\|^2 \mathbb{E}_{i-1} \|\boldsymbol{\nu}_i\|_2^2 \\
&\quad + \varkappa^2 \beta^2 \mathbb{E}_{i-1} \|\boldsymbol{\eta}_i\|_2^2 \\
&\lesssim \alpha^2 \left(1 + \varkappa^2 (1 + \gamma \lambda_\Sigma \rho_{\max})^2 \right) \cdot \rho_{\max}^2 \\
&\quad \left[4 \left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right) + 1 + \gamma^2 \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right] \\
&\quad + \varkappa^2 \beta^2 \cdot \left\{ \rho_{\max}^2 \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right) + 1 \right] + \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 \right\} \\
&\lesssim \varkappa^2 \beta^2 \rho_{\max}^2 \left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + \frac{1}{\varkappa^2} \|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}}^2 + 1 \right),
\end{aligned} \tag{138}$$

and

$$\begin{aligned}
&\left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\zeta}_i \right\|_2 \\
&\lesssim \alpha \|\widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\nu}_i\|_2 + \alpha \varkappa \|\boldsymbol{I} - \gamma \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Pi}\| \|\widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\nu}_i\|_2 \\
&\quad + \varkappa \beta \left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \eta_i \right\|_2 \\
&\lesssim \alpha \left(1 + \varkappa (1 + \gamma \lambda_\Sigma \rho_{\max}) \right) \cdot \rho_{\max} \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \\
&\quad \cdot \left[2 \left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right) + \gamma \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}} + 1 \right] \\
&\quad + \varkappa \beta \cdot \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \\
&\quad \cdot \left\{ \rho_{\max} \left[\left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \|\widetilde{\boldsymbol{\Delta}}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right) + 1 \right] + \|\boldsymbol{w}_i\|_{\widetilde{\boldsymbol{\Sigma}}} \right\} \\
&\lesssim \varkappa \beta \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} \left(\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{2}{\varkappa} \|\boldsymbol{x}_i\|_{\widetilde{\boldsymbol{\Sigma}}} + 1 \right)
\end{aligned} \tag{139}$$

b) Control of $\widetilde{W}$ and $\widetilde{B}$: With the norm of $\widetilde{\boldsymbol{\zeta}}_i$ bounded, we can apply similar techniques as in equations (105), (108), (109) and (112) of Section C-A to construct the following bound for $\widetilde{W}$:

$$\begin{aligned}
\widetilde{W} &\le \|\widetilde{\boldsymbol{\Sigma}}\| \sum_{i=0}^{t-1} \|\boldsymbol{\Psi}^{t-i-1}\|^2 \cdot \mathbb{E}_{i-1} \left[\|\boldsymbol{\zeta}_i\|_2^2 \mathbb{1} \left\{ \widetilde{\mathcal{H}}_i \right\} \right] \\
&\lesssim \|\widetilde{\boldsymbol{\Sigma}}\| \sum_{i=0}^{t-1} (1 - \frac{1}{2} \alpha \lambda_\theta)^{2t-2i-2} \\
&\quad \cdot \varkappa^2 \beta^2 \rho_{\max}^2 (2 \|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + 2\widetilde{R} + 1)^2 \\
&\lesssim \frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha \lambda_\theta} \varkappa^2 \beta^2 \rho_{\max}^2 (\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{1}{\varkappa} \widetilde{R} + 1)^2,
\end{aligned} \tag{140}$$

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

$$\begin{aligned}
\widetilde{B} &\le \|\widetilde{\boldsymbol{\Sigma}}\| \max_{i:0 \le i \le t-1} \left\| \widetilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\zeta}_i \mathbb{1} \{ \mathcal{H}_i \} \right\|_2 \\
&\lesssim \|\widetilde{\boldsymbol{\Sigma}}\| \varkappa \beta \rho_{\max} \max_s \left\{ \boldsymbol{\phi}(s) \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\phi}(s) \right\} (\|\widetilde{\boldsymbol{\theta}}^\star\|_2 + \widetilde{R} + 1) \\
&=: \widetilde{B}_{\max}.
\end{aligned} \tag{141}$$

c) Invoking the matrix Freedman's inequality: With $\widetilde{W}$ and $\widetilde{B}$ bounded, we again invoke the matrix Freedman's inequality [68, Corollary 1.3] to assert that

$$\begin{aligned}
&\left\| \sum_{i=0}^{t-1} \boldsymbol{\Psi}^{t-i-1} \widetilde{\boldsymbol{\zeta}}_i \right\|_2 \\
&\le 2 \sqrt{\widetilde{W}_{\max} \log \frac{2dT}{\delta}} + \frac{4}{3} \widetilde{B}_{\max} \log \frac{2dT}{\delta} \\
&\lesssim \sqrt{\frac{\|\widetilde{\boldsymbol{\Sigma}}\|}{\alpha \lambda_\theta} \log \frac{2dT}{\delta}} \varkappa \beta \rho_{\max} (\|\widetilde{\boldsymbol{\theta}}^\star\|_{\widetilde{\boldsymbol{\Sigma}}} + \frac{1}{\varkappa} \widetilde{R} + 1)
\end{aligned} \tag{142}$$

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

E. Proof of Lemma 6 and Lemma 7

a) 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 [69]. Note that

$$\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\widehat{\boldsymbol{A}}\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}=\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\boldsymbol{A}_t\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}}_{=:\boldsymbol{Z}_t}. \tag{143}$$

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]\\
&=\mathbb{E}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\boldsymbol{A}_t\big)\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{A}-\boldsymbol{A}_t\big)^\top\big(\boldsymbol{A}^\top\big)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&\preceq\mathbb{E}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_t^\top\big(\boldsymbol{A}^\top\big)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&=\mathop{\mathbb{E}}_{s\sim\mu,\,s'\sim P(\cdot|s)}\Big[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)^\top\boldsymbol{\Sigma}^{-1}\\
&\qquad\qquad\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)\boldsymbol{\phi}(s)^\top\big(\boldsymbol{A}^\top\big)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big]\\
&\preceq\max_{s,s'}\Big\{\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)^\top\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\phi}(s)-\gamma\boldsymbol{\phi}(s')\big)\Big\}\\
&\cdot\mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\big(\boldsymbol{A}^\top\big)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\
&\preceq\max_{s,s'}\big\{2\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)+2\gamma^2\boldsymbol{\phi}(s')^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s')\big\}\\
&\cdot\Big\{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\big(\boldsymbol{A}^\top\big)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big\}\\
&\preceq\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\mathbf{I},
\end{aligned}\tag{144}$$

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

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

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 $\max_s\big\{\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\big\}$, which can be accomplished as follows

$$\begin{aligned}
&\boldsymbol{\phi}(s)^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\phi}(s)\\
&=\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\Big\{\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big\}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\\
&\le\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2\cdot\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{A}^\top)^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\big\|\\
&\le\frac{\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}.
\end{aligned}$$

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

$$\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]\preceq\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{(1-\gamma)^2}\mathbf{I}. \tag{145}$$

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

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

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

$$\begin{aligned}
\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|&\le\max_s\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|\\
&\quad+\gamma\max_{s,s'}\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s')\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|\\
&\le 2\max_s\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2.
\end{aligned}\tag{146}$$

This combined with the preceding bounds yields

$$\begin{aligned}
\|\boldsymbol{Z}_t\|&\le\frac{2\max_s\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2}{1-\gamma}+1\\
&\le\frac{4\max_s\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2}{1-\gamma}\\
&=\frac{4\max_s\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{1-\gamma}.
\end{aligned}\tag{147}$$

Here, the inequality follows since

$$\begin{aligned}
\max_s\big\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\big\|_2^2&\ge\mathop{\mathbb{E}}_{s\sim\mu}\left[\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\right]\\
&=\mathop{\mathbb{E}}_{s\sim\mu}\left[\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^\top\big)\right]\\
&=\mathrm{tr}(\mathbf{I}_d)=d\ge 1.
\end{aligned}\tag{148}$$

Step 4: invoking the matrix Bernstein inequality. With the above bounds in mind, we are ready to apply the matrix Bernstein inequality [69] to obtain that: with probability at least $1-\delta$ one has

$$\big\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\widehat{\boldsymbol{A}}\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}\big\|$$

$$
\begin{aligned}
&\lesssim \sqrt{\frac{1}{T^2}\sum_{t=0}^{T-1}\max\left\{\left\|\mathbb{E}\left[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\right]\right\|,\ \left\|\mathbb{E}\left[\boldsymbol{Z}_t^\top\boldsymbol{Z}_t\right]\right\|\right\}\log\left(\frac{d}{\delta}\right)} \\
&+\frac{\max_t\|\boldsymbol{Z}_t\|\log(\frac{d}{\delta})}{T}. \\
&\overset{\text{(i)}}{\lesssim} \sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\left(\frac{d}{\delta}\right)} \\
&+\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log(\frac{d}{\delta})}{T(1-\gamma)} \\
&\overset{\text{(ii)}}{\asymp} \sqrt{\frac{\max_s \boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)}{T(1-\gamma)^2}\log\left(\frac{d}{\delta}\right)}. \qquad (149)
\end{aligned}
$$

Here, (i) results from the bounds (144), (145) and (147), 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 $\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\big(\boldsymbol{A}-\widehat{\boldsymbol{A}}\big)\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\| < 1$. 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 $\boldsymbol{A}^{-1}\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{u}=\boldsymbol{0}$, which in turn implies

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

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

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

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

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

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

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,

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

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 [69] yields

$$
\begin{aligned}
&\left\|\boldsymbol{A}^{-1}\big(\widehat{\boldsymbol{b}}-\boldsymbol{b}\big)\right\|_{\boldsymbol{\Sigma}} \\
&=\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_t\right\|_2 \\
&\lesssim\sqrt{\frac{1}{T^2}\sum_{t=0}^{T-1}\mathbb{E}\big[\boldsymbol{z}_t^\top\boldsymbol{z}_t\big]\log\left(\frac{d}{\delta}\right)}+\frac{1}{T}\max_t\left\|\boldsymbol{z}_t\right\|_2\log\left(\frac{d}{\delta}\right) \\
&\lesssim\frac{\max_{s\in\mathcal{S}}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2}{1-\gamma}\sqrt{\frac{1}{T}\log\left(\frac{d}{\delta}\right)} \\
&+\frac{\max_{s\in\mathcal{S}}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2}{1-\gamma}\cdot\frac{1}{T}\log\left(\frac{d}{\delta}\right) \\
&\asymp\frac{\max_{s\in\mathcal{S}}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\phi}(s)\right\|_2}{1-\gamma}\sqrt{\frac{1}{T}\log\left(\frac{d}{\delta}\right)} \qquad (150)
\end{aligned}
$$

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

F. 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 $\left\|\frac{1}{u-l+1}\sum_{i=l}^{u}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\right\|_2$ via matrix Freedman's inequality. The key is to control the following quantities (here, we abuse notation whenever it is clear from context):

$$
\begin{aligned}
W&:=\sum_{i=l}^{u}\mathbb{E}_{i-1}\left[\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\right\|_2^2\right]\quad\text{and} \\
B&:=\max_{i:l\leq i\leq u}\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i'\right\|_2. \qquad (151)
\end{aligned}
$$

a) Control of B: To begin with, observe that

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

where the second to last inequality comes from (147) and the triangle inequality, and the last line is due to the definition of $\mathcal{H}_i$.

b) Control of W: Moreover, one can derive

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

where the first inequality arises from (145), 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 [68, 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 $u-l+1 \ge \frac{4\max_s \boldsymbol{\phi}(s)^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s)\log\frac{2d}{\delta}}{9}$.

Appendix D
Comparisons with previous works

A. Comparisons with [21]

[21] 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.

a) Iterative relation on $\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\Sigma}^2$: 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

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

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:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\Delta}_{t+1}\|_{\boldsymbol{\Sigma}}^2 &= \mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2-\underbrace{2\eta_t\mathbb{E}[\boldsymbol{\Delta}_t^{\top}(\boldsymbol{A}^{\top}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{A})\boldsymbol{\Delta}_t]}_{I_1} \\
&\quad +\underbrace{2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2}_{I_2}+\underbrace{2\eta_t^2\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^{\star}-\boldsymbol{b}_t\|_{\boldsymbol{\Sigma}}^2}_{I_3}. \quad (152)
\end{aligned}$$

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

b) 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 (88e) 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

$$I_1 \ge 2\eta_t(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2. \quad (153)$$

c) 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 $I_2 \le \frac{1}{2}I_1$.

d) 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

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

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

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

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

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

Consequently, I_3 can be upper bounded by

$$I_3 \le 6\eta_t^2\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 1\right). \tag{154}$$

e) Bounding $\mathbb{E}\|\boldsymbol{\Delta}_T\|_{\boldsymbol{\Sigma}}^2$: By combining (152), (153) and (154) and recalling that $I_2 \le \frac{1}{2}I_1$ when η_t is sufficiently small, we obtain

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Delta}_{t+1}\|_{\boldsymbol{\Sigma}}^2 \le\ & (1-(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta_t)\mathbb{E}\|\boldsymbol{\Delta}_t\|_{\boldsymbol{\Sigma}}^2 \\ & + 6\eta_t^2\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 1\right).\end{aligned} \tag{155}$$

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

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

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

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

This implies the following upper bound for the sample complexity:

$$T \asymp \frac{\kappa\|\boldsymbol{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}, \tag{156}$$

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

B. Comparisons with [9]

Theorem 2(c) in [9] shows that with decaying stepsizes $\eta_t = \frac{\beta}{\lambda+t}$ where

$$\beta = \frac{2\|\boldsymbol{\Sigma}^{-1}\|}{(1-\gamma)}, \quad \lambda = \frac{16\|\boldsymbol{\Sigma}^{-1}\|}{(1-\gamma)^2}, \tag{157}$$

the expected ℓ_2 norm of TD estimation error is bounded by

$$\mathbb{E}\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^\star\|_2^2 \le \frac{\nu}{\lambda+T}, \tag{158}$$

where

$$\nu = \max\left\{\frac{8\sigma^2\|\boldsymbol{\Sigma}^{-2}\|}{(1-\gamma)^2}, \frac{16\|\boldsymbol{\theta}^\star\|_2^2\|\boldsymbol{\Sigma}^{-1}\|}{(1-\gamma)^2}\right\}. \tag{159}$$

- Suppose the maximum is attained at the second term for ν and T is sufficiently large, (158) is simplified as

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

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

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

  which implies the following sample complexity:

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

- Suppose that the first term on the right hand side of expression (159) is larger, (158) can be simplified as

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

  Then similarly, the sample complexity is

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

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

In the worst-case scenario, it satisfies $\sigma^2 \asymp \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + 1$. Therefore, the sample complexity implied by Theorem 2(c) of [9] scales as

$$T \asymp \frac{\kappa\|\boldsymbol{\Sigma}^{-1}\|\left(\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)^2}\frac{1}{\varepsilon^2}. \tag{160}$$

C. Comparison with [63] and [64]

[63] studied a more general problem of linear approximation for fixed point equations in Hilbert spaces, and considered its application to TD learning with linear function approximation and *i.i.d.* samples. A similar result was reached by [64] in their Theorem 2 and Theorem 3. While these works explored both the *approximation error*, which measures the difference between $\boldsymbol{\Phi}\boldsymbol{\theta}^\star$ and $\boldsymbol{V}^\star$ under our notation, and the *statistical error*, which measures the convergence of $\bar{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}^\star$, it is the latter that is directly comparable to our results. Therefore, we hereby provide a comparison between the statistical error term

in their Corollary 5 and the sample complexity result of ours as shown in Theorem 1. Translated to our notation, [63] proved that with a sufficiently large sample size T and a stepsize of $\eta \asymp \frac{1}{\sqrt{T}}$, the estimation error of the averaged TD learning algorithm satisfies

$$\begin{aligned}&\mathbb{E}_{s\sim\mu}[V_{\bar{\boldsymbol{\theta}}_T}(s) - V_{\boldsymbol{\theta}^\star}(s)]^2\\ &= \|\bar{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}\\ &\lesssim \frac{1}{T}\mathsf{Tr}\left[(\boldsymbol{I}-\boldsymbol{M})^{-1}(\boldsymbol{\Sigma}_L+\boldsymbol{\Sigma}_b)(\boldsymbol{I}-\boldsymbol{M})^{-\top}\right], \qquad (161)\end{aligned}$$

in which $\boldsymbol{M}$, $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_b$ are defined as

$$\begin{aligned}&\boldsymbol{M} = \gamma\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}_{s\sim\mu,s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top]\boldsymbol{\Sigma}^{-\frac{1}{2}},\\ &\boldsymbol{\Sigma}_L = \mathbf{Cov}_{s_t\sim\mu,s_t'\sim P(\cdot|s_t)}[\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{A}_t\boldsymbol{\theta}^\star], \quad \text{and}\\ &\boldsymbol{\Sigma}_b = \mathbf{Cov}_{s_t\sim\mu}[\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}_t].\end{aligned}$$

a) Translation into our notation: We firstly translate the upper bound (161) into our notation. By definition,

$$\mathbb{E}_{s\sim\mu,s'\sim P(\cdot|s)}[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s')^\top] = \boldsymbol{\Phi}^\top\boldsymbol{D}_{\boldsymbol{\mu}}\boldsymbol{P}\boldsymbol{\Phi}.$$

Therefore, the term $\boldsymbol{I}-\boldsymbol{M}$ can be expressed as

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

Furthermore, the terms $\boldsymbol{\Sigma}_L$ and $\boldsymbol{\Sigma}_b$ can be expressed in our notation as

$$\begin{aligned}\boldsymbol{\Sigma}_L &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Cov}_{s_t\sim\mu,s_t'\sim P(\cdot|s_t)}[\boldsymbol{A}_t\boldsymbol{\theta}^\star]\boldsymbol{\Sigma}^{-\frac{1}{2}}\\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}\left[[(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star][(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star]^\top\right]\boldsymbol{\Sigma}^{-\frac{1}{2}},\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_b &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Cov}_{s_t\sim\mu}[\boldsymbol{b}_t]\boldsymbol{\Sigma}^{-\frac{1}{2}}\\ &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}_{s_t\sim\mu}[(\boldsymbol{b}_t-\boldsymbol{b})(\boldsymbol{b}_t-\boldsymbol{b})^\top]\boldsymbol{\Sigma}^{-\frac{1}{2}}.\end{aligned}$$

For simplicity, we will omit the subscript $s_t\sim\mu, s_t'\sim P(\cdot\,|\,s_t)$ in the following. Combining these terms, the upper bound in (161) can be expressed as

$$\begin{aligned}&\frac{1}{T}\mathsf{Tr}\left[(\boldsymbol{I}-\boldsymbol{M})^{-1}(\boldsymbol{\Sigma}_L+\boldsymbol{\Sigma}_b)(\boldsymbol{I}-\boldsymbol{M})^{-\top}\right]\\ &= \frac{1}{T}\mathsf{Tr}\Big[\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}\Big[[(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star][(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star]^\top\\ &+ (\boldsymbol{b}_t-\boldsymbol{b})(\boldsymbol{b}_t-\boldsymbol{b})^\top\Big]\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-\top}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\Big]\\ &= \frac{1}{T}\mathsf{Tr}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\mathbb{E}\left[[(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star][(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star]^\top\right]\boldsymbol{A}^{-\top}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\ &+ \frac{1}{T}\mathsf{Tr}\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}^{-1}\mathbb{E}\left[(\boldsymbol{b}_t-\boldsymbol{b})(\boldsymbol{b}_t-\boldsymbol{b})^\top\right]\boldsymbol{A}^{-\top}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\\ &= \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})\|_{\boldsymbol{\Sigma}}^2\end{aligned}$$

So in summary, [63] bounds the estimation error by

$$\begin{aligned}&\|\bar{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}\\ &\lesssim \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})\|_{\boldsymbol{\Sigma}}^2. \quad (162)\end{aligned}$$

b) Comparison to our results: In the following, we show that the upper bound (162) can be directly deducted from our proof of Theorem 1. Specifically, our analysis in (62), (115) and (117) reveals that $\|\bar{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2$ is bounded by

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 &\lesssim \frac{1}{T^2}\left\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\right\|_{\boldsymbol{\Sigma}}^2\\ &+ \frac{1}{T^2}\left\|\sum_{i=0}^{T-1}\boldsymbol{A}^{-1}(\boldsymbol{b}_i-\boldsymbol{b})\right\|_{\boldsymbol{\Sigma}}^2 + o\left(\frac{1}{T}\right).\end{aligned}$$

Taking expectations on both sides and applying the martingale property, we obtain

$$\begin{aligned}&\mathbb{E}\|\bar{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2\\ &\lesssim \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}_i\right\|_{\boldsymbol{\Sigma}}^2 + \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{b}_i-\boldsymbol{b})\right\|_{\boldsymbol{\Sigma}}^2\\ &\lesssim \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\theta}^\star\right\|_{\boldsymbol{\Sigma}}^2 + \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{b}_i-\boldsymbol{b})\right\|_{\boldsymbol{\Sigma}}^2\\ &+ \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2\\ &= \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{A}_t-\boldsymbol{A})\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2 + \frac{1}{T}\mathbb{E}\|\boldsymbol{A}^{-1}(\boldsymbol{b}_t-\boldsymbol{b})\|_{\boldsymbol{\Sigma}}^2\\ &+ \frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

Notice here that the first two terms are exactly the same as the right-hand-side of (162). Hence, it now boils down to showing that

$$\frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2 = o\left(\frac{1}{T}\right). \qquad (163)$$

Towards this end, we firstly observe that

$$\begin{aligned}&\frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2\\ &\lesssim \frac{\|\boldsymbol{\Sigma}\|^2\|\boldsymbol{\Sigma}^{-1}\|^2}{(1-\gamma)^2T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

For the expectation of $\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}^2$, we again apply the iterative relation deducted in (155) and obtain

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}^2 &\le (1-(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta)^i\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2\\ &+ \frac{6\eta\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})}.\end{aligned}$$

Summing from $i=0$ through $i=T-1$ yields

$$\begin{aligned}\sum_{i=0}^{T-1}\mathbb{E}\|\boldsymbol{\Delta}_i\|_{\boldsymbol{\Sigma}}^2 &\le \frac{1}{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})\eta}\|\boldsymbol{\Delta}_0\|_{\boldsymbol{\Sigma}}^2\\ &+ \frac{6\eta T\|\boldsymbol{\Sigma}\|\left(2\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}^2+1\right)}{(1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})}.\end{aligned}$$

By setting $\eta \asymp T^{-1/2}$ as suggested by [63], this immediately implies

$$\frac{1}{T^2}\sum_{i=0}^{T-1}\mathbb{E}\left\|\boldsymbol{A}^{-1}(\boldsymbol{A}_i-\boldsymbol{A})\boldsymbol{\Delta}_i\right\|_{\boldsymbol{\Sigma}}^2 \lesssim \frac{1}{T^2}\cdot T^{1/2} = o\left(\frac{1}{T}\right).$$

In summary, we have shown that the upper bound proposed by [63] follows directly from our analysis. Our result, as is shown in Theorem 1, improves upon theirs in the sense that we use a stepsize η that only depends on the logarithm of T, provide a bound with high probability instead of in expectation, and reveal a clearer dependence on the problem-related parameters.

D. Comparison with [31]

It is difficult to place the corresponding instance dependent results in comparison, so, we focus our attention on the minimax results. In the following, we make use of the relations that $\|\boldsymbol{A}(\overline{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star)\|_2 \geq \|\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}\|\|\overline{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} \gtrsim (1-\gamma)\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}\|\overline{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}}$, and $\mathbb{E}\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2^2 \lesssim \frac{1}{(1-\gamma)^2}$, $\sup\|\boldsymbol{A}_t\boldsymbol{\theta}^\star-\boldsymbol{b}_t\|_2 \lesssim \frac{1}{1-\gamma}$. We also consider the situations when $\|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} \lesssim \frac{1}{1-\gamma}$, and $\boldsymbol{\phi}(s)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(s) \lesssim \lambda_{\min}(\boldsymbol{\Sigma})^{-1}$. Notice that there exists an MDP instance such that equality can be attained in all these bounds simultaneously. For ease of presentation, let us first rephrase the result [31, Corollary 1] in terms of our notation[2]. It is shown therein that for

$$\eta \lesssim \frac{(1-\gamma)^3\lambda_{\min}(\boldsymbol{\Sigma})}{\kappa\sqrt{T}},$$

with probability at least $1-\delta$, the averaged TD estimation error is bounded by

$$\|\overline{\boldsymbol{\theta}}_T-\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}} \lesssim \sqrt{\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})(1-\gamma)^4 T}}, \tag{164}$$

when $T \gtrsim \frac{1}{c_{\boldsymbol{A}}^2} \gtrsim \frac{\kappa^2}{(1-\gamma)^6\lambda_{\min}(\boldsymbol{\Sigma})^2}$. Here, we omit the dependency of log factors. In comparison, our result delivers the same bound as long as $T \gtrsim \frac{\kappa^2}{(1-\gamma)^4\lambda_{\min}(\boldsymbol{\Sigma})}$. We incur a lower born-in cost for the relation (164) to hold.

[2]We take $C_{\boldsymbol{A}} \lesssim (1-\gamma)^{-1}$, $a \asymp \|\boldsymbol{Q}\|^{-1} \lesssim (1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma})$ and then $c_{\boldsymbol{A}} \lesssim \kappa^{-1}(1-\gamma)^3\lambda_{\min}(\boldsymbol{\Sigma})$ (see the definitions of these parameters in [31]).

REFERENCES

[1] S. A. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 331–355, 2003.

[2] I. Bojinov and N. Shephard, "Time series experiments and causal estimands: exact randomization tests and trading," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1665–1682, 2019.

[3] S. Tang and J. Wiens, "Model selection for offline reinforcement learning: Practical considerations for healthcare settings," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 2–35.

[4] C. Dann, G. Neumann, J. Peters *et al.*, "Policy evaluation with temporal differences: A survey and comparison," *Journal of Machine Learning Research*, vol. 15, pp. 809–883, 2014.

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

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

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

[8] J. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.

[9] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," *Operations Research*, vol. 69, no. 3, pp. 950–973, 2021.

[10] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep q-learning," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 486–489.

[11] A.-m. Farahmand, M. Ghavamzadeh, C. Szepesvári, and S. Mannor, "Regularized policy iteration with nonparametric function spaces," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4809–4874, 2016.

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

[13] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: an overview," in *Proceedings of 1995 34th IEEE conference on decision and control*, vol. 1. IEEE, 1995, pp. 560–564.

[14] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *arXiv preprint arXiv:1708.05866*, 2017.

[15] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Conference on Learning Theory*. PMLR, 2020, pp. 2137–2143.

[16] B. Wang, Y. Yan, and J. Fan, "Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model," *Advances in neural information processing systems*, vol. 34, pp. 23 009–23 022, 2021.

[17] G. Li, Y. Chen, Y. Chi, Y. Gu, and Y. Wei, "Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 671–16 685, 2021.

[18] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

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

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

[21] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and TD learning," in *Conference on Learning Theory*, 2019, pp. 2803–2830.

[22] C. Lakshminarayanan and C. Szepesvari, "Linear stochastic approximation: How far does constant step-size and iterate averaging go?" in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1347–1355.

[23] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.

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

[25] G. Dalal, G. Thoppe, B. Szörényi, and S. Mannor, "Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning," in *Conference On Learning Theory*, 2018, pp. 1199–1233.

[26] T. Xu and Y. Liang, "Sample complexity bounds for two timescale value-based reinforcement learning algorithms," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 811–819.

[27] Y. Wang, S. Zou, and Y. Zhou, "Non-asymptotic analysis for two time-scale tdc with general smooth function approximation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9747–9758, 2021.

[28] G. Dalal, B. Szorenyi, and G. Thoppe, "A tale of two-timescale reinforcement learning with the tightest finite-time bound," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3701–3708.

[29] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with Markovian noise," in *Conference on Learning Theory*. PMLR, 2020, pp. 2144–2203.

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

[31] A. Durmus, E. Moulines, A. Naumov, and S. Samsonov, "Finite-time high-probability bounds for polyak-ruppert averaged iterates of linear stochastic approximation," *arXiv preprint arXiv:2207.04475*, 2022.

[32] C. Szepesvári, "The asymptotic convergence-rate of Q-learning," in *Advances in Neural Information Processing Systems*, 1998, pp. 1064–1070.

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

[34] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Advances in Neural Information Processing Systems*, 2018, pp. 4863–4873.

[35] J. A. Boyan, "Least-squares temporal difference learning," in *ICML*, 1999, pp. 49–56.

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

[37] A. Agarwal, S. Kakade, and L. F. Yang, "Model-based reinforcement learning with a generative model is minimax optimal," in *Conference on Learning Theory*. PMLR, 2020, pp. 67–83.

[38] A. Pananjady and M. J. Wainwright, "Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 566–585, 2021.

[39] G. Li, Y. Wei, Y. Chi, and Y. Chen, "Breaking the sample size barrier in model-based reinforcement learning with a generative model," *Operations Research*, 2023.

[40] T. L. Lai, "Stochastic approximation," *The Annals of Statistics*, vol. 31, no. 2, pp. 391–406, 2003.

[41] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[42] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[43] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.

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

[45] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," *Advances in neural information processing systems*, vol. 24, 2011.

[46] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[47] T. Xu, S. Zou, and Y. Liang, "Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 633–10 643.

[48] Y. F. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite-time analysis of two time-scale actor-critic methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 617–17 628, 2020.

[49] D. Precup, "Eligibility traces for off-policy policy evaluation," *Computer Science Department Faculty Publication Series*, p. 80, 2000.

[50] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 652–661.

[51] C. Ma, B. Zhu, J. Jiao, and M. J. Wainwright, "Minimax off-policy evaluation for multi-armed bandits," *IEEE Transactions on Information Theory*, vol. 68, no. 8, pp. 5314–5339, 2022.

[52] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2139–2148.

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

[54] N. Kallus and M. Uehara, "Double reinforcement learning for efficient off-policy evaluation in markov decision processes," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 6742–6804, 2020.

[55] M. Yang, O. Nachum, B. Dai, L. Li, and D. Schuurmans, "Off-policy evaluation via the regularized Lagrangian," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6551–6561, 2020.

[56] Y. Duan, Z. Jia, and M. Wang, "Minimax-optimal off-policy evaluation with linear function approximation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2701–2709.

[57] Y. Jin, Z. Yang, and Z. Wang, "Is pessimism provably efficient for offline RL?" in *International Conference on Machine Learning*. PMLR, 2021, pp. 5084–5096.

[58] T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai, "Policy finetuning: Bridging sample-efficient offline and online reinforcement learning," *Advances in neural information processing systems*, vol. 34, pp. 27 395–27 407, 2021.

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

[60] L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi, "Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity," in *International Conference on Machine Learning*. PMLR, 2022, pp. 19 967–20 025.

[61] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 702–11 716, 2021.

[62] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction," *IEEE Transactions on Information Theory*, vol. 68, no. 1, pp. 448–473, 2021.

[63] W. Mou, A. Pananjady, and M. J. Wainwright, "Optimal oracle inequalities for solving projected fixed-point equations," *arXiv preprint arXiv:2012.05299*, 2020.

[64] T. Li, G. Lan, and A. Pananjady, "Accelerated and instance-optimal policy evaluation with linear function approximation," *arXiv preprint arXiv:2112.13109*, 2021.

[65] R. S. Sutton, C. Szepesvári, and H. R. Maei, "A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation," *Advances in neural information processing systems*, vol. 21, no. 21, pp. 1609–1616, 2008.

[66] E. N. Gilbert, "A comparison of signalling alphabets," *The Bell system technical journal*, vol. 31, no. 3, pp. 504–522, 1952.

[67] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer, 2009, vol. 11.

[68] J. Tropp, "Freedman's inequality for matrix martingales," *Electronic Communications in Probability*, vol. 16, pp. 262–270, 2011.

[69] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, no. 1-2, pp. 1–230, May 2015. [Online]. Available: http://dx.doi.org/10.1561/2200000048

Gen Li (Member, IEEE) received his Ph.D. degree from the Department of Electronic Engineering at Tsinghua University, advised by Professor Yuantao Gu, and bachelor's degree from the Department of Electronic Engineering and Department of Mathematics at Tsinghua University in 2016. He is currently an Assistant Professor at the Department of Statistic, the Chinese University of Hong Kong. Prior to that, he was a post-doctoral researcher at the Department of Statistics and Data Science, Wharton School, University of Pennsylvania. His research interests include diffusion based generative model, reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization.

Weichen Wu received the B.S. degree in Data Science and Big Data Technology from Peking University, Beijing, China, in 2015, and is currently a Ph.D. candidate from the Department of Statistics and Data Science at Carnegie Mellon University, co-advised by Alessandro Rinaldo and Yuting Wei. His research interests lie in Reinforcement Learning and Topological Data Analysis.

Yuejie Chi (Fellow, IEEE; S'09–M'12–SM'17–F'23) received Ph.D. and M.A. in Electrical Engineering from Princeton University in 2012 and 2009, and B.E. (Hon.) in Electrical Engineering from Tsinghua University, Beijing, China, in 2007. She is currently the Sense of Wonder Group Endowed Professor of Electrical and Computer Engineering in AI Systems at Carnegie Mellon University, with courtesy appointments in the Machine Learning Department and CyLab. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems, with applications in sensing, imaging, decision making, and AI systems. Among others, she is a recipient of Presidential Early Career Award for Scientists and Engineers (PECASE), SIAM Activity Group on Imaging Science Best Paper Prize, IEEE Signal Processing Society Young Author Best Paper Award, and the inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing. She was named a Goldsmith Lecturer by IEEE Information Theory Society, a Distinguished Lecturer by IEEE Signal Processing Society, and a Distinguished Speaker by ACM. She currently serves or served as an Associate Editor for IEEE Trans. on Information Theory, IEEE Trans. on Signal Processing, IEEE Trans. on Pattern Recognition and Machine Intelligence, Information and Inference: A Journal of the IMA, and SIAM Journal on Mathematics of Data Science.

Cong Ma (Member, IEEE) received the B.Eng. degree from Tsinghua University in 2015 and the Ph.D. degree from Princeton University in 2020. He was a Post-Doctoral Researcher with the Department of Electrical Engineering and Computer Sciences, UC Berkeley. He is currently an Assistant Professor with the Department of Statistics, The University of Chicago. His research interests include mathematics of data science, machine learning, high-dimensional statistics, convex, and nonconvex optimization. He has received the Student Paper Award from the International Chinese Statistical Association in 2017, the School of Engineering and Applied Science Award for Excellence from Princeton University in 2019, the AI Labs Fellowship from Hudson River Trading in 2019, the Hannan Graduate Student Travel Award from IMS in 2020, as well as the Best Paper Prize from SIAM Activity Group on Imaging Science in 2024.

Alessandro Rinaldo is a Professor of Statistics and Data Science at the University of Texas at Austin. He received his Ph.D. in Statistics at Carnegie Mellon University (2005), and remained there as a faculty member until 2023. His research interests revolve around the theoretical properties of statistical and machine learning models for high-dimensional data under various structural assumptions, such as sparsity or intrinsic low dimensionality. An IMS fellow, he has served on the editorial boards of various journals, including the Annals of Statistics, the Journal of the American Statistics Association, and the Electronic Journal of Statistics.

Yuting Wei (Member, IEEE) received the B.S. degree (Hons.) in statistics from Peking University, Beijing, China, in 2013, and the Ph.D. degree in statistics from the University of California, Berkeley in 2018, advised by Martin Wainwright and Aditya Guntuboyina. She is currently an Assistant Professor at the Department of Statistics and Data Science, Wharton School, University of Pennsylvania. Prior to that, She was with Carnegie Mellon University from 2019 to 2021 as an Assistant Professor of Statistics and Data Science, and with Stanford University as a Stein Fellow from 2018 to 2019. Her research interests include high-dimensional and non-parametric statistics, statistical machine learning, and reinforcement learning. She was a recipient of the NSF CAREER award, the Google Research Award, the ICSA Junior Research Award, honorable mention for Bernoulli Society's New Researcher Award, and the Erich L. Lehmann Citation from Berkeley Statistics (awarded to the best dissertation in theoretical statistics).