A UNIFYING VIEW OF COVERAGE IN LINEAR OFF-POLICY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Off-policy evaluation (OPE) is a fundamental task in reinforcement learning (RL). In the classic setting of *linear OPE*, finite-sample guarantees often take the form

$$\text{Evaluation error} \leq \text{poly}(C^\pi, d, 1/n, \log(1/\delta)),$$

where d is the dimension of the features and C^π is a ***feature coverage parameter*** that characterizes the degree to which the visited features lie in the span of the data distribution. Though such guarantees are well-understood for several popular algorithms under the Bellman-completeness assumption, this form of guarantee has not yet been achieved in the minimal setting where it is only assumed that the target value function is linearly realizable in the features. Despite recent interest in tight characterizations for this setting, the right notion of coverage remains unclear, and candidate definitions from prior analyses have undesirable properties and are starkly disconnected from more standard quantities in the literature.

In this paper, we provide a novel finite-sample analysis of a canonical algorithm for this setting, LSTDQ. Inspired by an instrumental-variable (IV) view, we develop error bounds that depend on a novel coverage parameter, the feature-dynamics coverage, which can be interpreted as feature coverage in an induced feature-compressed MDP. With further assumptions—such as Bellman-completeness—our definition successfully recovers the coverage parameters specialized to those settings, finally yielding a unified understanding for coverage in linear OPE.

1 INTRODUCTION

Coverage is a foundational concept in reinforcement learning (RL) theory. In off-policy evaluation (OPE), the task of evaluating a target policy based on data collected from a different behavior policy, coverage characterizes the degree to which the data distribution contains relevant information about the target policy. The relevance of coverage extends beyond OPE, and the concept plays important roles in offline policy learning (Jin et al., 2021; Xie et al., 2021), *online* RL (Xie et al., 2023; Amortila et al., 2024a;b), or even statistical-computational trade-off in LLMs (Foster et al., 2025), as it provides a mathematical characterization of distribution shift which is a central challenge in RL.

Mathematically, coverage is manifested as *coverage parameters* in finite-sample guarantees: for example, standard OPE guarantees often take the form

$$\text{Evaluation error} \leq \text{poly}(C^\pi, d, 1/n, \log(1/\delta)), \tag{1}$$

where n is sample size, δ is the failure probability, d is the statistical dimension of the function class, δ is the failure probability, and C^π is the coverage parameter. The definition of C^π can take many different forms depending on the algorithm and the assumptions, as well as how the proof handles *error propagation* through the dynamics of the MDP (Farahmand et al., 2010). The most naïve definition is $\|\mu^\pi/\mu^D\|_\infty$, the boundedness of density ratio between the target policy's discounted occupancy μ^π and the data distribution μ^D. More refined definitions often take advantage of the structure of the underlying MDP or the function approximation scheme. Comparisons between these definitions offer connections and unified understanding across different learning settings, such as offline vs. online (Xie et al., 2023), tabular vs. function approximation (Yin & Wang, 2021), Markovian vs. partially observed (Zhang & Jiang, 2024), and single-agent vs. multi-agent RL (Cui & Du, 2022; Zhang et al., 2023).

While a unified understanding of RL through the lens of coverage is emerging, one of the most fundamental settings—where the target value function is linearly realizable in a given feature map—eludes such understanding and remains starkly disconnected from the rest of the literature. The most natural algorithm for this setting is arguably LSTD(Q) (Boyan, 1999; Lagoudakis & Parr, 2003). Despite recent statistical results for this method (Duan et al., 2021; Mou et al., 2022a; Perdomo et al., 2023), the error bounds often come with obscure conditions and are hard to interpret, with little or no understanding of which quantities play the role of coverage and how they connect to coverage parameters in related settings and algorithms.

On the other hand, simpler analyses do provide more interpretable candidates, such as $1/\sigma_{\min}(A)$ with $A = \mathbb{E}_{\mu^D}[\phi(s,a)(\phi(s,a)^\top - \gamma\phi(s',\pi)^\top)]$ being the key matrix estimated in LSTDQ (Section 2.1). Given that LSTDQ approximates $Q^\pi \approx \phi^\top\theta$ by solving a linear equation in the form of $A\theta = b$, $1/\sigma_{\min}(A)$ is a very natural candidate as it determines the invertibility of A and the solution's numerical stability. While bounds in the form of Eq.(1) can be established with $C^\pi = 1/\sigma_{\min}(A)$, the quantity $1/\sigma_{\min}(A)$ is unsatisfactory in many aspects as a coverage parameter:

1. **Lacking scale invariance.** The value of $\sigma_{\min}(A)$ can change arbitrarily if we simply redefine the features as $\phi_{\text{new}} = c\phi$.[1] While seemingly unrelated, this issue is mathematically tied to the fact that $\sigma_{\min}(A)$ as a coverage parameter has no concern over the initial state distribution of the MDP which should play an important role in the definition of coverage.
2. **Lacking off-policy characterization.** Coverage parameters provide important understanding for when data contains relevant information about the target policy. For $1/\sigma_{\min}(A)$, however, the only thing we know is its boundedness in a strict on-policy case, and it is hard to interpret for general off-policy distributions.
3. **Lacking unification with other analyses.** State abstractions are a special case of linear function approximation, under which LSTDQ coincides with the model-based solution. Prior works have established *aggregated concentrability* (Jia et al., 2024) as the coverage parameter for this setting, which cannot be recovered by specializing $1/\sigma_{\min}(A)$. Moreover, both concepts differ significantly from standard definitions of coverage in linear OPE when analyzed under the Bellman-completeness assumption: standard definitions measure coverage by analyzing how errors propagate under the ***groundtruth dynamics***, whereas aggregated concentrability does so under the ***compressed dynamics*** determined by the abstraction scheme.

In this paper, we provide a novel finite-sample analysis of LSTDQ inspired by an instrument-variable (IV) view, which comes with a new coverage parameter that we call ***feature-dynamics coverage***, C^π_ϕ. Feature-dynamics coverage replaces $1/\sigma_{\min}(A)$ and elegantly addresses the above problems. Furthermore, it corresponds to feature coverage in a linear dynamical system induced by the features (first studied by Parr et al. (2008)). The system is the transition dynamics of the true MDP compressed through the given features, and naturally subsumes the χ^2 version of aggregated concentrability as a special case. Furthermore, given Bellman-completeness as an additional assumption, feature-dynamics coverage recovers the standard notion of linear coverage, successfully unifying the previously fragmented understanding.

2 PRELIMINARIES

Markov Decision Process (MDP). We consider the groundtruth environment modeled as an infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma, \mu_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamics ($\Delta(\cdot)$ is the probability simplex), $R : \mathcal{S}\times\mathcal{A} \to \Delta([0, R_{\max}])$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution. We assume $\mathcal{S}$, $\mathcal{A}$ are finite, but their cardinalities can be prohibitively large and thus should not appear in sample-complexity guarantees. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a distribution over random trajectories, generated as $s_0 \sim \mu_0$, $a_t \sim \pi(\cdot|s_t)$ (or simply $a_t \sim \pi$), $r_t = R(s_t, a_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$. Let $\mathbb{P}_\pi[\cdot]$ and $\mathbb{E}_\pi[\cdot]$ denote the probability and expectation under such a distribution. The expected return of a policy is $J(\pi) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_t]$, which falls in the range of $[0, V_{\max}]$ with $V_{\max} := R_{\max}/(1-\gamma)$. The discounted occupancy of π is defined as

$$\mu^\pi(s,a) = (1-\gamma)\sum_{t\geq 0}\gamma^t\mu^\pi_t(s,a) \coloneqq (1-\gamma)\sum_{t\geq 0}\gamma^t\mathbb{P}_\pi[s_t = s, a_t = a]. \tag{2}$$

[1] Perdomo et al. (2023) provide a bound that depends on a term related to our coverage parameter, which is scale invariant. We will compare and connect to their results in Section 4.

Value Function and Bellman Operator. The Q-function $Q^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the fixed point of Bellman operator $\mathcal{T}^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, i.e., $Q^\pi = \mathcal{T}^\pi Q^\pi$, where $\forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $(\mathcal{T}^\pi f)(s,a) := R(s,a) + \gamma \mathbb{E}_{s\sim P(\cdot|s,a)}[f(s',\pi)]$. Here $f(s',\pi)$ is a shorthand for $\mathbb{E}_{a'\sim\pi(\cdot|s')}[f(s',a')]$. Given any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as an approximation of Q^π, we can induce an estimate of $J(\pi)$ as:

$$J_f(\pi) := E_{s_0\sim\mu_0, a_0\sim\pi}[f(s_0,a_0)], \tag{3}$$

since $J(\pi) = J_{Q^\pi}(\pi)$.

Linear Off-policy Evaluation (OPE). OPE is the task of estimating the performance of a given *target policy* π based on an offline dataset $\mathcal{D}$ sampled from a *behavior policy* π_b. As a standard simplification, We assume that $\mathcal{D}$ consists of n i.i.d. tuples (s,a,r,s',a') generated as

$$(s,a) \sim \mu^D, r \sim R(s,a), s' \sim P^\star(\cdot|s,a), a' \sim \pi(\cdot \mid s').$$

We use $\mathbb{E}_{\mu^D}[\cdot]$ to denote the expectation of functions of (s,a,r,s',a') under the data distribution, and $\mathbb{E}_{\mathcal{D}}[\cdot]$ denotes the empirical approximation from $\mathcal{D}$. For most of the paper we are concerned with *return* estimation via linear function approximation, i.e., estimating the scalar $J(\pi)$ as $J_{\widehat{Q}^\pi}(\pi)$ where $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top \widehat{\theta}$ for some given feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$. We make the following standard assumptions throughout the paper:

Assumption 1 (Feature boundedness and realizability)**.** *We assume that there exists $\theta^\star \in \mathbb{R}^d$ such that $Q^\pi(s,a) = \phi(s,a)^\top\theta^\star$. Furthermore, assume that $\|\phi(s,a)\|_2 \le B_\phi, \forall s,a$.*

Mathematical Notation. We use $\sigma_{\min}(\cdot)$ and $\lambda_{\min}(\cdot)$ to denote the smallest singular value of an asymmetric matrix and the smallest eigenvalue of a symmetric matrix, respectively. Let $\rho(\cdot)$ denote the spectral radius of a matrix. For functions over $\mathcal{S}\times\mathcal{A}$ such as Q^π and d^π, we also view them interchangeably as vectors in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ whenever convenient. We use $a \lesssim b$ as a shorthand for $a = O(b)$. Given two square and possibly asymmetric matrices Σ and Σ', $\Sigma \preceq \Sigma'$ means $v^\top(\Sigma - \Sigma')v \le 0$ for all v. We let $\|v\|_\Sigma = \sqrt{v^\top \Sigma v}$ denote the Mahalanobis norm.

2.1 LSTDQ

The LSTDQ algorithm estimates the following moments from data:

$$\Sigma = \mathbb{E}_{\mu^D}\left[\phi(s,a)\phi(s,a)^\top\right], \qquad \Sigma_{\rm cr} = \mathbb{E}_{\mu^D}\left[\phi(s,a)\phi(s',a')^\top\right],$$
$$A = \Sigma - \gamma\Sigma_{\rm cr}, \qquad b = \mathbb{E}_{\mu^D}[\phi(s,a)r].$$

Throughout the paper, we assume:

Assumption 2 (Invertibility)**.** *Σ and A are invertible.*

These moments satisfy

$$\begin{aligned} A\theta^\star - b &= \mathbb{E}_{\mu^D}[\phi(s,a)(\phi(s,a)^\top\theta^\star - r - \phi(s',\pi)^\top\theta^\star)] \\ &= \mathbb{E}_{\mu^D}[\phi(s,a)(Q^\pi(s,a) - (\mathcal{T}^\pi Q^\pi)(s,a)] = \mathbf{0}, \end{aligned}$$

which implies $\theta^\star = A^{-1}b$ if A is invertible, and LSTDQ is simply the plug-in estimate of this inverse: let $\widehat{\Sigma}, \widehat{\Sigma}_{\rm cr}, \widehat{A}, \widehat{b}$ be the empirical estimates from $\mathcal{D}$, and

$$\widehat{\theta}_{\rm lstd} = \widehat{A}^{-1}\widehat{b}, \quad \widehat{Q}_{\rm lstd}(s,a) = \phi(s,a)^\top\widehat{\theta}_{\rm lstd}. \tag{4}$$

We do not explicitly assume the empirical matrices $\widehat{\Sigma}$ and $\widehat{A}$ are invertible, with the understanding that algorithms based on these quantities have degenerate behaviors when they are singular, and our guarantees hold under such convention.[2]

[2]That is, either the high-probability event guarantees that the empirical matrix is invertible, or such matrices appears in the bound and makes the guarantee vacuous (as in e.g., Eq. (10)).

3 Related Works

Coverage Parameters in Offline RL. Early research in offline RL identifies the boundedness of density ratios, such as $\|\mu^\pi/\mu^D\|_\infty$, as the coverage parameter (Munos, 2007; Munos & Szepesvári, 2008; Antos et al., 2008). Later works point out that they can be tightened by leveraging the structure of the function class $\mathcal{F}$ used to approximate the value function, such as $\sup_{f\in\mathcal{F}} \frac{(\mathbb{E}_{\mu^\pi}[f-\mathcal{T}^\pi f])^2}{\mathbb{E}_{\mu^D}[(f-\mathcal{T}^\pi f)^2]}$. In our setting, the linear feature ϕ induces a linear $\mathcal{F}_\phi = \{\phi^\top\theta : \theta\in\mathbb{R}^d\}$, and these parameters often have simplified forms. Among them, the tightest known coverage parameter for off-policy return estimation is (Zanette et al., 2021; Yin et al., 2022; Gabbianelli et al., 2024; Jiang & Xie, 2024):

$$C^\pi_{\text{lin}} = (\phi^\pi)^\top\Sigma^{-1}\phi^\pi, \quad \text{where } \phi^\pi := \mathbb{E}_{(s,a)\sim\mu^\pi}[\phi(s,a)]. \tag{5}$$

C^π_{lin} only requires the *mean* feature under π, ϕ^π, to lie in the span of data features, whereas alternatives such as $\mathbb{E}_{\mu^\pi}[\|\phi\|_{\Sigma^{-1}}]$ require coverage of the distribution $\phi(s,a), (s,a)\sim\mu^\pi$ in a point-wise manner; see Jiang & Xie (2024) for further discussions. The majority of the study on coverage, however, crucially relies on the following assumption on $\mathcal{F}$ which is substantially stronger than realizability ($Q^\pi\in\mathcal{F}$), and we ***do not assume*** it in our main results.

Assumption 3 (Bellman-completeness)**.** *Let $\mathcal{F}\subset(\mathcal{S}\times\mathcal{A}\to\mathbb{R})$ be a function class for approximating Q^π. We say that it is Bellman-complete if*

$$\mathcal{T}^\pi f\in\mathcal{F}, \forall f\in\mathcal{F}.$$

Indeed, a major potential of LSTDQ is that it is one of the few algorithms who enjoy theoretical guarantees under only realizability $Q^\pi\in\mathcal{F}$, which enables its important role in the challenging problem of offline model selection (Xie & Jiang, 2020a; Liu et al., 2025). Unfortunately, coverage in the absence of Bellman-completeness is poorly understood even in the linear setting, as discussed in Section 1, which we address in this paper.

LSTD(Q) Analyses. LSTD methods are initially derived as the fixed point solution of TD methods (Sutton & Barto, 2018), and its closed-form nature separates it from typical dynamical-programming-style RL algorithms that often suffer from divergence. While we focus on LSTDQ, our result naturally extend to other variants such as LSTD (which uses state-feature to approximate V^π) or off-policy LSTD (up to handling importance sampling via concentration inequalities) (Nedić & Bertsekas, 2003; Bertsekas & Yu, 2009; Dann et al., 2014).

Early finite-sample analyses of LSTD are mostly in the on-policy setting and depend on quantities like $\sigma_{\min}(A)$ (Bertsekas, 2007; Lazaric et al., 2010; 2012). There is a recent surge of interest in providing tight statistical characterizations of LSTD, including in the off-policy setting, with results suggesting the necessity of $1/\sigma_{\min}(A)$ for this setting (Amortila et al., 2020; Mou et al., 2022b; Amortila et al., 2023; Perdomo et al., 2023). These results, however, offer very little discussion on coverage, and sometimes require additional regularity assumptions. Perdomo et al. (2023) recently provides a sharp analysis that largely subsumes the earlier results, which we will compare to in Section 5.1.

4 Finite-sample Analysis of LSTDQ

In this section we will present the finite-sample analysis of LSTDQ, which depends on our proposed coverage parameter.

Special case of $\gamma = 0$. It is instructive to start with the special case of $\gamma = 0$ where tight guarantees and the definition of coverage are well understood. Recall that $A\theta^\star = b$ can be rewritten as:

$$\mathbb{E}[ZX^\top]\theta^\star = \mathbb{E}[ZY], \tag{6}$$

where $Z = \phi(s,a)\in\mathbb{R}^d$, $X = \phi(s,a) - \gamma\phi(s',a')\in\mathbb{R}^d$, $Y = r\in\mathbb{R}$, and the expectation $\mathbb{E}[\cdot]$ is under μ^D. Below we will go back and forth between the linear regression (LR) notation system ($X, Y, \mathbb{E}[XX^\top], \ldots$) and the RL notation system ($\phi, r, \Sigma, \ldots$), where we obtain concentration bounds from the LR literature and meaningful guarantees for OPE in the RL setting, respectively.

When $\gamma = 0$, we essentially face a contextual bandit problem with linear reward, and Eq. (6) becomes

$$\mathbb{E}[XX^\top]\theta^\star = \mathbb{E}[XY], \tag{7}$$

which is a classic linear regression problem, with $A = \Sigma = \mathbb{E}[XX^\top]$. In this special case, LSTDQ simply performs LR to fit the parameter $\theta^\star$. While parameter identification guarantees in LR (i.e., error bounds on $\|\widehat{\theta} - \theta^\star\|$) inevitably have to depend on $\sigma_{\min}(A) = \lambda_{\min}(\mathbb{E}[XX^\top])$, the key is in how we use $\widehat{\theta}_{\text{lstd}}$ to form the final estimation in OPE:

$$J_{\widehat{Q}_{\text{lstd}}}(\pi) = \mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}[\phi(s_0,a_0)^\top \widehat{\theta}_{\text{lstd}}] = \phi_0^\top \widehat{\theta}_{\text{lstd}},$$

where

$$\phi_0 := \mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}[\phi(s_0,a_0)]. \tag{8}$$

That is, for the purpose of estimating $J(\pi)$, we only need $\widehat{\theta}_{\text{lstd}}$ to be accurate in the direction of ϕ_0, and a high-probability bound can be established if ϕ_0 is well *covered* by the distribution of $X = \phi(s,a)$ where $(s,a) \sim \mu^D$ (Abbasi-Yadkori et al., 2011; Jin et al., 2020): informally, with probability at least $1-\delta$,[3]

$$|J_{\widehat{Q}_{\text{lstd}}}(\pi) - J(\pi)| = |\phi_0^\top(\widehat{\theta}_{\text{lstd}} - \theta^\star)| \lesssim \|\phi_0\|_{\widehat{\Sigma}^{-1}} \sqrt{\frac{d\log(1/\delta)}{n}} V_{\max}. \tag{9}$$

Here $\|\phi_0\|_{\widehat{\Sigma}^{-1}}$ plays the role of coverage, characterizing how well the expected feature under the target policy π is covered by the random features observed in the data (which determines Σ). Similar bounds with the population version of coverage $\|\phi_0\|_{\Sigma^{-1}}$ also hold under additional regularity assumptions (Hsu et al., 2011). The quantity is also consistent with the standard notion of linear coverage in MDPs (Eq. (5)) under the Bellman-completeness assumption (Assumption 3), which is equivalent to realizability in the bandit setting ($\gamma = 0$).

Extending to $\gamma > 0$ with Instrumental-Variable inspiration. Given that the $\gamma = 0$ case is well-understood and does not suffer the issues mentioned in the introduction, we therefore seek to extend the above framework to $\gamma > 0$. When $\gamma > 0$, however, we have $Z \neq X$, and Eq. (6) is a form of Instrumental Variable (IV) problem induced by "error-in-the-variable" issues: it is known that

$$R(s,a) = \phi_{\text{td}}(s,a)^\top \theta^\star, \quad \text{with } \phi_{\text{td}}(s,a) := \phi(s,a) - \gamma\, \mathbb{E}_{\mu^D}[\phi(s',a')|s,a],$$

that is, the expected temporal-difference feature, ϕ_{td}, can linearly predict reward, which LSTDQ leverages to recover $\theta^\star$. However, in the data we do not observe the expected TD feature but its random realization, $X = \phi(s,a) - \gamma\phi(s',a')$, and $X - \phi_{\text{td}}(s,a)$ is zero-mean (conditioned on (s,a)) noise. Given such "error in the variable", $\mathbb{E}[XX^\top]\theta^\star \neq \mathbb{E}[XY]$, so a straightforward linear regression from X to Y does not work. LSTDQ solves this problem by introducing $Z = \phi(s,a)$ as an *instrumental variable*, which is independent of the noise $X - \phi_{\text{td}}(s,a)$ given (s,a) and thus helps marginalizes out the said noise. Based on this view, we extend the LR analysis of $\gamma = 0$ to the $\gamma > 0$ case by consulting the IV literature (e.g., Della Vecchia & Basu, 2025), which leads to our main finite-sample error bounds (see Appendix A for the proof):

Theorem 1 (Main Theorem)**.** *Under Assumptions 1 and 2, with probability at least* $1-\delta$*,*

$$\left|J_{\widehat{Q}_{\text{lstd}}}(\pi) - J(\pi)\right| \lesssim \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{\widehat{C}_\phi^\pi \cdot d\log(1/\delta)}{n}} \tag{10}$$

where

$$\widehat{C}_\phi^\pi := (1-\gamma)^2 \phi_0^\top \widehat{A}^{-1}\widehat{\Sigma}\widehat{A}^{-\top}\phi_0. \tag{11}$$

Here we take the convention that $\widehat{C}_\phi^\pi = +\infty$ *if* $\widehat{A}$ *or* $\widehat{\Sigma}$ *is not invertible.*

Corollary 1. *There exists* n_0 *such that when* $n \geq n_0$*, w.p.* $\geq 1-\delta$*,*

$$\left|J_{\widehat{Q}_{\text{lstd}}}(\pi) - J(\pi)\right| \lesssim \frac{V_{\max}}{1-\gamma}\sqrt{\frac{C_\phi^\pi \cdot d\log(1/\delta)}{n}} + o(\sqrt{1/n}), \tag{12}$$

where

$$C_\phi^\pi := (1-\gamma)^2 \phi_0^\top A^{-1}\Sigma A^{-\top}\phi_0 \tag{13}$$

and n_0 *and the* $o(\sqrt{1/n})$ *term may additionally depend on* $1/\sigma_{\min}(A)$*.*

[3]Most analyses require ridge regression (i.e., adding λI to $\widehat{\Sigma}$), and our analysis shows that this guarantee holds even without ridge.

Coverage Parameter and Tightness. Our main results consist of two guarantees where $\widehat{C}_\phi^\pi$ and C_ϕ^π play the role of the coverage parameters. The $(1-\gamma)^2$ are normalization constants, whose roles will become clear in Section 5. Eq. (10) gives a clean bound which exactly recovers that of linear regression in Eq. (9): when $\gamma = 0$, we have $\widehat{A} = \widehat{\Sigma}$, and therefore $\widehat{C}_\phi^\pi = \phi_0^\top \widehat{\Sigma}^{-1}\phi_0 = \|\phi_0\|^2_{\widehat{\Sigma}^{-1}}$. Since Eq. (9) is the well-established bound for linear regression, this demonstrates the tightness of our bound on its dependence on $\widehat{C}_\phi^\pi$, d, and n.

That said, a caveat of Eq. (10) is that $\widehat{C}_\phi^\pi$ is a random variable. While such a situation is fairly common in offline RL theory (Jin et al., 2021), ideally we would like to have its population version C_ϕ^π (Eq. (13)), which does not depend on data randomness and is more interpretable; our interpretation in Section 5 will also focus on C_ϕ^π. This is exactly what we offer in Corollary 1, where the asymptotically dominating term $O(1/\sqrt{n})$ is identical to Eq. (10) with $\widehat{C}_\phi^\pi$ replaced by C_ϕ^π, but the spectral properties of Σ and A may enter the burn-in term (n_0) and the fast-rate term ($o(\sqrt{1/n})$). The fact that Eq. (12) is less clean and requires additional assumptions than Eq. (10) is somewhat expected, as this is also the case for linear regression, where a bound that depends on $\|\phi_0\|_{\Sigma^{-1}}$ (instead of $\|\phi_0\|_{\widehat{\Sigma}^{-1}}$) also requires additional assumptions (Hsu et al., 2011; Oliveira, 2016; Mourtada, 2022). In Appendix D, we provide another result that eliminates the dependence on $1/\sigma_{\min}(A)$ in the C_ϕ^π bound, in exchange for a $1/\lambda_{\min}(\Sigma)$ dependence, which we expect can be further sharpened to leverage-score-type conditions (Hsu et al., 2011; Perdomo et al., 2023). Finally, as we will see in Section 5.3, our bound is also tight when comparing to existing OPE guarantees analyzed under Bellman completeness.

5 Understanding the Coverage Parameter

In this section we provide interpretations of C_ϕ^π as a coverage parameter and discuss how it addresses the issues mentioned in Section 1. First, it is clear that C_ϕ^π is invariant to feature rescaling, thanks to the introduction of ϕ_0. That said, the expression $C_\phi^\pi = (1-\gamma)^2\phi_0^\top A^{-1}\Sigma A^{-\top}\phi_0$ does not lend itself to easy intuition, let alone how it connects to and unifies existing results.

Warm-up: the tabular case. We start with the tabular setting and show that C_ϕ^π becomes something familiar, offering some basic intuitions as well as assurance that C_ϕ^π, a quantity that falls out of the IV concentration analyses, holds meaningful interpretations in RL. The key is to rewrite C_ϕ^π as

$$\phi_0^\top A^{-1}\Sigma A^{-\top}\phi_0 = \phi_0^\top (I-\gamma B^\pi)^{-1}\Sigma^{-1}(I-\gamma B^\pi)^{-\top}\phi_0, \quad \text{where } B^\pi := \Sigma^{-1}\Sigma_{\mathrm{cr}}. \tag{14}$$

The tabular setting can be viewed as a special case of linear function approximation with $d = |\mathcal{S}\times\mathcal{A}|$, and $\phi(s,a) = \mathbf{e}_{s,a}$ is the unit vector with the (s,a)-th coordinate being 1 and all other coordinates being 0. In this case, ϕ_0 is simply the vector representation of the initial state-action distribution μ_0^π, where $(s_0,a_0) \sim \mu_0^\pi \;\Leftrightarrow\; s_0 \sim \mu_0, a_0 \sim \pi$; B^π is an $|\mathcal{S}\times\mathcal{A}| \times |\mathcal{S}\times\mathcal{A}|$ matrix with $[B^\pi]_{(s,a),(s',a')} = P^\pi(s',a'|s,a) = P(s'|s,a)\pi(a'|s')$, i.e., the transition kernel of the Markov chain over $\mathcal{S}\times\mathcal{A}$ induced by policy π. Put together, we have the textbook identity

$$(1-\gamma)\phi_0^\top (I-\gamma B^\pi)^{-1} = (\mu^\pi)^\top,$$

where we recall the definition of the discounted occupancy μ^π from Eq. (2). Plugging it into C_ϕ^π,

$$C_\phi^\pi = (\mu^\pi)^\top\Sigma^{-1}\mu^\pi = \textstyle\sum_{s,a}\mu^\pi(s,a)^2/\mu^D(s,a) = \mathbb{E}_{\mu^D}[(\mu^\pi/\mu^D)^2],$$

which is the χ^2-divergence between μ^π and μ^D up to a constant shift and has appeared as a tight coverage parameter (especially when compared to $\|\mu^\pi/\mu^D\|_\infty$ (Xie & Jiang, 2020b)) when coverage is measured based on density ratios.

5.1 General Interpretation

We now offer the interpretation for the general setting. Note that B^π can be viewed as the multi-variate linear-regression solution of the regression problem $\phi(s,a) \mapsto \phi(s',\pi)$, thus

$$\mathbb{E}_{s'\sim P(\cdot|s,a)}[\phi(s',\pi)] \approx B^\pi\phi(s,a). \tag{15}$$

In general, the above relationship is only approximate (in Section 5.3 we will see that it becomes *exact* under an additional assumption), although B^π is the best linear predictor. This leads to the following interpretation of C_ϕ^π (see Appendix B.1 for the proof):

Proposition 1. *Define a deterministic linear dynamical system* $\{x_t\}_{t\geq 0}$*, with* $x_0 := \phi_0$*, and* $\forall t \geq 0$*,*

$$x_{t+1} = (B^\pi)^\top x_t.$$

When $\rho(B^\pi) < 1/\gamma$*, define the feature occupancy in* B^π *as* $\mu_\phi^\pi := (1-\gamma)\sum_{t\geq 0}\gamma^t x_t$*, then*

$$C_\phi^\pi = (\mu_\phi^\pi)^\top \Sigma^{-1} \mu_\phi^\pi.$$

The proposition rewrites C_ϕ^π in a form that closely resembles the standard notion of linear coverage in the literature, where we see the expected feature occupancy under the target policy (μ_ϕ^π here) measured under the data-covariance norm Σ^{-1}; see Section 5.3. Accordingly, we call C_ϕ^π the ***feature-dynamics coverage***. The difference is that here the feature occupancy is defined in a deterministic dynamical system B^π instead of the true MDP. Furthermore, while the latter, $\phi^\pi := \mathbb{E}_{(s,a)\sim\mu^\pi}[\phi(s,a)]$, is always bounded, μ_ϕ^π, on the other hand, may not be bounded in general and $\{x_t\}_{t\geq 0}$ may actually diverge. The connection between LSTD and the linear dynamical system B^π was first identified by Parr et al. (2008) (see also Duan & Wang (2020)), though they focused on the algebraic equivalence between LSTD and the model-based solution in B^π, and did not perform finite-sample analyses or connect this to the notion of coverage.

When is feature-dynamics coverage well-behaved? Our bound sharpens and generalizes existing understanding of when linear OPE using only realizability is possible. We provide a comparison to Perdomo et al., whose analysis was shown to be sharp and subsume many prior conditions known in the literature. They establish that, under some regularity assumptions, $\|\Sigma^{1/2}(\theta^\star - \widehat{\theta})\|_2 \lesssim \frac{1}{\sigma_{\min}(I-\gamma\Sigma^{-1/2}\Sigma_{\rm cr}\Sigma^{-1/2})}\cdot\varepsilon_{\rm stat}$, for some $\varepsilon_{\rm stat}$ which is polynomial in $d, 1/n, \log(1/\delta)$, and spectral properties of Σ. While they only show function-estimation guarantee on μ^D (c.f. Appendix C), this intermediate result immediately implies a return-estimation guarantee comparable to ours:

$$\left|J_{\widehat{Q}_{\rm lstd}}(\pi) - J(\pi)\right| \leq \frac{\|\phi_0\|_{\Sigma^{-1}}}{\sigma_{\min}(I-\gamma\Sigma^{-1/2}\Sigma_{\rm cr}\Sigma^{-1/2})}\cdot\varepsilon_{\rm stat}.$$

As we have already shown that our statistical rate is tight, it suffices to compare our C_ϕ^π to their multiplicative factor in front of $\varepsilon_{\rm stat}$. In particular, we establish the following relationship (see Appendix B.2 for the proof).

Proposition 2.

$$\sqrt{C_\phi^\pi} = (1-\gamma)\|(I-\gamma\Sigma^{-1/2}\Sigma_{\rm cr}\Sigma^{-1/2})^{-T}\Sigma^{-1/2}\phi_0\|_2 \leq (1-\gamma)\frac{\|\phi_0\|_{\Sigma^{-1}}}{\sigma_{\min}(I-\gamma\Sigma^{-1/2}\Sigma_{\rm cr}\Sigma^{-1/2})}.$$

This demonstrates that our coverage parameter provides a tighter return-estimation guarantee compared to the approach of Perdomo et al. (2023). As an immediate consequence, we also subsume other known conditions for this setting that were captured by Perdomo et al., including on-policy sampling (Tsitsiklis & Van Roy, 1997), Bellman completeness, low distribution shift (Wang et al., 2021), symmetric stability (Mou et al., 2022a), and contractivity (Kolter, 2011) (see the discussion in Perdomo et al. for formal definitions). Furthermore, we consider the $1/\sigma_{\min}$-type bound to provide little intuition about necessary coverage conditions for this fundamental task, and the unification of C_ϕ^π with existing concepts in the literature to be a major contribution.

5.2 Recovering Aggregated Concentrability

State abstractions are a special case of linear function approximation, where each state s is mapped to one of the K abstract states, $\psi(s) \in \{1,\dots,K\}$, effectively treating states with the same $\psi(s)$ as aggregated and equivalent to reduce the size of the state space. Under the abstraction scheme, the natural model-based solution coincides with LSTDQ with $\phi(s,a) = \mathbf{e}_{k,a}$. When we only assume realizability (Assumption 1), this has been analyzed by Xie & Jiang (2020a); Zhang & Jiang (2021); Jia et al. (2024) with the following notion of *aggregated concentrability* as its coverage parameter.

Definition 1 (Aggregated concentrability)**.** *Given* $\psi : \mathcal{S} \to \{1,\dots,K\}$*, define the abstract MDP* $M_\phi = (\mathcal{S}_\phi, \mathcal{A}, P_\phi, R_\phi, \gamma, \mu_0)$ *where*[4] $\mathcal{S}_\phi = \{1,\dots,K\}$*, and*

$$P_\phi(k'|k,a) = \frac{\mu^D(s,a)\cdot\sum_{s:\psi(s)=k}\left(\sum_{s':\psi(s')=k'}P(s'|s,a)\right)}{\sum_{s:\psi(s)=k}\mu^D(s,a)}.$$

[4]The definition of R_ϕ is irrelevant for our purpose and thus omitted.

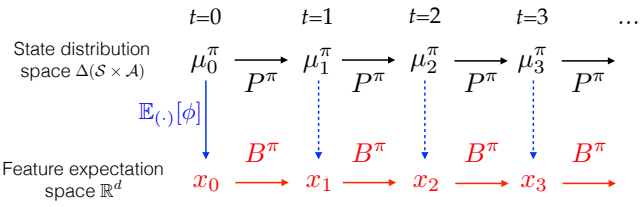

Figure 1: Illustration of the evolution of occupancies under the true dynamics P^π (top row) and that of features under the compressed dynamics B^π (bottom row). Under Bellman completeness, the dashed blue arrows hold and two routes ($\to \ldots \to \downarrow$ vs. $\downarrow \to \ldots \to$) yield the same expected feature vectors, but they are generally different without such an assumption.

For any π that only depends on s through $\psi(s)$, aggregated concentrability *refers to measures of $\mu^\pi_{M_\phi}/\mu^D_\phi$, either in $\|\cdot\|_\infty$ or χ^2 form, where $\mu^\pi_{M_\phi}$ is discounted occupancy in MDP M_ϕ, and $\mu^D_\phi(k,a) = \sum_{s:\psi(s)=k} \mu^D(s,a)$.*

In this definition, P_ϕ is the dynamics over the abstract state space, and it is easy to see that the transition kernel of abstract-state pairs under π is $P^\pi_\phi = B^\pi$, and $\Sigma = \mathrm{diag}(\mu^D_\phi)$. As a result, C^π_ϕ recovers the χ^2 version of aggregated concentrability (see Appendix B.3 for the proof):

Proposition 3. *When ϕ is induced by a state abstraction ψ and π depends on s only through $\psi(s)$,*

$$C^\pi_\phi = \mathbb{E}_{(k,a)\sim\mu^D_\phi}[(\mu^\pi_{M_\phi}/\mu^D_\phi)^2].$$

5.3 Recovering Standard Linear Coverage under Bellman-Completeness

Prior results on abstractions leave an intriguing question open: they measure coverage by analyzing error propagation in M_ϕ, which a lower-dimensional and approximate model **compressed** from M by ϕ, as evidenced by $\mu^\pi_{M_\phi}$ in the definition of aggregated concentrability; this is also consistent with our results in Section 5.1 where occupancy is measured in the compressed linear dynamical system B^π. On the other hand, the mainstream notion of coverage in linear OPE, obtained under the Bellman-completeness, is $C^\pi_{\mathrm{lin}} = (\phi^\pi)^\top \Sigma^{-1} \phi^\pi$ (Eq. (5)), which is concerned with error propagation in the **true dynamics** M since ϕ^π is defined w.r.t. the occupancy μ^π in M. This begs the question:

*Is error-propagation in **compressed** models,*
as in (Q^π-irrelevant) abstractions, an exception and outlier?

While anecdotally this has been the general perception from the community, our results below suggest otherwise, and *the results that are seemingly disconnected with each other can be elegantly unified* through the following proposition (see Appendix B.4 for the proof):

Proposition 4. *Let $\mathcal{F}_\phi := \{\phi^\top\theta : \theta \in \mathbb{R}^d\}$ be the space of functions linear in ϕ. Assume $\mathcal{F}_\phi$ satisfies Bellman-completeness (Assumption 3). Then, (1) B^π becomes an exact model for next-feature prediction, i.e., $\mathbb{E}_{s'\sim P(\cdot|s,a)}[\phi(s',\pi)] = (B^\pi)^\top \phi(s,a)$, (2) $\mu^\pi_\phi = \phi^\pi$, (3) $\rho(B^\pi) \le 1$, and (4)*

$$C^\pi_\phi = C^\pi_{\mathrm{lin}} = (\phi^\pi)^\top \Sigma^{-1} \phi^\pi.$$

The essence of the proposition is illustrated in Figure 1, showing that the expected features produced by the groundtruth dynamics (ϕ^π) and the compressed dynamics ($\mu^\pi_\phi = (1-\gamma)\sum_t \gamma^t x_t$) coincide under Bellman-completeness, thus demonstrating that **error propagation through true dynamics is a special case of and thus unified with that through compressed dynamics.**

Connection to Bellman Residual Minimization (BRM). Many (if not most) algorithms for learning Q^π with general function approximation coincide with LSTDQ under linear function approximation (Antos et al., 2008; Xie et al., 2021; Uehara et al., 2020), and this fact allows us to compare our bound to the more general analyses in the literature. Among those algorithms, BRM

is a well-investigated example, which approximates Q^π by solving the following minimax problem (Antos et al., 2008):

$$\widehat{f}^\pi = \arg\min_{f\in\mathcal{F}} \sup_{f'\in\mathcal{F}} \left(\mathbb{E}_\mathcal{D}[(f(s,a)-r-f(s',\pi)^2] - \mathbb{E}_\mathcal{D}[(f'(s,a)-r-f(s',\pi)^2]\right), \tag{16}$$

whose finite-sample guarantee can be established under Bellman completeness (Assumption 3). Antos et al. (2008); Xie et al. (2021) show that when $\mathcal{F}$ is linear, the solution coincides with LSTDQ, so we can compare the guarantee of BRM under linear $\mathcal{F}$ with our Theorem 1. Jiang & Xie (2024) show that BRM's error bound is (see their Eq. (18))

$$\left|J_{\widehat{f}^\pi}(\pi) - J(\pi)\right| \lesssim \frac{V_{\max}}{1-\gamma}\cdot\sqrt{\frac{C^\pi \log(|\mathcal{F}|/\delta)}{n}}.$$

In the linear setting, their C^π is C^π_{lin} (see their Eq. (22)), and $\log|\mathcal{F}| \approx d$ based on a standard covering-number argument. Under such translation, the main $O(n^{-\frac{1}{2}})$ term in our Eq. (12) match the guarantee of BRM, not only in coverage, but also in horizon and d dependence.

5.4 Unification with Marginalized Importance Sampling

In Section 5.3 we mentioned that many algorithms designed for general function approximation reduce to LSTDQ when linear classes are used. Another example is Minimax Weight Learning (MWL; Uehara et al., 2020), a representative method for marginalized importance sampling, whose key idea is illustrated by the following inequality: given $\mathcal{F}$ such that $Q^\pi \in \mathcal{F}$, $\forall w : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$,

$$\left|\frac{1}{1-\gamma}\mathbb{E}_{\mu^D}[w(s,a)r] - J(\pi)\right| \le \sup_{f\in\mathcal{F}}\left|J_f(\pi) + \tfrac{1}{1-\gamma}\mathbb{E}_{\mu^D}[w(s,a)\cdot(\gamma f(s',\pi)-f(s,a))]\right|, \tag{17}$$

so learning w from some $\mathcal{W}$ class that minimizes (the empirical estimate of) the RHS to ≈ 0 ensures that $\frac{1}{1-\gamma}\mathbb{E}_{\mu^D}[w(s,a)r]$ is a good estimation of $J(\pi)$. Theoretically, if some $w^\star\in\mathcal{W}$ sets the RHS of Eq. (17) to 0, finite-sample guarantees can be established, where coverage is reflected by the magnitude of $w^\star$. As an example, $w^\star(s,a) = \mu^\pi(s,a)/\mu^D(s,a)$ always sets the RHS to 0, and we pay the size of $w^\star$ as the coverage parameter (e.g., $\|\mu^\pi/\mu^D\|_\infty$) through concentration inequalities; see Xie & Jiang (2020b, Section 6.2) for further discussions on this.

When both $\mathcal{W}$ and $\mathcal{F}$ are linear, Uehara et al. (2020) show that the MWL algorithm is equivalent to LSTDQ. We now show that their coverage parameters and guarantees, when improved with insights from follow-up works, coincide with our analyses in the linear setting. In particular, Zhang & Jiang (2024) point out that the $w^\star$ that minimizes the population objective Eq. (17) takes a different form in the linear case: $w^\star(s,a) = (1-\gamma)\phi_0^\top A^{-1}\phi(s,a)$. An immediate implication is that

$$\mathbb{E}_{\mu^D}[w^\star(s,a)^2] = C^\pi_\phi.$$

That is, the second moment of $w^\star$ on data is precisely our coverage parameter. While Uehara et al. (2020) measures the size of $w^\star$ by $\|w^\star\|_\infty$ due to the use of Hoeffding's inequality, replacing it with Bernstein's will improve $\|w^\star\|_\infty$ to $\mathbb{E}_{\mu^D}[w^\star(s,a)^2]$ in the main $O(n^{-1/2})$ term, which matches our bound in Eq.(12).

6 Conclusion and Discussion

We tackled the fundamental problem of linear off-policy evaluation under the minimal assumption of realizability. We re-analyzed a canonical algorithm for this setting, LSTDQ, and developed error bounds that introduced the feature-dynamics coverage, a new notion of coverage that tightens and sharpens our understanding of this setting. This parameter admits a natural interpretation as coverage in a feature-induced dynamical system, while simultaneously generalizing special cases such as aggregated concentrability with state abstraction features and linear coverage with Bellman-complete features. Altogether, our results serve as clearer and more unified foundation for the theory of linear OPE.

DISCLOSURE OF LLM USAGE

In the initial phase of the project, the authors had a vague conjecture and rough road-map of the main results in the paper, and used an LLM to execute the plan further to verify the feasibility of the project. We also subsequently used LLMs to help with literature review and proofs with some elementary linear-algebraic lemmas.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting. In *arXiv preprint arXiv:2011.01075*, 2020.

Philip Amortila, Nan Jiang, and Csaba Szepesvári. The optimal approximation factors in misspecified off-policy value function estimation. In *International Conference on Machine Learning*, 2023.

Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. In *International Conference on Learning Representations*, 2024a.

Philip Amortila, Dylan J Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability. In *International Conference on Machine Learning*, 2024b.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. In *Machine Learning*. Springer, 2008.

Dimitri Bertsekas. *Dynamic programming and optimal control*, volume II. Athena scientific, 2007.

Dimitri P Bertsekas and Huizhen Yu. Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227(1):27–50, 2009.

Justin A Boyan. Least-squares temporal difference learning. In *ICML*, pp. 49–56, 1999.

Qiwen Cui and Simon S Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Advances in Neural Information Processing Systems*, 35:11739–11751, 2022.

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.

Riccardo Della Vecchia and Debabrota Basu. Stochastic online instrumental variable regression: Regrets for endogeneity and bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16190–16198, 2025.

Yaqi Duan and Mengdi Wang. Minimax-optimal Off-Policy Evaluation with Linear Function Approximation. In *International Conference on Machine Learning*, 2020.

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

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 3169–3177. PMLR, 2024.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Daniel Hsu, Sham M Kakade, and Tong Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 6, 2011.

Audrey Huang and Nan Jiang. Beyond the return: Off-policy function estimation under user-specified error-measuring distributions. *Advances in Neural Information Processing Systems*, 35:6292–6303, 2022.

Zeyu Jia, Alexander Rakhlin, Ayush Sekhari, and Chen-Yu Wei. Offline reinforcement learning: Role of state aggregation and trajectory data. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2644–2719. PMLR, 2024.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 2021.

J Kolter. The fixed points of off-policy td. In *Advances in Neural Information Processing Systems*, 2011.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. In *Journal of Machine Learning Research*, 2003.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of lstd. In *ICML-27th International Conference on Machine Learning*, pp. 615–622, 2010.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *The Journal of Machine Learning Research*, 13(1):3041–3074, 2012.

Pai Liu, Lingfeng Zhao, Shivangi Agarwal, Jinghan Liu, Audrey Huang, Philip Amortila, and Nan Jiang. Model selection for off-policy evaluation: New algorithms and experimental protocol. *arXiv preprint arXiv:2502.08021*, 2025.

Diego Martinez-Taboada and Aaditya Ramdas. Empirical bernstein in smooth banach spaces. *arXiv preprint arXiv:2409.06060*, 2024.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

Wenlong Mou, Ashwin Pananjady, and Martin J Wainwright. Optimal oracle inequalities for solving projected fixed-point equations. In *Mathematics of Operations Research*. INFORMS, 2022a.

Wenlong Mou, Martin J. Wainwright, and Peter L. Bartlett. Off-policy estimation of linear functionals: non-asymptotic theory for semi-parametric efficiency, September 2022b. URL `http://arxiv.org/abs/2209.13075`. arXiv preprint arXiv:2209.13075 [cs, math, stat].

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. In *SIAM Journal on Control and Optimization*. SIAM, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. In *Journal of Machine Learning Research*, 2008.

A Nedić and Dimitri P Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110, 2003.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 752–759, 2008.

Juan C. Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. A complete characterization of linear estimators for offline policy evaluation. In *Journal of Machine Learning Research*, 2023.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. In *The Annals of Probability*, volume 22. Institute of Mathematical Statistics, 1994.

Gilbert W Stewart and Ji-guang Sun. Matrix perturbation theory. *(No Title)*, 1990.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. In *Foundations of computational mathematics*, volume 12. Springer, 2012.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. In *IEEE Transactions on Automatic Control*. IEEE, 1997.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 2020.

J. L. van Hemmen and T. Ando. An inequality for trace ideals. *Communications in Mathematical Physics*, 76(2):143–148, 1980.

Roman Vershynin. *High-dimensional Probability: an Introduction With Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham Kakade. Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*, pp. 10948–10960. PMLR, 2021.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability, 2020a.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020b.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Siyuan Zhang and Nan Jiang. Towards hyperparameter-free policy selection for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12864–12875, 2021.

Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. *Advances in Neural Information Processing Systems*, 37:124756–124790, 2024.

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *International Conference on Machine Learning*, pp. 40804–40829. PMLR, 2023.

A Proofs of Section 4

A.1 Proof of Theorem 1

Theorem 1 (Main Theorem)**.** *Under Assumptions 1 and 2, with probability at least* $1-\delta$*,*

$$\left|J_{\widehat{Q}_{\mathrm{lstd}}}(\pi) - J(\pi)\right| \lesssim \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{\widehat{C}_\phi^\pi \cdot d\log(1/\delta)}{n}} \tag{10}$$

where

$$\widehat{C}_\phi^\pi := (1-\gamma)^2 \phi_0^\top \widehat{A}^{-1} \widehat{\Sigma} \widehat{A}^{-\top} \phi_0. \tag{11}$$

Here we take the convention that $\widehat{C}_\phi^\pi = +\infty$ *if* $\widehat{A}$ *or* $\widehat{\Sigma}$ *is not invertible.*

Proof of Theorem 1. We start by writing:

$$\begin{aligned}\left|J_{\widehat{Q}_{\mathrm{lstd}}}(\pi) - J(\pi)\right| &= \left|\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}\left[Q^\pi(s_0,a_0) - \hat{Q}_{\mathrm{lstd}}(s_0,a_0)\right]\right| \\ &= \left|\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}\left[\phi(s_0,a_0)^\top\left(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\right)\right]\right|,\end{aligned}$$

where in the second line we have used realizability (Assumption 1) and the definition of $\widehat{Q}_{\mathrm{lstd}}$. We can continue with simple algebra to find that:

$$\begin{aligned}\left|\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}\left[\phi(s_0,a_0)^\top\left(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\right)\right]\right| &= \left|\phi_0^\top\left(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\right)\right| \\ &= \left|\phi_0^\top \hat{A}^{-1}\left(\hat{A}\theta^\star - \hat{A}\widehat{\theta}_{\mathrm{lstd}}\right)\right| \\ &= \left|\phi_0^\top \hat{A}^{-1}\widehat{\Sigma}^{1/2}\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{A}\widehat{\theta}_{\mathrm{lstd}}\right)\right| \\ &\le \left\|\widehat{\Sigma}^{1/2}\hat{A}^{-\top}\phi_0\right\|_2 \left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{A}\widehat{\theta}_{\mathrm{lstd}}\right)\right\|_2,\end{aligned}$$

where in the last line we have used Cauchy-Schwartz. To proceed, we note that, when $\hat{A}$ is invertible, we have $\hat{A}\widehat{\theta}_{\mathrm{lstd}} - \hat{b} = \hat{A}\left(\hat{A}^{-1}\hat{b}\right) - \hat{b} = 0$, and thus, $\left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\widehat{\theta}_{\mathrm{lstd}} - b\right)\right\|_2 = 0$ whenever $\widehat{\Sigma}$ is invertible. Thus,

$$\begin{aligned}\left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{A}\widehat{\theta}_{\mathrm{lstd}}\right)\right\|_2 &\le \left(\left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{b}\right)\right\|_2 + \left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\widehat{\theta}_{\mathrm{lstd}} - \hat{b}\right)\right\|_2\right) \\ &\le \left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{b}\right)\right\|_2.\end{aligned}$$

We then note that

$$\left\|\widehat{\Sigma}^{1/2}\hat{A}^{-\top}\phi_0\right\|_2 = \sqrt{\phi_0^\top \hat{A}^{-1}\widehat{\Sigma}\hat{A}^{-T}\phi_0} = \frac{1}{1-\gamma}\sqrt{\widehat{C}_\phi^\pi},$$

which yields that

$$\left|J_{\widehat{Q}_{\mathrm{lstd}}}(\pi) - J(\pi)\right| \le \frac{1}{1-\gamma}\sqrt{\widehat{C}_\phi^\pi}\left\|\widehat{\Sigma}^{-1/2}\left(\hat{A}\theta^\star - \hat{b}\right)\right\|_2.$$

The proof will be concluded by establishing the following concentration lemma.

Lemma 1. *With probability* $1-\delta$ *over the randomness of the rewards and sampled transitions, we have:*

$$\|\widehat{\Sigma}^{-1/2}(\hat{A}\theta^\star - \hat{b})\|_2 \le \mathcal{O}\left(V_{\max}\sqrt{\frac{d\log(1/\delta)}{n}}\right).$$

□

Proof of Lemma 1. We firstly note that

$$\begin{aligned}\hat{A}\theta^\star - \hat{b} &= \frac{1}{n}\sum_{i=1}^n \phi(s_i,a_i)\big(\phi(s_i,a_i)^\top\theta^\star - \gamma\phi(s_i',a_i')^\top\theta^\star - r_i\big)\\ &= \frac{1}{n}\sum_{i=1}^n \phi(s_i,a_i)\underbrace{\big(Q^\pi(s_i,a_i) - r_i - \gamma Q^\pi(s_i',a_i')\big)}_{:=\varepsilon_i}.\end{aligned}$$

Thus, $\hat{A}\theta^\star - \hat{b}$ is a random variable that is conditionally zero-mean, when taking conditional expectations over the $r_i \sim R(s_i,a_i)$, $s_i' \sim P(s_i,a_i)$, and $a_i' \sim \pi(\cdot \mid s_i')$ (keeping the design over s_i, a_i fixed). Note that $|\varepsilon_i| \le 2V_{\max}$.

We apply a vector martingale Bernstein inequality (Lemma 10) on the random variables $Z_i = \widehat{\Sigma}^{-1/2}\phi(s_i,a_i)\varepsilon_i$. For $i \in [n]$, we let $\mathcal{H}_i = \{s_1,a_1,\ldots,s_n,a_n\} \cup \{r_1,s_1',a_1',\ldots,r_i,s_i',a_i'\}$ denote the histories including the entire design over s_i, a_i but only the first i samples from r_i, s_i', a_i'. Note that Z_i is adapted to the filtration generated by $\mathcal{H}_i$, and is a martingale difference sequence, since

$$\mathbb{E}_{i-1}[\widehat{\Sigma}^{-1/2}\phi(s_i,a_i)\varepsilon_i] = \mathbb{E}_{i-1}[\varepsilon_i]\|\phi(s_i,a_i)\|_{\widehat{\Sigma}^{-1}} = 0.$$

In the sequel we establish the following simple technical lemma.

Lemma 2. *Let $x_1,\ldots,x_n \in \mathbb{R}^d$, and assume $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ is invertible. Then for all $i \in [n]$ we have:*

$$x_i^\top \widehat{\Sigma}^{-1} x_i \le n$$

Using, Lemma 2, we then have that

$$\|\widehat{\Sigma}^{-1/2}\phi(s_i,a_i)\varepsilon_i\|_2 \le 2V_{\max}\|\phi(s_i,a_i)\|_{\widehat{\Sigma}^{-1}} \le 2V_{\max}\sqrt{n},$$

which establishes the norm bound. Lastly, for the variance term, we have:

$$\sum_{i=1}^n \mathbb{E}_{i-1}[\|X_i\|^2] = \sum_{i=1}^n \mathbb{E}_{i-1}[\varepsilon_i^2\|\phi_i\|^2_{\widehat{\Sigma}^{-1}}] \le 4V_{\max}^2\sum_{i=1}^n \|\phi_i\|^2_{\widehat{\Sigma}^{-1}}.$$

Note that the summation is equivalent to:

$$\sum_{i=1}^n \phi_i^\top\widehat{\Sigma}^{-1}\phi_i = \sum_{i=1}^n \mathrm{tr}(\widehat{\Sigma}^{-1}\phi_i\phi_i^\top) = n\mathrm{tr}(\widehat{\Sigma}^{-1}\frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top) = nd,$$

since the trace of the identity matrix is d. Plugging these observations into Lemma 10 gives that

$$\begin{aligned}\left\|\widehat{\Sigma}^{-1/2}(\hat{A}\theta^\star - \hat{b})\right\|_2 = \frac{1}{n}\left\|\sum_{i=1}^n Z_i\right\|_2 &\le \frac{1}{n}\left(V_{\max}\sqrt{8nd\log(2/\delta)} + \frac{2}{3}V_{\max}\sqrt{n}\log(2/\delta)\right)\\ &= \mathcal{O}\left(V_{\max}\sqrt{\frac{d\log(1/\delta)}{n}}\right),\end{aligned}$$

as desired. □

Proof of Lemma 2. Consider the un-normalized empirical covariance matrix $\widehat{\Sigma}_{\mathrm{un}} = \sum_{i=1}^n x_i x_i^\top$. Let $v = x_i^\top\Sigma_{\mathrm{un}}^{-1}x_i$. For each i, let $\Sigma = S_i + x_i x_i^\top$. Note that $S_i = \sum_{j\neq i} x_j x_j^\top$ is a PSD matrix, as is $\widehat{\Sigma}_{\mathrm{un}}^{-1}S_i\widehat{\Sigma}_{\mathrm{un}}^{-1}$. Then, we have

$$0 \le x_i^\top\widehat{\Sigma}_{\mathrm{un}}^{-1}S_i\widehat{\Sigma}_{\mathrm{un}}^{-1}x_i = x_i^\top\Sigma_{\mathrm{un}}^{-1}x_i - (x_i^\top\Sigma_{\mathrm{un}}^{-1}x_i)^2.$$

This implies that $v(1-v) \ge 0$, thus $0 \le v \le 1$. □

A.2 Proof of Corollary 1

Corollary 1. *There exists n_0 such that when $n \geq n_0$, w.p. $\geq 1-\delta$,*

$$\left|J_{\widehat{Q}_{\mathrm{lstd}}}(\pi) - J(\pi)\right| \lesssim \frac{V_{\max}}{1-\gamma}\sqrt{\frac{C_\phi^\pi \cdot d\log(1/\delta)}{n}} + o(\sqrt{1/n}), \tag{12}$$

where

$$C_\phi^\pi := (1-\gamma)^2 \phi_0^\top A^{-1} \Sigma A^{-\top} \phi_0 \tag{13}$$

and n_0 and the $o(\sqrt{1/n})$ term may additionally depend on $1/\sigma_{\min}(A)$.

Proof of Corollary 1. We begin by noting that it is sufficient to provide a high-probability bound on $\left|C_\phi^\pi - \widehat{C}_\phi^\pi\right| \leq \varepsilon_n$ for some $\varepsilon_n = o(1)$, since by Theorem 1 and the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we will then obtain

$$\begin{aligned}
\left|J_{\widehat{Q}_{\mathrm{lstd}}}(\pi) - J(\pi)\right| &\lesssim \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{\left(C_\phi^\pi + \varepsilon_n\right) \cdot d\log(1/\delta)}{n}} \\
&\leq \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{C_\phi^\pi \cdot d\log(1/\delta)}{n}} + \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{\varepsilon_n \cdot d\log(1/\delta)}{n}} \\
&= \frac{V_{\max}}{1-\gamma} \cdot \sqrt{\frac{C_\phi^\pi \cdot d\log(1/\delta)}{n}} + o(\sqrt{1/n}).
\end{aligned}$$

We now proceed to bound $\left|C_\phi^\pi - \widehat{C}_\phi^\pi\right|$ with high probability. Towards this, we note that:

$$\begin{aligned}
\left|C_\phi^\pi - \widehat{C}_\phi^\pi\right| &= (1-\gamma)^2 \left| \|\Sigma^{1/2} A^{-\top} \phi_0\|_2 - \|\widehat{\Sigma}^{1/2} \hat{A}^{-\top} \phi_0\|_2 \right| \\
&\leq (1-\gamma)^2 \left\| \Sigma^{1/2} A^{-\top} \phi_0 - \widehat{\Sigma}^{1/2} \hat{A}^{-\top} \phi_0 \right\|_2 \\
&\leq (1-\gamma)^2 \left(\left\| \widehat{\Sigma}^{1/2} \left(A^{-\top} - \hat{A}^{-\top} \right) \phi_0 \right\|_2 + \left\| \left(\widehat{\Sigma}^{1/2} - \Sigma^{1/2} \right) A^{-\top} \phi_0 \right\|_2 \right) \\
&\leq (1-\gamma)^2 B_\phi \left(\left\| \widehat{\Sigma}^{1/2} \right\|_2 \|A^{-1} - \hat{A}^{-1}\|_2 + \left\| \widehat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_2 \|A^{-1}\|_2 \right),
\end{aligned}$$

via applications of the triangle inequality and operator norm bounds. Let $\varepsilon(\Sigma^{1/2}) = \|\Sigma^{1/2} - \widehat{\Sigma}^{1/2}\|_2$ and $\varepsilon(A^{-1}) = \|A^{-1} - \hat{A}^{-1}\|_2$. Note that the above inequalities imply

$$\left|C_\phi^\pi - \widehat{C}_\phi^\pi\right| \leq (1-\gamma)^2 B_\phi \left(\left(\lambda_{\max}(\Sigma^{1/2}) + \varepsilon(\Sigma^{1/2}) \right) \varepsilon(A^{-1}) + \varepsilon(\Sigma^{1/2}) \frac{1}{\sigma_{\min}(A)} \right). \tag{18}$$

We conclude by bounding $\varepsilon(\Sigma^{1/2})$ and $\varepsilon(A^{-1})$. We first establish a concentration lemma for $\|\Sigma - \widehat{\Sigma}\|_2$ and $\|A - \hat{A}\|_2$, and then show how this can be converted to bounds for $\|\Sigma^{1/2} - \widehat{\Sigma}^{1/2}\|_2$ and $\|A^{-1} - \hat{A}^{-1}\|_2$. The following concentration lemma will be proved in the sequel. □

Lemma 3. *With probability at least $1-\delta$, we have:*

$$\|\Sigma - \widehat{\Sigma}\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma)) \log(d/\delta)}{n}} \right) := \epsilon(\Sigma).$$

and

$$\|A - \hat{A}\|_2 \leq \mathcal{O}\left(\left(B_\phi^2 + \sigma_{\max}(A)\right) \sqrt{\frac{\log(d/\delta)}{n}} \right) := \epsilon(A).$$

We use the above bound on $\|\Sigma - \widehat{\Sigma}\|_2$ in combination with the inequality $\|\Sigma^{1/2} - \widehat{\Sigma}^{1/2}\|_2 \leq \sqrt{\|\Sigma - \widehat{\Sigma}\|_2}$ (van Hemmen & Ando, 1980)[5], to obtain:

$$\|\Sigma^{1/2} - \widehat{\Sigma}^{1/2}\|_2 \leq \sqrt{\|\Sigma - \widehat{\Sigma}\|_2} \leq \sqrt{\epsilon(\Sigma)} = \mathcal{O}\left(\left(\frac{\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(d/\delta)}{n}\right)^{1/4}\right)$$

Then, to bound $\|A - \hat{A}^{-1}\|_2$, we note the following lemma.

Lemma 4 ((Stewart & Sun, 1990)). *Let $A \in \mathbb{R}^{m\times n}$, with $m \geq n$ and let $\widetilde{A} = A + E$. Then,*

$$\epsilon(A^{-1}) \leq \frac{1+\sqrt{5}}{2}\max\Big\{\|A^{-1}\|_2, \|\widetilde{A}^{-1}\|_2\Big\}\|E\|_2.$$

Furthermore, if $\|E\|_2 \leq \sigma_{\min}(A)/2$, then

$$\|\widetilde{A}^{-1} - A^{-1}\|_2 \lesssim \|A^{-1}\|_2^2\|E\|_2.$$

This immediately implies that, for $\|A - \hat{A}\|_2 \leq \sigma_{\min}(A)/2$, we have

$$\|\widetilde{A}^{-1} - A^{-1}\|_2 \leq \frac{1}{\sigma_{\min}(A)^2}\|A - \hat{A}\|_2 = \mathcal{O}\left(\frac{B_\phi^2 + \sigma_{\max}(A)}{\sigma_{\min}(A)^2}\sqrt{\frac{\log(d/\delta)}{n}}\right)$$

This latter condition is equivalent to

$$\big(B_\phi^2 + \sigma_{\max}(A)\big)\sqrt{\frac{\log(d/\delta)}{n}} \lesssim \frac{\sigma_{\min}(A)}{2} \implies n \gtrsim \underbrace{\left(\frac{B_\phi^2 + \sigma_{\max}(A)}{\sigma_{\min}(A)}\right)^2 \log(d/\delta)}_{:=n_0}.$$

We set this latter quantity as our burn-in time n_0. Returning to Eq. (18) and combining everything, we have:

$$\begin{aligned}\left|C_\phi^\pi - \widehat{C}_\phi^\pi\right| &\leq \frac{(1-\gamma)^2 B_\phi\Big(B_\phi^2 + \sigma_{\max}(A)\Big)}{\sigma_{\min}(A)^2}\sqrt{\frac{\log(d/\delta)}{n}}\Big(\Big(\frac{\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(d/\delta)}{n}\Big)^{1/4} + \sqrt{\lambda_{\max}(\Sigma)}\Big) \\ &\quad + \frac{(1-\gamma)^2 B_\phi}{\sigma_{\min}(A)}\Big(\frac{\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(d/\delta)}{n}\Big)^{1/4} \\ &= o(1)\end{aligned}$$

Proof of Lemma 3. We firstly establish that the bound on $\|\widehat{\Sigma} - \Sigma\|_2$. To do this, we use Matrix Bernstein (Lemma 8). Abbreviate $X_i \coloneqq \phi(s_i, a_i)$, and let $Z_i = X_i X_i^\top - \Sigma$ be the centered matrices. For the almost sure bound, we have

$$\|Z_i\|_2 \leq \|X_i X_i^\top\|_2 + \|\Sigma\|_2 \leq \|X_i\|_2^2 + \lambda_{\max}(\Sigma) \leq B_\phi^2 + \lambda_{\max}(\Sigma).$$

For the variance term, we have:

$$\begin{aligned}\big\|\mathbb{E}\big[(X_i X_i^\top - \Sigma)^2\big]\big\|_2 &= \big\|\mathbb{E}\big[(X_i X_i^\top)^2\big] - \Sigma^2\big\|_2 \\ &\leq \big\|B_\phi^2\,\mathbb{E}\big[X_i X_i^\top\big] - \Sigma^2\big\|_2 \\ &\leq B_\phi^2\lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma)^2.\end{aligned}$$

This yields

$$\begin{aligned}\|\hat{\Sigma} - \Sigma\|_2 &\leq \sqrt{\frac{2\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{n}} + \frac{2(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{3n} \\ &= \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(d/\delta)}{n}}\right).\end{aligned}$$

[5] See also this answer by user "jlewk" on Math StackExchange: https://math.stackexchange.com/a/3968174

We now establish the bound on $\|\hat{A} - A\|_2$ again via Matrix Bernstein (Lemma 8). Define the notation $X_i = \phi(s_i, a_i)$ and $X_i' = \phi(s_i, a_i) - \gamma\phi(s_i', a_i')$. Then we let $Z_i = X_i(X_i')^\top - A$ denote the centered matrices. For the almost sure norm bound, we have

$$\|Z_i\|_2 \leq \|X_i\|_2\|X_i'\|_2 + \|A\|_2 \leq B_\phi^2(1+\gamma) + \sigma_{\max}(A) \leq 2B_\phi^2 + \sigma_{\max}(A).$$

For the variance terms, we have:

$$\begin{aligned}\left\|\mathbb{E}\Big[\big(X_i(X_i')^\top - A\big)\big(X_i(X_i')^\top - A\big)^\top\Big]\right\|_2 &= \left\|\mathbb{E}\big[X_i(X_i')^\top X_i' X_i^\top - AX_i'X_i^\top - X_i'X_i^\top A + AA^\top\big]\right\|_2 \\ &\leq 4B_\phi^2\left\|\mathbb{E}\big[X_iX_i^\top\big]\right\|_2 + \left\|AA^\top\right\|_2 \\ &= 4B_\phi^2\lambda_{\max}(\Sigma) + \sigma_{\max}(A)^2,\end{aligned}$$

as well as

$$\begin{aligned}\left\|\mathbb{E}\Big[\big(X_i(X_i')^\top - A\big)^\top\big(X_i(X_i')^\top - A\big)\Big]\right\|_2 &= \left\|\mathbb{E}\big[X_i'(X_i)^\top X_i(X_i')^\top - A^\top X_i(X_i')^\top - X_i'X_i^\top A + A^\top A\big]\right\|_2 \\ &\leq 4B_\phi^4 + \left\|AA^\top\right\|_2 \\ &= 4B_\phi^4 + \sigma_{\max}(A)^2.\end{aligned}$$

With the latter, the variance term and the norm bound are of the same order, which gives

$$\|\hat{A} - A\|_2 \leq \mathcal{O}\left(\big(B_\phi^2 + \sigma_{\max}(A)\big)\sqrt{\frac{\log(d/\delta)}{n}}\right).$$

□

B Proofs of Section 5

B.1 Proof of Proposition 1

Proposition 1. *Define a deterministic linear dynamical system $\{x_t\}_{t\geq 0}$, with $x_0 := \phi_0$, and $\forall t \geq 0$,*

$$x_{t+1} = (B^\pi)^\top x_t.$$

When $\rho(B^\pi) < 1/\gamma$, define the feature occupancy in B^π as $\mu_\phi^\pi := (1-\gamma)\sum_{t\geq 0}\gamma^t x_t$, then

$$C_\phi^\pi = (\mu_\phi^\pi)^\top \Sigma^{-1}\mu_\phi^\pi.$$

Proof of Proposition 1. Recall that we defined $B^\pi = \Sigma^{-1}\Sigma_{\text{cr}}$. We note that

$$A = \Sigma - \gamma\Sigma_{\text{cr}} = \Sigma(I - \gamma B^\pi).$$

Substituting this into C_ϕ^π, we arrive at the expression:

$$C_\phi^\pi = (1-\gamma)^2\phi_0^\top A^{-1}\Sigma A^{-T}\phi_0 = (1-\gamma)^2\phi_0^\top(I-\gamma B^\pi)^{-1}\Sigma^{-1}(I-\gamma B^\pi)^{-T}\phi_0. \tag{19}$$

Note that when $\rho(B^\pi) < 1/\gamma$, the matrix $(I-\gamma B^\pi)^{-\top}$ has the series expansion:

$$(I-\gamma B^\pi)^{-\top} = \sum_{t=0}^{\infty}\gamma^t((B^\pi)^\top)^t.$$

Thus, we notice that

$$(I-\gamma B^\pi)^{-\top}\phi_0 = \sum_{t=0}^{\infty}\gamma^t((B^\pi)^\top)^t\phi_0 = \sum_{t=0}^{\infty}\gamma^t x_t = \frac{1}{1-\gamma}\mu_\phi^\pi.$$

Substituting this into Eq. (19) gives the result. □

B.2 Proof of Proposition 2

Proposition 2.

$$\sqrt{C_\phi^\pi} = (1-\gamma)\|(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})^{-T}\Sigma^{-1/2}\phi_0\|_2 \le (1-\gamma)\frac{\|\phi_0\|_{\Sigma^{-1}}}{\sigma_{\min}(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})}.$$

Proof of Proposition 2. The derivation is as follows:

$$\begin{aligned}
\phi_0^\top A^{-1}\Sigma A^{-T}\phi_0 &= \phi_0(\Sigma-\gamma\Sigma_{\mathrm{cr}})^{-1}\Sigma(\Sigma-\gamma\Sigma_{\mathrm{cr}})^{-T}\phi_0 \\
&= \phi_0^\top\Big(\Sigma^{1/2}(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})\Sigma^{1/2}\Big)^{-1}\Sigma\Big(\Sigma^{1/2}(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})\Sigma^{1/2}\Big)^{-T}\phi_0 \\
&= \phi_0^\top\Sigma^{-1/2}(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})^{-1}(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})^{-T}\Sigma^{-1/2}\phi_0 \\
&= \|(I-\gamma\Sigma^{-1/2}\Sigma_{\mathrm{cr}}\Sigma^{-1/2})^{-T}\Sigma^{-1/2}\phi_0\|_2^2.
\end{aligned}$$

□

B.3 Proof of Proposition 3

Proposition 3. *When ϕ is induced by a state abstraction ψ and π depends on s only through $\psi(s)$,*

$$C_\phi^\pi = \mathbb{E}_{(k,a)\sim\mu_\phi^D}[(\mu_{M_\phi}^\pi/\mu_\phi^D)^2].$$

Proof of Proposition 3. Let $\phi(s,a) = e_{\psi(s),a}$, where ψ is the state abstraction function. We compute the A matrix. Below, we define $P^\pi(s',a' \mid s,a) = P(s' \mid s,a)\pi(a' \mid s')$, $P(k' \mid s,a) = \sum_{s':\psi(s')=k} P(s' \mid s,a)$, and

$$P^\pi(k',a' \mid s,a) = P(k' \mid s,a)\pi(a' \mid k'),$$

which is valid since π is consistent with the state abstraction. To start, the covariance matrix Σ becomes

$$\begin{aligned}
\Sigma = \mathbb{E}_{s,a\sim\mu^D}\left[\phi(s,a)\phi(s,a)^\top\right] &= \sum_{k\in[K],a\in[A]}\sum_{s\in\mathcal{S}:\psi(s)=k}\mu^D(s,a)e_{k,a}e_{k,a}^\top \\
&= \sum_{k\in[K],a\in[A]} e_{k,a}e_{k,a}^\top\left(\sum_{s\in\mathcal{S}:\psi(s)=k}\mu^D(s,a)\right) \\
&= \sum_{k\in[K],a\in[A]} e_{k,a}e_{k,a}^\top\mu_\phi^D(k,a) \coloneqq \bar{D}_{\mathrm{data}} \in \mathbb{R}^{[K]\cdot[A]\times[K]\cdot[A]},
\end{aligned}$$

where we recalled the definition of $\mu_\phi^D(k,a) = \sum_{s\in\mathcal{S}:\psi(s)=k} \mu^D(s,a)$, and introduced the diagonal matrix $\bar{D}_{\text{data}}$ with elements $\mu_\phi^D(k,a)$ along the diagonal. Let's examine the cross-covariance Σ_{cr}.

$$
\begin{aligned}
\Sigma_{\text{cr}} &= \mathbb{E}_{s,a\sim\mu^D}\left[\phi(s,a)\phi(s',a')^\top\right] \\
&= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mu^D(s,a)\phi(s,a) \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P^\pi(s',a' \mid s,a)\phi(s',a')^\top \\
&= \sum_{k\in[K],a\in[A]} e_{k,a} \sum_{s\in\mathcal{S}:\psi(s)=k} \mu^D(s,a) \left(\sum_{k'\in[K],a'\in[A]} \sum_{s'\in\mathcal{S}:\psi(s')=k'} P^\pi(s',a' \mid s,a) e_{k',a'}^\top \right) \\
&= \sum_{k\in[K],a\in[A]} e_{k,a} \sum_{s\in\mathcal{S}:\psi(s)=k} \mu^D(s,a) \left(\sum_{k'\in[K],a'\in[A]} e_{k',a'}^\top P^\pi(k',a' \mid s,a) \right) \\
&= \sum_{k\in[K],a\in[A]} e_{k,a} \sum_{k'\in[K],a'\in[A]} e_{k',a'}^\top \sum_{s\in\mathcal{S}:\psi(s)=k} \mu^D(s,a) P^\pi(k',a' \mid s,a) \\
&= \sum_{k\in[K],a\in[A]} e_{k,a} \sum_{k'\in[K],a'\in[A]} e_{k',a'}^\top \mu_\phi^D(k,a) \left(\frac{\sum_{s\in\mathcal{S}:\psi(s)=k} \mu^D(s,a) P^\pi(k',a' \mid s,a)}{\mu_\phi^D(k,a)} \right) \\
&:= \sum_{k\in[K],a\in[A]} e_{k,a} \sum_{k'\in[K],a'\in[A]} e_{k',a'}^\top \mu_\phi^D(k,a) P_\phi^\pi(k',a' \mid k,a) \\
&= \bar{D}_{\text{data}} P_\phi^\pi,
\end{aligned}
$$

where we recall the definition of the aggregated transition matrix P_ϕ^π with elements

$$
P_\phi^\pi(k',a' \mid k,a) = \frac{\sum_{s:\psi(s)=k} \mu^D(s,a) P^\pi(k',a' \mid s,a)}{\mu_\phi^D(k,a)}.
$$

Putting our expressions for Σ and Σ_{cr} together, we conclude that

$$
A = \bar{D}_{\text{data}}(I - \gamma P_\phi^\pi).
$$

Note that P_ϕ^π is the π-dependent transition kernel of the MDP M_ϕ over the state space $[K]$ with action space $[A]$. We assign the MDP an initial state-action distribution $\mu_{0,\phi}$ in the canonical way:

$$
\mu_{0,\phi}^\pi(k,a) = \sum_{s:\psi(s)=k} \mu_0(s)\pi(a \mid s) = \mu_0(k)\pi(a \mid k),
$$

again using the fact that π is consistent with the abstraction ψ. Note that in the state abstraction setting, we have

$$
\phi_0 = \mathbb{E}_{s_0\sim\mu_0,a\sim\pi}[\phi(s,a)] = \sum_{k,a} e_{k,a} \sum_{s:\psi(s)=k} \mu_0(s)\pi(a \mid s) = \mu_{0,\phi}^\pi,
$$

Finally, our coverage coefficient becomes

$$
C_\phi^\pi = (1-\gamma)^2 \phi_0 A^{-1} \Sigma A^{-T} \phi_0 = (1-\gamma)^2 \mu_{0,\phi}^\pi (I - \gamma P_\phi^\pi)^{-1} \bar{D}_{\text{data}}^{-1} (I - \gamma P_\phi^\pi)^{-T} \mu_{0,\phi}^\pi.
$$

Since P_ϕ^π is a stochastic kernel with spectral radius less than 1, we have

$$
(I - \gamma P_\phi^\pi)^{-T} \mu_{0,\phi} = \sum_{t\geq 0} \gamma^t ((P_\phi^\pi)^t))^\top \mu_{0,\phi} = \frac{1}{1-\gamma} \mu_{M_\phi}^\pi,
$$

i.e. this is precisely the discounted occupancy of policy π in the abstract MDP M_ϕ. Thus,

$$
C_\phi^\pi = (\mu_{M_\phi}^\pi)^\top \bar{D}_{\text{data}}^{-1} (\mu_{M_\phi}^\pi) = \mathbb{E}_{k,a\sim\mu_\phi^D}\left[(\mu_{M_\phi}^\pi / \mu_\phi^D)^2 \right],
$$

as desired.

□

B.4 Proof of Proposition 4

Proposition 4. *Let $\mathcal{F}_\phi := \{\phi^\top \theta : \theta \in \mathbb{R}^d\}$ be the space of functions linear in ϕ. Assume $\mathcal{F}_\phi$ satisfies Bellman-completeness (Assumption 3). Then, (1) B^π becomes an exact model for next-feature prediction, i.e., $\mathbb{E}_{s' \sim P(\cdot|s,a)}[\phi(s', \pi)] = (B^\pi)^\top \phi(s, a)$, (2) $\mu^\pi_\phi = \phi^\pi$, (3) $\rho(B^\pi) \leq 1$, and (4)*

$$C^\pi_\phi = C^\pi_{\text{lin}} = (\phi^\pi)^\top \Sigma^{-1} \phi^\pi.$$

Proof of Proposition 4.

(1) B^π is an exact next-feature predictor: $\mathbb{E}_{s' \sim P(\cdot|s,a)}[\phi(s, a)] = B^\pi \phi(s, a)$ for all (s, a). First, we show that under Bellman completeness $\mathcal{F}_\phi$ is also closed under the transition operator $\mathcal{P}^\pi := \mathcal{T}^\pi - R$, that is, $\mathcal{P}^\pi f \in \mathcal{F}_\phi$ for all $f \in \mathcal{F}_\phi$.

The linearity of $\mathcal{F}_\phi$ together with Bellman completeness immediately imply that the reward function is linear, or $R \in \mathcal{F}_\phi$. Define $f_0 \in \mathcal{F}_\phi$ to be the function corresponding to the parameter $\theta = \mathbf{0}_d$, so that $f_0(s, a) = 0$ for all (s, a); we have $\mathcal{T}^\pi f_0 = R \in \mathcal{F}_\phi$.

Next, fix any $f \in \mathcal{F}_\phi$ and observe that $\mathcal{T}^\pi f$ is also linear, since $\mathcal{T}^\pi f \in \mathcal{F}_\phi$ under Bellman completeness. It follows that $\mathcal{T}^\pi f - R = \mathcal{P}^\pi f \in \mathcal{F}_\phi$ because the difference of two functions linear in the same features is also linear in those features, which proves that $\mathcal{F}_\phi$ is closed under $\mathcal{P}^\pi$. This closure implies that for any $f \in \mathcal{F}_\phi$, there exists some $\theta_f \in \mathbb{R}^d$ such that

$$\phi(s, a)^\top \theta_f = (\mathcal{P}^\pi f)(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s', \pi)].$$

To prove the stated claim we will utilize choice instantations of such functions and their corresponding parameters. For $i \in [d]$, define the function $f_i := \langle \phi, \mathbf{e}_i \rangle \in \mathcal{F}_\phi$, and let $\theta_i \in \mathbb{R}^d$ be such that

$$\phi(s, a)^\top \theta_i = \mathbb{E}_{s' \sim P(\cdot|s,a)}[f_i(s', \pi)], \ \forall (s, a).$$

Then for all (s, a),

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}[\phi(s', \pi)] = \begin{bmatrix} \mathbb{E}_{s' \sim P(\cdot|s,a)}[f_1(s', \pi)] \\ \mathbb{E}_{s' \sim P(\cdot|s,a)}[f_2(s', \pi)] \\ \vdots \\ \mathbb{E}_{s' \sim P(\cdot|s,a)}[f_d(s', \pi)] \end{bmatrix} = \underbrace{\begin{bmatrix} — & \theta_1^\top & — \\ — & \theta_2^\top & — \\ & \vdots & \\ — & \theta_d^\top & — \end{bmatrix}}_{(*)} \phi(s, a).$$

Lastly, we will show that the above system of equations is satisfied by setting

$$(*) = \Sigma_{\text{cr}}^{-\top} \Sigma^{-1} = (B^\pi)^\top.$$

Right-multiplying both sides by $\phi(s, a)^\top$ then taking the expectation over $(s, a) \sim \mu^D$, we obtain

$$\mathbb{E}_{(s,a,s',a') \sim \mu^D \times P \times \pi}\left[\phi(s', a')\phi(s, a)^\top\right] = (B^\pi)^\top \mathbb{E}_{(s,a) \sim \mu^D}\left[\phi(s, a)\phi(s, a)^\top\right].$$

Solving for $(B^\pi)^\top$ and rearranging gives

$$\begin{aligned} (B^\pi)^\top &= \left(\mathbb{E}_{(s,a,s',a') \sim \mu^D \times P \times \pi}\left[\phi(s, a)\phi(s', a')^\top\right]\right)^\top \Sigma^{-1} \\ &= \Sigma_{\text{cr}}^\top \Sigma^{-1}, \end{aligned}$$

which confirms that $B^\pi = \Sigma^{-1} \Sigma_{\text{cr}}$ satisfies for all (s, a) the equivalence

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}[\phi(s', \pi)] = (B^\pi)^\top \phi(s, a).$$

(2) Showing $\mu_\phi^\pi = \phi^\pi$. Recall that $\phi^\pi = \mathbb{E}_{(s,a)\sim\mu^\pi}[\phi(s,a)]$. Using the Bellman flow equations for μ^π, we obtain a recursive system of equations for the dynamics of ϕ^π:

$$\begin{aligned}
\phi^\pi &= \sum_{s,a} \phi(s,a)\mu^\pi(s,a) \\
&= \sum_{s,a} \phi(s,a)\left((1-\gamma)\mu_0^\pi(s,a) + \gamma \sum_{s',a'} P^\pi(s,a \mid s',a')\mu^\pi(s',a')\right) \\
&= (1-\gamma)\phi_0 + \gamma\, \mathbb{E}_{(s,a)\sim\mu^\pi}\left[\mathbb{E}_{s'\sim P(\cdot|s,a)}[\phi(s',\pi)]\right] \\
&= (1-\gamma)\phi_0 + \gamma\, \mathbb{E}_{(s,a)\sim\mu^\pi}\left[(B^\pi)^\top \phi(s,a)\right] \\
&= (1-\gamma)\phi_0 + \gamma (B^\pi)^\top \phi^\pi,
\end{aligned}$$

where we invoke the result from **(1)** in the second-to-last line. Repeatedly expanding the RHS of the equation with the recursion,

$$\begin{aligned}
\phi^\pi &= (1-\gamma)\phi_0 + \gamma (B^\pi)^\top \phi^\pi, \\
&= (1-\gamma)\phi_0 + \gamma (B^\pi)^\top \left((1-\gamma)\phi_0 + \gamma (B^\pi)^\top \phi^\pi\right) \\
&= (1-\gamma)\left(\phi_0 + \gamma (B^\pi)^\top \phi_0 + \gamma^2 \left((B^\pi)^\top\right)^2 \phi^\pi\right) \\
&\cdots \\
&= (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \left((B^\pi)^\top\right)^t \phi_0,
\end{aligned}$$

which is exactly the definition of μ_ϕ^π from Proposition 1.

(3) Showing $\rho(B^\pi) \leq 1$. The proof of (2) implies that for any $(s_0,a_0) \in \mathcal{S}\times\mathcal{A}$, $((B^\pi)^\top)^t\phi(s_0,a_0) = \mathbb{E}_{(s,a)\sim\mu_t^{\pi,s_0,a_0}}[\phi(s,a)]$, where μ_t^{π,s_0,a_0} is the t-th step state-action distribution under π when the initial state-action pair is the given (s_0,a_0). Given $\|\phi(s,a)\| \leq B_\phi, \forall (s,a)$, we have

$$((B^\pi)^\top)^t\phi(s_0,a_0) \leq B_\phi, \forall t.$$

Given that Σ is full-rank, we can always find $\{(s_0^{(i)}, a_0^{(i)})\}_{i=1}^d$ such that $\{u_i := \phi(s_0^{(i)}, a_0^{(i)})\}_{i=1}^d$ forms a basis of $\mathbb{R}^d$. Then we have $\|((B^\pi)^\top)^t u_i\| \leq B_\phi, \forall t$.

Now we show that $\|((B^\pi)^\top)^t\|_{\text{op}}$, where $\|\cdot\|_{\text{op}}$ is the operator norm, also has a finite bound that is independent of t. Recall that operator norm is the largest singular value; let the corresponding singular vector be u, and we express $u = \sum_{i=1}^d \alpha_i u_i$. We have

$$\|((B^\pi)^\top)^t\|_{\text{op}} = \|((B^\pi)^\top)^t u\| = \|\sum_{i=1}^d \alpha_i ((B^\pi)^\top)^t u_i\| \leq \sum_{i=1}^d |\alpha_i| B_\phi =: v.$$

The key here is that the upper bound $v < \infty$ is independent of t. Plugging into the Gelfand's formula, we have

$$\rho(B^\pi) = \rho((B^\pi)^\top) = \lim_{t\to\infty} \|((B^\pi)^\top)^t\|_{\text{op}}^{1/t} \leq \lim_{t\to\infty} v^{1/t} = 1.$$

(4) Proving equivalence $C_\phi^\pi = C_{\text{lin}}^\pi$. Recalling Eq. (14) and the definition of μ_ϕ^π, following the proof of Proposition 1, when $\sigma_{\max}(B^\pi) < 1/\gamma$ we may write

$$\begin{aligned}
C_\phi^\pi &= (1-\gamma)^2 \phi_0^\top A^{-1}\Sigma A^{-\top}\phi_0 \\
&= (1-\gamma)^2 \phi_0^\top (I-\gamma B^\pi)^{-1}\Sigma^{-1}(I-\gamma B^\pi)^{-\top}\phi_0. \\
&= \left(\mu_\phi^\pi\right)^\top \Sigma^{-1}\mu_\phi^\pi.
\end{aligned}$$

Substituting the previously derived identity that $\mu_\phi^\pi = \phi^\pi$ in the last line,

$$\left(\mu_\phi^\pi\right)^\top \Sigma^{-1} \mu_\phi^\pi = (\phi^\pi)^\top \Sigma^{-1} \phi^\pi = C_{\text{lin}}^\pi.$$

□

C Function Estimation Guarantees

For most of the paper we have focused on providing return estimation guarantees, i.e., error bounds for estimating $J(\pi)$. In some scenarios, it is desirable to obtain stronger *function estimation* guarantees (Huang & Jiang, 2022; Perdomo et al., 2023), that $\widehat{Q}_{\text{lstd}}$ and Q^π are close as functions, typically measured by weighted 2-norm. Indeed, our proof of Theorem 1 can be easily adapted to provide the following guarantee:

Theorem 2 (Function Estimation)**.** *Under the same assumptions as Theorem 1, w.p. $\geq 1-\delta$, for any $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$,*

$$\sqrt{\mathbb{E}_{(s,a)\sim\nu}[(Q^\pi(s,a) - \hat{Q}_{\text{lstd}}(s,a))^2]} \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{\widehat{C}_{\text{fn}}^\pi \cdot d \log(1/\delta)}{n}},$$

where $\widehat{C}_{\text{fn}}^\pi := (1-\gamma)^2\, \mathbb{E}_{(s_0,a_0)\sim\nu}\left[\|\widehat{\Sigma}^{1/2} \widehat{A}^{-\top} \phi(s_0,a_0)\|_2^2\right]$.

When $\nu = \mu_0 \circ \pi$ is a point-mass, the LHS of Theorem 2 coincides with that of Theorem 1, and the guarantees on the RHS are identical, too. Also recall that the naïve analysis based on $1/\sigma_{\min}(A)$ (Section 1) provides parameter identification (i.e., bounded $\|\widehat{\theta}_{\text{lstd}} - \theta^\star\|$), which immediately provides ℓ_∞ function-estimation guarantee. This result is directly implied by our Theorem 2, where the coverage parameter can be bounded as a function of $\sigma_{\min}(A)$ and B_ϕ.

Remark on C_{fn}^π. Similar to Corollary 1 we can induce a corollary that depends on the population version of $\widehat{C}_{\text{fn}}^\pi$, which we denote as C_{fn}^π. It is interesting to compare it to standard coverage parameters that enable function-estimation guarantees under completeness (Section 3). Note that the term inside $C_{\text{fn}}^\pi = \mathbb{E}_{(s_0,a_0)\sim\nu}[\cdot]$ is simply C_ϕ^π but for a deterministic initial state-action pair (s_0, a_0). Applying Proposition 4, we have

$$C_{\text{fn}}^\pi = \mathbb{E}_{(s_0,a_0)\sim\nu}\left[(\phi_{s_0,a_0}^\pi)^\top \Sigma^{-1} \phi_{s_0,a_0}^\pi\right],$$

where $\phi_{s_0,a_0}^\pi = \mathbb{E}_{(s,a)\sim\mu_{s_0,a_0}^\pi}[\phi(s,a)]$ is the expected feature under the occupancy induced from deterministic s_0, a_0 as the initial state-action pair. In comparison, the standard coverage in the literature is

$$C_{\text{lin,fn}}^\pi = \mathbb{E}_{(s,a)\sim\mu^\pi}[\phi(s,a)^\top \Sigma^{-1} \phi(s,a)].$$

As can be seen, our C_{fn}^π is in between C_ϕ^π and $C_{\text{lin,fn}}^\pi$, since we partially marginalize out the portion of μ^π that can be attribute to each initial state-action pair, instead of measuring every single $(s,a) \sim \mu^\pi$ under Σ^{-1} in a completely point-wise manner.

Proof of Theorem 2. We repeat a similar derivation to Eq. (21), noting that the proof holds when the initial state-action distribution $s_0 \sim \mu_0, a_0 \sim \pi$ changes to an arbitrary distribution ν.

$$\begin{aligned}
&\mathbb{E}_{(s_0,a_0)\sim\nu}\left[\left(Q^\pi(s_0,a_0) - \hat{Q}_{\text{lstd}}(s_0,a_0)\right)^2\right] \\
&= \mathbb{E}_{(s_0,a_0)\sim\nu}\left[\left(\phi(s_0,a_0)^\top\left(\theta^\star - \widehat{\theta}_{\text{lstd}}\right)\right)^2\right] \\
&= \mathbb{E}_{(s_0,a_0)\sim\nu}\left[\left(\phi(s_0,a_0)^\top \hat{A}^{-1}\widehat{\Sigma}^{1/2}\widehat{\Sigma}^{-1/2}\hat{A}\left(\theta^\star - \widehat{\theta}_{\text{lstd}}\right)\right)^2\right] \\
&\leq \mathbb{E}_{(s_0,a_0)\sim\nu}\left[\left\|\widehat{\Sigma}^{1/2}\hat{A}^{-T}\phi(s_0,a_0)\right\|_2^2 \left\|\widehat{\Sigma}^{-1/2}\hat{A}\left(\theta^\star - \widehat{\theta}_{\text{lstd}}\right)\right\|_2^2\right]
\end{aligned}$$

As in the proof of Theorem 1, we note that

$$\left\|\widehat{\Sigma}^{-1/2}\hat{A}\left(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\right)\right\|_2 \le \left\|\widehat{\Sigma}^{-1/2}(\hat{A}\theta^\star - \hat{b})\right\|_2 + \left\|\widehat{\Sigma}^{-1/2}(\hat{A}\widehat{\theta}_{\mathrm{lstd}} - \hat{b})\right\|_2$$
$$\le \left\|\widehat{\Sigma}^{-1/2}(\hat{A}\theta^\star - \hat{b})\right\|_2,$$

since $\hat{A}\widehat{\theta}_{\mathrm{lstd}} - \hat{b} = 0$ and $\widehat{\Sigma}$ is invertible. To conclude, we recall that the concentration bound from Lemma 1, which implies that

$$\|\widehat{\Sigma}^{-1/2}(\hat{A}\theta^\star - \hat{b})\|_2^2 = \mathcal{O}\left(V_{\max}^2 \frac{d\log(1/\delta)}{n}\right).$$

Plugging this in yields the proof. □

D Loss Minimization Algorithm

Here we provide an alternative analysis to Corollary 1, where we are able to eliminate the dependence on $1/\sigma_{\min}(A)$, but the rates still depend on $1/\sigma_{\min}(\Sigma)$. The analysis also requires a slight change of the LSTDQ algorithm to a loss-minimization form (Liu et al., 2025):

$$\widehat{\theta}_{\mathrm{lstd}} = \arg\min_{\theta\in\Theta} \|\widehat{\Sigma}^{-1/2}(\widehat{A}\theta - \widehat{b})\|_2. \tag{20}$$

In practice, when $\widehat{A}$ is near-singular, the inverse solution $\widehat{A}^{-1}\widehat{b}$ may have a very large norm which is clearly problematic, demanding some regularization to control the norm of the solution. The loss-minimization formulation of Eq. (20) is a natural abstraction of this process, where we search for $\widehat{\theta}$ in a pre-defined parameter space with bounded norm. If $\widehat{A}^{-1}\widehat{b} \in \Theta$, it is easy to see that the loss-minimization solution coincides with the inverse solution; when $\widehat{A}^{-1}\widehat{b} \notin \Theta$, Eq. (20) still outputs a bounded solution to ensure generalization and good statistical properties.

We will need the following boundedness assumption on Θ.

Assumption 4 (Boundedness of Θ)**.** *Assume* $\|\theta\|_2 \le B_\Theta,\ \forall\, \theta \in \Theta$.

Additional linear algebraic notation. For symmetric Σ, let $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ be the condition number, where $\lambda_{\max}(\cdot)$ is the largest eigenvalue. Let $\mathrm{tr}(A)$ be the trace of a matrix A.

Theorem 3. *Assume that* $n \gtrsim \log(d/\delta)\kappa(\Sigma)B_\phi^2/\lambda_{\min}(\Sigma)$. *Let* $\phi_0 = \mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}[\phi(s_0,a_0)]$ *denote the initial feature. Under Assumptions 1, 2 and 4, the estimator in Eq. (20) satisfies that*

$$\left|J(\pi) - J_{\widehat{Q}_{\mathrm{lstd}}}(\pi)\right| \lesssim \frac{\sqrt{C_\phi^\pi}}{1-\gamma}\max\{B_\phi B_\Theta, R_{\max}\}^2\kappa(\Sigma)\sqrt{\frac{d\log(B_\Theta n\delta^{-1})}{\lambda_{\min}(\Sigma)n}}$$

with probability at least $1-\delta$.

Proof of Theorem 3. Let $\hat{\ell}(\theta)$ and $\ell(\theta)$ denote the empirical and population vectors:

$$\hat{\ell}(\theta) = \hat{A}\theta - \hat{b} \quad \text{and} \quad \ell(\theta) = A\theta - b.$$

Recall that $A\theta^\star = b$ and thus $\ell(\theta^\star) = 0$. We establish in the sequel the following concentration lemma.

Lemma 5. *With probability at least* $1-\delta$*, we have that for all* $\theta \in \Theta$*:*

$$\left|\|\Sigma^{-1/2}\ell(\theta)\|_2 - \|\Sigma^{-1/2}\hat{\ell}(\theta)\|_2\right| \le \max\{B_\phi B_\Theta, R_{\max}\}^2\sqrt{\frac{288d\log(864B_\Theta n\delta^{-1})}{\lambda_{\min}(\Sigma)n}} := \varepsilon_{\mathrm{stat}}.$$

We also note the following simple technical lemma.

Lemma 6. *For all $v \in \mathbb{R}^d$, we have*

$$v^\top \Sigma^{-1} v \le \frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})} v^\top \widehat{\Sigma}^{-1} v, \quad \textit{and} \quad v^\top \widehat{\Sigma}^{-1} v \le \frac{\lambda_{\max}(\widehat{\Sigma}^{-1})}{\lambda_{\min}(\Sigma^{-1})} v^\top \Sigma^{-1} v.$$

Recall that $\widehat{\theta}$ satisfies $\arg\min_{\theta \in \Theta} \|\widehat{\Sigma}^{-1/2} \hat{\ell}(\theta)\|_2$. We now show that Lemma 5 and Lemma 6 imply that $\|\Sigma^{-1/2} \ell(\widehat{\theta})\|_2$ is small. This follows since:

$$\begin{aligned}
\|\Sigma^{-1/2}\ell(\widehat{\theta})\|_2 &\le \|\Sigma^{-1/2}\hat{\ell}(\widehat{\theta})\|_2 + \varepsilon_{\mathrm{stat}} \\
&\le \sqrt{\frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})}} \|\widehat{\Sigma}^{-1/2}\hat{\ell}(\widehat{\theta})\|_2 + \varepsilon_{\mathrm{stat}} \\
&\le \sqrt{\frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})}} \|\widehat{\Sigma}^{-1/2}\hat{\ell}(\theta^\star)\|_2 + \varepsilon_{\mathrm{stat}} \\
&\le \sqrt{\frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})} \cdot \frac{\lambda_{\max}(\widehat{\Sigma}^{-1})}{\lambda_{\min}(\Sigma^{-1})}} \|\Sigma^{-1/2}\hat{\ell}(\theta^\star)\|_2 + \varepsilon_{\mathrm{stat}} \\
&\le \sqrt{\frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})} \cdot \frac{\lambda_{\max}(\widehat{\Sigma}^{-1})}{\lambda_{\min}(\Sigma^{-1})}} \|\Sigma^{-1/2}\ell(\theta^\star)\|_2 + \left(1 + \sqrt{\frac{\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\widehat{\Sigma}^{-1})} \cdot \frac{\lambda_{\max}(\widehat{\Sigma}^{-1})}{\lambda_{\min}(\Sigma^{-1})}}\right)\varepsilon_{\mathrm{stat}} \\
&= \left(1 + \sqrt{\kappa(\Sigma)\kappa(\widehat{\Sigma})}\right)\varepsilon_{\mathrm{stat}} \\
&\le 2\sqrt{\kappa(\Sigma)\kappa(\widehat{\Sigma})}\varepsilon_{\mathrm{stat}}.
\end{aligned}$$

In the sequel, we also show concentration for the condition number of $\widehat{\Sigma}$ to Σ.

Lemma 7. *Let $n \ge 32 \log(6d/\delta)\kappa(\Sigma)\big(B_\phi^2/\lambda_{\min}(\Sigma) + \kappa(\Sigma)\big)$. Then, with probability at least $1-\delta$, we have:*

$$\kappa(\widehat{\Sigma}) \le 3\kappa(\Sigma)$$

This implies that, under the condition on sample size, we have $\left\|\Sigma^{-1/2}\ell(\widehat{\theta})\right\|_2 \le \sqrt{12}\kappa(\Sigma)\varepsilon_{\mathrm{stat}}$ with high-probability. We can now conclude the proof. Under the conditions and events stated above, we have:

$$\begin{aligned}
\left|\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}\Big[Q^\pi(s_0,a_0) - \hat{Q}_{\mathrm{lstd}}(s_0,a_0)\Big]\right| &= \left|\mathbb{E}_{s_0\sim\mu_0, a_0\sim\pi}\Big[\phi(s_0,a_0)^\top\Big(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\Big)\Big]\right| \qquad (21)\\
&= \left|\phi_0^\top\Big(\theta^\star - \widehat{\theta}_{\mathrm{lstd}}\Big)\right| \\
&= \left|\phi_0^\top\Big(A^{-1}b - \widehat{\theta}_{\mathrm{lstd}}\Big)\right| \\
&= \left|\phi_0^\top A^{-1}\Big(b - A\widehat{\theta}_{\mathrm{lstd}}\Big)\right| \\
&= \left|\phi_0^\top A^{-1}\Sigma^{1/2}\Sigma^{-1/2}\Big(b - A\widehat{\theta}_{\mathrm{lstd}}\Big)\right| \\
&= \left|\phi_0^\top A^{-1}\Sigma^{1/2}\Sigma^{-1/2}\Big(b - A\widehat{\theta}_{\mathrm{lstd}}\Big)\right| \\
&\le \left\|\Sigma^{1/2}A^{-T}\phi_0\right\|_2 \left\|\Sigma^{-1/2}\Big(A\widehat{\theta}_{\mathrm{lstd}} - b\Big)\right\|_2 \\
&= \left\|\Sigma^{1/2}A^{-T}\phi_0\right\|_2 \left\|\Sigma^{-1/2}\Big(A\widehat{\theta}_{\mathrm{lstd}} - b\Big)\right\|_2 \\
&= \sqrt{\phi_0^\top A^{-1}\Sigma A^{-T}\phi_0}\|\Sigma^{-1/2}\ell(\widehat{\theta})\|_2 \\
&\le \frac{1}{1-\gamma}\sqrt{C_\phi^\pi}\sqrt{12}\kappa(\Sigma)\varepsilon_{\mathrm{stat}}, \qquad (22)
\end{aligned}$$

as desired. We now establish Lemmas 5 to 7.

□

Proof of Lemma 5. Let θ be fixed for now, and $\Delta(\theta) = \hat{\ell}(\theta) - \ell(\theta)$. Note that by the reverse triangle inequality,

$$\left| \|\Sigma^{-1/2}\ell(\theta)\|_2 - \|\Sigma^{-1/2}\hat{\ell}(\theta)\|_2 \right| \leq \left\|\Sigma^{-1/2}\Delta(\theta)\right\|_2 = \left\|\Sigma^{-1/2}(\hat{A} - A)\theta - \Sigma^{-1/2}(\hat{b} - b)\right\|_2.$$

We use Vector Bernstein (Lemma 9) to show that this is small. Let $X_i = \phi(s_i, a_i)$, $Y_i = \phi(s_i, a_i) - \gamma\phi(s_i', a_i')$, and $\Delta_i(\theta) = X_i(Y_i^\top\theta - r_i) - (A\theta - b)$ denote the centered vectors. Note that

$$\|\Sigma^{-1/2}\Delta_i(\theta)\|_2 \leq \frac{1}{\sqrt{\lambda_{\min}(\Sigma)}} 2\max\big\{\|X_i(Y_i^\top\theta - r_i)\|_2, \|A\theta - b\|_2\big\}.$$

We have the following bound:

$$\begin{aligned}
\|A\theta - b\|_2 &= \left\|\mathbb{E}\Big[\phi(s,a)\Big((\phi(s,a) - \gamma\phi(s',a'))^\top\theta - r(s,a)\Big)\Big]\right\|_2 \\
&\leq \left\|\mathbb{E}\big[\phi(s,a)\phi(s,a)^\top\theta\big]\right\|_2 + \gamma\left\|\mathbb{E}\big[\phi(s,a)\phi(s',a')^\top\theta\big]\right\|_2 + \|\mathbb{E}[\phi(s,a)r(s,a)]\|_2 \\
&\leq (1+\gamma)\max_{s,a}\|\phi(s,a)\|_2^2\|\theta\|_2 + \max_{s,a}\|\phi(s,a)\|_2 R_{\max} \\
&\leq 3B_\phi \max\{B_\phi B_\Theta, R_{\max}\}. \qquad (23)
\end{aligned}$$

We remark that with a similar derivation, this bound applies just as well to $\|X_i(Y_i^\top\theta - r_i)\|_2$, so in fact we have

$$\|\Sigma^{-1/2}\Delta_i(\theta)\|_2 \leq \frac{6}{\sqrt{\lambda_{\min}(\Sigma)}} B_\phi \max\{B_\phi B_\Theta, R_{\max}\}.$$

For the variance bound, we simply use that

$$\mathbb{E}\Big[\|\Sigma^{-1/2}\Delta_i(\theta)\|_2^2\Big] \leq \left(\frac{6}{\sqrt{\lambda_{\min}(\Sigma)}} B_\phi \max\{B_\phi B_\Theta, R_{\max}\}\right)^2.$$

Then, we conclude via Lemma 9 that

$$\|\Sigma^{-1/2}\Delta(\theta)\|_2 \leq B_\phi \max\{B_\phi B_\Theta, R_{\max}\}\sqrt{\frac{32\log(288\delta^{-1})}{\lambda_{\min}(\Sigma)n}}.$$

We now apply a covering argument over $\theta \in \Theta$. Let $\Theta_0 \subseteq \Theta$ be an L_2-covering of Θ of size $\mathcal{N}(\varepsilon)$, satisfying for for each $\theta \in \Theta$ there exists a covering member $\rho(\theta) \in \Theta_0$ satisfying $\|\theta - \rho(\theta)\|_2 \leq \varepsilon$. Via a simple triangle inequality:

$$\left\|\Sigma^{-1/2}\Delta(\theta)\right\|_2 \leq \left\|\Sigma^{-1/2}\Delta(\rho(\theta))\right\|_2 + \left\|\Sigma^{-1/2}(\Delta(\theta) - \Delta(\rho(\theta)))\right\|_2.$$

We bound the latter term as a function of ε.

$$\begin{aligned}
\left\|\Sigma^{-1/2}(\Delta(\theta) - \Delta(\rho(\theta)))\right\|_2 &= \left\|\Sigma^{-1/2}\Big(A - \hat{A}\Big)(\theta - \rho(\theta))\right\|_2 \\
&\leq \frac{1}{\sqrt{\lambda_{\min}(\Sigma)}} 2\max\Big\{\sigma_{\max}(A), \sigma_{\max}(\hat{A})\Big\}\varepsilon.
\end{aligned}$$

We notice that $\max\Big\{\sigma_{\max}(A), \sigma_{\max}(\hat{A})\Big\} \leq 2B_\phi^2$ via a similar reasoning to Eq. (23). This leaves us with:

$$\begin{aligned}
\left\|\Sigma^{-1/2}\Delta(\theta)\right\|_2 &\leq B_\phi \max\{B_\phi B_\Theta, R_{\max}\}\sqrt{\frac{32\log(288|\Theta_0|\delta^{-1})}{\lambda_{\min}(\Sigma)n}} + \frac{2B_\phi^2}{\sqrt{\lambda_{\min}(\Sigma)}}\varepsilon, \\
&\leq B_\phi \max\{B_\phi B_\Theta, R_{\max}\}\sqrt{\frac{32d\log(864\delta^{-1}/\varepsilon)}{\lambda_{\min}(\Sigma)n}} + \frac{2B_\phi^2}{\sqrt{\lambda_{\min}(\Sigma)}}\varepsilon,
\end{aligned}$$

where we have applied a union bound over the set Θ_0, which is of size at most $(3B_\Theta/\varepsilon)^d$ for $\varepsilon \in (0, 1]$ by standard covering number bounds (Vershynin, 2018), since $\Theta \subset \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B_\Theta\right\}$. Picking $\varepsilon = 1/\sqrt{n}$ lets us conclude that, with probability at least $1 - \delta$, for all $\theta \in \Theta$,

$$\left\|\Sigma^{-1/2}\Delta(\theta)\right\|_2 \leq B_\phi \max\{B_\phi B_\Theta, R_{\max}\}\sqrt{\frac{288d\log(864B_\Theta n\delta^{-1})}{\lambda_{\min}(\Sigma)n}},$$

as desired. □

Proof of Lemma 6. Follows from the fact that for any positive semi-definite matrix $M \in \mathbb{R}^{d\times d}$ and for any $v \in \mathbb{R}^d$, we have the inequalities

$$\lambda_{\min}(M)v^\top v \leq v^\top Mv \leq \lambda_{\max}(M)v^\top v.$$

□

Proof of Lemma 7. We firstly establish that

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq \sqrt{\frac{8\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(6d/\delta)}{n}} =: \varepsilon_{\text{op}}. \tag{24}$$

To do this, we use Matrix Bernstein (Lemma 8). Abbreviate $X_i \coloneqq \phi(s_i, a_i)$, and let $Z_i = X_iX_i^\top - \Sigma$ be the centered matrices. Note that $\|Z_i\|_2 \leq \|X_iX_i^\top\|_2 + \|\Sigma\|_2 \leq \|X_i\|_2^2 + \lambda_{\max}(\Sigma) \leq B_\phi^2 + \lambda_{\max}(\Sigma)$. For the variance term, we have:

$$\begin{aligned}\left\|\mathbb{E}\left[(X_iX_i^\top - \Sigma)^2\right]\right\|_2 &= \left\|\mathbb{E}\left[(X_iX_i^\top)^2\right] - \Sigma^2\right\|_2 \\ &\leq \left\|B_\phi^2\,\mathbb{E}\left[X_iX_i^\top\right] - \Sigma^2\right\|_2 \\ &\leq B_\phi^2\lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma)^2.\end{aligned}$$

This yields

$$\|\hat{\Sigma} - \Sigma\|_2 \leq \sqrt{\frac{2\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{n}} + \frac{2(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{3n}$$

The slow term dominates when n is large enough:

$$n \geq \frac{2(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{\lambda_{\max}(\Sigma)} = 2\log(2d/\delta)\left(\frac{B_\phi^2}{\lambda_{\max}(\Sigma)} + 1\right). \tag{25}$$

Note that this is implied by our assumption on n, since $\lambda_{\max}(\Sigma) \geq \lambda_{\min}(\Sigma)$ and $\kappa(\Sigma) \geq 1$. Thus, under this condition we have

$$\|\hat{\Sigma} - \Sigma\|_2 \leq 2\sqrt{\frac{2\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{n}} = \varepsilon_{\text{op}},$$

as desired. Now, by Weyl's theorem (Horn & Johnson, 2012, Theorem 4.3.1), we have

$$|\lambda_{\min}(\hat{\Sigma}) - \lambda_{\min}(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_2 \leq \varepsilon_{\text{op}},$$

which implies that

$$\lambda_{\min}(\hat{\Sigma}) \geq \frac{\lambda_{\min}(\Sigma)}{2} \tag{26}$$

using the condition that $\varepsilon_{\text{op}} \leq \frac{\lambda_{\min}(\Sigma)}{2}$. This latter condition is equivalent to

$$\begin{aligned}&\sqrt{\frac{8\lambda_{\max}(\Sigma)(B_\phi^2 + \lambda_{\max}(\Sigma))\log(2d/\delta)}{n}} \leq \frac{\lambda_{\min}(\Sigma)}{2} \\ &\iff n \geq 32\log(2d/\delta)\kappa(\Sigma)\left(\frac{B_\phi^2}{\lambda_{\min}(\Sigma)} + \kappa(\Sigma)\right),\end{aligned} \tag{27}$$

which is precisely our assumption on n. Similarly, an application of the reverse triangle inequality (or of Weyl's theorem again) yields,

$$|\lambda_{\max}(\hat{\Sigma}) - \lambda_{\max}(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_2 \leq \varepsilon_{\text{op}},$$

which implies that

$$\lambda_{\max}(\widehat{\Sigma}) \leq \lambda_{\max}(\Sigma) + \varepsilon_{\text{op}} \leq \frac{3}{2}\lambda_{\max}(\Sigma), \tag{28}$$

using the condition that $\varepsilon_{\text{op}} \leq \frac{\lambda_{\min}(\Sigma)}{2} \leq \frac{\lambda_{\max}(\Sigma)}{2}$. Combining Eqs. (28) and (26), we have:

$$\kappa(\widehat{\Sigma}) = \frac{\lambda_{\max}(\widehat{\Sigma})}{\lambda_{\min}(\widehat{\Sigma})} \leq \frac{3}{2}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\widehat{\Sigma})} \leq 3\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} = 3\kappa(\Sigma).$$

□

E TECHNICAL TOOLS

Lemma 8 (Matrix Bernstein, Tropp (2012)). *Let $S_1, \ldots, S_n \in \mathbb{R}^{d_1 \times d_2}$ be random, independent matrices satisfying $\mathbb{E}[S_k] = 0$, $\max\{\|\mathbb{E}[S_k S_k^\top]\|_{\text{op}}, \|\mathbb{E}[S_k^\top S_k]_{\text{op}}\|\} \leq \sigma^2$, and $\|S_k\|_{\text{op}} \leq L$ almost surely for all k. Then, with probability at least $1-\delta$ for any $\delta \in (0,1)$,*

$$\left\|\frac{1}{n}\sum_{k=1}^{n} S_k\right\|_{\text{op}} \leq \sqrt{\frac{2\sigma^2 \log((d_1+d_2)/\delta)}{n}} + \frac{2L\log((d_1+d_2)/\delta)}{3n}$$

Lemma 9 (Vector Bernstein, Minsker (2017)). *Let $v_1, \ldots, v_n$ be independent vectors in $\mathbb{R}^d$ such that $\mathbb{E}[v_k] = 0$, $\mathbb{E}[\|v_k\|_2^2] \leq \sigma^2$, and $\|v_k\|_2 \leq L$ almost surely for all k. Then, with probability at least $1-\delta$ for any $\delta \in (0,1)$,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} v_i\right\|_2 \leq \sqrt{\frac{2\sigma^2 \log(28/\delta)}{n}} + \frac{2L\log(28/\delta)}{3n}.$$

Lemma 10 (Vector Martingale Bernstein (Pinelis, 1994; Martinez-Taboada & Ramdas, 2024)). *Let $(X_t)_{t\leq T}$ be a martingale sequence of vectors in $\mathbb{R}^d$ adapted to a filtration $(\mathcal{F}_t)_{t\leq T}$, such that $\mathbb{E}_{t-1}[X_t] = 0$, and $\|X_t\|_2 \leq B$, and $\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\|X_t\|^2\right] \leq \sigma^2$. Then, with probability at least $1-\delta$, we have:*

$$\left\|\sum_{t=1}^{T} X_t\right\|_2 \leq \sqrt{2\sigma^2 \log(2/\delta)} + \frac{2}{3}B\log(2/\delta).$$