A Complete Characterization of Linear Estimators for Offline Policy Evaluation

Juan C. Perdomo
jcperdomo@berkeley.edu
University of California, Berkeley

Akshay Krishnamurthy
akshaykr@microsoft.com
Microsoft Research

Peter Bartlett
peter@berkeley.edu
University of California, Berkeley

Sham Kakade
sham@seas.harvard.edu
Harvard University

December 20, 2022

Abstract

Offline policy evaluation is a fundamental statistical problem in reinforcement learning that involves estimating the value function of some decision-making policy given data collected by a potentially different policy. In order to tackle problems with complex, high-dimensional observations, there has been significant interest from theoreticians and practitioners alike in understanding the possibility of *function approximation* in reinforcement learning. Despite significant study, a sharp characterization of when we might expect offline policy evaluation to be tractable, even in the simplest setting of *linear* function approximation, has so far remained elusive, with a surprising number of strong negative results recently appearing in the literature.

In this work, we identify simple control-theoretic and linear-algebraic conditions that are necessary and sufficient for classical methods, in particular Fitted Q-iteration (FQI) and least squares temporal difference learning (LSTD), to succeed at offline policy evaluation. Using this characterization, we establish a precise hierarchy of regimes under which these estimators succeed. We prove that LSTD works under strictly weaker conditions than FQI. Furthermore, we establish that if a problem is not solvable via LSTD, then it cannot be solved by a broad class of linear estimators, even in the limit of infinite data. Taken together, our results provide a complete picture of the behavior of linear estimators for offline policy evaluation, unify previously disparate analyses of canonical algorithms, and provide significantly sharper notions of the underlying statistical complexity of offline policy evaluation.

1 Introduction

A central component of a practical sequential decision making system is its ability to cope with high-dimensional and complex data sources. While feature engineering or discretization techniques can in principle be used to address the challenges associated with complex data, these approaches require significant domain expertise and suffer from a curse-of-dimensionality phenomenon that limit their practical relevance. Instead, the use of more general *function approximation methods* for reinforcement learning (RL) promises to avoid these drawbacks. Consequently, understanding these methods has long been a topic of interest to theoreticians and practitioners alike.

While the use of nonlinear methods is by now common in the empirical reinforcement learning literature, the much simpler linear function approximation setting remains somewhat poorly understood theoretically, despite decades of study. Indeed, recently there has been a surge of research effort focusing on necessary and sufficient conditions for reinforcement learning with linear function approximation, including the first provably efficient algorithms for online exploration [Yang and Wang, 2020, Jin et al., 2020] and a number of

surprising statistical lower bounds that hold even under strong assumptions [Wang et al., 2021c, Weisz et al., 2021a,b]. This line of work represents substantial progress, yet we still lack a clear picture as to precisely when and why RL with linear function approximation is tractable.

As a step towards providing this clarity, in this paper we focus on the simpler *offline policy evaluation* problem (OPE) in infinite horizon, discounted MDPs, under the assumption that the action-value function is *linearly realizable* in some known features. Here, rather than interacting with an environment to maximize reward as in the standard RL formulation, the goal is to estimate the performance of a given decision-making policy by leveraging an observational dataset collected by a potentially different policy. OPE is perhaps the simplest, non-trivial setting in which to study function approximation in RL. It is also practically relevant in its own right: both OPE and the closely-related offline policy *optimization* problem represent a promising avenue toward applying RL in safety-critical domains where active exploration is infeasible. Moreover, the principles developed for OPE are routinely used in online RL algorithms.

Fitted Q-iteration (FQI) [Ernst et al., 2005, Riedmiller, 2005] and least squares temporal difference learning (LSTD) [Bradtke and Barto, 1996, Boyan, 1999, Nedić and Bertsekas, 2003] are canonical algorithms for offline policy evaluation with function approximation. These simple, moment-based methods are some of the most popular approaches in practice and have served as inspiration for recent empirical breakthroughs in RL [Mnih et al., 2015]. They have also been the subject of intense theoretical investigation, with early results on convergence and instability described by Bertsekas and Tsitsiklis [1995], Tsitsiklis and Van Roy [1996] as well as several more recent results [Antos et al., 2008, Chen and Jiang, 2019, Lazaric et al., 2012]. Nevertheless, a sharp finite sample characterization of the behavior of FQI and LSTD, even in the linear realizability setting, remains undeveloped.

In this paper, we identify necessary and sufficient conditions for FQI and LSTD to succeed at offline policy evaluation under linear realizability. In doing so, we establish a precise hierarchy of conditions under which these methods work; in particular, we prove that LSTD succeeds under strictly weaker assumptions than FQI. Moreover, if an offline policy evaluation problem is not solvable via LSTD, then it cannot be solved by any linear, moment-based method (see Definition 4.1) even in the limit of infinite data. Our characterization draws upon ideas from the theory of Lyapunov stability and provides a new, unifying perspective on the statistical complexity of offline policy evaluation. In particular, we show how traditional quantities, such as the "effective horizon", fail to capture the true complexity of the problem (Sections 3.1 and 4.1) and propose instance-dependent measures which are significantly sharper. Furthermore, our results unify previously disparate analyses for FQI and LSTD as our conditions are implied by prior assumptions (Sections 3.2 and 4.2). Taken together, our results provide a complete picture of the possibilities and limitations of linear estimators for offline policy evaluation under linear realizability.

1.1 Linear estimators & the offline policy evaluation problem

Let $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ denote an infinite horizon, γ-discounted MDP where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $R : \mathcal{S} \times \mathcal{A} \to \Delta([-1, 1])$ is the random reward function, and $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator, which defines a distribution over states for every pair (s, a). The action-value function Q^π captures the expected total reward achieved by a randomized policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ from an initial state-action pair (s, a) when the trajectory is generated such that for each time step h, $a_h \sim \pi(s_h)$ and $s_{h+1} \sim P(\cdot \mid s_h, a_h)$.

$$Q^\pi(s, a) := \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \mid (s_0, a_0) = (s, a), \pi\right]. \tag{1.1}$$

In the offline policy evaluation problem, we are given a policy π and a dataset $\{(s_i, a_i, r_i(s_i, a_i), s'_i, a'_i)\}_{i=1}^n$ of observed transitions and rewards, where the initial pair (s_i, a_i) is sampled from some *arbitrary* distribution $\mathcal{D}$, $r_i(s_i, a_i) \sim R(s_i, a_i)$, the next state is sampled from the transition operator $s'_i \sim P(\cdot \mid s_i, a_i)$, and the next action $a'_i \sim \pi(s'_i)$ is sampled according to π.[1] Our goal is to return an estimate $\widehat{Q}^\pi$ of Q^π. For concreteness, we measure performance via $\mathbb{E}_{(s,a)\sim\mathcal{D}}|\widehat{Q}^\pi(s, a) - Q^\pi(s, a)|$ and we ask that this quantity is vanishingly small

[1] We "augment" the dataset to include the next state action $a' \sim \pi(s')$ purely for notational convenience.

with high probability over the draw of the dataset. For simplicity, we assume that samples are drawn i.i.d. via the procedure described above.[2]

As we would like to develop methods that scale to settings where the cardinalities of the sets $\mathcal{S}$ and $\mathcal{A}$ are large or infinite, our focus is on understanding policy evaluation using linear function approximation, as per the following definition:

Assumption 1 (Linear Realizability)**.** Q^π is linearly realizable[3] in a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exists a vector $\theta^\star_\gamma \in \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q^\pi(s, a) = \phi(s, a)^\top \theta^\star_\gamma$.

Fitted Q-iteration. As mentioned previously, fitted Q-iteration is one of the most popular algorithms for policy evaluation in practice and can in principle work with any function approximation method. In the linear case, given a dataset $\{(s_i, a_i, r_i(s_i, a_i), s'_i, a'_i)\}_{i=1}^n$ and an initial vector $\widehat{\theta}_0$, FQI iteratively solves least squares regression problems of the form

$$\widehat{\theta}_{t+1} \in \arg\min_\theta \sum_{i=1}^n \left(\phi(s_i, a_i)^\top \theta - r(s_i, a_i) - \gamma \phi(s'_i, a'_i)^\top \widehat{\theta}_t\right)^2, \tag{1.2}$$

for some number of rounds T and returns the estimator $\widehat{Q}^\pi(s, a) := \phi(s, a)^\top \widehat{\theta}_T$.

Least squares temporal difference learning. In the linear function approximation setting, the vector $\theta^\star_\gamma$ which realizes Q^π in the feature mapping ϕ satisfies the fixed point equation,[4]

$$\Sigma_{\text{cov}} \theta^\star_\gamma = \gamma \Sigma_{\text{cr}} \theta^\star_\gamma + \theta_{\phi,r}. \tag{1.3}$$

Here, Σ_{cov} if the offline feature covariance matrix, Σ_{cr} is the cross-covariance matrix between time-adjacent features, and $\theta_{\phi,r}$ is the mean feature-reward vector. (see Eqs. (1.5) and (2.3) for formal definitions). LSTD tries to approximate $\theta^\star_\gamma$ by computing the plug-in estimate to the closed-form solution to the equation above,

$$\widehat{\theta}_{\text{LS}} := (I - \gamma \widehat{\Sigma}_{\text{cov}}^{-1} \widehat{\Sigma}_{\text{cr}})^\dagger \widehat{\Sigma}_{\text{cov}}^{-1} \widehat{\theta}_{\phi,r} = (\widehat{\Sigma}_{\text{cov}} - \gamma \widehat{\Sigma}_{\text{cr}})^\dagger \widehat{\theta}_{\phi,r}. \tag{1.4}$$

and returns $\widehat{Q}^\pi(s, a) := \phi(s, a)^\top \widehat{\theta}_{\text{LS}}$ [Bradtke and Barto, 1996]. We focus on the unregularized variant of both of these algorithms. However, similar insights apply to the regularized cases (see Appendix A.7).

1.2 Our contributions

The main result of our work is that we identify simple linear algebraic conditions which exactly characterize when (and why) linear estimators will succeed at offline policy evaluation under linear realizability of Q^π. Under these conditions, which we introduce below, we establish upper bounds on the sample complexity of offline policy evaluation which scale with: (1) for FQI, the operator norm of the solution to a particular discrete-time Lyapunov equation, and (2) for LSTD, the minimum singular value of an instance-dependent matrix. In both cases, we illustrate how our results unify previously disparate analyses of these algorithms, and demonstrate how our new instance-dependent quantities provide sharper notions of the statistical complexity of OPE when compared to bounds that explicitly depend on traditional parameters such as the "effective horizon", i.e., $1/(1-\gamma)$.

[2] In particular, extensions to Markovian data, where samples are drawn from an ergodic chain, are fairly well-understood, see e.g., Mou et al. [2021], Nagaraj et al. [2020]. Overall, the statistical rates in the Markovian setting mimic those obtained under i.i.d assumptions, up to mixing time factors.

[3] Note that realizability of Q^π does not imply that the rewards are linearly realizable. We say that rewards are linearly realizable in a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exists $\theta^\star_r \in \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\phi(s, a)^\top \theta^*_r = \mathbb{E} r(s, a)$.

[4] This fixed point relationship comes from examining the definition of $Q^\pi(s, a)$ which satisfies, $Q^\pi(s, a) = \mathbb{E} r(s, a) + \gamma \cdot \mathbb{E} Q^\pi(s', a')$ point-wise over (s, a). The precise equation follows from substituting in $Q^\pi = \phi(s, a)^\top \theta^\star_\gamma$.

Our conditions can be introduced rather succinctly. For FQI, the key definitions and assumptions are:

$$\Sigma_{\text{cov}} := \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}}\left[\phi(s,a)\phi(s,a)^\top\right], \quad \Sigma_{\text{cr}} := \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D} \\ s'\sim P(\cdot|s,a),\ a'\sim\pi(s')}}\left[\phi(s,a)\phi(s',a')^\top\right]. \tag{1.5}$$

Assumption 2 (Stability)**.** The matrix Σ_{cov} is full rank and $\rho(\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}) < 1$.

Here, Σ_{cov} is the offline state-action covariance, Σ_{cr} is the cross-covariance, $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is the *whitened cross-covariance*,[5] and $\rho(A) = \max_i |\lambda_i(A)|$ is the *spectral radius* of the matrix A. The assumption that Σ_{cov} is full rank is not fundamental and is included primarily to simplify the presentation.[6] If Assumption 2 holds, we let P_γ be the unique solution (over X) to the Lyapunov equation,

$$X = (\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^\top X(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) + I.$$

Our first main result is that, under stability, FQI satisfies the following error guarantee:

Theorem 1 (Informal). *Let $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top\widehat{\theta}_T$, where $\widehat{\theta}_T$ is the T-step FQI solution. Under Assumptions 1 and 2, as well as standard regularity assumptions for linear regression, for n large enough,*

$$\mathbb{E}_{\mathcal{D}}|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| \lesssim \mathsf{cond}(P_\gamma)\|P_\gamma\|_{\text{op}}^2\sqrt{\frac{d\log(1/\delta)}{n}} + \mathcal{O}(\exp(-T)),$$

with probability $1-\delta$. Here, $\mathsf{cond}(\cdot)$ and $\|\cdot\|_{\text{op}}$ denote the condition number and operator norm.

For the sake of clarity, we have suppressed dependence on universal constants and other quantities which arise from standard analysis of linear regression in the informal statement of the upper bound. Since $P_\gamma \succeq I$, $\mathsf{cond}(P_\gamma)$ can always be crudely upper bounded by the operator norm, so that the primary factor, beyond the standard $\sqrt{d/n}$ term for linear regression, is the dependence on $\|P_\gamma\|_{\text{op}}$. We show in Section 3.2 that, for settings where FQI was previously shown to succeed (e.g., under low distribution shift or Bellman completeness [Wang et al., 2021a]), stability always holds and $\|P_\gamma\|_{\text{op}}$ is never much larger than $1/(1-\gamma)$, demonstrating how our bound recovers and unifies prior results. However, we also find that, in general, this quantity provides a much *sharper* notion of complexity for OPE. Indeed, there are simple instances where $\|P_\gamma\|_{\text{op}}$ is $\mathcal{O}(1)$ for all $\gamma \in (0,1)$, but of course, $1/(1-\gamma)$ can be arbitrarily large.

The key insight behind this result is that, in the linear setting, FQI can be written as a power series in the empirical versions of the second moment matrices described in Eq. (1.5). More precisely, $\widehat{\theta}_T = \sum_{k=0}^T(\gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})^k\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\theta}_{\phi,r}$ where $\widehat{\theta}_{\phi,r}$ is obtained by solving a regression for the rewards. The behavior of the algorithm is governed by the growth of these matrix powers. Using ideas from Lyapunov theory, we show that if stability holds, then these decay at a geometric rate governed by $\|P_\gamma\|_{\text{op}}$ and FQI succeeds. On the other hand, if the spectral radius is greater than one, then these matrix powers grow exponentially, and FQI will drastically amplify any estimation errors. This leads to the *necessity* of stability for FQI:

Proposition 3.4 (Informal). *If $\rho(\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}) > 1$, the variance of the FQI solution grows exponentially with the number of regression rounds T.*

Turning to LSTD, while the solution is defined in terms of similar moment quantities to those relevant for FQI, it solves for $\theta_\gamma^\star$ in a more direct manner and hence its behavior is somewhat different. We prove that LSTD succeeds if the following condition holds:

Assumption 3 (Invertibility)**.** The matrices Σ_{cov} and $I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$ are both full rank.

[5] For any matrix A and invertible matrix L, the eigenvalues of A and $L^{-1}AL$ are identical. Therefore, one could equivalently state Assumptions 2 and 3 in terms of $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$.

[6] For example, the results carry over if all features $\phi(s,a)$ lie in a low dimensional subspace.

Our main result for LSTD is that under invertibility, $\theta^\star_\gamma$ is identifiable via LSTD as per the following informal theorem statement:

Theorem 2 (Informal). *Let $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top \widehat{\theta}_{\mathrm{LS}}$, where $\widehat{\theta}_{\mathrm{LS}}$ is the LSTD solution. Under Assumptions 1 and 3, as well as standard regularity assumptions for linear regression, if n is large enough,*

$$\mathbb{E}_{\mathcal{D}}|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| \lesssim \frac{1}{\sigma_{\min}(I - \gamma \Sigma_{\mathrm{cov}}^{-1/2} \Sigma_{\mathrm{cr}} \Sigma_{\mathrm{cov}}^{-1/2})^2} \sqrt{\frac{d \log(1/\delta)}{n}}$$

with probability $1-\delta$. Here, $\sigma_{\min}(\cdot)$ denotes the minimum singular value of a matrix.

This result follows somewhat directly from a perturbation analysis of approximate solutions to the fixed point equation Eq. (1.3). Perhaps surprisingly, we will see that invertibility is strictly weaker than stability (Assumption 2), which highlights a fundamental distinction between these two methods. This comparison also reveals that stability cannot be a necessary condition in any algorithm-independent sense, since LSTD can succeed without stability. However, complementing Theorem 2, we prove that invertibility is necessary for a large class of natural estimators, specifically those that rely on low-order moments of the features and the regression function between features and the rewards (this includes FQI and LSTD). The following lower bound shows that the value function is *unidentifiable* by these linear estimators if invertibility does not hold.

Theorem 3 (Informal). *Even in the limit of infinite data, any OPE problem for which invertibility does not hold cannot be solved by a broad class of linear estimators, including FQI and LSTD.*

Together with our previous results, this result completes our analysis of linear estimators for offline policy evaluation under linear realizability. We remark that our results are sharp in the sense that they stipulate exactly *which* problems are solvable by linear estimators. They are not necessarily sharp in the sense that the associated statistical rates for each problem are optimal. We believe that establishing appropriate lower bounds for these problems is an important direction for future work.

1.3 Related work

RL with function approximation. Analyses of function approximation in reinforcement learning can be traced to the seminal papers of Bellman and Dreyfus [1959], Bellman [1961], as well as Reetz [1977] and Whitt [1978]. Schweitzer and Seidmann [1985] were one of the first to consider approximating value functions using linear combinations of some known set of features. More recently, a number of modeling assumptions—typically involving strong representational conditions on both the MDP and the features—that enable statistically efficient online RL with linear function approximation have been proposed, along with corresponding algorithms [Zanette et al., 2020, Yang and Wang, 2020, Jin et al., 2020].

FQI. Introduced by Ernst et al. [2005] and extended by Riedmiller [2005], fitted Q-iteration has been analyzed several times in the context of offline policy evaluation. Building off previous studies of approximate methods in dynamic programming [Antos et al., 2008, Munos, 2007, Gordon, 1999], Chen and Jiang [2019] establish sample complexity upper bounds for FQI assuming that the corresponding distributions and MDP satisfy concentrability [Munos, 2003] and Bellman completeness [Szepesvári and Munos, 2005]. While concentrability conditions are orthogonal to realizability assumptions, completeness is significantly stronger than mere realizability of value functions. More recent work by Wang et al. [2021a,b] adapts these results to the linear setting and additionally shows that a "low distribution shift" condition suffices for linear FQI.

LSTD. Initial analysis of least squares temporal difference learning (LSTD) date back to the work of Baird [1995], Bradtke and Barto [1996], Boyan [1999] and Nedić and Bertsekas [2003]. Since then, the finite sample performance of the algorithm has been analyzed by Lazaric et al. [2012], Bhandari et al. [2018], Duan et al. [2021] and its behavior in the offline setting studied by Yu [2010], Li et al. [2021], Mou et al. [2020, 2021], Pires and Szepesvari [2012]. Tu and Recht [2018] analyze on-policy LSTD for the LQR setting. Miyaguchi

[2021] studies the behavior of LSTD for OPE in settings where the value function is only approximately linearly realizable in a known feature mapping ϕ. We evaluate our contributions in light of these previous works in Section 4.2.

Other OPE estimators. Apart from these methods, researchers have studied "min-max" algorithms for OPE which estimate the value of the underlying policy using ideas from the importance sampling literature [Liu et al., 2018, Uehara et al., 2020, Yin and Wang, 2020]. Xie and Jiang [2021] establish formal guarantees for the BVFT algorithm which carries out policy evaluation for general *nonlinear* function classes assuming realizability, albeit under stronger notions of data coverage (see Assumption 8). Recent work by Zhan et al. [2022] extends this line of research. They introduce a new algorithm which works under weaker data coverage assumptions than those in Xie and Jiang [2021]. However, to do so they require additional assumptions on the expressivity of the underlying class of function approximators. In particular, Zhan et al. [2022], and the class of minimax algorithms more broadly, rely on a function class that can (at a minimum) realize the state-occupancy density ratio between the distribution induced by the policy π and the offline distribution $\mathcal{D}$, which is a distinct condition from linear realizability of Q^π.

Lower bounds under linear realizability. For the finite horizon, policy evaluation setting, Wang et al. [2021a] illustrate how exponential dependence on the horizon is unavoidable, even if the offline covariance matrix is robustly full rank. Since then, these bounds have been extended to the discounted, infinite horizon case by Amortila et al. [2020] and Zanette [2021]. Importantly, Amortila et al. [2020] establish that OPE can be information-theoretically intractable, even if: 1) all features are bounded, 2) Σ_{cov} is full rank, and 3) the learner has access to infinitely many samples drawn as in Section 1.1. Analogous negative results for online or generative-model settings have been shown to hold even in the presence of a constant suboptimality gap [Wang et al., 2021c] or polynomially large action sets [Weisz et al., 2021a,b]. Duan et al. [2020] prove lower bounds for OPE which hold for general function classes. Foster et al. [2021] illustrate that polynomially many samples in the size of the state space are necessary for offline policy evaluation, even if concentrability and realizability both hold. In summary, a clean characterization of when offline policy evaluation is tractable using linear function approximation has, so far, proven to be quite elusive.

2 Preliminaries

Before delving into our main results, we review some of the relevant definitions and preliminaries.

Notation. We use $s \in \mathcal{S}$ and $a \in \mathcal{A}$ to denote states and actions, $\top$ to denote vector or matrix transposes, and $\dagger$ to denote pseudoinverses. For a matrix X, we let $\mathsf{cond}(X) := \sigma_{\max}(X)/\sigma_{\min}(X)$ denote its condition number, the ratio between the largest and smallest singular values $\sigma(\cdot)$. For symmetric matrices, A and B, we use $A \succeq B$ if $A - B$ is positive semidefinite. We let $\rho(X) := \max_i |\lambda_i(X)|$ be the spectral radius of a matrix X where λ_i are the eigenvalues.[7] We say that a matrix is *stable* if its spectral radius is strictly smaller than 1. For square, stable matrices A, we let $\mathsf{dlyap}(A)$ be the solution, over X, to the discrete-time Lyapunov equation: $X = A^\top X A + I$. This equation has a solution if and only if $\rho(A) < 1$ [Callier and Desoer, 2012]. If the solution exists, it admits the closed-form expression $\mathsf{dlyap}(A) = \sum_{j=0}^{\infty} (A^\top)^j A^j$. Lastly, we say $a \lesssim b$ if $a \leq c \cdot b$ for some universal constant c.

We define the next state-action covariance Σ_{next} and the distribution shift coefficient $\mathcal{C}_{\text{ds}}$ as

$$\Sigma_{\text{next}} := \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D} \\ s'\sim P(\cdot|s,a),\, a'\sim\pi(s')}} \left[\phi(s',a')\phi(s',a')^\top\right], \quad \mathcal{C}_{\text{ds}} := \inf\{\beta > 0 : \Sigma_{\text{next}} \preceq \beta\Sigma_{\text{cov}}\}. \tag{2.1}$$

Note that $\mathcal{C}_{\text{ds}}$ is guaranteed to be finite if Σ_{cov} is full rank. Given a dataset $\{(s_i, a_i, r(s_i, a_i), s_i', a_i')\}_{i=1}^n$ of n i.i.d. data points drawn according to the data generating process described in Section 1.1, we define the

[7] Recall that for square, but non-symmetric matrices A, it is in general *not* true that $\rho(A) = \sigma_{\max}(A)$. However, $\rho(A) \leq \sigma_{\max}(A)$ does always hold.

empirical counterparts of the second-moment matrices defined in Eq. (1.5),

$$\widehat{\Sigma}_{\text{cov}} := \frac{1}{n}\sum_{i=1}^{n} \phi(s_i, a_i)\phi(s_i, a_i)^\top, \quad \widehat{\Sigma}_{\text{cr}} := \frac{1}{n}\sum_{i=1}^{n} \phi(s_i, a_i)\phi(s_i', a_i')^\top, \tag{2.2}$$

as well as the true, and empirical, mean feature-reward vectors:

$$\theta_{\phi,r} := \mathbb{E}_{\mathcal{D}}\phi(s, a)r(s, a), \quad \widehat{\theta}_{\phi,r} := \frac{1}{n}\sum_{i=1}^{n} \phi(s_i, a_i)r(s_i, a_i). \tag{2.3}$$

Linear regression. Next, we introduce moment-type quantities that arise in our analysis of linear regression. Here, we adopt the approach from Hsu et al. [2012], however, other approaches for analyzing linear regression will yield the same qualitative results. In particular, we make use of the statistical leverages ρ_s and $\rho_{s'}$. These quantities correspond to the maximum length of features, $\phi(s, a)$ and $\phi(s', a')$, when measured in the (inverse) covariance norm. Intuitively, they capture the worst-case coverage of the offline distribution $\mathcal{D}$ over directions in feature space.

$$\rho_s := \sup_{(s,a)\in\text{supp}(\mathcal{D})} \|\Sigma_{\text{cov}}^{-1/2}\phi(s, a)\|_2, \quad \rho_{s'} := \sup_{\substack{(s,a)\in\text{supp}(\mathcal{D}),\\ s'\in\text{supp}(P(\cdot|(s,a)),\ a'\in\text{supp}(\pi(s'))}} \|\Sigma_{\text{cov}}^{-1/2}\phi(s', a')\|_2. \tag{2.4}$$

In addition, we define the variances $\sigma_{\text{cov}}^2, \sigma_r^2$, and σ_{cr}^2 where,

$$\sigma_{\text{cov}}^2 := \|\mathbb{E}(\Sigma_{\text{cov}}^{-1/2}\phi(s, a)\phi(s, a)^\top\Sigma_{\text{cov}}^{-1/2})^2 - I\|_{\text{op}}, \quad \sigma_r^2 := \mathbb{E}\|\Sigma_{\text{cov}}^{-1/2}\phi(s, a)r(s, a)\|_2^2 - \|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2^2, \tag{2.5}$$

and σ_{cr}^2 is the maximum of the following two quantities,

$$\sup_{\|v\|_2=1} \mathbb{E}\left(v^\top\Sigma_{\text{cov}}^{-1/2}\phi(s', a')\right)^2 \|\Sigma_{\text{cov}}^{-1/2}\phi(s, a)\|_2^2 - \|\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}^\top\Sigma_{\text{cov}}^{1/2}v\|_2^2 \tag{2.6}$$

$$\sup_{\|v\|_2=1} \mathbb{E}\left(v^\top\Sigma_{\text{cov}}^{-1/2}\phi(s, a)\right)^2 \|\Sigma_{\text{cov}}^{-1/2}\phi(s', a')\|_2^2 - \|\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{1/2}v\|_2^2. \tag{2.7}$$

In Appendix C.3, we prove that σ_{cr}^2 and σ_{cov}^2 can always be upper bounded in terms of the statistical leverages and the coefficient $\mathcal{C}_{\text{ds}}$.[8] However, they can be much smaller in some settings.[9] Therefore, for the sake of generality, we opt to state our bounds in terms of these quantities. Informally, these variance terms measure how much the corresponding matrices or vectors vary from their means, in the $\Sigma_{\text{cov}}^{-1/2}$ geometry.

Throughout our analysis of methods for offline policy evaluation, we will repeatedly make use of the following concentration result:

Lemma 2.1. *For all $n \gtrsim \rho_s^2 \log(d/\delta)$, define the estimation errors,*

$$\varepsilon_{\text{op}} := \|\Sigma_{\text{cov}}^{1/2}(\gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})\Sigma_{\text{cov}}^{-1/2} - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}, \quad \varepsilon_r := \|\Sigma_{\text{cov}}^{1/2}(\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\theta}_{\phi,r} - \Sigma_{\text{cov}}^{-1}\theta_{\phi,r})\|_2. \tag{2.8}$$

With probability $1-\delta$, $\widehat{\Sigma}_{\text{cov}}$ is full rank and ε_r, ε_{op} satisfy the following inequalities:

$$\varepsilon_{\text{op}} \lesssim \sqrt{\frac{\max(\sigma_{\text{cr}}^2, \sigma_{\text{cov}}^2\mathcal{C}_{\text{ds}})\log(d/\delta)}{n}} + \frac{\max(\mathcal{C}_{\text{ds}}^{1/2}\rho_s^2, \rho_s\rho_{s'})\log(d/\delta)}{n}$$

$$\varepsilon_r \lesssim \sqrt{\frac{\max(\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2^2\sigma_{\text{cov}}^2,\ \sigma_r^2)\log(d/\delta)}{n}} + \frac{\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2\rho_s^2\log(d/\delta)}{n}.$$

Later on, we state our upper bounds on the policy evaluation error of FQI and LSTD in terms of these regression errors ε_{op}, ε_r, with the understanding that they satisfy the high probability upper bounds above.

[8] On the other hand, σ_r^2 is always upper bounded by d.

[9] For example, tighter bounds can be achieved if the distributions are hypercontractive, see Appendix C.3.

3 Fitted Q-Iteration

In this section, we present our first set of results illustrating how stability (Assumption 2) characterizes the success of fitted Q-iteration for OPE under linear realizability of Q^π. Following some initial remarks regarding the functional form of the FQI solution, in Section 3.1, we present our upper bound on the estimation error of FQI. Later on, in Section 3.2, we illustrate how our Lyapunov stability analysis unifies previous studies of when FQI succeeds and conclude by discussing lower bounds and limitations of the algorithm in Section 3.3.

FQI preliminaries. From examining the definition of FQI in Eq. (1.2), we see that, at the population level, the algorithm develops the recursion:

$$\theta_{t+1} = \gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}} \theta_t + \Sigma_{\text{cov}}^{-1} \theta_{\phi,r}.$$

Unrolling the recursion above, and setting $\theta_0 = 0$, the T-step regression vector is equal to:[10]

$$\theta_T = \sum_{k=0}^{T} (\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}})^k \Sigma_{\text{cov}}^{-1} \theta_{\phi,r}. \tag{3.1}$$

Linear realizability of Q^π (Assumption 1) implies that the true weight vector $\theta_\gamma^\star$ satisfies the equation,

$$\Sigma_{\text{cov}} \theta_\gamma^\star = \theta_{\phi,r} + \gamma \Sigma_{\text{cr}} \theta_\gamma^\star. \tag{3.2}$$

Hence, if $I - \gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}}$ is invertible, then $\theta_\gamma^\star = (I - \gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}})^{-1} \Sigma_{\text{cov}}^{-1} \theta_{\phi,r}$. We now recall the following fact:

Fact 3.1. *If $\rho(A) < 1$, then the matrix $(I - A)$ is invertible. Moreover, $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$.*

Using this, along with the observation that the spectrum of a matrix is invariant to the choice of basis, we see that if stability (Assumption 2) holds, then the vector $\theta_\gamma^\star$ can also be written as a power series:

$$\theta_\gamma^\star = \sum_{k=0}^{\infty} (\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}})^k \Sigma_{\text{cov}}^{-1} \theta_{\phi,r}. \tag{3.3}$$

One of the key insights tying stability and FQI is that, regardless of whether $\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}}$ is stable, the FQI solution at the population level is *always* equal to the power series in Eq. (3.1). If stability holds, performing infinitely many exact FQI updates converges to $\theta_\gamma^\star$. However, $\theta_\gamma^\star$ is (in general) *only* equal to this power series if stability holds, which hints at the necessity of this condition. With these connections between stability and the functional forms of FQI and $\theta_\gamma^\star$ in mind, we now present our upper bounds on the performance of this algorithm.

3.1 Stability is sufficient for fitted Q-iteration

Theorem 1. *Assume that Q^π is linearly realizable (Assumption 1) and that stability holds (Assumption 2). For $\varepsilon_{\text{op}}, \varepsilon_r$ defined as in Eq. (2.8), if $n \gtrsim \rho_s^2 \log(d/\delta)$ and $\varepsilon_{\text{op}} \leq 1/(6\|P_\gamma\|_{\text{op}}^2)$, T-step FQI satisfies,*

$$\begin{aligned} \|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2 \lesssim\ & \mathsf{cond}(P_\gamma)^{1/2} \|P_\gamma\|_{\text{op}} \cdot \varepsilon_r + \mathsf{cond}(P_\gamma) \|P_\gamma\|_{\text{op}}^2 \cdot \|\Sigma_{\text{cov}}^{-1/2} \theta_{\phi,r}\|_2 \cdot \varepsilon_{\text{op}} \\ & + \mathsf{cond}(P_\gamma) \|P_\gamma\|_{\text{op}} \cdot \|\Sigma_{\text{cov}}^{1/2} \theta_{\phi,r}\|_2 \cdot \exp\left(-\frac{T+1}{2\|P_\gamma\|_{\text{op}}} \right). \end{aligned} \tag{3.4}$$

Let $\widehat{Q}^\pi(s,a) := \phi(s,a)^\top \widehat{\theta}_T$. Much like in standard analyses of linear regression, from Theorem 1 we immediately obtain: (1) a bound on $\mathbb{E}_{\mathcal{D}}|Q^\pi(s,a) - \widehat{Q}(s,a)|$ via Jensen's inequality since $\mathbb{E}_{\mathcal{D}}(Q^\pi(s,a) -$

[10] We initialize at 0 for simplicity, but this is not fundamental for the overall analysis of FQI.

$\widehat{Q}^\pi(s,a))^2 = \|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2^2$ and (2) a bound on $|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)|$ for any (s,a) pair since $|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| \le \|\Sigma_{\text{cov}}^{-1/2}\phi(s,a)\|_2 \|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2$ via Cauchy-Schwarz.

We defer the full proof to Appendix A.1 and instead summarize the key steps here. The theorem is essentially a perturbation bound which distinguishes between two sources of error in policy evaluation for FQI: ε_r which captures errors in learning the rewards, and the dominant error, ε_{op}, which comes from estimating the transitions. Since under stability, we can write the true vector $\theta_\gamma^\star$ as a power series in second moment matrices (see Eq. (3.3)), and since $\widehat{\theta}_T$ is by definition a truncated power series in the empirical counterparts of these matrices, we can show that the error between $\theta_\gamma^\star$ and $\widehat{\theta}_T$ is bounded by the operator norm of two power series: one in $(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^k$ and the other in $(\gamma\widehat{\Sigma}_{\text{cov}}^{-1/2}\widehat{\Sigma}_{\text{cr}}\widehat{\Sigma}_{\text{cov}}^{-1/2})^k$. Lyapunov arguments directly show that the powers of $(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$ decay exponentially in k since the matrix is stable. For the empirical version, we use the fact that any stable matrix A has nontrivial *stability margin*: for small enough perturbations Δ, matrices of the form $A + \Delta$ satisfy similar decay rates to A. Thus, we can bound the two power series by simple geometric series and the perturbation bound follows.

We now highlight some of the salient aspects of the bound.

Coordinate invariance. The bound in Theorem 1 is *coordinate-free*, in the sense that all problem quantities are invariant to the basis in which one chooses to represent the features. Linear realizability states that $Q^\pi(s,a) = \phi(s,a)^\top\theta_\gamma^\star$. Consequently, for any invertible matrix L, it also holds that $Q^\pi(s,a) = \widetilde{\phi}(s,a)^\top\widetilde{\theta}_\gamma^\star$ where,

$$\widetilde{\phi}(\cdot) = L\phi(\cdot) \text{ and } \widetilde{\theta}_\gamma^\star = L^{-1}\theta_\gamma^\star.$$

Observe that the regression errors (ε_r and ε_{op}) in the data norm, the geometry induced by Σ_{cov}, do not depend on the choice of matrix L, since the variances and statistical leverages are invariant to the coordinate system (see Lemma 2.1). The invariance of $\|P_\gamma\|_{\text{op}}$ and $\mathsf{cond}(P_\gamma)$ is perhaps less straightforward, but verified in the following proposition:

Proposition 3.2. *Let $L \in \mathbb{R}^{d\times d}$ be an invertible matrix and let $\widetilde{\phi}(\cdot) = L\phi(\cdot)$ be the feature mapping in the new coordinates. Now, define $\widetilde{P}_\gamma := \mathsf{dlyap}(\gamma\widetilde{\Sigma}_{\text{cov}}^{-1/2}\widetilde{\Sigma}_{\text{cr}}\widetilde{\Sigma}_{\text{cov}}^{-1/2})$, where*

$$\widetilde{\Sigma}_{\text{cov}} := \mathbb{E}_{(s,a)\sim\mathcal{D}}\widetilde{\phi}(s,a)\widetilde{\phi}(s,a)^\top, \text{ and } \widetilde{\Sigma}_{\text{cr}} := \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D}\\ s'\sim P(\cdot|s,a),\, a'\sim\pi(s')}} \widetilde{\phi}(s,a))\widetilde{\phi}(s',a')^\top. \tag{3.5}$$

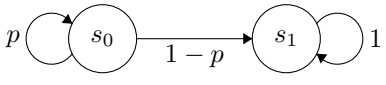
Then, $\|P_\gamma\|_{\text{op}} = \|\widetilde{P}_\gamma\|_{\text{op}}$ and $\mathsf{cond}(P_\gamma) = \mathsf{cond}(\widetilde{P}_\gamma)$. Furthermore,

$$\gamma\widetilde{\Sigma}_{\text{cov}}^{-1/2}\widetilde{\Sigma}_{\text{cr}}\widetilde{\Sigma}_{\text{cov}}^{-1/2} = \gamma U\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}U^\top,$$

where $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix.

Sharpness of $\|P_\gamma\|_{\text{op}}$ vs $1/(1-\gamma)$. Apart from showing how stability is sufficient for offline policy evaluation under linear realizability, another highlight of Theorem 1 is that it introduces a new measure of problem complexity, $\|P_\gamma\|_{\text{op}}$, which is in general significantly sharper than previous complexity measures traditionally considered in the literature, such as the effective horizon, $1/(1-\gamma)$. The difference between these two quantities is evident even in very simple settings:

Consider the following MDP (with no actions), where arrows denote transition probabilities:



(3.6)

If $\mathbb{E}r(s_0) \neq 0$ and $\mathbb{E}r(s_1) = 0$, realizability holds with 1 dimensional features: $\phi(s_0) = 1$ and $\phi(s_1) = 0$. For $\mathcal{D}$ supported just on s_0, then $\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}} = p\gamma$, and $P_\gamma = 1/(1-(p\gamma)^2)$. If $p \le 0.7$, then for all $\gamma \in (0,1)$, $\|P_\gamma\|_{\text{op}} \le 2$, but $(1-\gamma)^{-1}$ can be arbitrarily large as $\gamma \to 1$.

This example illustrates how there are problems for which $\|P_\gamma\|_{\text{op}}$ is significantly smaller than $1/(1-\gamma)$. In the next subsection, we complement this result by illustrating how for settings where FQI was previously shown to succeed, $\|P_\gamma\|_{\text{op}}$ is in fact never much worse than $1/(1-\gamma)$. Taken together, these results demonstrate how $\|P_\gamma\|_{\text{op}}$ provides a sharper notion of the statistical complexity of OPE than $1/(1-\gamma)$.

3.2 Contextualizing Lyapunov stability

Having presented our analysis of fitted Q-iteration through the lens of Lyapunov stability, we now illustrate how this perspective unifies previously disparate analyses of FQI for offline policy evaluation. The central message of this subsection is that the previously proposed conditions which guarantee that FQI will succeed at offline policy evaluation directly imply our key assumption that $\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$ is stable.

Before discussing these connections, we present the following lemma which is closely related to Theorem 1. It upper bounds the error of FQI assuming particular decay rates on the powers of the whitened cross-covariance matrix. Although the proof is essentially identical to the previous result, we can obtain sharper results assuming particular rates of decay, which will be helpful for later comparisons.

Lemma 3.3. *Assume $n \gtrsim \rho_s^2 \log(d/\delta)$ and let ε_{op} and ε_r be defined as in Eq. (2.8). Under the same assumptions as Theorem 1, if there exist $\alpha > 0$ and $\beta \in (0,1)$ such that for all $k \geq 0$,*

$$\|(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^k\|_{\text{op}} \leq \alpha \cdot \beta^k, \tag{3.7}$$

then the T-step FQI solution satisfies the following guarantee. With probability $1-\delta$, if $\varepsilon_{\text{op}} \leq \frac{(1-\beta)}{2\alpha}$,

$$\|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2 \lesssim \varepsilon_r \cdot \frac{\alpha}{1-\beta} + \varepsilon_{\text{op}} \cdot \|\Sigma_{\text{cov}}^{1/2}\theta_{\phi,r}\|_2 \frac{\alpha^2}{(1-\beta)^2} + \|\Sigma_{\text{cov}}^{1/2}\theta_{\phi,r}\|_2 \frac{\alpha}{1-\beta} \cdot \beta^{T+1}. \tag{3.8}$$

Throughout this section, we will present corollaries of this result, which can be viewed as specializations of Theorem 1 to particular settings. In each case, we will focus on discussing variants of the perturbation bound (Eq. (3.8)) which hold under the specific assumptions.

3.2.1 Low distribution shift implies stability

Recent work by Wang et al. [2021b] shows that FQI succeeds at OPE for infinite horizon, discounted problems if there is low distribution shift. More formally, they prove offline evaluation is tractable if the offline covariance Σ_{cov} has good coverage over the next state covariance Σ_{next} as per the following assumption.

Assumption 4 (Low Distribution Shift)**.** There is low distribution shift if $\mathcal{C}_{\text{ds}} < 1/\gamma^2$.

Note that if $\mathcal{D}$ is the stationary measure for π, then $\Sigma_{\text{cov}} = \Sigma_{\text{next}}$ and Assumption 4 holds with $\mathcal{C}_{\text{ds}} = 1$ (recall the definition of $\mathcal{C}_{\text{ds}}$ in Eq. (2.1)). Under this low distribution shift condition, we prove:

Corollary 3.1. *If there is low distribution shift (Assumption 4) and if Σ_{cov} is full rank, then for all $j \geq 0$,*

$$\|(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^j\|_{\text{op}} \leq (\sqrt{\mathcal{C}_{\text{ds}}\gamma^2})^j. \tag{3.9}$$

Hence, $\|P_\gamma\|_{\text{op}} \leq 1/(1-\gamma\sqrt{\mathcal{C}_{\text{ds}}})$ and Assumption 2 holds.[11] *Furthermore, for $\gamma_{\text{ds}} := \gamma\sqrt{\mathcal{C}_{\text{ds}}}$, if Q^π is linearly realizable (Assumption 1), $n \gtrsim \rho_s^2 \log(d/\delta)$, and $\varepsilon_{\text{op}} \leq 1/2(1-\gamma_{\text{ds}})$, then T-step FQI satisfies:*

$$\|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2 \lesssim \frac{1}{1-\gamma_{\text{ds}}}\varepsilon_r + \frac{1}{(1-\gamma_{\text{ds}})^2}\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \cdot \varepsilon_{\text{op}} + \|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \frac{1}{1-\gamma_{\text{ds}}}\gamma_{\text{ds}}^{T+1}.$$

While low distribution shift implies stability, the converse is not true. It is not hard to come up with examples where $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is stable, yet the distribution shift coefficient is larger than $1/\gamma^2$. We present such an example later on in Proposition 4.4.

[11] A matrix A is stable if and only if $\lim_{k\to\infty} A^k = 0$.

3.2.2 Bellman completeness implies stability

In addition to the low-distribution shift setting, FQI is known to succeed in both finite horizon and discounted, infinite horizon settings under a representational condition known as Bellman completeness [Szepesvári and Munos, 2005, Wang et al., 2021a,b]:

Assumption 5 (Bellman completeness)**.** A feature map ϕ is Bellman complete for an MDP $\mathcal{M}$, if for all $\theta \in \mathbb{R}^d$, there exists a vector θ' such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\phi(s,a)^\top \theta' = \mathbb{E}\left[r(s,a)\right] + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a), a' \sim \pi(s')} \phi(s',a')^\top \theta.$$

Intuitively, completeness asserts that Bellman backups of linear functions of the features again lie in the span of the features. It has previously been observed [Wang et al., 2021a,b] that completeness implies a certain "non-expansiveness" of Bellman backups. This non-expansiveness is the key step towards establishing the connection to stability and is formalized in the following result:

Corollary 3.2. *If ϕ is Bellman complete (Assumption 5) and Σ_{cov} is full rank, then for all $j \geq 0$,*

$$\|(\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2})^j\|_{\text{op}} \leq \rho_s \gamma^j. \tag{3.10}$$

Hence, $\|P_\gamma\|_{\text{op}} \leq \rho_s/(1-\gamma)$, and Assumption 2 holds. Furthermore, if Q^π is linearly realizable (Assumption 1), $n \gtrsim \rho_s^2 \log(d/\delta)$, and $\varepsilon_{\text{op}} \leq (1-\gamma)/(2\rho_s)$, T-step FQI satisfies: [12]

$$\|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta^\star_\gamma)\|_2 \lesssim \frac{\rho_s}{1-\gamma} \cdot \varepsilon_r + \frac{\rho_s^2}{(1-\gamma)^2}\varepsilon_{\text{op}} + \frac{\rho_s}{1-\gamma}\gamma^{T+1}.$$

Again, as with low distribution shift setting, the converse statement is not true. There are OPE instances which are stable, but not Bellman complete (Proposition 4.4)

The tabular case. To help contextualize this result, and build some intuition between Bellman completeness and stability, we can consider the case of the tabular MDP. The tabular MDP is perhaps the simplest setting in which the Bellman completeness holds. In our setup, it means that $\mathcal{S}$ and $\mathcal{A}$ are both finite sets and that the feature mapping is equal to $\phi(s,a) = e_{sa} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for all s and a (each input maps to a distinct standard basis vector). The matrix Σ_{cov} being full rank means that every pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is in the support of the offline distribution $\mathcal{D}$. A direct calculation shows that

$$\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1} = \gamma P^\pi,$$

where $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is a row-stochastic matrix with nonnegative entries. Each row in this matrix is indexed by a pair (s,a). Entries in each row describe the probability that the next state action pair is (s',a') given that the current pair is (s,a). Because the spectral radius of any stochastic matrix is 1, when we multiply by γ, we get that $\rho(\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2}) < 1$ and stability holds.

3.3 Stability is necessary for fitted Q-iteration

We conclude our analysis of FQI by showing that our characterization of when the algorithm succeeds is exactly sharp, in an instance-dependent sense. If stability fails that is, $\rho(\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}}) > 1$, then estimation procedures of this sort are guaranteed to have exponentially large variance.

Proposition 3.4. *Let $\mathcal{M}$ be any infinite horizon, discounted MDP with corresponding offline distribution $\mathcal{D}$ which satisfies the following properties: Σ_{cov} is full rank and $\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}}$ has an eigenvalue λ with $|\lambda| > 1$. Then, approximations of the T-step FQI solution, $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top \widehat{\theta}_T$ where,*

$$\widehat{\theta}_T := \sum_{k=0}^{T} (\gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}})^k \Sigma_{\text{cov}}^{-1} \widehat{\theta}_{\phi,r}, \quad \widehat{\theta}_{\phi,r} := \theta^\star_{\phi,r} + z,$$

[12] Completeness implies realizability of rewards which in turn implies $\|\Sigma_{\text{cov}}^{-1/2} \theta_{\phi,r}\|_2^2 = \mathbb{E} r(s,a)^2 \leq 1$, see Lemma A.6.

and z is a zero-mean, random vector satisfying $\Lambda := \mathbb{E}zz^\top \succ 0$, have exponentially large variance,

$$\mathbb{E}\|\widehat{\theta}_T - \mathbb{E}\widehat{\theta}_T\|_2^2 \;\geq\; \sigma_{\min}(\Lambda) \cdot \left(\frac{\lambda^{T+1}-1}{\lambda-1}\right)^2.$$

This proposition corroborates empirical findings on the instability of FQI by Wang et al. [2021b] and shows that an idealized variant of FQI incurs exponentially large variance (in the number of rounds T) for an instance that results in an unstable "backup operator" $\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$. By standard bias-variance decomposition, this directly implies exponentially large error for estimating $\theta_\gamma^\star$. Although, note that since stability does not hold, there is no guarantee that $\theta_\gamma^\star$ can be written as a power series, so it may not even be the limiting solution of population FQI as discussed at the beginning of this section.

The algorithm is idealized in two senses, both of which are relatively minor. First, it has perfect knowledge of Σ_{cov} and Σ_{cr} which does not happen in practice, but is favorable to the algorithm, resulting in a stronger lower bound. Second, the error in estimating the reward is assumed to have a full-rank covariance; this arises naturally whenever rewards are perturbed with centered Gaussian noise since Σ_{cov} is full rank. Thus, the result shows that even when the dynamics are known, errors in estimating the rewards will be exponentially magnified, resulting in overall divergence of the algorithm.

While the theorem does not consider the marginally stable case where $\rho(\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}) = 1$, we note in the proof that if the spectral radius is exactly one, the variance can grow at least linearly with T. However, marginal stability introduces other issues as we illustrate later on.

At this point, it is natural to wonder whether stability is necessary not just for the success of this algorithm, but rather for the success of *any* algorithm at offline policy evaluation. It turns out that this is is not the case. As we will show in the following section, least squares temporal difference learning works under strictly weaker conditions than fitted Q-iteration.

4 Least Squares Temporal Difference Learning

Building on our analysis of FQI, we now analyze how a closely related algorithm, least squares temporal difference learning, overcomes some of its shortcomings in the context of offline policy evaluation. Similarly to the previous section, we start by illustrating how invertibility is sufficient for LSTD in Section 4.1, and discuss connections to previous sufficient conditions in Section 4.2. Lastly, we conclude in Section 4.3 by presenting lower bounds which show that if invertibility does not hold, then the offline policy evaluation problem cannot be solved using linear estimators (FQI and LSTD being special cases), even asymptotically.

4.1 Invertibility is sufficient for LSTD

Theorem 2. *Assume that realizability and invertibility (Assumptions 1 and 3) both hold and let ε_r, ε_{op} be defined as in Eq. (2.8). If $n \gtrsim \rho_s^2 \log(d/\delta)$ and $\varepsilon_{\text{op}} \leq \sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})/2$, then the LSTD solution,*

$$\widehat{\theta}_{\text{LS}} := (I - \gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})^\dagger\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\theta}_{\phi,r},$$

satisfies the following error guarantee:

$$\|\Sigma_{\text{cov}}^{1/2}(\theta_\gamma^\star - \widehat{\theta}_{\text{LS}})\|_2 \lesssim \frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \cdot \varepsilon_r + \frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^2}\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \cdot \varepsilon_{\text{op}}. \tag{4.1}$$

As per our discussion immediately following Theorem 1, the upper bound on $\|\Sigma_{\text{cov}}^{1/2}(\theta_\gamma^\star - \widehat{\theta}_\gamma)\|_2$ again directly implies guarantees on $|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)|$, both pointwise and in expectation, where now $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top\widehat{\theta}_{\text{LS}}$. On a technical level, the proof follows from standard perturbation bounds on matrix inverses.

Our upper bound for LSTD has qualitatively similar properties to that presented for FQI in Theorem 1.

A sharper notion of problem complexity. Much like $\|P_\gamma\|_{\text{op}}$ for FQI, the magnitude of our upper bound for the policy evaluation error of LSTD is determined by an instance-dependent quantity: $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$. As before this term is: (1) never much larger than $1/(1-\gamma)$ for settings where OPE was previously shown to be tractable (see the next subsection for further discussion of this point), and (2) is often significantly smaller. For example, for the OPE instance detailed in (3.6), if $p \leq .7$, then $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) \leq 4$ for all $\gamma \in (0,1)$.

Coordinate invariance. From Proposition 3.2, we know that for any choice of full rank matrix L and features $\widetilde{\phi}(\cdot) = L\phi(\cdot)$, the whitened cross-covariance in these new features, $\gamma\widetilde{\Sigma}_{\text{cov}}^{-1/2}\widetilde{\Sigma}_{\text{cr}}\widetilde{\Sigma}_{\text{cov}}^{-1/2}$ (see definition in Eq. (3.5)) is equal to $\gamma U\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}U^\top$ for some orthogonal matrix U. Since conjugating by an orthogonal matrix preserves singular values, $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$ is invariant to the choice of coordinates.

4.2 Contextualizing Invertibility

Paralleling our discussion of stability for FQI, we now discuss how our notion of invertibility relates to previous conditions analyzed in the literature. Furthermore, we will present how stability implies invertibility, establishing a precise "nesting" between the classes of OPE problems which satisfy each condition.

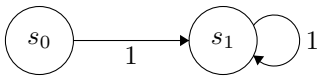
4.2.1 Stability $\subsetneq$ Invertibility

Proposition 4.1. *If Σ_{cov} is full rank and $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is stable (Assumption 2), then $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible (Assumption 3). Furthermore,*

$$\frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \lesssim \mathsf{cond}(P_\gamma)^{1/2}\|P_\gamma\|_{\text{op}}. \tag{4.2}$$

The main message of this proposition is twofold. First, for the case of linear function approximation, any OPE problem that is solvable via FQI, must also be solvable via LSTD. Second, from Eq. (4.2) we see that main complexity measure for Theorem 2, $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$ is never larger than the corresponding upper bound for FQI in Theorem 1, $\mathsf{cond}(P_\gamma)^{1/2}\|P_\gamma\|_{\text{op}}$.

Interestingly enough, while stability implies invertibility, the converse is not true. There exist problems for which $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible, but $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is not stable. For example, consider the following 2 state MDP, with no actions:



If we set $R(s_0) = R(s_1) = \text{Unif}(\{\pm 1\})$, and $\phi(s_0) = 1$, $\phi(s_1) = 2$, then this OPE instance is trivially linearly realizable with $\theta^\star_\gamma = 0$. If the offline distribution $\mathcal{D}$ places mass p on s_0 and $1-p$ on s_1, it is easy to see that $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible for all $p, \gamma \in (0,1)$. However, for $p = \gamma = .9$, $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is at least $3/2$, hence stability does not hold and FQI will necessarily diverge. Together, these results establish a separation between the set of problems solvable via FQI and those solvable via LSTD.[13]

Moreover, for the set of previously analyzed settings where stability holds, we can establish quantitative upper bounds on $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$ illustrating how this quantity is comparable to $1/(1-\gamma)$.

Corollary 4.1. *Assume $\Sigma_{\text{cov}} \succ 0$. If there is low distribution shift (Assumption 4), then for $\gamma_{\text{ds}} := \gamma\sqrt{\mathcal{C}_{\text{ds}}}$,*

$$\frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \leq \frac{1}{1-\gamma_{\text{ds}}}.$$

[13] The careful reader might observe that the main reason why FQI fails in this example is that the algorithm is sensitive to the scale of the next state features. For instance, stability (and realizability) would hold if $|\phi(s_1)| < 1$.

Moreover, if Bellman completeness holds (Assumption 5), then

$$\frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \leq \frac{\rho_s}{1-\gamma}.$$

This result follows from observing that $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) = \|(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\|_{\text{op}}$. Since stability holds for both of these settings, we can use Fact 3.1 to write $(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}$ as an infinite power series in $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$. Applying the triangle inequality and the bounds from Eqs. (3.9) and (3.10) on the powers of $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ finishes the proof of this corollary.

4.2.2 Other connections

Recent work by Mou et al. [2020] analyzes oracle inequalities for solving projected fixed point equations, of which the Bellman equation (Eq. (3.2)) is a special case. For the offline policy evaluation setting, they prove that a stochastic approximation variant of LSTD succeeds if the following condition holds:

Assumption 6 (Symmetric Stability)**.** The matrix Σ_{cov} is full rank, and $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ satisfies

$$\kappa := \frac{1}{2}\lambda_{\max}(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2} + (\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^\top) < 1.$$

Here, $\lambda_{\max}$ denotes the maximal eigenvalue of a matrix.[14] In their paper, the authors remark how Assumption 6 directly implies that $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible. Amongst other quantities, their bounds scale with $1/(1-\kappa)$. This quantity is always at least as large as $1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$.

Proposition 4.2. *If Assumption 6 holds, then $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible and*

$$\frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \leq \frac{1}{1-\kappa}.$$

Recent work by Li et al. [2021] extends the stochastic approximation analysis from Mou et al. [2020] to incorporate variance reduction techniques. Their upper bounds directly assume invertibility, but also have explicit dependence $1/(1-\gamma)$ which can be quite loose in certain settings as detailed earlier.

Apart from these analyses, Kolter [2011] proves that LSTD succeeds in the offline setting if a certain linear matrix inequality holds:

Assumption 7 (Contractivity)**.** The matrix Σ_{cov} is full rank and together with Σ_{cr} satisfies,

$$\begin{bmatrix} \Sigma_{\text{cov}} & \Sigma_{\text{cr}} \\ \Sigma_{\text{cr}}^\top & \Sigma_{\text{cov}} \end{bmatrix} \succeq 0.$$

A simple Schur complement argument illustrates that this assumption from Kolter [2011] implies that the whitened cross covariance has *operator norm* strictly less than 1. Since the spectral radius of a matrix is always smaller than its operator norm, this condition directly implies that $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is stable (Assumption 2) and that $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is invertible (Assumption 3).

Proposition 4.3. *If Assumption 7 holds, then $\|\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} < 1$ and stability holds.*

As in the case of FQI, we see how our characterization of LSTD in terms of invertibility neatly unifies previous analyses of when this algorithm succeeds in the offline setting. Furthermore, our invertibility-based analysis strictly subsumes these previous studies. There exist problems for which stability and invertibility hold but these other conditions (e.g., low distribution shift, Bellman completeness, etc.) do not.

[14] The matrix in Assumption 6 is symmetric so all eigenvalues are real and the maximum is well defined.

Proposition 4.4. *For each of the following cases, there exists an offline policy evaluation problem defined by an MDP $\mathcal{M}$, an offline distribution $\mathcal{D}$, and a target policy π such that Q^π is linearly realizable in a feature mapping ϕ (Assumption 1 holds) where:*

- *Stability and invertibility both hold, yet low distribution shift (Assumption 4) does not.*
- *Stability and invertibility both hold, yet Bellman completeness (Assumption 5) does not.*
- *Stability and invertibility both hold, yet symmetric stability (Assumption 6) does not.*
- *Stability and invertibility both hold, yet contractivity (Assumption 7) does not.*

In short, there is a nontrivial gap between the problems we knew could be solved via previous analyses and the ones we know we can solve in light of our work.

4.3 Invertibility is necessary for all linear estimators

We finish our presentation of LSTD by proving that invertibility is not just sufficient, it is also strictly necessary for LSTD, as well as for a broad class of "linear" estimators. To do so, we first formally define what we mean by linear estimators:

Definition 4.1 (Population Linear Estimator)**.** Let Alg be a deterministic algorithm which given an infinite horizon, discounted MDP $\mathcal{M}$, a distribution $\mathcal{D}$ over $\mathcal{S}\times\mathcal{A}$, and a policy π returns a function $\widehat{Q}^\pi : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$. Furthermore, let $(\mathcal{M},\mathcal{D},\pi)$ and $(\overline{\mathcal{M}},\overline{\mathcal{D}},\overline{\pi})$ be two OPE instances such that:

- The corresponding action value functions $Q^\pi, \bar{Q}^{\overline{\pi}}$ are both linearly realizable in a feature map ϕ.
- The covariance, cross-covariance and mean feature-reward vectors (as defined in Eqs. (1.5) and (2.3)) are identical in $(\mathcal{M},\mathcal{D},\pi)$ and $(\overline{\mathcal{M}},\overline{\mathcal{D}},\overline{\pi})$:

$$\bar{\Sigma}_{\text{cov}} = \Sigma_{\text{cov}}, \quad \bar{\Sigma}_{\text{cr}} = \Sigma_{\text{cr}}, \quad \bar{\theta}_{\phi,r} = \theta_{\phi,r}, \quad \mathbb{E}_{\mathcal{D}}\overline{r}(s,a) = \mathbb{E}_{\mathcal{D}} r(s,a).$$

We say that Alg is a *population linear estimator* if $\mathsf{Alg}(\mathcal{M},\mathcal{D},\pi) = \mathsf{Alg}(\overline{\mathcal{M}},\overline{\mathcal{D}},\overline{\pi})$.

While our focus has been on studying the finite sample performance of estimators for OPE, in this definition we choose to catalogue algorithms based on their asymptotic behavior so as to neatly abstract technical modifications like variance reduction. These techniques introduce differences in finite sample behaviors, but are not essential to the overall *identifiability* concerns that are the focus of this subsection.

Intuitively, linear estimators are those whose population-level solution depends on the low-order moments of the data. These moments correspond to the quantities which appear in the solution to the projected Bellman equation:

$$\Sigma_{\text{cov}}\theta^\star_\gamma = \theta_{\phi,r} + \gamma\Sigma_{\text{cr}}\theta^\star_\gamma.$$

From their definitions in Eqs. (1.4) and (3.1), we see that common estimators such as LSTD and FQI both satisfy this definition. Interestingly, not all known, or least-squares-like, estimators are linear (e.g Bellman Residual Minimization). We will discuss these after presenting the lower bound.

Theorem 3. *Let $\mathcal{M} = (\mathcal{S},\mathcal{A},P,R,\gamma)$ be any MDP with associated offline distribution $\mathcal{D}$ with rewards uniformly bounded by 1 (i.e., $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}|r(s,a)| \le 1$) such that:*

- *$Q^\pi(s,a)$ is linearly realizable in ϕ.*
- *Σ_{cov} is full rank.*
- *$I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$ is rank deficient.*

Then, there exists a different MDP $\overline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, P, \overline{R}, \gamma)$*, with identical states, actions, and transitions, and whose reward distribution* $\overline{R}$ *is uniformly bounded by 2, such that for the same offline distribution* $\mathcal{D}$*:*

- *The Q-function for* π *in* $\overline{\mathcal{M}}$*,* $\bar{Q}^\pi$*, is linearly realizable in the same feature mapping* ϕ*.*
- *The covariance, cross-covariance, next state covariance, and mean feature-reward vector in* $\overline{\mathcal{M}}$ *are identical to their counterparts in* $\mathcal{M}$*:*

$$\bar{\Sigma}_{\text{cov}} = \Sigma_{\text{cov}}, \quad \bar{\Sigma}_{\text{cr}} = \Sigma_{\text{cr}}, \quad \bar{\Sigma}_{\text{next}} = \Sigma_{\text{next}}, \quad \bar{\theta}_{\phi,r} = \theta_{\phi,r}.$$

- *However, the Q functions are different:*

$$\mathbb{E}_{\mathcal{D}}(Q^\pi(s,a) - \bar{Q}^\pi(s,a))^2 \gtrsim \sigma_{\min}(\Sigma_{\text{cov}}) \,/ \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|_2^2.$$

Consequently, if we define LE *as the set of population linear estimators which satisfy Definition 4.1, we have that*

$$\inf_{\mathsf{Alg}\in\mathsf{LE}} \sup_{(\mathcal{M}',\mathcal{D}',\pi')\in\mathcal{N}} \mathbb{E}_{\mathcal{D}}(Q'^\pi(s,a) - \widehat{Q}^\pi(s,a))^2 \gtrsim \sigma_{\min}(\Sigma_{\text{cov}}) \,/ \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\|_2^2.$$

where $\widehat{Q}^\pi = \mathsf{Alg}(\mathcal{M}', \mathcal{D}', \pi')$ *and* $\mathcal{N} = \{(\mathcal{M}, \mathcal{D}, \pi), (\overline{\mathcal{M}}, \mathcal{D}, \pi)\}$

In other words, this theorem states that for *any* OPE instance where $I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$, or equivalently, $I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$, is rank deficient, we can perturb the rewards to construct an alternative instance with matching low order moments. Consequently, any population linear estimator, such as LSTD or FQI, will return the same estimate $\widehat{Q}^\pi$ in both cases. Yet, since the Q-functions are distinct, they will necessarily converge to the wrong answer in one case. Note that the alternative instance $\overline{\mathcal{M}}$ has identical states, actions, and transitions. Therefore, any function of these quantities, not just the ones explicitly listed above, will be the same in $\mathcal{M}$ and $\overline{\mathcal{M}}$. Together with Theorem 2, this result illustrates how our characterization of the settings where LSTD succeeds is exactly sharp in an instance-dependent (local) sense.

4.3.1 Going beyond linear estimators

Bellman residual minimization. Bellman residual minimization attempts to estimate the value of a decision making policy by solving the following optimization problem, defined here at the population level:

$$\theta_{\text{BRM}} \in \arg\min_\theta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}, s'\sim P(\cdot|s,a), a'\sim\pi(s')} (\phi(s,a)^\top\theta - r(s,a) - \gamma\cdot\phi(s',a')^\top\theta)^2.$$

In the linear function approximation setting, the BRM solution is equal to

$$\theta_{\text{BRM}} = (\Sigma_{\text{cov}} - \gamma\Sigma_{\text{cr}} - \gamma\Sigma_{\text{cr}}^\top + \gamma^2\Sigma_{\text{next}})^\dagger(\theta_{\phi,r} - \gamma\mathbb{E}\phi(s',a')r(s,a)).$$

The key difference with regards to previously analyzed estimators is that BRM depends on the correlation between the *next* state feature vector $\phi(s',a')$ and the reward. However, FQI and LSTD only depend on the correlation $\theta_{\phi,r} = \mathbb{E}\phi(s,a)r(s,a)$ between the *current* state and the reward.

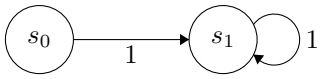
To the best of our knowledge, there is no exact characterization of when BRM succeeds at offline policy evaluation under linear realizability. In particular, it is not sufficient for the matrix,

$$\Sigma_{\text{cov}} - \gamma\Sigma_{\text{cr}} - \gamma\Sigma_{\text{cr}}^\top + \gamma^2\Sigma_{\text{next}},$$

to be invertible. On the other hand, it is well-known that BRM can be inconsistent if the dynamics of the MDP are not deterministic. In general, this algorithm requires use of the *double sampling trick* and the ability to reset the environment to particular states via a simulator. We provide a more detailed discussion of these issues in Appendix B.6 and refer the interested reader to [Baird, 1995, Saleh and Jiang, 2019].

Algorithm independent limits of OPE Given the negative result from Theorem 3, a natural question to ask is: what are the algorithm-independent limits for OPE under linear realizability? We close this section with a brief discussion of how our work provides insight into this question.

We start by pointing out that there are settings where invertibility fails and for which offline policy evaluation is information-theoretically impossible. That is, OPE is not solvable regardless of the choice of estimator or the number of samples observed. This observation follows from the construction in Amortila et al. [2020]. We reproduce their result for the sake of completeness:



There are 2 states and no actions. The feature map is defined as $\phi(s_0) = \gamma$ and $\phi(s_1) = 1$. The rewards are $\mathbb{E}r(s_0) = 0$ and $\mathbb{E}r(s_1) = r_\star \neq 0$. Realizability holds for any choice $r_\star$ with $\theta^\star_\gamma = r_\star/(1-\gamma)$. If the offline distribution $\mathcal{D}$ is supported just on s_0, then $\Sigma_{\text{cov}} = \gamma^2$, $\Sigma_{\text{cr}} = \gamma$ and $I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}} = 0$. Hence, invertibility fails for this problem. Furthermore, because the nonzero reward $r_\star$ is never observed under the offline distribution $\mathcal{D}$, OPE is impossible even in the limit of infinite data. [15] In short, this example shows that if invertibility fails, then OPE cannot be solved in the worst case. However, there are problems where invertibility fails, yet offline policy evaluation is still possible via nonlinear estimators.

Introduced by Xie and Jiang [2021], the BVFT algorithm is a statistically, but not computationally, efficient algorithm for offline policy evaluation using a *general* function class $\mathcal{F}$ under two assumptions: (1) Q^π is realizable by a function in the class $\mathcal{F}$ and (2) the offline distribution $\mathcal{D}$ and the MDP dynamics satisfy a strong data coverage condition referred to as *pushforward concentrability*.

Assumption 8 (Pushforward Concentrability, Xie and Jiang [2021])**.** An MDP $\mathcal{M}$ and offline distribution $\mathcal{D}$ satisfy pushforward concentrability if:

- The offline distribution $\mathcal{D}$ has strictly positive mass on all $(s,a) \in \mathcal{S} \times \mathcal{A}$: $P_{\mathcal{D}}(s,a) > 0$.
- There exists a constant $1 \leq C_A < \infty$ such that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $P_{\mathcal{D}}(a \mid s) \geq 1/C_A$.
- The exists a constant $0 < C_S < \infty$ such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.[16]

$$\frac{P(s' \mid s,a)}{P_{\mathcal{D}}(s')} \leq C_S.$$

In the linear function approximation setting, realizability of Q^π in $\mathcal{F}$ reduces to our realizability condition (Assumption 1). However, pushforward concentrability is in general distinct from stability or invertibility. That is, for problems that are linearly realizable, pushforward concentrability does not imply, nor is implied by, the assumption that $\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) > 0$. Therefore, there exist settings where linear estimators may fail, yet BVFT can succeed and vice versa.

To see this, we consider a variation of the MDP defined just above. The dynamics are identical, but we alter the reward function and the feature mapping. In particular, here we choose the feature map $\phi(s_0) = 1$ and $\phi(s_1) = 2/\gamma$. If we set the rewards to have nonzero variance and satisfy $\mathbb{E}r(s_1) = r^\star$, $\mathbb{E}r(s_0) = \frac{-\gamma}{2(1-\gamma)}r^\star$, then this MDP is linearly realizable with $\theta^\star_\gamma = \frac{\gamma}{2(1-\gamma)}r^\star$. For any $\gamma \in (0,1)$, a simple continuity argument proves that there always exists a $p \in (0,1)$ such that if the offline distribution places mass p on s_0 and $1-p$ on s_1, Σ_{cov} is full rank and $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2} = 1$. Therefore, realizability and pushforward concentrability both hold, but invertibility does not. For the converse direction, it is not hard to see how one might construct examples where linear realizability and invertibility both hold, but Assumption 8 does not. The first condition asserting that $\mathcal{D}$ be supported on all states and actions is particularly stringent.[17]

[15] We can check that invertibility holds if the distribution $\mathcal{D}$ places nonzero mass on the second state s_1.

[16] We omit the last assumption on the initial state distribution from Xie and Jiang [2021] as it is not essential for the purposes of our discussion.

[17] In this construction, we have departed from our assumption that $\sup_{s,a}|r(s,a)| < 1$ since $\mathbb{E}r(s,a)$ is on the order of $\Omega((1-\gamma)^{-1})$. However, the magnitude of the rewards should not affect the *identifiability* of Q^π, only the estimation rate for quantities like ε_{op} and ε_r.

Recall from the construction in Theorem 3, that for any OPE instance where invertibility fails, the alternative $\overline{\mathcal{M}}$ has exactly the same states and transitions. Therefore, any estimator that outperforms linear methods must necessarily consider nonlinear or higher-order interactions between features and rewards. Interestingly enough, a simple tabular method, which ignores the feature mapping ϕ and directly estimates the rewards, successfully approximates the value function in this example.

5 Offline Policy Evaluation without Realizability

Throughout our presentation thus far, our main focus has been on understanding exactly when and why various popular estimators succeed at offline policy evaluation, under the assumption that the action value function *exactly* satisfies the linear realizability condition. Of course, in practice, we might not expect linear realizability to hold exactly, but rather only approximately.

As a sanity check, we therefore investigate how the performance of FQI and LSTD degrade if the relevant function approximation guarantees are weakened. Relative to previous results in this paper, the results in this section are more exploratory and speculative. We leave the problem of generating a more complete understanding of OPE under misspecification to future work. For simplicity, here we analyze the behavior of these estimators under an ℓ_∞ guarantee on the error of the feature mapping ϕ.

Definition 5.1 (Approximate Realizability)**.** We define $\theta^\star_\infty$ as the vector that minimizes the worst-case error with respect to Q^π. Formally, $\theta^\star_\infty$ is the solution to the following optimization problem, where $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$:

$$\theta^\star_\infty \in \arg\min_{\theta\in\mathbb{R}^d} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q^\pi(s,a) - \phi(s,a)^\top\theta| \tag{5.1}$$

We define the approximation error of $\theta^\star_\infty$ as, $\varepsilon_\infty := \min_{\theta\in\mathbb{R}^d} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q^\pi(s,a) - \phi(s,a)^\top\theta|$.

Since the rewards are always bounded, ε_∞ is trivially always bounded by $1/(1-\gamma)$. On the other hand, if $\varepsilon_\infty = 0$, Assumption 1 holds, and we recover the linear realizability setting that has been the main focus of this paper. Values of ε_∞ interpolating between these two extremes measure the extent to which the value function Q^π can be expressed as a linear function of the features ϕ, in a worst case sense.

Using this definition, we prove the following proposition which, together with Theorems 1 and 2, bounds the error of FQI and LSTD under misspecification.

Proposition 5.1. *Assume that invertibility (Assumption 3) holds and let $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top\widehat{\theta}$ be an estimator satisfying,*

$$\|\Sigma_{\text{cov}}^{1/2}(\theta^\star_{\text{fp}} - \widehat{\theta})\|_2 \;\le\; \varepsilon_{\text{fp}} \text{ for } \theta^\star_{\text{fp}} := (I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\Sigma_{\text{cov}}^{1/2}\theta_{\phi,r}.$$

Then, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| \lesssim \|\Sigma_{\text{cov}}^{-1/2}\phi(s,a)\|_2(\varepsilon_{\text{fp}} + \frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})}\rho_s\varepsilon_\infty) + \varepsilon_\infty. \tag{5.2}$$

The main message of this proposition, is that if linear realizability fails, but invertibility still holds, then the performance of LSTD and other linear estimators degrades gracefully with the level of misspecification.

To help parse the result, we can walk through each of the terms appearing on the right hand side of Eq. (5.2). The first source of error, captured in ε_{fp}, is statistical in nature. It arises from bounding the statistical error inherent in estimating Q^π by approximating the fixed point solution to the (projected) Bellman equation, $\theta^\star_{\text{fp}}$. Note that controlling this term is the precisely the main focus on the previous results upper bounding the error of LSTD and FQI.

Because $\theta^\star_\gamma$, as defined in Eq. (1.3), equals $\theta^\star_{\text{fp}}$, if invertibility holds, then for large enough n, Theorem 2 proves that LSTD return a vector such that, with probability $1-\delta$,

$$\varepsilon_{\text{fp}} \lesssim \frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})} \cdot \varepsilon_r + \frac{1}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^2}\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \cdot \varepsilon_{\text{op}}$$

Likewise, Theorem 1 shows that if stability holds, then for large enough n, performing T-steps of FQI return a solution $\widehat{\theta}_T$ such that with probability $1-\delta$,

$$\varepsilon_{\text{fp}} \lesssim \mathsf{cond}(P_\gamma)^{1/2}\|P_\gamma\|_{\text{op}} \cdot \varepsilon_r + \mathsf{cond}(P_\gamma)\|P_\gamma\|_{\text{op}}^2 \cdot \|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \cdot \varepsilon_{\text{op}} + \mathcal{O}(\exp(-T)).$$

As we might expect, this statistical error, ε_{fp} becomes vanishingly small as the number of samples goes to infinity, regardless of whether Q^π is linearly realizable.

The second set of terms in Proposition 5.1, depending on ε_∞, come from the fact that Q^π cannot be expressed as a linear function of the features ϕ. Consequently, this term does *not* go to zero as the number of samples becomes large. This approximation error is amplified by a factor of $\rho_s/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$. Since this is only an upper bound, we cannot assert that these multiplicative factors are necessary. However, the dependence on the statistical leverage ρ_s is reminiscent of previous upper bounds from the linear bandits literature [Lattimore et al., 2020] where the approximation error is also amplified by a factor of $\sqrt{d}$.[18] Du et al. [2019] and Van Roy and Dong [2019] provide similar lower bounds under approximate misspecification of the relevant feature mappings.

In any case, beyond the specific scaling on the various error sources, the main take away message from this result is that FQI and LSTD are reasonable estimators to use beyond the linear realizability setting. Under the necessary assumption that invertibility (or stability) hold, the extent to which these methods estimate the underlying value functions is only mildy affected by the approximation error ε_∞. As alluded to previously, the results in this section are not the focus of our work. We primarily view them as a first step towards a more complete understanding of offline policy evaluation in the absence of realizability.

6 Discussion

In this work, we characterize the exact limits of linear estimators for offline policy evaluation, under the assumption that the value function is linearly realizable in some known set of features. Our stability and invertibility based analyses introduce new, sharper notions of complexity for this classical setting and provide a simple, unifying perspective which brings together previously disparate analysis of popular algorithms.

Two extensions to our results pertain to the finite horizon setting and to policy optimization. As a starting point, we have focused on the infinite horizon, discounted setting as the conditions there are cleaner than in the finite horizon case. Nevertheless, we conjecture that Lyapunov stability and invertibility can be used to analyze finite horizon problems as well. Regarding policy optimization, understanding when this task is possible under linear realizability is an important direction for future work. We hope that our characterization of linear estimators for policy evaluation provides a useful perspective on this closely related problem.

Apart from these extensions, it would be valuable to study quantitative, instance-dependent lower bounds on the sample complexity necessary for offline policy evaluation under linear realizability. In particular, our characterization of linear estimators is sharp in the sense that we precisely determine when the value function of a policy is *identifiable* (alternatively, *learnable*) using classical methods. Having established that a problem is learnable, it is interesting to understand whether the estimation rates for the various algorithms are sharp in a worst case or instance dependent sense.

7 Acknowledgments

We gratefully acknowledge the support of Microsoft through the BAIR Open Research Commons. JCP was in part supported by an NSF Graduate Research Fellowship. Sham Kakade acknowledges funding from the National Science Foundation under award #CCF-1703574 and the Office of Naval Research under award N00014-21-1-2822. We would also like to thank the anonymous reviewers whose insightful comments greatly improved the resulting manuscript.

[18] The statistical leverage ρ_s is exactly $\sqrt{d}$ in the best case.

References

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the Wang-Foster-Kakade lower bound for the discounted setting. *arXiv:2011.01075*, 2020.

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

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning*, 1995.

Richard Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 1961.

Richard Bellman and Stuart Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 1959.

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: An overview. In *IEEE Conference on Decision and Control*, 1995.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, 2018.

Justin A Boyan. Least-squares temporal difference learning. In *International Conference on Machine Learning*, 1999.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 1996.

Frank M Callier and Charles A Desoer. *Linear system theory*. Springer Science & Business Media, 2012.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Symposium on Theory of Computing*, 2020.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Yaqi Duan, Zeyu Jia, 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. *arXiv:2109.12002*, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 2005.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value functionaapproximation. *arXiv:2111.10919*, 2021.

Geoffrey J Gordon. *Approximate solutions to Markov decision processes*. PhD thesis, Carnegie Mellon University, 1999.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory*, 2012.

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

Zico Kolter. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 2012.

Tianjiao Li, Guanghui Lan, and Ashwin Pananjady. Accelerated and instance-optimal policy evaluation with linear function approximation. *arxiv:2112.13109*, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 2018.

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

Kohei Miyaguchi. Asymptotically exact error characterization of offline policy evaluation with misspecified linear models. *Advances in Neural Information Processing Systems*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersin, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassibis. Human-level control through deep reinforcement learning. *Nature*, 2015.

Wenlong Mou, Ashwin Pananjady, and Martin J. Wainwright. Optimal oracle inequalities for solving projected fixed-point equations. *arXiv:2012.05299*, 2020.

Wenlong Mou, Ashwin Pananjady, Martin J Wainwright, and Peter L Bartlett. Optimal and instance-dependent guarantees for markovian linear stochastic approximation. *arXiv:2112.12770*, 2021.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 2007.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 2020.

A Nedić and Dimitri P Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 2003.

Juan Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in Neural Information Processing Systems*, 2021.

Bernardo Ávila Pires and Csaba Szepesvari. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *International Conference on Machine Learning*, 2012.

Dieter Reetz. Approximate solutions of a discounted markovian decision process. *Bonner Mathematische Schriften*, 1977.

Martin Riedmiller. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, 2005.

Ehsan Saleh and Nan Jiang. Deterministic bellman residual minimization. In *Optimization Foundations for Reinforcement Learning Workshop, Neural Information Processing Systems*, 2019.

Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of Mathematical Analysis and Applications*, 1985.

Gilbert W Stewart. *Matrix perturbation theory*. Citeseer, 1990.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*, 2005.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2012.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 1996.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

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

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Ruosong Wang, Dean Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021a.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham M. Kakade. Instabilities of offline RL with pre-trained neural representation. In *International Conference on Machine Learning*, 2021b.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable MDPs with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 2021c.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, 2021a.

Gellért Weisz, Csaba Szepesvári, and András György. Tensorplan and the few actions lower bound for planning in MDPs under linear realizability of optimal value functions. *arXiv:2110.02195*, 2021b.

Ward Whitt. Approximations of dynamic programs, I. *Mathematics of Operations Research*, 1978.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2020.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Huizhen Yu. Convergence of least squares temporal difference methods under general conditions. In *International Conference on Machine Learning*, 2010.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv:2202.04634*, 2022.

A Supporting Arguments for Section 3: FQI

A.1 Proof of Theorem 1: stability is sufficient for FQI

The existence of $\widehat{\Sigma}_{\text{cov}}$ and the upper bounds on the regression errors ε_r and ε_{op} are guaranteed by Lemmas C.3 and C.4. To analyze the error of FQI, we introduce the shorthand,

$$A := \gamma \Sigma_{\text{cov}}^{-1} \Sigma_{\text{cr}}, \quad \widehat{A} = \gamma \widehat{\Sigma}_{\text{cov}}^{-1} \widehat{\Sigma}_{\text{cr}}, \quad \theta_t^\star := \sum_{k=0}^{t} A^k \theta_0^\star, \quad \widehat{\theta}_t := \sum_{k=0}^{t} \widehat{A}^k \widehat{\theta}_0,$$

$$w_t := \widehat{\theta}_t - \theta_t^\star, \quad \Delta := \widehat{A} - A,$$

where $\widehat{\theta}_0 = \widehat{\Sigma}_{\text{cov}}^{-1} \widehat{\theta}_{\phi,r}$ and $\theta_0^\star = \Sigma_{\text{cov}}^{-1} \theta_{\phi,r}$. Using this notation, by stability, we observe that $\theta_\gamma^\star = \theta_\infty^\star$, and we can write the errors vectors of the t-step FQI solution as,

$$\Sigma_{\text{cov}}^{1/2}(\theta_\gamma^\star - \widehat{\theta}_t) = \Sigma_{\text{cov}}^{1/2} \sum_{k=t+1}^{\infty} A^k \theta_0^\star + \Sigma_{\text{cov}}^{1/2} w_t. \tag{A.1}$$

Next, we develop the recursion in w_t,

$$\begin{aligned} w_{t+1} &= \sum_{j=0}^{t+1} \widehat{A}^j \widehat{\theta}_0 - \sum_{j=0}^{t+1} A^j \theta_0^\star \\ &= \widehat{A}\widehat{\theta}_t + \widehat{\theta}_0 - A\theta_t^\star - \theta_0^\star \\ &= \widehat{A} w_t + \Delta \theta_t^\star + w_0. \end{aligned}$$

Unrolling the recursion and multiplying on the left by $\Sigma_{\text{cov}}^{1/2}$, we get that

$$\Sigma_{\text{cov}}^{1/2} w_{t+1} = \sum_{j=0}^{t+1} \left(\Sigma_{\text{cov}}^{1/2} \widehat{A} \Sigma_{\text{cov}}^{-1/2}\right)^j \Sigma_{\text{cov}}^{1/2} w_0 + \sum_{j=0}^{t} \left(\Sigma_{\text{cov}}^{1/2} \widehat{A} \Sigma_{\text{cov}}^{-1/2}\right)^j \left(\Sigma_{\text{cov}}^{1/2} \Delta \Sigma_{\text{cov}}^{-1/2}\right) \Sigma_{\text{cov}}^{1/2} \theta_{t-j}^\star.$$

Note that $\varepsilon_r = \|\Sigma_{\text{cov}}^{1/2} w_0\|_2$ and $\varepsilon_{\text{op}} = \|\Sigma_{\text{cov}}^{1/2} \Delta \Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}$. Therefore, taking the norm of both sides and applying the triangle inequality,

$$\begin{aligned} \|\Sigma_{\text{cov}}^{1/2} w_{t+1}\|_2 \;\le\; & \sum_{k=0}^{t+1} \| \left(\Sigma_{\text{cov}}^{1/2} \widehat{A} \Sigma_{\text{cov}}^{-1/2}\right)^k \|_{\text{op}} \|\Sigma_{\text{cov}}^{1/2} w_0\|_2 \qquad (A.2) \\ & + \sum_{k=0}^{t} \| \left(\Sigma_{\text{cov}}^{1/2} \widehat{A} \Sigma_{\text{cov}}^{-1/2}\right)^k \|_{\text{op}} \|\Sigma_{\text{cov}}^{1/2} \Delta \Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \sup_{0 \le h \le t} \|\Sigma_{\text{cov}}^{1/2} \theta_h^\star\|_2 \\ = \; & (\varepsilon_r + \varepsilon_{\text{op}} \sup_{0 \le h \le t} \|\Sigma_{\text{cov}}^{1/2} \theta_h^\star\|_2) \cdot \sum_{k=0}^{t+1} \| \left(\Sigma_{\text{cov}}^{1/2} \widehat{A} \Sigma_{\text{cov}}^{-1/2}\right)^k \|_{\text{op}}. \qquad (A.3) \end{aligned}$$

Now, recalling the definition of $\theta_h^\star$, we bound:

$$\sup_{0 \le h \le t} \|\Sigma_{\text{cov}}^{1/2} \theta_h^\star\|_2 \;\le\; \sum_{j=0}^{t} \| \left(\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2}\right)^j \|_{\text{op}} \|\Sigma_{\text{cov}}^{1/2} \theta_0^\star\|_2. \tag{A.4}$$

Therefore, combining these last two inequalities (A.4), (A.3), and the identity from Eq. (A.1),

$$\begin{aligned}\|\Sigma_{\text{cov}}^{1/2}(\theta_\gamma^\star - \widehat{\theta}_t)\|_2 &\leq \sum_{k=t+1}^{\infty} \|\left(\Sigma_{\text{cov}}^{1/2} A \Sigma_{\text{cov}}^{-1/2}\right)^k\|_{\text{op}} \|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 + \|\Sigma_{\text{cov}}^{1/2} w_t\|_2 \\ &\leq \|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 \sum_{k=t+1}^{\infty} \alpha_k + \left(\varepsilon_r + \varepsilon_{\text{op}}\|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 \sum_{k=0}^{t-1}\alpha_k\right) \sum_{k=0}^{t} \widehat{\alpha}_k, \end{aligned} \tag{A.5}$$

where $\widehat{\alpha}_k := \|\left(\Sigma_{\text{cov}}^{1/2}\widehat{A}\Sigma_{\text{cov}}^{-1/2}\right)^k\|_{\text{op}}$ and $\alpha_k := \|\left(\Sigma_{\text{cov}}^{1/2} A \Sigma_{\text{cov}}^{-1/2}\right)^k\|_{\text{op}}$. Since $\varepsilon_{\text{op}} \leq 1/(6\|P_\gamma\|_{\text{op}})$ and $\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ is stable, Lemma A.1 tells us that

$$\begin{aligned}\widehat{\alpha}_j &= \|P_\gamma^{-1/2} P_\gamma^{1/2}\left(\Sigma_{\text{cov}}^{1/2}\widehat{A}\Sigma_{\text{cov}}^{-1/2}\right)^j\|_{\text{op}} \\ &\leq \|P_\gamma^{-1/2}\|_{\text{op}}\|P_\gamma^{1/2}\left(\Sigma_{\text{cov}}^{1/2}\widehat{A}\Sigma_{\text{cov}}^{-1/2}\right)^j\|_{\text{op}} \\ &\leq \|P_\gamma^{1/2}\|_{\text{op}}\|P_\gamma^{1/2}\|_{\text{op}}\left(1 - \frac{1}{2\|P_\gamma\|_{\text{op}}}\right)^{j/2} = \mathsf{cond}(P_\gamma)^{1/2}\left(1 - \frac{1}{2\|P_\gamma\|_{\text{op}}}\right)^{j/2}.\end{aligned}$$

Using similar reasoning, we get that

$$\alpha_j \leq \mathsf{cond}(P_\gamma)^{1/2}\left(1 - \frac{1}{\|P_\gamma\|_{\text{op}}}\right)^{j/2}.$$

In conclusion, $\|\Sigma_{\text{cov}}^{1/2}(\theta_\gamma^\star - \widehat{\theta}_t)\|_2$ is bounded by,

$$\|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 \mathsf{cond}(P_\gamma)^{1/2}\left(1 - \frac{1}{\|P_\gamma\|_{\text{op}}}\right)^{(t+1)/2}\sum_{k=0}^{\infty}\alpha_k + \left(\varepsilon_r + \varepsilon_{\text{op}}\|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2\sum_{k=0}^{\infty}\alpha_k\right)\sum_{k=0}^{\infty}\widehat{\alpha}_k.$$

The final bound comes from summing the geometric series, $\sum_{j=0}^{\infty}(1-c)^{j/2} = (1-\sqrt{1-c})^{-1}$, for $c \in (0,1)$ and applying the numerical inequality,

$$\left(1 - \sqrt{1 - \frac{1}{2z}}\right)^{-1} \leq 10z,$$

which holds for all $z \geq 1$.

Lemma A.1. *Let A be a square, stable matrix and let $P = \mathsf{dlyap}(A)$ Then, for all $k \geq 0$,*

$$\|A^k\|_{\text{op}}^2 \leq \mathsf{cond}(P)\left(1 - \frac{1}{\|P\|_{\text{op}}}\right)^k.$$

Furthermore, for any matrix Δ such that $\|\Delta\|_{\text{op}} \leq 1/(6\|P\|_{\text{op}}^2)$,

$$\|(A+\Delta)^k\|_{\text{op}}^2 \leq \mathsf{cond}(P)\left(1 - \frac{1}{2\|P\|_{\text{op}}}\right)^k.$$

Proof. This particular lemma is almost identical to the one from Perdomo et al. [2021]. However, we include the proof for the sake of providing a self-contained presentation. For the first result, by definition of the solution to the Lyapunov equation, for any unit vector x,

$$\begin{aligned}x^\top A^\top P A x &= x^\top P x - x^\top I x \\ &= x^\top P x\left(1 - \frac{\|x\|_2^2}{x^\top P x}\right) \\ &\leq x^\top P x\left(1 - \frac{1}{\|P\|_{\text{op}}}\right).\end{aligned}$$

Hence, $A^\top PA \preceq P(1 - \|P\|_{\text{op}}^{-1})$. By iterating $(A^k)^\top PA^k \preceq P(1 - \|P\|_{\text{op}}^{-1})^k$ and

$$\|P^{1/2}A^k\|_{\text{op}}^2 \;\leq\; \|P\|_{\text{op}}(1 - \|P\|_{\text{op}}^{-1})^k.$$

Therefore,

$$\|A^k\|_{\text{op}} = \|P^{-1/2}P^{1/2}A^k\|_{\text{op}} \;\leq\; \|P^{-1/2}\|_{\text{op}}\|P^{1/2}A^k\|_{\text{op}} \;\leq\; \mathsf{cond}(P)^{1/2}\left(1 - \frac{1}{\|P\|_{\text{op}}}\right)^{k/2}.$$

For the second result, using the insights from above,

$$(A+\Delta)^\top P(A+\Delta) = A^\top PA + A^\top P\Delta + \Delta^\top PA + \Delta^\top P\Delta.$$

Now, $A^\top PA \preceq P(1 - \|P\|_{\text{op}}^{-1})$ and

$$\|A^\top P\Delta\|_{\text{op}} = \|\Delta^\top PA\|_{\text{op}} \;\leq\; \|\Delta P^{1/2}\|_{\text{op}}\|P^{1/2}A\|_{\text{op}} \;\leq\; \|\Delta P^{1/2}\|_{\text{op}}\|P^{1/2}\|_{\text{op}} \;\leq\; \|\Delta\|_{\text{op}}\|P\|_{\text{op}}.$$

Bounding, $\|\Delta^\top P\Delta\|_{\text{op}} \;\leq\; \|P\|_{\text{op}}\|\Delta\|_{\text{op}}^2$, and using the fact that $P \succeq I$ we get that for,

$$\|\Delta\|_{\text{op}} \;\leq\; 1/(6\|P\|_{\text{op}}^2),$$

the following relationship holds:

$$A^\top P\Delta + \Delta^\top PA + \Delta^\top P\Delta \preceq P\frac{1}{2\|P\|_{\text{op}}}.$$

Therefore,

$$(A+\Delta)^\top P(A+\Delta) \preceq P\left(1 - \frac{1}{2\|P\|_{\text{op}}}\right),$$

and the second result follows by using the same steps as the first. □

A.2 Proof of Proposition 3.2: coordinate invariance of P_γ

If we define the whitened features, $\phi_w(\cdot) = \Sigma_{\text{cov}}^{-1/2}\phi(\cdot)$, then $\widetilde{\phi}(\cdot) = L'\phi_w(\cdot)$ where $L' = L\Sigma_{\text{cov}}^{1/2}$. Now, let $USV^\top$ be the singular value decomposition of L'. Then,

$$\widetilde{\Sigma}_{\text{cov}} = \mathbb{E}_{x\sim\mathcal{D}}\widetilde{\phi}(x)\widetilde{\phi}(x)^\top = L'\mathbb{E}_{x\sim\mathcal{D}}\phi_w(x)\phi_w(x)^\top L'^\top = L'L'^\top = US^2U^\top,$$

where we have used the fact that the whitened features have identity covariance. By this calculation, we have that $\widetilde{\Sigma}_{\text{cov}}^{1/2} = USU^\top$. Using similar substitutions, we can also deduce that $\widetilde{\Sigma}_{\text{cr}} = L'\Sigma_{\text{cr}}^{(w)}L'^\top$ where $\Sigma_{\text{cr}}^{(w)} = \Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$. Therefore,

$$\widetilde{\Sigma}_{\text{cov}}^{-1/2}\widetilde{\Sigma}_{\text{cr}}\widetilde{\Sigma}_{\text{cov}}^{-1/2} = (US^{-1}U^\top)(USV^\top)\Sigma_{\text{cr}}^{(w)}(VSU^\top)(US^{-1}U^\top) = (UV^\top)\Sigma_{\text{cr}}^{(w)}(UV^\top)^\top.$$

Since $(UV^\top)$ is an orthogonal matrix, the equality of condition numbers follows by the fact that for any matrix A and orthogonal matrix M, $MAM^\top = A$ have the same singular values. On the other hand, the invariance of the operator norm of P_γ follows from the following lemma:

Lemma A.2. *Let A be a stable matrix and M be any orthogonal matrix, then*

$$\|\mathsf{dlyap}(A^\top)\|_{\text{op}} = \|\mathsf{dlyap}(MA^\top M^\top)\|_{\text{op}}.$$

Proof. Let $P = \mathsf{dlyap}(A)$ be the unique solution over X to the matrix equation:

$$X = A^\top XA + I.$$

Likewise, let $P' = \mathsf{dlyap}(MAM^\top)$ be the unique solution (over X') to the equation:

$$X' = MA^\top M^\top X'MAM^\top + I.$$

From this, we can deduce that $M^\top X'M = A^\top M^\top X'MA + I$. Therefore, $P = M^\top X'M = M^\top P'M$. The conclusion follows from the fact that singular values are invariant to conjugation by an orthogonal matrix. □

A.3 Proof of Lemma 3.3: FQI under specific growth rates

As discussed in the main body, the proof is identical to that of Theorem 1 except that we specialize to the particular assumptions on the growth of matrix powers. We recall the key inequality from the proof of the main theorem, Eq. (A.5):

$$\|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_t - \theta_\gamma^\star)\|_2 \;\le\; \|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 \sum_{k=t+1}^{\infty} \alpha_k + \left(\varepsilon_r + \varepsilon_{\text{op}}\|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2 \sum_{k=0}^{t-1}\alpha_k\right) \sum_{k=0}^{t} \widehat{\alpha}_k.$$

Here, $\alpha_k = \| \left(\gamma \Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\right)^k \|_{\text{op}}$ and $\widehat{\alpha}_k := \| \left(\Sigma_{\text{cov}}^{1/2}(\gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})\Sigma_{\text{cov}}^{-1/2}\right)^k \|_{\text{op}}$. By assumption, $\alpha_k \le \alpha\beta^k$ hence, $\sum_{k=0}^{\infty}\alpha_k \le \alpha/(1-\beta)$. Now, by Lemma A.3 since

$$\varepsilon_{\text{op}} = \|\Sigma_{\text{cov}}^{1/2}(\gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})\Sigma_{\text{cov}}^{-1/2} - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}},$$

we have that:

$$\widehat{\alpha}_k \;\le\; \alpha(\beta + \varepsilon_{\text{op}}\alpha)^k.$$

Therefore, as long as $\varepsilon_{\text{op}} < \frac{9}{10}\frac{(1-\beta)}{\alpha}$,

$$\sum_{k=0}^{\infty}\widehat{\alpha}_k \;\le\; \alpha\sum_{k=0}^{\infty}(\beta+\varepsilon_{\text{op}}\alpha)^k = \alpha\frac{1}{1-\beta-\alpha\varepsilon_{\text{op}}} \;\le\; 10\frac{\alpha}{1-\beta}.$$

Putting everything together,

$$\|\Sigma_{\text{cov}}^{1/2}(\widehat{\theta}_T - \theta_\gamma^\star)\|_2 \lesssim \|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2\frac{\alpha}{1-\beta}\cdot\beta^{T+1} + \left(\varepsilon_r + \varepsilon_{\text{op}}\|\Sigma_{\text{cov}}^{1/2}\theta_0^\star\|_2\frac{\alpha}{1-\beta}\right)\frac{\alpha}{1-\beta}.$$

Lemma A.3. *Let A be a square matrix such that for all nonnegative integers j, $\|A^j\|_{\text{op}} \le a\cdot b^j$ for scalars $a > 0$ and $b \in (0,1)$. Then, for any square matrix Δ if we let $\varepsilon := \|\Delta\|_{\text{op}}$ then,*

$$\|(A+\Delta)^n\|_{\text{op}} \;\le\; a(b+\varepsilon\cdot a)^n.$$

Proof. We begin by expanding $(A+\Delta)^n$ into monomials $T_{k,j}$,

$$(A+\Delta)^n = \sum_{k=0}^{n}\sum_{j=1}^{\binom{n}{k}} T_{k,j}, \tag{A.6}$$

where each $T_{k,j}$ has k factors of Δ and $n-k$, A factors. Now, by the submultiplicative property of the operator norm,

$$\|T_{k,j}\|_{\text{op}} \;\le\; \varepsilon^k \prod_{s_i\in S_{k,j}} \|A^{s_i}\|_{\text{op}},$$

where $S_{k,j}$ is a set of positive integers s_i satisfying $\sum_i s_i = n-k$ and $|S| \le k+1$. Using our assumption on the growth of $\|A^k\|_{\text{op}}$, we get that,

$$\|T_{k,j}\|_{\text{op}} \;\le\; \varepsilon^k \prod_{s_i\in S_{k,j}} (a\cdot b^{s_i}) \;\le\; a^{k+1}\varepsilon^k b^{n-k}.$$

Going back to the original expansion into monomials, and using the identity,

$$\sum_{k=0}^{n}\binom{n}{k}x^k = (1+x)^n.$$

We conclude:

$$\|(A+\Delta)^n\|_{\text{op}} \;\le\; a\cdot b^n\sum_{k=0}^{n}\binom{n}{k}\left(\frac{a\varepsilon}{b}\right)^k = ab^n(1+\frac{a\cdot\varepsilon}{b})^n = a(b+a\varepsilon)^n.$$

□

A.4 Proof of Corollary 3.1: low distribution shift implies stability

Consider the augmented covariance matrix,

$$\mathbb{E}\begin{bmatrix}\phi(s,a)\\ \phi(s',a')\end{bmatrix}\begin{bmatrix}\phi(s,a)\\ \phi(s',a')\end{bmatrix}^\top = \begin{bmatrix}\Sigma_{\text{cov}} & \Sigma_{\text{cr}}\\ \Sigma_{\text{cr}}^\top & \Sigma_{\text{next}}\end{bmatrix} \succeq 0.$$

By a Schur complement argument, $\Sigma_{\text{cr}}^\top \Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}} \preceq \Sigma_{\text{next}}$. After conjugating by $\Sigma_{\text{cov}}^{-1/2}$ and multiplying by γ^2, we get that:

$$(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^\top(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) \preceq \gamma^2\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{next}}\Sigma_{\text{cov}}^{-1/2}.$$

Now, by the low distribution shift assumption, $\gamma^2\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{next}}\Sigma_{\text{cov}}^{-1/2} \preceq \gamma^2\Sigma_{\text{cov}}^{-1/2}(\mathcal{C}_{\text{ds}}\Sigma_{\text{cov}})\Sigma_{\text{cov}}^{-1/2} = \mathcal{C}_{\text{ds}}\gamma^2 I$. Therefore, $(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^\top(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) \preceq \mathcal{C}_{\text{ds}}\gamma^2 I$. Iterating for $j \geq 0$ gives the first part of the result. The rest follows from Lemma 3.3 by observing that Eq. (3.7) holds with $\alpha = 1, \beta = \sqrt{\mathcal{C}_{\text{ds}}\gamma^2} \in (0,1)$.

A.5 Proofs of Corollary 3.2: Bellman completeness implies stability

To take advantage of matrix notation, for this result we assume that the state-action space is finite, $|\mathcal{S}||\mathcal{A}| < \infty$. In particular, we introduce the following quantities.

1. Feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$.
2. Offline distribution vector $\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

With this, we have that $\Sigma_{\text{cov}} = \Phi^\top \text{diag}(\mu)\Phi$ and $\Sigma_{\text{cr}} = \Phi^\top \text{diag}(\mu)P^{(\pi)}\Phi$ where $P^{(\pi)}$ is a row stochastic matrix representing the transition operator. Corollary 3.2 follows from the following lemma and Lemma 3.3.

Lemma A.4. *If ϕ is complete (Assumption 5) and Σ_{cov} is full rank, then for $j \geq 0$,*

$$\|(\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^j\|_{\text{op}} \leq \rho_s.$$

Proof. First, we rewrite the relevant matrix as follows,

$$\begin{aligned}(\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^j &= \Sigma_{\text{cov}}^{1/2}(\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})^j\Sigma_{\text{cov}}^{-1/2}\\ &= \Sigma_{\text{cov}}^{-1/2}\Phi^\top\text{diag}(\mu)\Phi^\top(\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})^j\Sigma_{\text{cov}}^{-1/2}.\end{aligned}$$

Therefore,

$$\|(\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^j\|_{\text{op}} \leq \underbrace{\|\Sigma_{\text{cov}}^{-1/2}\Phi^\top\text{diag}(\mu)^{1/2}\|_{\text{op}}}_{:=T_1}\underbrace{\|\text{diag}(\mu)^{1/2}\Phi(\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})^j\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}}_{:=T_2}.$$

To bound T_1, we observe that

$$\|\Sigma_{\text{cov}}^{-1/2}\Phi^\top\text{diag}(\mu)^{1/2}\|_{\text{op}}^2 = \|(\Phi^\top\text{diag}(\mu)\Phi)^{-1/2}\Phi^\top\text{diag}(\mu)^{1/2}\|_{\text{op}}.$$

Letting $A := \text{diag}(\mu)^{1/2}\Phi$, the above expression satisfies,

$$\|(A^\top A)^{-1/2}A^\top\|_{\text{op}}^2 = \sup_{\|v\|_2=1} v^\top A(A^\top A)^{-1}A^\top v \leq 1,$$

since $A(A^\top A)^{-1}A^\top$ is a projection matrix. Moving onto T_2, we recall that

$$\|\text{diag}(\mu)^{1/2}\Phi(\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})^j\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} = \sup_{\|v\|_2=1}\|\text{diag}(\mu)^{1/2}\Phi(\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})^j\Sigma_{\text{cov}}^{-1/2}v\|_2.$$

For any fixed vector v, since the entries of μ form a probability measure,

$$\|\mathrm{diag}(\mu)^{1/2}v\|_2 = \sqrt{\sum_{i=1}^{d}\mu_i v_i^2} \;\le\; \max_i v_i = \|v\|_\infty.$$

Therefore,

$$\|\mathrm{diag}(\mu)^{1/2}\Phi(\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^j\Sigma_{\mathrm{cov}}^{-1/2}\|_{\mathrm{op}} \;\le\; \sup_{\|v\|_2=1}\|\Phi(\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^j\Sigma_{\mathrm{cov}}^{-1/2}v\|_\infty.$$

Then, by repeatedly applying Lemma A.5, we get that

$$\|\Phi(\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^j\Sigma_{\mathrm{cov}}^{-1/2}v\|_\infty \;\le\; \|\Phi\Sigma_{\mathrm{cov}}^{-1/2}v\|_\infty.$$

Lastly,

$$\|\Phi\Sigma_{\mathrm{cov}}^{-1/2}v\|_\infty = \sup_{(s,a)}|\phi(s,a)^\top\Sigma_{\mathrm{cov}}^{-1/2}v| \;\le\; \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\Sigma_{\mathrm{cov}}^{-1/2}\phi(s,a)\|_2 = \rho_s.$$

□

Lemma A.5. *If ϕ is complete (Assumption 5) and Σ_{cov} is full rank, then for all θ,*

$$\|\Phi\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}}\theta\|_\infty \;\le\; \|\Phi\theta\|_\infty.$$

Proof. If we denote the vector of expected rewards by $\vec{r}\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, then completeness implies that for all θ, there exists a θ' such that

$$\Phi\theta' = \vec{r} + \gamma P^{(\pi)}\Phi\theta.$$

Choosing $\theta = 0$, this means that there exists a vector θ_r such that $\vec{r} = \Phi\theta_r$. Consequently, we deduce that for all θ, there always exists a θ' such that $\Phi\theta' = \gamma P^{(\pi)}\Phi\theta$. Using this realizability condition, for a given distribution μ, θ' must satisfy

$$\begin{aligned}\theta' &= \arg\min_{\bar\theta}\mathbb{E}_{(s,a)\sim\mu,s'\sim P(\cdot|s,a)}\left(\phi(s,a)^\top\bar\theta - \gamma\phi(s',a')^\top\theta\right)^2\\ &= \gamma\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}}\theta.\end{aligned}$$

Together with the previous equation, this implies that for all θ, $\gamma\Phi\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}}\theta = \gamma P^{(\pi)}\Phi\theta$. Thus, we conclude that

$$\begin{aligned}\|\Phi\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}}\theta\|_\infty &= \|P^{(\pi)}\Phi\theta\|_\infty\\ &\le \|\Phi\theta\|_\infty,\end{aligned}$$

where we have used the fact that $P^{(\pi)}$ is row stochastic so $\|P^{(\pi)}\|_1 \le 1$. □

Lemma A.6. *Assume that the rewards are linearly realizable in the feature mapping ϕ. That is, there exists a vector $\theta_r^\star\in\mathbb{R}^d$ such that for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $\mathbb{E}r(s,a) = \phi(s,a)^\top\theta_r^\star$. Then, $\|\Sigma_{\mathrm{cov}}^{-1/2}\theta_{\phi,r}\|_2 \le 1$.*

Otherwise, if reward realizability does not hold $\|\Sigma_{\mathrm{cov}}^{-1/2}\theta_{\phi,r}\|_2 \le \sqrt{d}$.

Proof. Expanding out the definition of $\theta_{\phi,r}$,

$$\|\Sigma_{\mathrm{cov}}^{-1/2}\theta_{\phi,r}\|_2^2 = \mathrm{tr}\left[\Sigma_{\mathrm{cov}}^{-1/2}\mathbb{E}[\phi(s,a)r(s,a)]\mathbb{E}\phi(s,a)^\top r(s,a)\Sigma_{\mathrm{cov}}^{-1/2}\right]$$

Under realizability, $\mathbb{E}[\phi(s,a)r(s,a)] = \mathbb{E}\phi(s,a)\phi(s,a)^\top\theta_r^\star$. Hence, the expression above can be rewritten as,

$$\mathrm{tr}\left[\Sigma_{\mathrm{cov}}^{-1}\mathbb{E}[\phi(s,a)\phi(s,a)^\top]\theta_r^\star\theta_r^{\star\top}\mathbb{E}\phi(s,a)\phi(s,a)^\top\right] = \mathbb{E}(\phi(s,a)^\top\theta_r^\star)^2 = \mathbb{E}r(s,a)^2 \;\leq\; 1.$$

If the rewards are not linearly realizable in ϕ, then by Jensen's inequality,

$$\begin{aligned}
\|\Sigma_{\mathrm{cov}}^{-1/2}\mathbb{E}\phi(s,a)r(s,a)\|_2^2 &\leq \mathbb{E}\|\Sigma_{\mathrm{cov}}^{-1/2}\phi(s,a)r(s,a)\|_2^2 \\
&= \mathbb{E}\mathrm{tr}\left[\Sigma_{\mathrm{cov}}^{-1/2}\mathbb{E}\phi(s,a)\phi(s,a)^\top r(s,a)^2\Sigma_{\mathrm{cov}}^{-1/2}\right] \\
&\leq \sup_{s,a} r(s,a)^2\mathrm{tr}\left[I\right] \\
&\leq d.
\end{aligned}$$

□

A.6 Proof of Proposition 3.4: FQI lower bound

Recall the functional form of the FQI approximation,

$$\widehat{\theta}_T = \sum_{k=0}^{T}(\gamma\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^k\Sigma_{\mathrm{cov}}^{-1}(\theta_{\phi,r}+z) = \mu + v,$$

where $\mathbb{E}\widehat{\theta}_T = \mu := \sum_{k=0}^{T}(\gamma\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^{-1}\Sigma_{\mathrm{cov}}^{-1}\theta_{\phi,r}$ and $v := \sum_{k=0}^{T}(\gamma\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}})^{-1}\Sigma_{\mathrm{cov}}^{-1}z$. Expanding out and using $\mathbb{E}v = 0$, we have that

$$\begin{aligned}
\mathbb{E}\|\widehat{\theta}_T - \mathbb{E}\widehat{\theta}_T\|_2^2 &= \mathbb{E}\|\widehat{\theta}_T\|_2^2 - \|\mathbb{E}\widehat{\theta}_T\|_2^2 \\
&= \mathbb{E}\|\mu\|_2^2 + \mathbb{E}\|v\|_2^2 - \|\mathbb{E}\widehat{\theta}_T\|_2^2 \\
&= \mathbb{E}\|v\|_2^2.
\end{aligned}$$

Now, letting $A = \gamma\Sigma_{\mathrm{cov}}^{-1}\Sigma_{\mathrm{cr}}$, we have that

$$\mathbb{E}\|v\|_2^2 = \mathrm{tr}\left[(\sum_{k=0}^{T}A^k)^\top\Lambda(\sum_{k=0}^{T}A^k)\right] \;\geq\; \sigma_{\min}(\Lambda)\|\sum_{k=0}^{T}A^k\|_{\mathrm{op}}^2 = \sigma_{\min}(\Lambda)\sup_{\|v\|_2=1} v^\top(\sum_{k=0}^{T}A^k)^\top(\sum_{k=0}^{T}A^k)v,$$

where we have used $\mathrm{tr}\left[A^\top A\right] = \|A\|_{\mathrm{F}}^2 \;\geq\; \|A\|_{\mathrm{op}}^2$ ($\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix) and the variational characterization of the operator norm for symmetric matrices. By assumption on the spectral radius, A has an eigenvector u with eigenvalue λ such that $|\lambda| > 1$. Therefore,

$$\sup_{\|v\|_2=1} v^\top(\sum_{k=0}^{T}A^k)^\top(\sum_{k=0}^{T}A^k)v \;\geq\; u^\top(\sum_{k=0}^{T}A^k)^\top(\sum_{k=0}^{T}A^k)u = \|u\|_2^2(\sum_{k=0}^{T}\lambda^k)^2 = \left(\frac{\lambda^{T+1}-1}{\lambda-1}\right)^2.$$

Note that if $|\lambda| = 1$, this series can grow linearly in T (e.g if $\lambda = 1$) or oscillate (if $\lambda = -1$). The last equality only holds for $\lambda \neq 1$.

A.7 Extensions to ridge regression

One might wonder whether adding ℓ_2 regularization, that is, an $\lambda\|\theta\|_2^2, \lambda > 0$ additive penalty to the FQI or LSTD objective in Eq. (1.2), could help mitigate the divergence phenomenon outlined in Proposition 3.4 or the limits of linear estimators from Theorem 3.

For finite-dimensional problems with full rank covariance, typical analyses of ridge regression set the regularizer λ to shrink with the number of samples n. In this case, the ridge estimator achieves consistent parameter recovery and asymptotically returns the same solution as just performing ordinary least squares.

Therefore, we can expect similar blowup if stability fails (in fact, this phenomenon is verified empirically by Wang et al. [2021b]). On the other hand, if the parameter λ is lower bounded by a constant, then ridge regression will have constant bias which will then be amplified by the number of rounds T. Hence, adding regularization does not avoid the need for stability when performing fitted Q-iteration. Similar arguments demonstrate why regularization is unlikely to overcomes the limitations of least squares temporal differencing learning (or other linear estimators) in settings where invertibility does not hold.

B Supporting Arguments for Section 4: LSTD

B.1 Proof of Theorem 2: invertibility is sufficient for LSTD

Recall the closed form expression of the empirical LSTD estimator:

$$\widehat{\theta}_{\mathrm{LS}} = (I - \gamma \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\Sigma}_{\mathrm{cr}})^{\dagger} \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\theta}_{\phi,r}.$$

Multiplying on the left by $\Sigma_{\mathrm{cov}}^{1/2}$,

$$\begin{aligned}
\Sigma_{\mathrm{cov}}^{1/2} \widehat{\theta}_{\mathrm{LS}} &= \Sigma_{\mathrm{cov}}^{1/2} (I - \gamma \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\Sigma}_{\mathrm{cr}})^{\dagger} \Sigma_{\mathrm{cov}}^{-1/2} \Sigma_{\mathrm{cov}}^{1/2} \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\theta}_{\phi,r} \\
&= \left(\Sigma_{\mathrm{cov}}^{1/2} (I - \gamma \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\Sigma}_{\mathrm{cr}}) \Sigma_{\mathrm{cov}}^{-1/2} \right)^{\dagger} (\Sigma_{\mathrm{cov}}^{1/2} \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\theta}_{\phi,r}),
\end{aligned}$$

where we have used the identity $(ABA^{-1})^{\dagger} = AB^{\dagger}A^{-1}$ for any invertible A and B. Similarly,

$$\Sigma_{\mathrm{cov}}^{1/2} \theta_{\gamma}^{\star} = \left(\Sigma_{\mathrm{cov}}^{1/2} (I - \gamma \Sigma_{\mathrm{cov}}^{-1} \Sigma_{\mathrm{cr}}) \Sigma_{\mathrm{cov}}^{-1/2} \right)^{-1} (\Sigma_{\mathrm{cov}}^{1/2} \Sigma_{\mathrm{cov}}^{-1} \theta_{\phi,r}).$$

Now defining the following quantities,

$$A := \Sigma_{\mathrm{cov}}^{1/2} (I - \gamma \Sigma_{\mathrm{cov}}^{-1} \Sigma_{\mathrm{cr}}) \Sigma_{\mathrm{cov}}^{-1/2}, \quad \widehat{A} := \Sigma_{\mathrm{cov}}^{1/2} (I - \gamma \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\Sigma}_{\mathrm{cr}}) \Sigma_{\mathrm{cov}}^{-1/2}$$

$$b := \Sigma_{\mathrm{cov}}^{1/2} \Sigma_{\mathrm{cov}}^{-1} \theta_{\phi,r}, \quad \widehat{b} := \Sigma_{\mathrm{cov}}^{1/2} \widehat{\Sigma}_{\mathrm{cov}}^{-1} \widehat{\theta}_{\phi,r}.$$

We can rewrite the above expression as:

$$\Sigma_{\mathrm{cov}}^{1/2} (\theta_{\gamma}^{\star} - \widehat{\theta}_{\gamma}) = (A^{-1} - \widehat{A}^{\dagger}) b + \widehat{A}^{\dagger} (b - \widehat{b}).$$

Therefore,

$$\|\Sigma_{\mathrm{cov}}^{1/2} (\theta_{\gamma}^{\star} - \widehat{\theta}_{\gamma})\|_2 \;\le\; \|A^{-1} - \widehat{A}^{\dagger}\|_{\mathrm{op}} \|b\|_2 + \|\widehat{A}^{\dagger}\|_{\mathrm{op}} \|b - \widehat{b}\|_2.$$

Using Lemma B.1, since $\varepsilon_{\mathrm{op}} \;\le\; \frac{1}{2} \sigma_{\min}(I - \gamma \Sigma_{\mathrm{cov}}^{-1} \Sigma_{\mathrm{cr}})$:

$$\|\Sigma_{\mathrm{cov}}^{1/2} (\theta_{\gamma}^{\star} - \widehat{\theta}_{\gamma})\|_2 \lesssim \frac{\varepsilon_{\mathrm{op}}}{\sigma_{\min}(I - \gamma \Sigma_{\mathrm{cov}}^{-1/2} \Sigma_{\mathrm{cr}} \Sigma_{\mathrm{cov}}^{-1/2})^2} \|\Sigma_{\mathrm{cov}}^{-1/2} \theta_{\phi,r}\|_2 + \frac{\varepsilon_r}{\sigma_{\min}(I - \gamma \Sigma_{\mathrm{cov}}^{-1/2} \Sigma_{\mathrm{cr}} \Sigma_{\mathrm{cov}}^{-1/2})}.$$

Lemma B.1 (Theorem 3.8 in Stewart [1990]). *Let $A \in \mathbb{R}^{m \times n}$, with $m \;\ge\; n$ and let $\widetilde{A} = A + E$. Then*

$$\|\widetilde{A}^{\dagger} - A^{\dagger}\|_{\mathrm{op}} \;\le\; \frac{1 + \sqrt{5}}{2} \max\{\|\widetilde{A}^{\dagger}\|_{\mathrm{op}}^2, \|A^{\dagger}\|_{\mathrm{op}}^2\} \|E\|_{\mathrm{op}}.$$

Furthermore, if $\|E\|_{\mathrm{op}} \;\le\; \frac{1}{2} \sigma_{\min}(A)$, then

$$\|\widetilde{A}^{\dagger} - A^{\dagger}\|_{\mathrm{op}} \lesssim \|A^{\dagger}\|_{\mathrm{op}}^2 \|E\|_{\mathrm{op}}.$$

B.2 Proof of Proposition 4.1: Relating stability and invertibility

The first part of the proposition follows directly from Fact 3.1. For the second, again using Fact 3.1:

$$\begin{aligned}
1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) &= \|(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\|_{\text{op}} \\
&= \|\sum_{k=0}^{\infty}(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^k\|_{\text{op}} \\
&\leq \sum_{k=0}^{\infty}\|(\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^k\|_{\text{op}}. \\
&\leq \sum_{k=0}^{\infty}\mathsf{cond}(P_\gamma)^{1/2}\left(1 - \frac{1}{\|P_\gamma\|_{\text{op}}}\right)^{k/2}
\end{aligned}$$

Here, we've used Lemma A.1 in the last line. The final bound follows from applying the final argument from the proof of Theorem 1.

B.3 Proof of Proposition 4.2: Relationship to Mou et al. [2020]

The result follows from the proof of Corollary 1 in Mou et al. [2020]. We include the calculation for the sake of completeness. For any unit vector u,

$$(1-\kappa)\|u\|_2^2 \;\leq\; u^\top(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})u \;\leq\; \|(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})u\|_{\text{op}}\|u\|_2.$$

Therefore, $\|(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\|_{\text{op}} = 1/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}) \;\leq\; 1/(1-\kappa)$.

B.4 Proof of Proposition 4.3: contractivity implies stability

By the Schur Complement Lemma, the contractivity condition implies that

$$\Sigma_{\text{cov}} - \Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}^\top \succeq 0.$$

Rearranging and multiplying on the left and the right by $\Sigma_{\text{cov}}^{-1/2}$,

$$I \succeq (\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})(\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^\top.$$

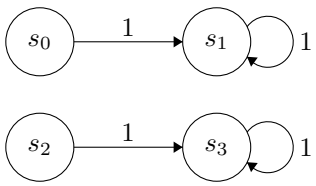
Using the fact that $\gamma \in (0,1)$ and the identity that for any matrix A, $\|A\|_{\text{op}}^2 = \|AA^\top\|_{\text{op}}$, we conclude

$$\|\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}^2 < 1.$$

Stability follows from the observation that the spectral radius of a matrix is always smaller than the operator norm.

B.5 Proof of Proposition 4.4: gaps between stability and other conditions

Consider the following MDP with 4 states and no actions,



The reward distribution at every state is a mean-zero coin toss: $R(s) = \text{Unif}(\{\pm 1\})$ for all $s \in \mathcal{S}$. Now, consider the two-dimensional feature mapping,

$$\phi(s_0) = [1, 0]^\top,\ \phi(s_1) = [0, 1/\varepsilon]^\top,\ \phi(s_2) = [0, 1]^\top,\ \phi(s_3) = [\varepsilon, 0]^\top,$$

where $\varepsilon > 0$ is a problem parameter to be determined later. This MDP is (trivially) linearly realizable with $\theta^\star_\gamma = 0$ because all rewards have 0 mean. If $\mathcal{D}$ place probability $1/2$ on s_0 and s_2, then

$$\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2} = \gamma \begin{bmatrix} 0 & 1/\varepsilon \\ \varepsilon & 0 \end{bmatrix}.$$

This matrix has eigenvalues equal to γ and $-\gamma$ for all values of $\varepsilon > 0$. Hence, its spectral radius of this matrix is always strictly smaller than 1 and the OPE instance is stable (and hence invertible).

For this problem, we can check that

$$\Sigma_{\text{next}} = \frac{1}{2} \begin{bmatrix} \varepsilon^2 & 0 \\ 0 & 1/\varepsilon^2 \end{bmatrix} \text{ and } \Sigma_{\text{cov}} = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore, $\mathcal{C}_{\text{ds}}$ is the smallest positive number β such that

$$0 \preceq \frac{1}{2} \begin{bmatrix} \beta - \varepsilon^2 & 0 \\ 0 & \beta - 1/\varepsilon^2 \end{bmatrix}$$

Low distribution shift. While stability holds for all values of $\varepsilon > 0$, as $\varepsilon \to 0$, $\mathcal{C}_{\text{ds}}$ goes to ∞ (because $1/\varepsilon^2$ becomes arbitrarily large). Hence, stability holds, but low distribution shift does not. This proves the first case.

Symmetric stability. Similarly, as $\varepsilon \to 0$, we can check that the two eigenvalues of

$$\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2} + (\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2})^\top,$$

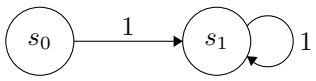
go to $\pm\infty$. Therefore, the symmetric stability condition (Assumption 6) also fails for this problem.

Contractivity. From the argument in Proposition 4.3, we know that if contractivity (Assumption 7) held, then

$$\|\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} < 1.$$

However, a direct calculation shows that as $\varepsilon \to 0$, then $\|\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \to \infty$. Therefore, while stability holds, contractivity does not.

Bellman completeness. To prove the last case, we use a different example. In particular, consider the following MDP (with no actions) presented in Amortila et al. [2020],



The rewards are $R(s_0) = 0$ (almost surely) and $\mathbb{E}R(s_1) = 1$. The value function of any policy is linearly realizable in the feature mapping $\phi(s_0) = \gamma$ and $\phi(s_1) = 1$ with $\theta^\star_\gamma = 1/(1-\gamma)$. If the offline distribution places mass $1/2$ on each state then,

$$\gamma \Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{cr}} \Sigma_{\text{cov}}^{-1/2} = \left(\frac{\gamma}{2}\right) \frac{\gamma^2 + 1}{\gamma + 1}.$$

This matrix (scalar) lies in the interval $(0, 1)$ and is hence clearly stable and invertible. However, Bellman completeness fails for this MDP. In particular, Bellman completeness asserts that for every θ there exist a θ' such that for all $s \in \mathcal{S}$,

$$\phi(s)\theta' = \mathbb{E}R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s)} \phi(s') \cdot \theta$$

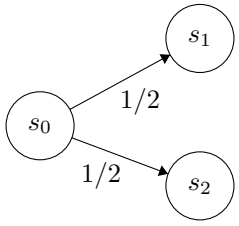
In this case, this means that for all θ, there exists a θ' such that

$$\phi(s_1) \cdot \theta' = 1 + \gamma \cdot \phi(s_1) \cdot \theta$$
$$\phi(s_0) \cdot \theta' = 0 + \gamma \cdot \phi(s_1) \cdot \theta.$$

Plugging in our choice of feature map, these equations become $\theta' = 1 + \gamma \cdot \theta$ and $\gamma \cdot \theta' = \gamma \cdot \theta$. They clearly cannot be satisfied if we pick any $\theta \neq 0$.

B.6 Bellman Residual Minimization Counterexample

Consider the following 3 state MDP with no actions and stochastic transitions:



The feature mapping is:

$$\phi(s_0) = \frac{\gamma}{4},\ \phi(s_1) = \frac{1}{2},\ \phi(s_2) = 0.$$

Rewards are exactly 0 everywhere except for s_1, where $r(s_1) = 1$ deterministically. We can check that this example is linearly realizable with $\theta^\star_\gamma = \frac{1}{1-\gamma}$. However, it also holds that

$$\Sigma_{\text{cov}} = \frac{\gamma^2}{16},\ \Sigma_{\text{cr}} = \frac{\gamma}{16},\ \Sigma_{\text{next}} = \frac{1}{8}.$$

Hence, $\Sigma_{\text{cov}} - \gamma\Sigma_{\text{cr}} - \gamma\Sigma_{\text{cr}}^\top . + \gamma^2\Sigma_{\text{next}} \succ 0$, but BRM returns the wrong answer,

$$\theta_{\text{BRM}} = (\Sigma_{\text{cov}} - \gamma\Sigma_{\text{cr}} - \gamma\Sigma_{\text{cr}}^\top + \gamma^2\Sigma_{\text{next}})^\dagger(\theta_{\phi,r} - \gamma\mathbb{E}\phi(s', a')r(s, a)) = 0,$$

since $\mathbb{E}\phi(s, a)r(s, a) = \mathbb{E}\phi(s', a')r(s, a) = 0$.

B.7 Proof of Theorem 3: necessity of invertibility for LSTD

We begin by proving two auxiliary claims and then move on to proving each part of the theorem separately.

Claim B.2. *If the matrix $I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$ is singular, then there exists a real vector $v \in \mathbb{R}^d$ such that:*

$$\mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}, s'\sim P(\cdot|s,a), a'\sim\pi(s')} \phi(s, a)\langle\gamma \cdot \phi(s', a') - \phi(s, a), v\rangle = 0.$$

Proof. The matrix being rank deficient implies that there exists a vector v such that $(I - \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}})v = 0$, or equivalently, that the matrix $\gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}$ has an eigenvector v with eigenvalue 1. Because the matrix and eigenvalue are both real, we can also take v to be real. From here, $v = \gamma\Sigma_{\text{cov}}^{-1}\Sigma_{\text{cr}}v$. Hence, $\Sigma_{\text{cov}}v = \gamma\Sigma_{\text{cr}}v$. Expanding out the definitions of these matrices,

$$\mathbb{E}\phi(s, a)\langle\phi(s, a), v\rangle = \gamma\mathbb{E}\phi(s, a)\langle\phi(s', a'), v\rangle.$$

Rearranging both terms to be on the same side we get the claim. □

Claim B.3. *For any* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\phi(s,a) = -\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(\gamma\phi(s_{t+1},a_{t+1}) - \phi(s_t,a_t)\right) \mid (s_0,a_0) = (s,a), \pi\right].$$

Proof. The sum telescopes and $\lim_{t\to\infty} \gamma^t \mathbb{E}\phi(s_t,a_t) = 0$. □

We conclude with the proof of Theorem 3:

Alternate reward. As per the presentation of theorem, the only difference between $\mathcal{M}$ and $\overline{\mathcal{M}}$ is the unknown reward. In particular, we define the new reward function $\overline{R}$ as

$$\overline{R}(s,a) = R(s,a) + \frac{1}{2B}\langle \gamma \cdot \phi(s',a') - \phi(s,a), v\rangle \tag{B.1}$$

where v is as in Claim B.2, $s' \sim P(\cdot \mid s,a)$, $a' \sim \pi(s')$, and $B = \sup_{s,a} \|\phi(s,a)\|_2$. Note that by Cauchy-Schwarz, and the definition of B, for any s, s', a, and a':

$$|\frac{1}{2B}\langle \gamma \cdot \phi(s',a') - \phi(s,a), v\rangle| \;\le\; \frac{1}{2B}\|v\|_2(\|\phi(s,a)\|_2 + \|\phi(s',a')\|_2) \;\le\; 1$$

Therefore, $|r(s,a)|$ is uniformly bounded by 2.

Proof of identical moments. Since the features, offline distribution, and transitions are all the same, then $\Sigma_{\text{cov}} = \bar{\Sigma}_{\text{cov}}, \Sigma_{\text{cr}} = \bar{\Sigma}_{\text{cr}}$, and $\Sigma_{\text{next}} = \bar{\Sigma}_{\text{next}}$. Next, by expanding out the new reward function:

$$\begin{aligned}
\bar{\theta}_{\phi,r} &= \mathbb{E}_{(s,a)\sim\mathcal{D}}\phi(s,a)\bar{r}(s,a) \\
&= \mathbb{E}\phi(s,a)r(s,a) + \frac{1}{2B}\mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}, s'\sim P(\cdot|s,a), a'\sim\pi(s')}\phi(s,a)\langle \gamma\cdot\phi(s',a') - \phi(s,a), v\rangle \\
&= \mathbb{E}\phi(s,a)r(s,a) + 0,
\end{aligned}$$

where the last line follows from Claim B.2.

Proof of realizability. Expanding out the definition of $\bar{Q}^\pi$,

$$\begin{aligned}
\bar{Q}^\pi(s,a) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\bar{r}(s_t,a_t) \mid (s_0,a_0) = (s,a), \pi\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\cdot r(s_t,a_t) \mid (s_0,a_0) = (s,a), \pi\right] \\
&+ \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\cdot\langle\gamma\cdot\phi(s_{t+1},a_{t+1}) - \phi(s_t,a_t), \frac{1}{2B}v\rangle \mid (s_0,a_0) = (s,a), \pi\right] \\
&= Q^\pi(s,a) - \phi(s,a)^\top v\frac{1}{2B} \\
&= \phi(s,a)^\top(\theta^\star_\gamma - \frac{1}{2B}v),
\end{aligned}$$

where in the 3rd line we have used Claim B.3 and in the last one used the assumption that Q^π is linearly realizable. In short, $\bar{Q}^\pi$ is linearly realizable with weight vector $\theta^\star_\gamma - (2B)^{-1}v$.

Proof of different Q functions By the previous part establishing the realizability of $\bar{Q}^\pi$,

$$\mathbb{E}_{\mathcal{D}}(Q^\pi(s,a) - Q'^\pi(s,a))^2 = \frac{1}{4B^2} v^\top \Sigma_{\text{cov}} v \;\geq\; \frac{\sigma_{\min}(\Sigma_{\text{cov}})}{4B^2}\|v\|_2^2.$$

The precise statement follows from the fact that v has unit length.

C Concentration Analysis: Proof of Lemma 2.1

Lemma C.1 (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 $\delta \in (0,1)$,*

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

Lemma C.2 (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 $1-\delta$ for any $\delta \in (0,1)$,*

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

To shorten the notation in our concentration analysis, we use $x_i = \phi(s_i, a_i)$ and $y_i = \phi(s_i', a_i')$, and $r_i = r(s_i, a_i)$. With this shorthand:

$$\Sigma_{\text{cov}} = \mathbb{E}xx^\top, \quad \widehat{\Sigma}_{\text{cov}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top, \quad \Sigma_{\text{cr}} = \mathbb{E}xy^\top, \quad \widehat{\Sigma}_{\text{cr}} = \frac{1}{n}\sum_{i=1}^n x_i y_i^\top, \tag{C.1}$$

$$\theta_{\phi,r} = \mathbb{E}xr, \quad \widehat{\theta}_{\phi,r} = \frac{1}{n}\sum_{i=1}^n x_i r_i.$$

C.1 Bounding ε_{op}

Lemma C.3. *If $n \gtrsim \rho_s^2 \log(d/\delta)$ then, with probability $1-\delta$,*

$$\|\Sigma_{\text{cov}}^{1/2}(\gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}})\Sigma_{\text{cov}}^{-1/2} - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \lesssim \sqrt{\frac{\max(\sigma_{\text{cr}}^2, \sigma_{\text{cov}}^2\mathcal{C}_{\text{ds}})\log(d/\delta)}{n}} + \frac{\max(\mathcal{C}_{\text{ds}}^{1/2}\rho_s^2, \rho_s\rho_{s'})\log(d/\delta)}{n}.$$

Proof. Let $\widehat{A} := \gamma\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\Sigma}_{\text{cr}}$. We start by using the following error decomposition,

$$\begin{aligned}
&\|\Sigma_{\text{cov}}^{1/2}\widehat{A}\Sigma_{\text{cov}}^{-1/2} - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \\
&\leq\; \gamma\|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}\cdot\Sigma_{\text{cov}}^{-1/2}\left(\widehat{\Sigma}_{\text{cr}} - \Sigma_{\text{cr}}\right)\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} + \gamma\|\Sigma_{\text{cov}}^{1/2}(\widehat{\Sigma}_{\text{cov}}^{-1} - \Sigma_{\text{cov}}^{-1})\Sigma_{\text{cov}}^{1/2}\cdot\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \\
&\leq\; \gamma\underbrace{\|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}\|_{\text{op}}}_{:=T_1}\cdot\underbrace{\|\Sigma_{\text{cov}}^{-1/2}\left(\widehat{\Sigma}_{\text{cr}} - \Sigma_{\text{cr}}\right)\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}}_{:=T_2} \\
&+ \underbrace{\|\Sigma_{\text{cov}}^{1/2}(\widehat{\Sigma}_{\text{cov}}^{-1} - \Sigma_{\text{cov}}^{-1})\Sigma_{\text{cov}}^{1/2}\|_{\text{op}}}_{:=T_3}\cdot\underbrace{\|\gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}}_{:=T_4}.
\end{aligned}$$

We now bound each of these terms separately.

Bound on T_2. We apply the Matrix Bernstein inequality on $\Sigma_{\text{cov}}^{-1/2}\left(\widehat{\Sigma}_{\text{cr}} - \Sigma_{\text{cr}}\right)\Sigma_{\text{cov}}^{-1/2}$. Here we define

$$S_k = \Sigma_{\text{cov}}^{-1/2}\left(x_k y_k^\top - \Sigma_{\text{cr}}\right)\Sigma_{\text{cov}}^{-1/2}$$

which is centered and satisfies:

$$\begin{aligned}\|S_k\| \;&\le\; \|\Sigma_{\text{cov}}^{-1/2} x_k y_k^\top \Sigma_{\text{cov}}^{-1/2}\| + \mathbb{E}_{\mathcal{D}}\|\Sigma_{\text{cov}}^{-1/2} x y^\top \Sigma_{\text{cov}}^{-1/2}\| \;\le\; 2\sup_{(x,y)\in\text{supp}(\mathcal{D})}\|\Sigma_{\text{cov}}^{-1/2} x y^\top \Sigma_{\text{cov}}^{-1/2}\| \\ &\le\; 2\sup_{(x,y)\in\text{supp}(\mathcal{D})}\|\Sigma_{\text{cov}}^{-1/2} x\|\cdot\|\Sigma_{\text{cov}}^{-1/2} y\| \;\le\; 2\rho_s\rho_{s'}.\end{aligned}$$

Therefore for σ_{cr}^2 defined as in Eq. (2.6), , we get that with probability $1-\delta$,

$$T_2 \;\le\; \sqrt{\frac{2\sigma_{\text{cr}}^2\log(2d/\delta)}{n}} + \frac{4\rho_s\rho_{s'}\log(2d/\delta)}{3n}.$$

Bound on T_1 and T_3. Essentially the same argument as for the bound on T_2 reveals that,

$$\|\Sigma_{\text{cov}}^{-1/2}(\widehat{\Sigma}_{\text{cov}} - \Sigma_{\text{cov}})\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \;\le\; \sqrt{\frac{2\sigma_{\text{cov}}^2\log(2d/\delta)}{n}} + \frac{2\rho_s^2\log(2d/\delta)}{3n} =: \tau. \tag{C.2}$$

This inequality directly implies that

$$1-\tau \;\le\; \lambda_{\min}(\Sigma_{\text{cov}}^{-1/2}\widehat{\Sigma}_{\text{cov}}\Sigma_{\text{cov}}^{-1/2}) \;\le\; \lambda_{\max}(\Sigma_{\text{cov}}^{-1/2}\widehat{\Sigma}_{\text{cov}}\Sigma_{\text{cov}}^{-1/2}) \;\le\; 1+\tau,$$

which in particular implies that $\Sigma_{\text{cov}}^{-1/2}\widehat{\Sigma}_{\text{cov}}\Sigma_{\text{cov}}^{-1/2}$ is invertible whenever $\tau < 1/2$, a fact that is ensured by our lower bound on n. Therefore:

$$T_1 = \|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}\| = \frac{1}{\lambda_{\min}(\Sigma_{\text{cov}}^{-1/2}\widehat{\Sigma}_{\text{cov}}\Sigma_{\text{cov}}^{-1/2})} \;\le\; \frac{1}{1-\tau}. \tag{C.3}$$

More generally, we have that:

$$1-2\tau \;\le\; \lambda_{\min}(\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}) \;\le\; \lambda_{\max}(\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}) \;\le\; 1+2\tau.$$

Using the fact that $1/(1+\tau) \;\ge\; 1-2\tau$ and $1/(1-\tau) \;\le\; 1+2\tau$ for $\tau \;\le\; 1/2$, this directly yields

$$T_3 = \|\Sigma_{\text{cov}}^{1/2}(\widehat{\Sigma}_{\text{cov}}^{-1} - \Sigma_{\text{cov}}^{-1})\Sigma_{\text{cov}}^{1/2}\| \;\le\; 2\tau. \tag{C.4}$$

Thus, we have bounded T_1 and T_3. In particular, for $\tau < 1/2$, $T_1 \;\le\; 2$, and $T_3 \;\le\; 2\tau$.

Bound on T_4. For T_4, no concentration argument is required. Instead, a Schur complement argument implies that,

$$\|\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}}^2 \;\le\; \|\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{next}}\Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \;\le\; \mathcal{C}_{\text{ds}},$$

where we've used $\Sigma_{\text{next}} \preceq \mathcal{C}_{\text{ds}}\Sigma_{\text{cov}}$. Hence, $T_4 \;\le\; \mathcal{C}_{\text{ds}}^{1/2}$.

Wrapping up. Taking a union bound, we obtain that

$$\varepsilon_{\text{op}} \lesssim \sqrt{\frac{\max(\sigma_{\text{cr}}^2, \sigma_{\text{cov}}^2\mathcal{C}_{\text{ds}})\log(d/\delta)}{n}} + \frac{\max(\mathcal{C}_{\text{ds}}^{1/2}\rho_s^2, \rho_s\rho_{s'})\log(d/\delta)}{n}.$$

□

C.2 Bounding ε_r

Lemma C.4. *If $n \gtrsim \rho_s^2 \log(d/\delta)$ then, with probability $1-\delta$,*

$$\|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\theta}_{\phi,r} - \Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \lesssim \sqrt{\frac{\max(\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2^2\sigma_{\text{cov}}^2, \sigma_r^2)\log(d/\delta)}{n}} + \frac{\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2\rho_s^2\log(d/\delta)}{n}.$$

Proof. The ideas are very similar to Lemma C.3. In this case, the relevant error decomposition is,

$$\begin{aligned}\varepsilon_r &= \|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\widehat{\theta}_{\phi,r} - \Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \\ &\le \underbrace{\|\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2}\|_{\text{op}}}_{:=T_1}\underbrace{\|\Sigma_{\text{cov}}^{-1/2}(\theta_{\phi,r} - \widehat{\theta}_{\phi,r})\|_2}_{:=T_2} + \underbrace{\|(\Sigma_{\text{cov}}^{1/2}\widehat{\Sigma}_{\text{cov}}^{-1}\Sigma_{\text{cov}}^{1/2} - I)\|_{\text{op}}}_{:=T_3}\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2.\end{aligned}$$

Bound on T_1 and T_3. Whenever τ, defined as in Eq. (C.2), is strictly less than $1/2$, the analysis therein (in particular, Eq. (C.4) and Eq. (C.3)) proves that $T_1 \le 2$ and $T_3 \le 2\tau$.

Bound on T_2. We apply the vector Bernstein inequality, Lemma C.2, on the vectors

$$v_i = \Sigma_{\text{cov}}^{-1/2}x_i r_i - \Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}.$$

Note that, since the rewards have magnitude bounded by 1,

$$\sup_i \|v_i\|_2 \le \sup_i \|\Sigma_{\text{cov}}^{-1/2}x_i r_i\|_2 + \|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2 \le \|\Sigma_{\text{cov}}^{-1/2}x_i r_i\|_2 + \mathbb{E}\|\Sigma_{\text{cov}}^{-1/2}xr\|_2 \le 2\rho_s,$$

and,

$$\mathbb{E}\|v_i\|_2^2 = \mathbb{E}\|\Sigma_{\text{cov}}^{-1/2}x_i r_i\|_2^2 - \|\Sigma_{\text{cov}}^{1/2}\theta_{\phi,r}\|_2^2 = \sigma_r^2.$$

Applying vector Bernstein,

$$T_2 \le \sqrt{\frac{2\sigma_r^2\log(28/\delta)}{n}} + \frac{4\rho_s\log(28/\delta)}{3n}.$$

Wrapping up. Combining these, we get that,

$$\varepsilon_r \lesssim \sqrt{\frac{\max(\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2^2\sigma_{\text{cov}}^2, \sigma_r^2)\log(d/\delta)}{n}} + \frac{\|\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}\|_2\rho_s^2\log(d/\delta)}{n}.$$

□

C.3 Bounding variances

Bounding σ_r^2 Since the rewards $r(s,a)$ satisfy $|r(s,a)| \le 1$, we have that

$$\mathbb{E}\|v_i\|_2^2 = \mathbb{E}\|\Sigma_{\text{cov}}^{-1/2}x_i r_i\|_2^2 - \|\Sigma_{\text{cov}}^{1/2}\theta_{\phi,r}\|_2^2 \le \text{tr}\left[\Sigma_{\text{cov}}^{-1/2}\mathbb{E}r_i^2 x_i x_i^\top \Sigma_{\text{cov}}^{-1/2}\right] \le d.$$

Bounding σ_{cr}^2. Again using the notation from Eq. (C.1), and letting

$$S_k = \Sigma_{\text{cov}}^{-1/2}\left(x_k y_k^\top - \Sigma_{\text{cr}}\right)\Sigma_{\text{cov}}^{-1/2}$$

bounding σ_{cr}^2 is equivalent to bounding the operator norms of:

$$\begin{aligned}\mathbb{E}[S_k S_k^\top] &= \mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2}y\|_2^2(\Sigma_{\text{cov}}^{-1/2}x)(\Sigma_{\text{cov}}^{-1/2}x)^\top] - \Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cr}}^\top\Sigma_{\text{cov}}^{-1/2} \\ \mathbb{E}[S_k^\top S_k] &= \mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2}x\|_2^2(\Sigma_{\text{cov}}^{-1/2}y)(\Sigma_{\text{cov}}^{-1/2}y)^\top] - \Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}^\top\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}.\end{aligned}$$

We will subsequently show that, for any vector $v \in \mathbb{R}^d$, we have

$$v^\top(\mathbb{E}[S_k S_k^\top])v \;\geq\; 0, \quad v^\top(\mathbb{E}[S_k^\top S_k])v \;\geq\; 0. \tag{C.5}$$

Additionally, for any random variables $(a, b) \in \mathbb{R} \times \mathbb{R}^d$ from some joint distribution, Holder's inequality implies that

$$\begin{aligned}\|\mathbb{E}[a^2 bb^\top]\|_{\text{op}} = \sup_{v, \|v\|_2=1} \mathbb{E}[a^2(v^\top b)^2] \;&\leq\; \min\{\sup\{a\} \sup_v \mathbb{E}[(v^\top b)^2], \sup_{b,v}\{(v^\top b)^2\}\mathbb{E}[a^2]\} \\ &= \min\{\sup\{a\}\|\mathbb{E}[bb^\top]\|_{\text{op}}, \sup\{\|b\|_2^2\}\mathbb{E}[a^2]\}.\end{aligned}$$

Using these two facts and positive semi-definiteness, we have that

$$\|\mathbb{E}[S_k S_k^\top]\|_{\text{op}} \;\leq\; \|\mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2} y\|_2^2 (\Sigma_{\text{cov}}^{-1/2} x)(\Sigma_{\text{cov}}^{-1/2} x)^\top]\|_{\text{op}} \;\leq\; \sup_y \|\Sigma_{\text{cov}}^{-1/2} y\|_2^2 \|\mathbb{E}[(\Sigma_{\text{cov}}^{-1/2} x)(\Sigma_{\text{cov}}^{-1/2} x)^\top]\| \;\leq\; \rho_{s'}^2.$$

Essentially the same proof yields a similar bound on $\|\mathbb{E}[S_k^\top S_k]\|$:

$$\|\mathbb{E}[S_k^\top S_k]\|_{\text{op}} \;\leq\; \|\mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2} x\|_2^2 (\Sigma_{\text{cov}}^{-1/2} y)(\Sigma_{\text{cov}}^{-1/2} y)^\top]\|_{\text{op}} \;\leq\; \rho_0^2 \|\Sigma_{\text{cov}}^{-1/2} \Sigma_{\text{next}} \Sigma_{\text{cov}}^{-1/2}\|_{\text{op}} \;\leq\; \rho_s^2 \mathcal{C}_{\text{ds}}.$$

Alternatively, we can get

$$\begin{aligned}\|\mathbb{E}[S_k^\top S_k]\|_{\text{op}} \;&\leq\; \|\mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2} x\|_2^2 (\Sigma_{\text{cov}}^{-1/2} y)(\Sigma_{\text{cov}}^{-1/2} y)^\top]\|_{\text{op}} \;\leq\; \mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2} x\|_2^2 \|(\Sigma_{\text{cov}}^{-1/2} y)(\Sigma_{\text{cov}}^{-1/2} y)^\top]\|] \\ &= \mathbb{E}[\|\Sigma_{\text{cov}}^{-1/2} x\|_2^2 \|\Sigma_{\text{cov}}^{-1/2} y\|_2^2] \;\leq\; \rho_{s'}^2 d.\end{aligned}$$

Let us now verify (C.5). Rebinding $\tilde{x} = \Sigma_{\text{cov}}^{-1/2} x, \tilde{y} = \Sigma_{\text{cov}}^{-1/2} y$, we have

$$v^\top(\mathbb{E}[S_k S_k^\top])v = \mathbb{E}[(v^\top \tilde{x})^2 \|\tilde{y}\|_2^2] - (\mathbb{E}(v^\top \tilde{x})\tilde{y})^\top (\mathbb{E}(v^\top \tilde{x})\tilde{y}) = \mathbb{E}\|(v^\top \tilde{x})\tilde{y}\|_2^2 - \|\mathbb{E}[(v^\top \tilde{x})\tilde{y}]\|_2^2 \;\geq\; 0,$$

where the last inequality is by convexity. In conclusion,

$$\sigma_{\text{cr}}^2 \;\leq\; \max(\rho_{s'}^2, \min(\rho_s^2 \mathcal{C}_{\text{ds}}, \rho_{s'}^2 d)).$$

Bounding σ_{cov}^2. For $\tilde{x} = \Sigma_{\text{cov}}^{-1/2} \phi(s, a)$, the variance σ_{cov}^2 is equal to

$$\sigma_{\text{cov}}^2 = \|\mathbb{E}\tilde{x}\tilde{x}^\top \tilde{x}\tilde{x}^\top - I\|_{\text{op}} = \|\mathbb{E}\|\tilde{x}\|_2^2 \tilde{x}\tilde{x}^\top - I\|_{\text{op}}.$$

While this quantity is always less that ρ_s^2, one can achieve tighter bounds if the offline distribution is *hypercontractive* as per the following definition:

Definition C.1. A distribution $\mathcal{D}$ over random vectors x is $L8$-$L2$ hypercontractive if there exists a positive constant L such that for all unit vectors u,

$$\mathbb{E}_{x \sim \mathcal{D}}((x - \mathbb{E}x)^\top u)^8 \;\leq\; L^2 \left(\mathbb{E}_{x \sim \mathcal{D}}((x - \mathbb{E}x)^\top u)^2\right)^4.$$

Gaussians or strongly log-concave distributions are some examples of probability measures that satisfy this condition. If $\Sigma_{\text{cov}}^{-1/2} \phi(s, a)$ is $L8$-$L2$ hypercontractive, then one can show that

$$\sigma_{\text{cov}}^2 \lesssim L \text{tr}\left[I + \mu\mu^\top\right] \|I + \mu\mu^\top\|_{\text{op}},$$

where $\mu := \Sigma_{\text{cov}}^{-1/2} \mathbb{E}_{(s,a) \sim \mathcal{D}} \phi(s, a)$. We point the interested reader to Lemma A.3 in Cherapanamjeri et al. [2020] for a more formal derivation.

D Analyzing the Misspecified Case: Proof of Proposition 5.1

By definition of $\theta^\star_\infty$, we have that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ we can write Q^π as

$$Q^\pi(s,a) = \phi(s,a)^\top \theta^\star_\gamma + f(s,a), \tag{D.1}$$

where $f(s,a) = Q^\pi(s,a) - \phi(s,a)^\top \theta^\star_\gamma$ and $\sup_{s,a} |f(s,a)| \leq \varepsilon_\infty$.

From the relationship above, we have that for $\widehat{Q}^\pi(s,a) = \phi(s,a)^\top \widehat{\theta}$

$$\begin{aligned} |Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| &= |\phi(s,a)^\top(\theta^\star_\infty - \widehat{\theta}) + f(s,a)| \\ &\leq \|\Sigma_{\text{cov}}^{-1/2}\phi(s,a)\|_2 \|\Sigma_{\text{cov}}^{1/2}(\theta^\star_\infty - \widehat{\theta})\|_2 + |f(s,a)|. \end{aligned} \tag{D.2}$$

Applying the triangle inequality again,

$$\|\Sigma_{\text{cov}}^{1/2}(\theta^\star_\infty - \widehat{\theta})\|_2 \leq \|\Sigma_{\text{cov}}^{1/2}(\theta^\star_\infty - \theta^\star_{\text{fp}})\|_2 + \|\Sigma_{\text{cov}}^{1/2}(\theta^\star_{\text{fp}} - \widehat{\theta})\|_2. \tag{D.3}$$

By assumption on $\widehat{\theta}$, $\|\Sigma_{\text{cov}}^{1/2}(\theta^\star_{\text{fp}} - \widehat{\theta})\|_2 \leq \varepsilon_{\text{fp}}$. Therefore, it remains to bound $\|\Sigma_{\text{cov}}^{1/2}(\theta^\star_\infty - \theta^\star_{\text{fp}})\|_2$. By Claim D.1, we have that

$$\begin{aligned} \Sigma_{\text{cov}}^{1/2}\theta^\star_\infty &= (I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r} \\ &\quad + (I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\Sigma_{\text{cov}}^{-1/2} \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D} \\ s'\sim P(\cdot|s,a), a'\sim\pi(s')}} \phi(s,a)(\phi(s,a)^\top\theta^\star_\infty - \gamma\phi(s',a')^\top\theta^\star_\infty - r(s,a)). \end{aligned}$$

Note that $\Sigma_{\text{cov}}^{1/2}\theta^\star_{\text{fp}}$ is exactly equal to $(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r}$. Furthermore, by the second part of Claim D.1, the ℓ_2 norm of the second term in the expression above is upper bounded by $\rho_s\varepsilon_\infty/\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})$. Consequently,

$$\|\Sigma_{\text{cov}}^{1/2}(\theta^\star_\infty - \theta^\star_{\text{fp}})\|_2 \leq \frac{\rho_s}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})}\varepsilon_\infty. \tag{D.4}$$

Combining Eqs. (D.2) to (D.4), we get that

$$|Q^\pi(s,a) - \widehat{Q}^\pi(s,a)| \leq \|\Sigma_{\text{cov}}^{-1/2}\phi(s,a)\|_2(\varepsilon_{\text{fp}} + \frac{\rho_s}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})}\varepsilon_\infty) + \varepsilon_\infty.$$

Claim D.1. *Let $\theta^\star_\infty$ be defined as in Eq. (5.1) and let $A := I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2}$ then,*

$$\Sigma_{\text{cov}}^{1/2}\theta^\star_\infty - \Sigma_{\text{cov}}^{1/2}\theta_\star = A^{-1}\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r} + A^{-1}\Sigma_{\text{cov}}^{-1/2} \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D} \\ s'\sim P(\cdot|s,a), a'\sim\pi(s')}} \phi(s,a)(\phi(s,a)^\top\theta^\star_\infty - \gamma\phi(s',a')^\top\theta^\star_\infty - r(s,a)).$$

where

$$\|A^{-1}\Sigma_{\text{cov}}^{-1/2} \mathop{\mathbb{E}}_{\substack{(s,a)\sim\mathcal{D} \\ s'\sim P(\cdot|s,a), a'\sim\pi(s')}} \phi(s,a)(\phi(s,a)^\top\theta^\star_\infty - \gamma\phi(s',a')^\top\theta^\star_\infty - r(s,a))\|_2 \;\le\; \frac{\varepsilon_\infty}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})}\rho_s.$$

Proof. By the Bellman equation, we have that,

$$Q^\pi(s,a) = \mathbb{E}r(s,a) + \gamma \cdot \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} Q^\pi(s',a').$$

Using the decomposition from Eq. (D.1), the following relationship holds for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\phi(s,a)^\top\theta^\star_\gamma = \mathbb{E}r(s,a) + \gamma \cdot \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} \phi(s',a')^\top\theta^\star_\gamma - f(s,a) + \gamma \cdot \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} f(s',a').$$

Now we do a couple of things, we multiply on the left by $\Sigma_{\text{cov}}^{-1/2}\phi(s,a)$ and take expectations with respect to $(s,a)\sim\mathcal{D}$. Rearranging, we get the following equation:

$$\Sigma_{\text{cov}}^{1/2}\theta^\star_\gamma = (I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1}\Sigma_{\text{cov}}^{-1/2}\theta_{\phi,r} + (I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1} \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} \Sigma_{\text{cov}}^{-1/2}\phi(s,a)(\gamma\cdot f(s',a') - f(s,a))$$

Focusing on the second term, we have that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, $|f(s,a)| \le \varepsilon_\infty$ and $\|\Sigma_{\text{cov}}^{-1/2}\phi(s,a)\|_2 \le \rho_s$. Therefore,

$$\|(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})^{-1} \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} \Sigma_{\text{cov}}^{-1/2}\phi(s,a)(\gamma\cdot f(s',a') - f(s,a))\|_2 \;\le\; \frac{\varepsilon_\infty\cdot\rho_s}{\sigma_{\min}(I - \gamma\Sigma_{\text{cov}}^{-1/2}\Sigma_{\text{cr}}\Sigma_{\text{cov}}^{-1/2})}.$$

Moreover, $f(s,a) = Q^\pi(s',a') - \phi(s,a)^\top\theta^\star_\infty$ and $Q^\pi(s,a) = \mathbb{E}r(s,a) + \gamma\cdot \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a) \\ a'\sim\pi(s')}} Q^\pi(s',a')$.

Using these identities, we have that:

$$\gamma\cdot f(s',a') - f(s,a) = \phi(s,a)^\top\theta^\star_\infty - \gamma\phi(s',a')^\top\theta^\star_\infty - r(s,a).$$

□