arXiv:2203.12786v2 [cs.LG] 27 Jun 2022

Bellman Residual Orthogonalization for Offline Reinforcement Learning

Andrea Zanette[†] zanette@berkeley.edu
Martin J. Wainwright[†,⋆] wainwrig@berkeley.edu

Department of Electrical Engineering and Computer Sciences[†]
Department of Statistics[⋆]
UC Berkeley, Berkeley, CA

Abstract

We propose and analyze a reinforcement learning principle that approximates the Bellman equations by enforcing their validity only along an user-defined space of test functions. Focusing on applications to model-free offline RL with function approximation, we exploit this principle to derive confidence intervals for off-policy evaluation, as well as to optimize over policies within a prescribed policy class. We prove an oracle inequality on our policy optimization procedure in terms of a trade-off between the value and uncertainty of an arbitrary comparator policy. Different choices of test function spaces allow us to tackle different problems within a common framework. We characterize the loss of efficiency in moving from on-policy to off-policy data using our procedures, and establish connections to concentrability coefficients studied in past work. We examine in depth the implementation of our methods with linear function approximation, and provide theoretical guarantees with polynomial-time implementations even when Bellman closure does not hold.

1 Introduction

Markov decision processes (MDP) provide a general framework for optimal decision-making in sequential settings (e.g., [Put94, Ber95a, Ber95b]). Reinforcement learning refers to a general class of procedures for estimating near-optimal policies based on data from an unknown MDP (e.g., [BT96, SB18]). Different classes of problems can be distinguished depending on our access to the data-generating mechanism. Many modern applications of RL involve learning based on a pre-collected or offline dataset. Moreover, the state-action spaces are often sufficiently complex that it becomes necessary to implement function approximation. In this paper, we focus on model-free offline reinforcement learning (RL) with function approximation, where prior knowledge about the MDP is encoded via the value function. In this setting, we focus on two fundamental problems: (1) offline policy evaluation—namely, the task of accurately predicting the value of a target policy; and (2) offline policy optimization, which is the task of finding a high-performance policy.

There are various broad classes of approaches to off-policy evaluation, including importance sampling [Pre00, TB16, JL16, LLTZ18], as well as regression-based methods [LP03, MS08, CJ19]. Many methods for offline policy optimization build on these techniques, with a line of recent papers including the addition of pessimism [JYW21, XCJ+21, ZWB21]. We provide a more detailed summary of the literature in Section 6.3.

In contrast, this work investigates a different model-free principle—different from importance sampling or regression-based methods—to learn from an offline dataset. It belongs to the class of weight learning algorithms, which leverage an auxiliary function class to either encode the marginalized importance weights of the target policy [LLTZ18, XJ20b], or estimates of the Bellman errors [ASM08, CJ19, XJ20b]. Some work has considered kernel classes [FRTL20] or other weight classes to construct off-policy estimators [UHJ20] as well as confidence intervals at the population level [JH20]. However, these works do not examine in depth the statistical aspects of the problem, nor elaborate upon the design of the weight function classes.[1] The last two considerations are essential to obtaining data-dependent procedures accompanied by rigorous guarantees, and to provide guidance on the choice of weight class, which are key contributions of this paper.

For space reasons, we motivate our approach in the idealized case where the Bellman operator is known in Section 6.1, and compare with the weight learning literature at the population level in Section 6.2. Let us summarize our main contributions in the following three paragraphs.

Conceptual contributions Our paper makes two novel contributions of conceptual nature:

1. We propose a method, based on *approximate empirical orthogonalization* of the Bellman residual along test functions, to construct confidence intervals and to perform policy optimization.
2. We propose a sample-based approximation of such principle, based on *self-normalization* and *regularization*, and obtain general guarantees for parametric as well as non-parametric problems.

The construction of the estimator, its statistical analysis, and the concrete consequences (described in the next paragraph) are the major distinctions with respect to past work on weight learning methods [UHJ20, JH20]. Our analysis highlights the statistical trade-offs in the choice of the test functions. (See Section 6.2 for comparison with past work at the population level.)

Domain-specific results In order to illustrate the broad effectiveness and applicability of our general method and analysis, we consider several domains of interest. We show how to recover various results from past work—and to obtain novel ones—by making appropriate choices of the test functions and invoking our main result. Among these consequences, we discuss the following:

1. When marginalized importance weights are available, they can be used as test class. In this case we recover a similar results as the paper [XJ20b]; however, here we only require concentrability with respect to a comparator policy instead of over all policies in the class.
2. When some knowledge of the Bellman error class is available, it can be used as test class. Similar results have appeared previously either with stronger concentrability [CJ19] or in the special case of Bellman closure [XCJ+21].
3. We provide a test class that projects the Bellman residual along the error space of the $\mathcal{Q}$ class. The resulting procedure is as an extension of the LSTD algorithm [BB96] to non-linear spaces, which makes it a natural approach if no domain-specific knowledge is available. A related result is the lower bound by [FKSLX21], which proves that without Bellman closure learning is hard

[1] For instance, the paper [FRTL20] only shows validity of ther intervals, not a performance bound; on the other hand, the paper [JH20] gives analyses at the population level, and so does not address the alignment of weight functions with respect to the dataset in the construction of the empirical estimator, which we do via self-normalization and regularization. This precludes obtaining the same type of guarantees that we present here.

even with small density ratios. In contrast, our work shows that learning is still possible even with large density ratios.

4. Finally, our procedure inherits some form of "multiple robustness". For example, the two test classes corresponding to Bellman completeness and marginalized importance weights can be used together, and guarantees will be obtained if *either* Bellman completeness holds or the importance weights are correct. We examine this issue in Section 4.4.

Linear setting We examine in depth an application to the linear setting, where we propose the first *computationally tractable* policy optimization procedure *without assuming Bellman completeness*. The closest result here is given in the paper [ZWB21], which holds under Bellman closure. Our procedure can be thought of making use of LSTD-type estimates so as to establish confidence intervals for the projected Bellman equations, and then using an iterative scheme for policy improvement.

2 Background and set-up

We begin with some notation used throughout the paper. For a given probability distribution ρ over a space $\mathcal{X}$, we define the $L^2(\rho)$-inner product and semi-norm as $\langle f_1, f_2 \rangle_\rho = \mathbb{E}_\rho[f_1 f_2]$, and $\|f_1\|_\rho = \sqrt{\langle f_1, f_1 \rangle_\rho}$. The identity function that returns one for every input is denoted by $\mathbb{1}$. We frequently use notation such as $c, c', \tilde{c}, c_1, c_2$ etc. to denote constants that can take on different values in different sections of the paper.

2.1 Markov decision processes and Bellman errors

We focus on infinite-horizon discounted Markov decision processes [Put94, BT96, SB18] with discount factor $\gamma \in [0, 1)$, state space $\mathcal{S}$, and an action set $\mathcal{A}$. For each state-action pair (s, a), there is a reward distribution $R(s, a)$ supported in $[0, 1]$ with mean $r(s, a)$, and a transition $\mathbb{P}(\cdot \mid s, a)$.

A (stationary) stochastic policy π maps states to actions. For a given policy, its Q-function is the discounted sum of future rewards based on starting from the pair (s, a), and then following the policy π in all future time steps $Q^\pi(s, a) = r(s, a) + \sum_{h=0}^{\infty} \gamma^h \mathbb{E}[r_h(S_h, A_h) \mid (S_0, A_0) = (s, a)]$, where the expectation is taken over trajectories with $A_h \sim \pi(\cdot \mid S_h)$, and $S_{h+1} \sim \mathbb{P}(\cdot \mid S_h, A_h)$ for $h = 1, 2, \ldots$. We also use $Q^\pi(s, \pi) = \mathbb{E}_{A \sim \pi(\cdot|s)} Q^\pi(s, A)$ and define the *Bellman evaluation operator* as $(\mathcal{T}^\pi Q)(s, a) = r(s, a) + \mathbb{E}_{S^+ \sim \mathbb{P}(\cdot|s,a)} Q(S^+, \pi)$. The value function satisfies $V^\pi(s) = Q^\pi(s, \pi)$. In our analysis, we assume that policies have action-value functions that satisfy the uniform bound $\sup_{(s,a)} |Q^\pi(s, a)| \leq 1$. We are also interested in approximating optimal policies, whose value and action-value functions are defined as $V^\star(s) = V^{\pi^\star}(s) = \sup_\pi V^\pi(s)$ and $Q^\star(s, a) = Q^{\pi^\star}(s, a) = \sup_\pi Q^\pi(s, a)$.

We assume that the starting state S_0 is drawn according to ν_{start} and study $V^\pi = \mathbb{E}_{S_0 \sim \nu_{\text{start}}}[V^\pi(S_0)]$. We define the *discounted occupancy measure* associated with a policy π as the distribution over the state action space $d_\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h[(S_h, A_h) = (s, a)]$. We adopt the shorthand notation $\mathbb{E}_\pi$ for expectations over d_π. For any functions $f, g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we make frequent use of the shorthands $\mathbb{E}_\pi[f] \stackrel{def}{=} \mathbb{E}_{(S,A) \sim d_\pi}[f(S, A)]$, and $\langle f, g \rangle_\pi \stackrel{def}{=} \mathbb{E}_{(S,A) \sim d_\pi}\big[f(S, A)\, g(S, A)\big]$. Note moreover that we have $\langle \mathbb{1}, f \rangle_\pi = \mathbb{E}_\pi[f]$ where $\mathbb{1}$ denotes the identity function.

For a given Q-function and policy π, let us define the *temporal difference error* (or TD error) associated with the sample $z = (s, a, r, s^+)$ and the *Bellman error* at (s, a)

$$(\delta^\pi Q)(z) \stackrel{def}{=} Q(s,a) - r - \gamma Q(s^+, \pi), \qquad (\mathcal{B}^\pi Q)(s,a) \stackrel{def}{=} Q(s,a) - r(s,a) - \gamma \mathbb{E}_{s^+ \sim \mathbb{P}(s,a)} Q(s^+, \pi).$$

The TD error is a random variable function of z, while the Bellman error is its conditional expectation with respect to the immediate reward and successor state at (s, a). Many of our bounds involve the quantity $\mathbb{E}_\pi \mathcal{B}^\pi Q = \mathbb{E}_{(S,A)\sim d_\pi}\big[\mathcal{B}^\pi Q(S, A)\big]$.

2.2 Function Spaces and Weak Representation

Our methods involve three different types of function spaces, corresponding to policies, action-value functions, and test functions. A test function f is a mapping $(s, a, o) \mapsto f(s, a, o)$ such that $\sup_{(s,a,o)}|f(s, a, o)| \leq 1$, where o is an optional identifier containing side information. Our methodology involves the following three function classes:

- a *policy class* Π that contains all policies π of interest (for evaluation or optimization);
- for each π, the *predictor class* $\mathcal{Q}^\pi$ of action-value functions Q that we permit; and
- for each π, the *test function class* $\mathcal{F}^\pi$ that we use to enforce the Bellman residual constraints.

We use the shorthands $\mathcal{Q} = \cup_{\pi\in\Pi}\mathcal{Q}^\pi$ and $\mathcal{F} = \cup_{\pi\in\Pi}\mathcal{F}^\pi$. We assume *weak realizability*:

Assumption 1 (Weak Realizability)**.** *For a given policy π, the predictor class $\mathcal{Q}^\pi$ is weakly realizable with respect to the test space $\mathcal{F}^\pi$ and the measure μ if there exists a predictor $Q^\pi_\star \in \mathcal{Q}^\pi$ such that*

$$\langle f, \mathcal{B}^\pi Q^\pi_\star\rangle_\mu = 0 \text{ for all } f \in \mathcal{F}^\pi \qquad \text{and} \qquad \langle \mathbb{1}, \mathcal{B}^\pi Q^\pi_\star\rangle_\pi = 0. \tag{1}$$

The first condition requires the predictor to satisfy the Bellman equations *on average*. The second condition amounts to requiring that the predictor returns the value of π at the start distribution: using Lemma 9 stated in the sequel, we have

$$\mathbb{E}_{S\sim\nu_{\text{start}}} Q^\pi_\star(S,\pi) - V^\pi = \mathbb{E}_{S\sim\nu_{\text{start}}}[Q^\pi_\star - Q^\pi](S,\pi) = \frac{1}{1-\gamma}\mathbb{E}_\pi \mathcal{B}^\pi Q^\pi_\star = \frac{1}{1-\gamma}\langle \mathbb{1}, \mathcal{B}^\pi Q^\pi_\star\rangle_\pi = 0.$$

This weak notion should be contrasted with *strong realizability*, which requires a function $Q^\pi \in \mathcal{Q}^\pi$ that satisfies the Bellman equation in all state-action pairs.

A stronger assumption that we sometime use is Bellman closure, which requires that $\mathcal{T}^\pi(Q) \in \mathcal{Q}^\pi$ for all $Q \in \mathcal{Q}^\pi$. The corresponding 'weak' version is given in Section 6.4.

3 Policy Estimates via the Weak Bellman Equations

In this section, we introduce our high-level approach, first at the population level and then in terms of regularized/normalized sample-based approximations.

3.1 Weak Bellman equations, empirical approximations and confidence intervals

We begin by noting that the predictor Q^π satisfies the Bellman equations everywhere in the state-action space, i.e., $\mathcal{B}^\pi Q^\pi = 0$. However, if our dataset is "small" relative to the complexity of (functions) on the state-action space, then it is unrealistic to enforce such a stringent condition.

Instead, the idea is to control the Bellman error in a weighted-average sense, where the weights are given by a set of *test functions*. At the idealized population level (corresponding to an infinite sample size), we consider predictors that satisfy the conditions

$$\langle f, \mathcal{B}^\pi Q \rangle_\mu = 0, \qquad \text{for all } f \in \mathcal{F}^\pi. \tag{2}$$

where $\mathcal{F}^\pi$ is a user-defined set of test functions. The two main challenges here are how to use data to enforce an approximate version of such constraints (along with rigorous data-dependent guarantees), and how to design the test function space. We begin with the former challenge.

Construction of the empirical set Given a dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i^+, o_i)\}_{i=1}^n$, we can approximate the Bellman errors by a linear combination of the temporal difference errors:

$$\int f(s,a) \underbrace{[Q(s,a) - (\mathcal{T}^\pi Q)(s,a)]}_{= \mathcal{B}^\pi Q(s,a)} d\mu \approx \frac{1}{n} \sum_{i=1}^n f(s_i, a_i) \underbrace{[Q(s_i, a_i) - r_i - \gamma Q(s_i^+, \pi)]}_{= \delta^\pi Q(s_i, a_i, r_i, s_i^+, o_i)}. \tag{3}$$

Note that the approximation (3) corresponds to a weighted linear combination of temporal differences. Written more compactly in inner product notation, equation (3) reads $\langle f, \mathcal{B}^\pi Q \rangle_\mu \approx \langle f, \delta^\pi Q \rangle_n$, where $\langle f, g \rangle_n = \frac{1}{n} \sum_{(s,a,r,s^+,o) \in \mathcal{D}} (fg)(s,a,r,s^+,o)$.

In general, the action value function Q^π does not satisfy $\langle f, \, \delta^\pi Q^\pi \rangle_n = 0$ because the empirical approximation (3) involves sampling error. For these reasons, in order to (approximately) identify Q^π, we impose only inequalities. Given a class of test functions $\mathcal{F}^\pi$, a radius parameter $\rho \geq 0$ and regularization parameter $\lambda \geq 0$, we define the set

$$\widehat{\mathcal{C}}_n^\pi(\rho, \lambda; \mathcal{F}^\pi) \overset{def}{=} \left\{ Q \in \mathcal{Q}^\pi \quad \text{such that} \quad \frac{|\langle f, \delta^\pi Q \rangle_n|}{\sqrt{\|f\|_n^2 + \lambda}} \leq \sqrt{\frac{\rho}{n}} \quad \text{for all } f \in \mathcal{F}^\pi \right\}. \tag{4}$$

When the choices of (ρ, λ) are clear from the context, we adopt the shorthand $\widehat{\mathcal{C}}_n^\pi(\mathcal{F}^\pi)$, or $\widehat{\mathcal{C}}_n^\pi$ when the function class $\mathcal{F}^\pi$ is also clear. If $\mathcal{F}^\pi$ and $\mathcal{Q}^\pi$ have finite cardinality, $\rho \approx \ln |\mathcal{F}^\pi| |\mathcal{Q}^\pi| + \ln 1/\delta$, where δ is a prescribed failure probability.

Our definition of the empirical constraint set (4) has two key components: first, the division by $\sqrt{\|f\|_n^2 + \lambda}$ corresponds to a form of *self-normalization*, whereas the addition of λ corresponds to a form of *regularization*. Self-normalization is needed so that the constraints remain suitably scale-invariant. More importantly—in conjunction with the regularization—it ensures that test functions that have poor coverage under the dataset do not have major effects on the solution. In particular, the empirical norm $\|f\|_n^2$ in the self-normalization measures how well the given test function is covered by the dataset. Any test function with poor coverage (i.e., $\|f\|_n^2 \approx 0$) will not yield useful information, and the regularization counteracts its influence. In our guarantees, the choices of λ and ρ are critical; as shown in our theory, we typically have $\lambda = \rho/n$, where ρ scales with the metric entropy of the predictor, test and policy spaces. Disregarding ρ, the right-hand side of the constraint decays as $1/\sqrt{n}$, so that the constraints are enforced more tightly as the sample size increases.

Confidence bounds and policy optimization: First, for any fixed policy π, we can use the feasibility set (4) to compute the lower and upper estimates

$$\widehat{V}^{\pi}_{\min} \overset{def}{=} \min_{Q \in \widehat{\mathcal{C}}^{\pi}_n(\rho,\lambda;\mathcal{F}^{\pi})} \mathbb{E}_{S\sim\nu_{\text{start}}}\big[Q(S,\pi)\big], \text{ and } \widehat{V}^{\pi}_{\max} \overset{def}{=} \max_{Q \in \widehat{\mathcal{C}}^{\pi}_n(\rho,\lambda;\mathcal{F}^{\pi})} \mathbb{E}_{S\sim\nu_{\text{start}}}\big[Q(S,\pi)\big], \tag{5}$$

corresponding to estimates of the minimum and maximum value that the policy π can take at the initial distribution. The family of lower estimates can be used to perform policy optimization over the class Π, in particular by solving the *max-min* problem

$$\max_{\pi\in\Pi}\Big[\min_{Q\in\widehat{\mathcal{C}}^{\pi}_n} \mathbb{E}_{S\sim\nu_{\text{start}}} Q(S,\pi)\Big], \qquad \text{or equivalently} \qquad \max_{\pi\in\Pi} \widehat{V}^{\pi}_{\min}. \tag{6}$$

Form of guarantees Let us now specify and discuss the types of guarantees that we establish for our estimators (5) and (6). All of our theoretical guarantees involve a μ-based counterpart $\mathcal{C}^{\pi}_n$ of the data-dependent set $\widehat{\mathcal{C}}^{\pi}_n$. More precisely, we define the population set

$$\mathcal{C}^{\pi}_n(4\rho,\lambda;\mathcal{F}^{\pi}) \overset{def}{=} \Big\{ Q \in \mathcal{Q}^{\pi} \quad \text{such that} \quad \frac{|\langle f, \mathcal{B}^{\pi}Q\rangle_{\mu}|}{\sqrt{\|f\|^2_{\mu}+\lambda}} \leq \sqrt{\frac{4\rho}{n}} \qquad \text{for all } f\in\mathcal{F}\Big\}, \tag{7}$$

where $\langle f, g\rangle_{\mu} \overset{def}{=} \int f(s,a)g(s,a)d\mu$ is the inner product induced by a distribution[2] μ over (s,a). As before, we use the shorthand notation $\mathcal{C}^{\pi}_n$ when the underlying arguments are clear from context. Moreover, in the sequel, we generally ignore the constant 4 in the definition (7) by assuming that ρ is rescaled appropriately—e.g., that we use a factor of $\frac{1}{4}$ in defining the empirical set.

It should be noted that in contrast to the set $\widehat{\mathcal{C}}^{\pi}_n$, the set $\mathcal{C}^{\pi}_n$ is *non-random* and it is defined in terms of the distribution μ and the input space $(\Pi,\mathcal{F},\mathcal{Q})$. It relaxes the orthogonality constraints in the weak Bellman formulation (2). Our guarantees for off-policy confidence intervals take the following form:

$$\underline{\text{Coverage guarantee:}} \qquad \big[\widehat{V}^{\pi}_{\min}, \widehat{V}^{\pi}_{\max}\big] \ni V^{\pi}. \tag{8a}$$

$$\underline{\text{Width bound:}} \qquad \max\Big\{|\widehat{V}^{\pi}_{\min}-V^{\pi}|,\ |\widehat{V}^{\pi}_{\max}-V^{\pi}|\Big\} \leq \frac{1}{1-\gamma}\max_{Q\in\mathcal{C}^{\pi}_n(\mathcal{F}^{\pi})} |\mathbb{E}_{\pi}\mathcal{B}^{\pi}Q|. \tag{8b}$$

Turning to policy optimization, let $\widetilde{\pi}$ be a solution to the max-min criterion (6). Then we prove a result of the following type:

$$\underline{\text{Oracle inequality:}} \qquad V^{\widetilde{\pi}} \geq \max_{\pi\in\Pi}\Big\{ \underbrace{V^{\pi}}_{\text{Value}} - \underbrace{\frac{1}{1-\gamma}\max_{Q\in\mathcal{C}^{\pi}_n(\mathcal{F})}|\mathbb{E}_{\pi}\mathcal{B}^{\pi}Q|}_{\text{Evaluation uncertainty}}\Big\}. \tag{9}$$

Note that this result guarantees that the estimator competes against an oracle that can search over all policies, and select one based on the optimal trade-off between its value and evaluation uncertainty.

[2]See Section 7.2.1 for a precise definition of the relevant μ for a fairly general sampling model.

3.2 High-probability guarantees

In this section, we present some high-probability guarantees. So as to facilitate understanding under space constraints, we state here results under simplifying assumptions: (a) the dataset originates from a fixed distribution, and (b) the classes $\Pi, \mathcal{F}$ and $\mathcal{Q}$ have finite cardinality. We emphasize that Section 7 provides a far more general version of this result, with an extremely flexible sampling model, and involving metric entropies of parametric or non-parametric function classes.

Assumption 2 (I.i.d. dataset)**.** *An i.i.d. dataset is a collection* $\mathcal{D} = \{(s_i, a_i, r_i, s_i^+, o_i)\}_{i=1}^n$ *such that for each* $i = 1, \ldots, n$ *we have* $(s_i, a_i, o_i) \sim \mu$ *and conditioned on* (s_i, a_i, o_i)*, we observe a noisy reward* $r_i = r(s_i, a_i) + \eta_i$ *with* $\mathbb{E}[\eta_i \mid \mathcal{F}_i] = 0$, $|r_i| \leq 1$ *and the next state* $s_i^+ \sim \mathbb{P}(s_i, a_i)$.

Theorem 1 (Guarantees for finite classes)**.** *Consider a triple* $(\Pi, \mathcal{F}, \mathcal{Q})$ *that is weakly Bellman realizable (Assumption 1); an i.i.d. dataset (Assumption 2); and the choices* $\rho = c\big\{\log(|\mathcal{F}||\Pi||\mathcal{Q}|) + \log(1/\delta)\big\}$ *and* $\lambda = c'\rho/n$ *for some constants* c, c'*. Then w.p. at least* $1 - \delta$*:*

- *Policy evaluation: For any* $\pi \in \Pi$*, the estimates* $(\widehat{V}^\pi_{min}, \widehat{V}^\pi_{max})$ *specify a confidence interval satisfying the coverage* (8a) *and width bounds* (8b)
- *Policy optimization: Any max-min policy* (6) $\widetilde{\pi}$ *satisfies the oracle inequality* (9)*.*

4 Concentrability Coefficients and Test Spaces

In this section, we develop some connections to concentrability coefficients that have been used in past work, and discuss various choices of the test class. Like the predictor class $\mathcal{Q}^\pi$, the test class $\mathcal{F}^\pi$ encodes domain knowledge, and thus its choice is delicate. Different from the predictor class, the test class does not require a 'realizability' condition. As a general principle, the test functions should be chosen as orthogonal as possible with respect to the Bellman residual, so as to enable rapid progress towards the solution; at the same time, they should be sufficiently "aligned" with the dataset, meaning that $\|f\|_\mu$ or its empirical counterpart $\|f\|_n$ should be large. Given a test class, each additional test function posits a new constraint which helps identify the correct predictor; at the same time, it increases the metric entropy (parameter ρ), which makes each individual constraints more loose. In summary, there are trade-offs to be made in the selection of the test class $\mathcal{F}$, much like $\mathcal{Q}$.

In order to assess the statistical cost that we pay for off-policy data, it is natural to define the *off-policy cost coefficient* (OPC) as

$$K^\pi(\mathcal{C}^\pi_n, \rho, \lambda) \overset{def}{=} \max_{Q \in \mathcal{C}^\pi_n} \frac{|\mathbb{E}_\pi \mathcal{B}^\pi Q|^2}{(1+\lambda)\frac{\rho}{n}} = \max_{Q \in \mathcal{C}^\pi_n} \frac{\langle \mathbb{1}, \mathcal{B}^\pi Q \rangle_\mu^2}{(1+\lambda)\frac{\rho}{n}}, \tag{10}$$

With this notation, our off-policy width bound (8b) can be re-expressed as

$$|\widehat{V}^\pi_{\text{min}} - \widehat{V}^\pi_{\text{max}}| \leq 2\frac{\sqrt{1+\lambda}}{1-\gamma}\sqrt{K^\pi \frac{\rho}{n}}, \tag{11a}$$

while the oracle inequality (9) for policy optimization can be re-expressed in the form

$$V^{\widetilde{\pi}} \geq \max_{\pi \in \Pi} \Big\{ V^\pi - \frac{\sqrt{1+\lambda}}{1-\gamma}\sqrt{K^\pi \frac{\rho}{n}} \Big\}, \tag{11b}$$

Since $\lambda \sim \rho/n$, the factor $\sqrt{1+\lambda}$ can be bounded by a constant in the typical case $n \geq \rho$. We now offer concrete examples of the OPC , while deferring further examples to Section 6.5.

4.1 Likelihood ratios

Our broader goal is to obtain small Bellman error along the distribution induced by π. Assume that one constructs a test function class $\mathcal{F}^\pi$ of possible likelihood ratios.

Proposition 1 (Likelihood ratio bounds). *Assume that for some constant b_π, the test function defined as $f^*(s,a) = \frac{1}{b_\pi}\frac{d_\pi(s,a)}{\mu(s,a)}$ belongs to $\mathcal{F}^\pi$ and satisfies $\|f^*\|_\infty \le 1$. Then the OPC coefficient satisfies*

$$K^\pi \overset{(i)}{\le} \frac{\mathbb{E}_\pi\Big[\frac{d_\pi(S,A)}{\mu(S,A)}\Big] + b_\pi^2\lambda}{1+\lambda} \overset{(ii)}{\le} \frac{b_\pi\big(1+b_\pi\lambda\big)}{1+\lambda} \tag{12}$$

The proof is in Section 9.1. Since $\lambda = \lambda_n \to 0$ as n increases, the OPC coefficient is bounded by a multiple of the expected ratio $\mathbb{E}_\pi\Big[\frac{d_\pi(S,A)}{\mu(S,A)}\Big]$. Up to an additive offset, this expectation is equivalent to the χ^2-distribution between the policy-induced occupation measure d_π and data-generating distribution μ. The concentrability coefficient can be plugged back into Eqs. (11a) and (11b) to obtain a concrete policy optimization bound. In this case, we recover a result similar to [XJ20b], but with a much milder concentrability coefficient that involves only the chosen comparator policy.

4.2 The error test space

We now turn to the discussion of a choice for the test space that extends the LSTD algorithm to non-linear spaces. A simplification to the linear setting is presented later in Section 5.

As is well known, the LSTD algorithm [BB96] can be seen as minimizing the Bellman error projected onto the linear prediction space Q. Define the transition operator $(\mathbb{P}^\pi Q)(s,a) = \mathbb{E}_{s^+\sim\mathbb{P}(s,a)}Q(s^+,\pi)$, and the prediction error $\epsilon = Q - Q^\pi_\star$, where $Q^\pi_\star$ is a Q-function from the definition of weak realizability. The Bellman error can be re-written as $\mathcal{B}^\pi Q = \mathcal{B}^\pi Q - \mathcal{B}^\pi Q^\pi_\star = (\mathcal{I}-\gamma\mathbb{P}^\pi)\epsilon$. When realizability holds, in the linear setting and at the population level, the LSTD solution seeks to satisfy the projected Bellman equations

$$\langle f, \mathcal{B}^\pi Q\rangle_\mu = 0, \quad \text{for all } f \in \mathcal{E}^\pi_\star. \tag{13}$$

In the linear case, $\mathcal{E}^\pi_\star$ is the class of linear functions $\mathcal{Q}^\pi$ used as predictors; when $\mathcal{Q}^\pi$ is non-linear, we can extend the LSTD method by using the (nonlinear) error test space $\mathcal{F}^\pi = \mathcal{E}^\pi_\star = \{Q - Q^\pi_\star\}$. Since $\mathcal{E}^\pi_\star$ is unknown (as it depends on the weak solution $Q^\pi_\star$), we choose instead the larger class

$$\mathcal{E}^\pi = \{Q - Q' \mid Q, Q' \in \mathcal{Q}^\pi\},$$

which contains $\mathcal{E}^\pi_\star$. The resulting approach can be seen as performing a projection of the Bellman operator $\mathcal{B}^\pi Q$ into the error space $\mathcal{E}^\pi_\star$, much like LSTD does in the linear setting. However, different from LSTD, our procedure returns confidence intervals as opposed to a point estimator. This choice of the test space is related to the Bubnov-Galerkin method [Rep17] for linear spaces; it selects the test space $\mathcal{F}^\pi$ to be identical to the trial space $\mathcal{E}^\pi_\star$ that contains all possible solution errors.

Lemma 1 (OPC coefficient from prediction error). *For any test function class $\mathcal{F}^\pi \supseteq \mathcal{E}^\pi$, we have*

$$K^\pi \le \max_{Q\in\mathcal{Q}^\pi}\Big\{\frac{\|\epsilon\|^2_\mu + \lambda}{\|\mathbb{1}\|^2_\pi + \lambda}\,\frac{\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle^2_\pi}{\langle \epsilon, \mathcal{B}^\pi Q\rangle^2_\mu}\Big\} = \max_{\epsilon\in\mathcal{E}^\pi_\star}\Big\{\frac{\|\epsilon\|^2_\mu + \lambda}{\|\mathbb{1}\|^2_\pi + \lambda}\,\frac{\langle \mathbb{1}, (\mathcal{I}-\gamma\mathbb{P}^\pi)\epsilon\rangle^2_\pi}{\langle \epsilon, (\mathcal{I}-\gamma\mathbb{P}^\pi)\epsilon\rangle^2_\mu}\Big\}. \tag{14}$$

The above coefficient measures the ratio between the Bellman error along the distribution of the target policy π and that projected onto the error space $\mathcal{E}^{\pi}_{\star}$ defined by $\mathcal{Q}^{\pi}$. It is a concentrability coefficient that *always* applies, as the choice of the test space does not require domain knowledge. See Section 9.2 for the proof, and Section 6.6 for further comments and insights, as well as a simplification in the special case of Bellman closure.

4.3 The Bellman test space

In the prior section we controlled the projected Bellman error. Another longstanding approach in reinforcement learning is to control the Bellman error itself, for example by minimizing the squared Bellman residual. In general, this cannot be done if only an offline dataset is available due to the well known *double sampling* issue. However, in some cases we can use an helper class to try to capture the Bellman error. Such class needs to be a superset of the class of *Bellman test functions* given by

$$\mathcal{F}^{\mathcal{B}}_{\pi} \stackrel{def}{=} \{\mathcal{B}^{\pi}Q \mid Q \in \mathcal{Q}^{\pi}\}. \tag{15}$$

Any test class that contains the above allows us to control the Bellman residual, as we show next.

Lemma 2 (Bellman Test Functions). *For any test function class $\mathcal{F}^{\pi}$ that contains $\mathcal{F}^{\mathcal{B}}_{\pi}$, we have*

$$\|\mathcal{B}^{\pi}Q\|_{\mu} \leq c_1\sqrt{\frac{\rho}{n}} \qquad \text{for any } Q \in \mathcal{C}^{\pi}_{n}(\mathcal{F}^{\pi}). \tag{16a}$$

Moreover, the off-policy cost coefficient is upper bounded as

$$K^{\pi} \stackrel{(i)}{\leq} c_1 \sup_{Q\in\mathcal{Q}^{\pi}} \frac{\langle \mathbb{1}, \mathcal{B}^{\pi}Q\rangle^2_{\pi}}{\|\mathcal{B}^{\pi}Q\|^2_{\mu}} \stackrel{(ii)}{\leq} c_1 \sup_{Q\in\mathcal{Q}^{\pi}} \frac{\|\mathcal{B}^{\pi}Q\|^2_{\pi}}{\|\mathcal{B}^{\pi}Q\|^2_{\mu}} \stackrel{(iii)}{\leq} c_1 \sup_{(s,a)} \frac{d_{\pi}(s,a)}{\mu(s,a)}. \tag{16b}$$

See Section 9.4 for the proof of this claim.

Consequently, whenever the test class includes the Bellman test functions, the off-policy cost coefficient is at most the ratio between the squared Bellman residuals along the data generating distribution and the target distribution. If Bellman closure holds, then the prediction error space $\mathcal{E}^{\pi}$ introduced in Section 4.2 contains the Bellman test functions: for $Q \in \mathcal{Q}^{\pi}$, we can write $\mathcal{B}^{\pi}Q = Q - \mathcal{T}^{\pi}Q \in \mathcal{E}^{\pi}$. This fact allows us to recover a result in the recent paper [XCJ+21] in the special case of Bellman closure, although the approach presented here is more general.

4.4 Combining test spaces

Often, it is natural to construct a test space that is a union of several simpler classes. A simple but valuable observation is that the resulting procedure inherits the best of the OPC coefficients. Suppose that we are given a collection $\{\mathcal{F}^{\pi}_{m}\}_{m=1}^{M}$ of M different test function classes, and define the union $\mathcal{F}^{\pi} = \bigcup_{m=1}^{M} \mathcal{F}^{\pi}_{m}$. For each $m = 1, \ldots, M$, let K^{π}_{m} be the OPC coefficient defined by the function class $\mathcal{F}^{\pi}_{m}$ and radius ρ, and let $K^{\pi}(\mathcal{F})$ be the OPC coefficient associated with the full class. Then we have the following guarantee:

Lemma 3 (Multiple test classes). $K^{\pi}(\mathcal{F}) \leq \min_{m=1,\ldots,M} K^{\pi}_{m}$.

This guarantee is a straightforward consequence of our construction of the feasibility sets: in particular, we have $\mathcal{C}_n^\pi(\mathcal{F}) = \cap_{m=1}^M \mathcal{C}_n^\pi(\mathcal{F}_m)$, and consequently, by the variational definition of the off-policy cost coefficient $K^\pi(\mathcal{F})$ as optimization over $\mathcal{C}_n^\pi(\mathcal{F})$, the bound (3) follows. In words, when multiple test spaces are combined, then our algorithms inherit the best (smallest) OPC coefficient over all individual test spaces. While this behavior is attractive, one must note that there is a statistical cost to using a union of test spaces: the choice of ρ scales as a function of $\mathcal{F}$ via its metric entropy. This increase in ρ must be balanced with the benefits of using multiple test spaces.[3]

5 Linear Setting

In this section, we turn to a detailed analysis of our estimators using function classes that are linear in a feature map. Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a given feature map, and consider linear expansions $g_w(s,a) \overset{def}{=} \langle w,\, \phi(s,a)\rangle = \sum_{j=1}^d w_j \phi_j(s,a)$. The class of *linear functions* takes the form

$$\mathcal{L} \overset{def}{=} \{(s,a) \mapsto g_w(s,a) \mid w \in \mathbb{R}^d,\ \|w\|_2 \leq 1\}. \tag{17}$$

Throughout our analysis, we assume that $\|\phi(s,a)\|_2 \leq 1$ for all state-action pairs.

Following the approach in Section 4.2, which is based on the LSTD method, we should choose the test function class $\mathcal{F}^\pi = \mathcal{L}$, as in the linear case the prediction error is linear.

In order to obtain a computationally efficient implementation, we need to use a test class that is a "simpler" subset of $\mathcal{L}$. In particular, for linear functions, it is not hard to show that the estimates $\widehat{V}^\pi_{\min}$ and $\widehat{V}^\pi_{\max}$ from equation (5) can be computed by solving a quadratic program, with two linear constraints for each test function. (See Section 6.8 for the details.) Consequently, the computational complexity scales linearly with the number of test functions. Thus, if we restrict ourselves to a finite test class contained within $\mathcal{L}$, we will obtain a computationally efficient approach.

5.1 A computationally friendly test class and OPC coefficients

Define the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \phi_i \phi_i^T$ where $\phi_i \overset{def}{=} \phi(s_i, a_i)$. Let $\{\widehat{u}_j\}_{j=1}^d$ be the eigenvectors of empirical covariance matrix $\widehat{\Sigma}$, and suppose that they are normalized to have unit ℓ_2-norm. We use these normalized eigenvectors to define the finite test class

$$\widetilde{\mathcal{F}}^\pi \overset{def}{=} \{f_j, j = 1, \ldots, d\} \quad \text{where } f_j(s,a) \overset{def}{=} \langle \widehat{u}_j,\, \phi(s,a)\rangle \tag{18}$$

A few observations are in order:

- This test class has only d functions, so that our QP implementation has $2d$ constraints, and can be solved in polynomial time. (Again, see Section 6.8 for details.)
- Since $\widetilde{\mathcal{F}}^\pi$ is a subset of $\mathcal{L}$ the choice of radius $\rho = c(\frac{d}{n} + \log 1/\delta)$ is valid for some constant c.

[3]For space reasons, we defer to Section 6.7 an application in which we construct a test function space as a union of subclasses, and thereby obtain a method that automatically leverages Bellman closure when it holds, falls back to importance sampling if closure fails, and falls back to a worst-case bound in general.

Concentrability When weak Bellman closure does not hold, then our analysis needs to take into account how errors propagate via the dynamics. In particular, we define the *next-state feature extractor* $\phi^{+\pi}(s,a) \overset{def}{=} \mathbb{E}_{s^+\sim\mathbb{P}(s,a)}\phi(s^+,\pi)$, along with the population covariance matrix $\Sigma \overset{def}{=} \mathbb{E}_\mu\big[\phi(s,a)\phi^\top(s,a)\big]$, and its λ-regularized version $\Sigma_\lambda \overset{def}{=} \Sigma + \lambda I$. We also define the matrices

$$\Sigma^{+\pi} \overset{def}{=} \mathbb{E}_\mu[\phi(\phi^{+\pi})^\top], \quad \Sigma^{+\pi}_{\lambda,\text{Boot}} \overset{def}{=} (\Sigma_\lambda^{\frac{1}{2}} - \gamma\Sigma_\lambda^{-\frac{1}{2}}\Sigma^{+\pi})^\top(\Sigma_\lambda^{\frac{1}{2}} - \gamma\Sigma_\lambda^{-\frac{1}{2}}\Sigma^{+\pi}).$$

The matrix $\Sigma^{+\pi}$ is the cross-covariance between successive states, whereas the matrix $\Sigma^{+\pi}_{\lambda,\text{Boot}}$ is a suitably renormalized and symmetrized version of the matrix $\Sigma^{\frac{1}{2}} - \gamma\Sigma^{-\frac{1}{2}}\Sigma^{+\pi}$, which arises naturally from the policy evaluation equation. We refer to quantities that contain evaluations at the next-state (e.g., $\phi^{+\pi}$) as bootstrapping terms, and now bound the OPC coefficient in the presence of such terms:

Proposition 2 (OPC bounds with bootstrapping)**.** *Under weak realizability, we have*

$$K^\pi(\widetilde{\mathcal{F}}^\pi) \le c\, d\|\mathbb{E}_\pi[\phi - \gamma\phi^{+\pi}]\|^2_{(\Sigma^{+\pi}_{\lambda,Boot})^{-1}} \qquad \textit{with probability at least } 1-\delta. \tag{19}$$

See Section 10.1 for the proof. The bound (19) takes a familiar form, as it involves the same matrices used to define the LSTD solution. This is expected, as our approach here is essentially equivalent to the LSTD method; the difference is that LSTD only gives a point estimate as opposed to the confidence intervals that we present here; however, they are both derived from the same principle, namely from the Bellman equations projected along the predictor (error) space.

The bound quantifies how the feature extractor ϕ together with the bootstrapping term $\phi^{+\pi}$, averaged along the target policy π, interact with the covariance matrix with bootstrapping $\Sigma^{+\pi}_{\lambda,\text{Boot}}$. It is an approximation to the OPC coefficient bound derived in Lemma 1. The bootstrapping terms capture the temporal difference correlations that can arise in reinforcement learning when strong assumptions like Bellman closure do not hold. As a consequence, such an OPC coefficient being small is a *sufficient* condition for reliable off-policy prediction. This bound on the OPC coefficient always applies, and it reduces to the simpler one (20) when weak Bellman closure holds, with no need to inform the algorithm of the simplified setting; see Section 10.3 for the proof.

Proposition 3 (OPC bounds under weak Bellman Closure)**.** *Under Bellman closure, we have*

$$K^\pi(\widetilde{\mathcal{F}}^\pi) \le c\, d\|\mathbb{E}_\pi\phi\|^2_{\Sigma_\lambda^{-1}} \qquad \textit{with probability at least } 1-\delta. \tag{20}$$

5.2 Actor-critic scheme for policy optimization

Having described a practical procedure to compute $\widehat{V}^\pi_{\min}$, we now turn to the computation of the max-min estimator for policy optimization. We define the *soft-max policy class*

$$\Pi_{\text{lin}} \overset{def}{=} \Big\{(s,a) \mapsto \frac{e^{\langle\phi(s,a),\theta\rangle}}{\sum_{a^+\in\mathcal{A}} e^{\langle\phi(s,a^+),\theta\rangle}} \mid \|\theta\|_2 \le T,\ \theta\in\mathbb{R}^d\Big\}. \tag{21}$$

In order to compute the max-min solution (6) over this policy class, we implement an actor-critic method, in which the actor performs a variant of mirror descent.[4]

[4] Strictly speaking, it is mirror ascent, but we use the conventional terminology.

- At each iteration $t = 1, \ldots, T$, the policy $\pi_t \in \Pi_{\text{lin}}$ can be identified with a parameter $\theta_t \in \mathbb{R}^d$. The sequence is initialized with $\theta_1 = 0$.
- Using the finite test function class (18) based on normalized eigenvectors, the pessimistic value estimate $\widehat{V}_{\min}^{\pi_t}$ is computed by solving a quadratic program, as previously described. This computation returns the weight vector w_t of the associated optimal action-value function.
- Using the action-value vector w_t, we update the actor's parameter as

$$\theta_{t+1} = \theta_t + \eta w_t \qquad \text{where } \eta = \sqrt{\tfrac{\log|\mathcal{A}|}{2T}} \text{ is a stepsize parameter.} \tag{22}$$

We now state a guarantee on the behavior of this procedure, based on two OPC coefficients:

$$K_{(1)}^{\widetilde{\pi}} = d\|\mathbb{E}_{\widetilde{\pi}}\phi\|^2_{\Sigma_\lambda^{-1}}, \quad \text{and} \quad K_{(2)}^{\widetilde{\pi}} = d \sup_{\pi \in \Pi} \Big\{ \|\mathbb{E}_{\widetilde{\pi}}[\phi - \gamma\phi^{+\pi}]\|^2_{(\Sigma^{+\pi}_{\lambda,\text{Boot}})^{-1}} \Big\}. \tag{23}$$

Moreover, in making the following assertion, we assume that every weak solution $Q_\star^\pi$ can be evaluated against the distribution of a comparator policy $\widetilde{\pi} \in \Pi$, i.e., $\langle \mathbb{1}, \mathcal{B}^\pi Q_\star^\pi \rangle_{\widetilde{\pi}} = 0$ for all $\pi \in \Pi$. (This assumption is still weaker than strong realizability).

Theorem 2 (Approximate Guarantees for Linear Soft-Max Optimization)**.** *Under the above conditions, running the procedure for T rounds returns a policy sequence $\{\pi_t\}_{t=1}^T$ such that, for any comparator policy $\widetilde{\pi} \in \Pi$,*

$$\frac{1}{T}\sum_{t=1}^{T}\big\{V^{\widetilde{\pi}} - V^{\pi_t}\big\} \le \frac{c_1}{1-\gamma}\Big\{ \underbrace{\sqrt{\frac{\log|\mathcal{A}|}{T}}}_{\textit{Optimization error}} + \underbrace{\sqrt{K_{(\cdot)}^{\widetilde{\pi}} \frac{d\log(nT) + \log\left(\frac{n}{\delta}\right)}{n}}}_{\textit{Statistical error}} \Big\}, \tag{24}$$

with probability at least $1-\delta$. This bound always holds with $K_{(\cdot)}^{\widetilde{\pi}} = K_{(2)}^{\widetilde{\pi}}$, and moreover, it holds with $K_{(\cdot)}^{\widetilde{\pi}} = K_{(1)}^{\widetilde{\pi}}$ when weak Bellman closure is in force.

See Section 11 for the proof. Whenever Bellman closure holds, the result automatically inherits the more favorable concentrability coefficient $K_{(2)}^{\widetilde{\pi}}$, as originally derived in Proposition 3. The resulting bound is only $\sqrt{d}$ worse than the lower bound recently established in the paper [ZWB21]. However, the method proposed here is robust, in that it provides guarantees even when Bellman closure does not hold. In this case, we have a guarantee in terms of the OPC coefficient $K_{(1)}^{\widetilde{\pi}}$. Note that it is a uniform version of the one derived previously in Proposition 2, in that there is an additional supremum over the policy class. This supremum arises due to the use of gradient-based method, which implicitly searches over policies in bootstrapping terms; see Section 6.9 for a more detailed discussion of this issue.

Acknowledgment

AZ was partially supported by NSF-FODSI grant 2023505. In addition, this work was partially supported by NSF-DMS grant 2015454, NSF-IIS grant 1909365, as well as Office of Naval Research grant DOD-ONR-N00014-18-1-2640 to MJW. The authors are grateful to Nan Jiang and Alekh Agarwal for pointing out further connections with the existing literature.

References

[AMS07] András Antos, Rémi Munos, and Csaba Szepesvári. Fitted Q-iteration in continuous action-space MDPs. 2007.

[ASM08] 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*, 71(1):89–129, 2008.

[ASN20] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020.

[BB96] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

[Ber95a] D. P. Bertsekas. *Dynamic programming and stochastic control*, volume 1. Athena Scientific, Belmont, MA, 1995.

[Ber95b] D.P. Bertsekas. *Dynamic programming and stochastic control*, volume 2. Athena Scientific, Belmont, MA, 1995.

[BGB20] Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

[BLL+11] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[BT96] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[CJ19] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051, 2019.

[CQ22] Xiaohong Chen and Zhengling Qi. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. *arXiv preprint arXiv:2201.06169*, 2022.

[DJL21] Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

[DW20] Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*, 2020.

[Eva10] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

[FCG18] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

[FGSM16] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

[FKSLX21] Dylan J. Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation, 2021.

[Fle84] Clive AJ Fletcher. Computational Galerkin methods. In *Computational Galerkin methods*, pages 72–85. Springer, 1984.

[FRTL20] Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR, 2020.

[FSM10] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[Gal15] Boris Grigoryevich Galerkin. Series solution of some problems of elastic equilibrium of rods and plates. *Vestnik inzhenerov i tekhnikov*, 19(7):897–908, 1915.

[Haz21] Elad Hazan. Introduction to online convex optimization, 2021.

[HJD+21] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping statistical inference for off-policy evaluation. *arXiv preprint arXiv:2102.03607*, 2021.

[JGS+19] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[JH20] Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

[JKA+17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[JL16] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[JYW21] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[K+03] Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

[KFTL19] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

[KHSL21] Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *Deep RL Workshop NeurIPS 2021*, 2021.

[KRNJ20] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

[KU19] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

[LLTZ18] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

[LP03] Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149, 2003.

[LSAB20] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

[LTDC19] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

[LTND21] Jonathan N Lee, George Tucker, Ofir Nachum, and Bo Dai. Model selection in batch policy optimization. *arXiv preprint arXiv:2112.12320*, 2021.

[MS08] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

[Mun03] Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

[Mun05] Rémi Munos. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

[ND20] Ofir Nachum and Bo Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[NDGL20] Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[NDK+19] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[Pre00] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[Rep17] Sergey Repin. One hundred years of the Galerkin method. *Computational Methods in Applied Mathematics*, 17(3):351–357, 2017.

[RZM+21] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

[SSB+20] Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

[TB16] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[TFL+19] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

[UHJ20] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[UIJ+21] Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

[US21] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage, 2021.

[VJY21] Cameron Voloshin, Nan Jiang, and Yisong Yue. Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1612–1620. PMLR, 2021.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[WFK20] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[WNŻ+20] Ziyu Wang, Alexander Novikov, Konrad Żołna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020.

[WTN19] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[XCJ+21] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

[XJ20a] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020.

[XJ20b] Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. volume 124 of *Proceedings of Machine Learning Research*, pages 550–559, Virtual, 03–06 Aug 2020. PMLR.

[XMW19] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9668–9678, 2019.

[YB10] H. Yu and D. P. Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010.

[YBW20] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*, 2020.

[YND+20] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

[YQCC21] Chao-Han Huck Yang, Zhengling Qi, Yifan Cui, and Pin-Yu Chen. Pessimistic model selection for offline deep reinforcement learning, 2021.

[YTY+20] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[YW20] Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

[YW21] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *arXiv preprint arXiv:2110.08695*, 2021.

[YWDW] Ming Yin, Yu-Xiang Wang, Yaqi Duan, and Mengdi Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism.

[Zan20] Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online RL. *arXiv preprint arXiv:2012.08005*, 2020.

[ZDLS20] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

[ZHH+22] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

[ZJZ21] Zihan Zhang, Xiangyang Ji, and Yuan Zhou. Almost optimal batch-regret tradeoff for batch linear contextual bandits, 2021.

[ZLW20] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pages 11194–11203. PMLR, 2020.

[ZSU+22] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block MDPs: A model-free representation learning approach, 2022.

[ZWB21] Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

Contents

6 Additional Discussion and Results

6.1 Bellman Residual Orthogonalization

Suppose that our goal is to estimate the action-value function Q^π of a given policy π. This function is known to be a fixed point of the Bellman evaluation operator $\mathcal{T}^\pi$ associated with the policy π. Thus, when the MDP is known, one option is to (approximately) solve the Bellman evaluation equations $Q(s,a) = (\mathcal{T}^\pi Q)(s,a)$ for all state-action pairs. However, even if function approximation for Q is implemented, it is still difficult to directly solve these equations if the state-action space is sufficiently complex.

This observation motivates the strategy taken in this paper: instead of enforcing the Bellman equations for all state-action pairs, suppose that we do so only in an average sense, and with respect to a certain set of functions. More formally, a *test function* is a mapping from the state-action space to the real line; any such function serves to enforce the Bellman equations in an average sense in the following way. Let $\mathcal{F}^\pi$ denote some user-prescribed class of test functions, which we refer to as the *test space*. Then for a given measure μ, we require only that the action-value function Q^π satisfy the integral constraints

$$\langle f, Q - \mathcal{T}^\pi(Q)\rangle_\mu \overset{def}{=} \int f(s,a)[Q(s,a) - (\mathcal{T}^\pi Q)(s,a)]d\mu = 0, \qquad \text{for all } f \in \mathcal{F}^\pi. \tag{25}$$

We refer to this design principle as *Bellman residual orthogonalization*, because it requires the Bellman error function to be orthogonal to a set of test functions, as measured under the $L^2(\mu)$ inner product. Of course, by enlarging the test space $\mathcal{F}^\pi$, the Bellman error is required to be orthogonal to more test functions, and it will ultimately be zero if enough test functions are added as constraints. But at the same time, as shown by our analysis, any such enlargement has both computational and statistical costs, so there are tradeoffs to be understood.

In numerical analysis, especially in solving partial differential equations, the design principle (25) is called the weak or variational formulation (e.g., [Eva10]), and its solutions are referred to as weak solutions. Here we are advocating a *weak formulation* of the Bellman equations. Of course, the constraints (25) are necessary but not sufficient: the *weak (Bellman) solutions* need not solve the Bellman equations. However, whenever we need to learn based on a limited dataset, it is unreasonable to satisfy the Bellman equations everywhere; instead, by choosing the test space appropriately, we can seek to satisfy the Bellman equations over regions of the state-action space that are most important. In some cases, the formulation (25) can be fruitfully viewed as a type of Galerkin approximation (e.g., [Gal15, Fle84]) to the Bellman equations. For example, when both the test functions and Q-value functions belong to some linear space (and the empirical constraints are enforced exactly), then the weak formulation and Galerkin approximation lead to the least-squares temporal difference (LSTD) estimator; this connection between Galerkin methods and LSTD has been noted in past work by Yu and Bertsekas [YB10]. In this paper, our goal is to understand the weak formulation (25) in a broader sense for general test and predictor classes.

6.2 Comparison with Weight Learning Methods

The work closest to ours is [JH20]. They also use an auxiliary weight function class, which is comparable to our test class. However, the test class is used in different ways; we compare them in this section at the population level.[5] Let us assume that weak realizability holds and that $\mathcal{F}$ is symmetric, i.e., if $f \in \mathcal{F}$ then $-f \in \mathcal{F}$ as well. At the population level, our program seeks to solve

$$\sup_{Q \in \mathcal{Q}^\pi} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) \quad \text{s.t.} \quad \sup_{f \in \mathcal{F}} \langle f, \mathcal{B}^\pi Q \rangle_\mu = 0, \tag{26}$$

which is equivalent for any $w \in \mathcal{F}$ to

$$\sup_{Q \in \mathcal{Q}^\pi} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) - \frac{1}{1-\gamma} \langle w, \mathcal{B}^\pi Q \rangle_\mu \quad \text{s.t.} \quad \sup_{f \in \mathcal{F}} \langle f, \mathcal{B}^\pi Q \rangle_\mu = 0.$$

Removing the constraints leads to the upper bound

$$\sup_{Q \in \mathcal{Q}^\pi} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) - \frac{1}{1-\gamma} \langle w, \mathcal{B}^\pi Q \rangle_\mu.$$

Since this is a valid upper bound for any $w \in \mathcal{F}$, minimizing over w must still yield an upper bound, which reads

$$\inf_{w \in \mathcal{F}} \sup_{Q \in \mathcal{Q}^\pi} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) - \frac{1}{1-\gamma} \langle w, \mathcal{B}^\pi Q \rangle_\mu.$$

This is the population program for "weight learning", as described in [JH20]. It follows that Bellman residual orthogonalization always produces tighter confidence intervals than "weight learning" at the population level.

Another interesting comparison is with "value learning", also described in [JH20]. In this case, assuming symmetric $\mathcal{F}$, we can equivalently express the population program (26) using a Lagrange multiplier as follows

$$\sup_{Q \in \mathcal{Q}^\pi} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) - \sup_{\lambda \geq 0, f \in \mathcal{F}} \lambda \langle f, \mathcal{B}^\pi Q \rangle_\mu. \tag{27}$$

Rearranging we obtain

$$\sup_{Q \in \mathcal{Q}^\pi} \inf_{\lambda \geq 0, f \in \mathcal{F}} \mathbb{E}_{s \sim \nu_{\text{start}}} Q(s, \pi) - \lambda \langle f, \mathcal{B}^\pi Q \rangle_\mu.$$

The "value learning" program proposed in [JH20] has a similar formulation to ours but differs in two key aspects. The first—and most important—is that [JH20] ignores the Lagrange multiplier; this means "value learning" is not longer associated to a constrained program. While the Lagrange multiplier could be "incorporated" into the test class $\mathcal{F}$, doing so would cause the entropy of $\mathcal{F}$ to be unbounded. Another point of difference is that "value learning" uses such expression with $\lambda = 1$ to derive the confidence interval *lower bound*, while we use it to construct the confidence interval *upper bound*. While this may seem like a contradiction, we notice that the expression is derived using different assumptions: we assume weak realizability of Q, while [JH20] assumes realizability of the density ratios between μ and the discounted occupancy measure π.

[5] The empirical estimator in [JH20] does not take into account the 'alignment' of each weight function with respect to the dataset, which we do through self-normalization and regularization in the construction of the empirical estimator. This precludes obtaining the same type of strong finite time guarantees that we are able to derive here.

6.3 Additional Literature

Here we summarize some additional literature. The efficiency of off-policy tabular RL has been investigated in the papers [YBW20, YW20, YW21]. For empirical studies on offline RL, see the papers [LTDC19, JGS+19, WTN19, ASN20, WNŻ+20, SSB+20, NDGL20, YQCC21, KHSL21, BGB20, KFTL19, KRNJ20, YTY+20].

Some of the classical RL algorithm are presented in the papers [Mun03, Mun05, AMS07, ASM08, FSM10, FGSM16]. For a more modern analysis, see [CJ19]. These works generally make additionally assumptions on top of realizability. Alternatively, one can use importance sampling [Pre00, TB16, JL16, FCG18]. A more recent idea is to look at the distributions themselves [LLTZ18, NDK+19, XMW19, ZDLS20, ZLW20, YND+20, KU19].

Offline policy optimization with pessimism has been studied in the papers [LSAB20, RZM+21, JYW21, XCJ+21, ZWB21, YWDW, US21]. There exists a fairly extensive literature on lower bounds with linear representations, including the two papers [Zan20, WFK20] that concurrently derived the first exponential lower bounds for the offline setting, and [FKSLX21] proves that realizability and coverage alone are insufficient.

In the context of off-policy optimization several works have investigated methods that assume only realizability of the optimal policy [XJ20a, XJ20b]. Related work includes the papers [DW20, DJL21, JH20, UHJ20, TFL+19, ND20, VJY21, HJD+21, ZSU+22, UIJ+21, CQ22, LTND21]. Among concurrent works, we note [ZHH+22].

6.4 Definition of Weak Bellman Closure

Definition 1 (Weak Bellman Closure)**.** *The Bellman operator* $\mathcal{T}^\pi$ *is* weakly closed *with respect to the triple* $\left(\mathcal{Q}^\pi, \mathcal{F}^\pi, \mu\right)$ *if for any* $Q \in \mathcal{Q}^\pi$*, there exists a predictor* $\mathcal{P}^\pi(Q) \in \mathcal{Q}^\pi$ *such that*

$$\langle f, \mathcal{P}^\pi(Q)\rangle_\mu = \langle f, \mathcal{T}^\pi(Q)\rangle_\mu. \tag{28}$$

6.5 Additional results on the concentrability coefficients

6.5.1 Testing with the identity function

Suppose that the identity function $\mathbb{1}$ belongs to the test class. Doing so amounts to requiring that the Bellman error is controlled in an average sense over all the data. When this choice is made, we can derive some generic upper bounds on K^π, which we state and prove here:

Lemma 4. *If $\mathbb{1} \in \mathcal{F}^\pi$, then we have the upper bounds*

$$K^\pi \overset{(i)}{\leq} \frac{\max_{Q \in \mathcal{C}_n^\pi} |\mathbb{E}_\pi \mathcal{B}^\pi Q|^2}{\max_{Q \in \mathcal{C}_n^\pi} |\mathbb{E}_\mu \mathcal{B}^\pi Q|^2} \overset{(ii)}{\leq} K_*^\pi \overset{def}{=} \max_{Q \in \mathcal{C}_n^\pi} \frac{|\mathbb{E}_\pi \mathcal{B}^\pi Q|^2}{|\mathbb{E}_\mu \mathcal{B}^\pi Q|^2}. \tag{29}$$

Proof. Since $\mathbb{1} \in \mathcal{F}$, the definition of $\mathcal{C}_n^\pi$ implies that

$$\max_{Q \in \mathcal{C}_n^\pi} |\mathbb{E}_\mu \mathcal{B}^\pi Q|^2 \leq \big(\| \mathbb{1} \|_\mu^2 + \lambda\big) \frac{\rho}{n} = \big(1 + \lambda\big) \frac{\rho}{n}.$$

The upper bound (i) then follows from the definition of K^π. The upper bound (ii) follows since the right hand side is the maximum ratio. □

Note that large values of K_*^π can arise when there exist Q-functions in the set $\mathcal{C}_n^\pi$ that have low average Bellman error under the data-generating distribution μ, but relatively large values under π. Of course, the likelihood of such unfavorable choices of Q is reduced when we use a larger test function class, which then reduces the size of $\mathcal{C}_n^\pi$. However, we pay a price in choosing a larger test function class, since the choice (40b) of the radius ρ needed for Theorem 3 depends on its complexity.

6.5.2 Mixture distributions

Now suppose that the dataset consists of a collection of trajectories collected by different protocols. More precisely, for each $j = 1, \ldots, m$, let μ_j be a particular protocol for generating a trajectory. Suppose that we generate data by first sampling a random index $J \in [m]$ according to a probability distribution $\{p_j\}_{j=1}^m$, and conditioned $J = j$, we sample (s, a, o) according to μ_j. The resulting data follows a mixture distribution, where we set $o = j$ to tag the protocol used to generate the data. To be clear, for each sample $i = 1, \ldots, n$, we sample J as described, and then draw a single sample $(s, a, o) \sim \mu_j$.

Following the intuition given in the previous section, it is natural to include test functions that code for the protocol—that is, the binary-indicator functions

$$f_j(s, a, o) = \begin{cases} 1 & \text{if } o = j \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

This test function, when included in the weak formulation, enforces the Bellman evaluation equations for the policy $\pi \in \Pi$ under consideration along the distribution induced by each data-generating policy μ_j.

Lemma 5 (Mixture Policy Concentrability)**.** *Suppose that μ is an m-component mixture, and that the indicator functions $\{f_j\}_{j=1}^m$ are included in the test class. Then we have the upper bounds*

$$K^\pi \overset{(i)}{\leq} \frac{1 + m\lambda}{1 + \lambda} \frac{\max\limits_{Q \in \mathcal{C}_n^\pi} [\mathbb{E}_\pi \mathcal{B}^\pi Q]^2}{\max\limits_{Q \in \mathcal{C}_n^\pi} \sum_{j=1}^m p_j^2 [\mathbb{E}_{\mu_j} \mathcal{B}^\pi Q]^2} \overset{(ii)}{\leq} \frac{1 + m\lambda}{1 + \lambda} \max_{Q \in \mathcal{C}_n^\pi} \left\{ \frac{[\mathbb{E}_\pi \mathcal{B}^\pi Q]^2}{\sum_{j=1}^m p_j^2 [\mathbb{E}_{\mu_j} \mathcal{B}^\pi Q]^2} \right\}. \tag{31}$$

Proof. From the definition of K^π, it suffices to show that

$$\max_{Q \in \mathcal{C}_n^\pi} \sum_{j=1}^m p_j^2 [\mathbb{E}_{\mu_j} \mathcal{B}^\pi Q]^2 \leq \frac{\rho}{n} \left(1 + m\lambda\right).$$

A direct calculation yields $\langle f_j, \mathcal{B}^\pi Q \rangle_\mu = \mathbb{E}_\mu \mathbb{I}\{o = j\} \mathcal{B}^\pi Q = p_j \mathbb{E}_{\mu_j} \mathcal{B}^\pi Q$. Moreover, since each f_j belongs to the test class by assumption, we have the upper bound $\left| p_j \mathbb{E}_{\mu_j} \mathcal{B}^\pi Q \right| \leq \sqrt{\frac{\rho}{n}} \sqrt{\|f_j\|_\mu^2 + \lambda}$. Squaring each term and summing over the constraints yields

$$\sum_{j=1}^m p_j^2 [\mathbb{E}_{\mu_j} \mathcal{B}^\pi Q]^2 \leq \frac{\rho}{n} \sum_{j=1}^m \left(\|f_j\|_\mu^2 + \lambda \right) = \frac{\rho}{n} \left(1 + m\lambda\right),$$

where the final equality follows since $\sum_{j=1}^m \|f_j\|_\mu^2 = 1$. □

As shown by the upper bound, the off-policy coefficient K^π provides a measure of how the squared-averaged Bellman errors along the policies $\{\mu_j\}_{j=1}^m$, weighted by their probabilities $\{p_j\}_{j=1}^m$, transfers to the evaluation policy π. Note that the regularization parameter λ decays as a function of the sample size—e.g., as $1/n$ in Theorem 3—the factor $(1 + m\lambda)/(1 + \lambda)$ approaches one as n increases (for a fixed number m of mixture components).

6.5.3 Bellman Rank for off-policy evaluation

In this section, we show how more refined bounds can be obtained when—in addition to a mixture condition—additional structure is imposed on the problem. In particular, we consider a notion similar to that of Bellman rank [JKA+17], but suitably adapted[6] to the off-policy setting.

Given a policy class $\widetilde{\Pi}$ and a predictor class $\widetilde{\mathcal{Q}}$, we say that it has Bellman rank is d if there exist two maps $\nu : \widetilde{\Pi} \to \mathbb{R}^d$ and $\xi : \widetilde{\mathcal{Q}} \to \mathbb{R}^d$ such that

$$\mathbb{E}_\pi \mathcal{B}^\pi Q = \langle \nu_\pi, \, \xi_Q \rangle_{\mathbb{R}^d}, \qquad \text{for all } \; \pi \in \widetilde{\Pi} \text{ and } Q \in \widetilde{\mathcal{Q}}. \tag{32}$$

In words, the average Bellman error of any predictor Q along any given policy π can be expressed as the Euclidean inner product between two d-dimensional vectors, one for the policy and one for the predictor. As in the previous section, we assume that the data is generated by a mixture of m different distributions (or equivalently policies) $\{\mu_j\}_{j=1}^m$. In the off-policy setting, we require that the policy class $\widetilde{\Pi}$ contains all of these policies as well as the target policy—viz. $\{\mu_j\} \cup \{\pi\} \subseteq \widetilde{\Pi}$. Moreover, the predictor class $\widetilde{\mathcal{Q}}$ should contain the predictor class for the target policy, i.e., $\mathcal{Q}^\pi \subseteq \widetilde{\mathcal{Q}}$. We also assume weak realizability for this discussion.

Our result depends on a positive semidefinite matrix determined by the mixture weights $\{p_j\}_{j=1}^m$ along with the embeddings $\{\nu_{\mu_j}\}_{j=1}^m$ of the associated policies that generated the data. In particular, we define

$$\Sigma_\nu = \sum_{j=1}^m p_j^2 \nu_{\mu_j} \nu_{\mu_j}^\top.$$

[6]The original definition essentially takes $\widetilde{\Pi}$ as the set of all greedy policies with respect to $\widetilde{\mathcal{Q}}$. Since a dataset need not originate from greedy policies, the definition of Bellman rank is adapted in a natural way.

Assuming that this is matrix is positive definite,[7] we define the norm $\|u\|_{\Sigma_\nu^{-1}} = \sqrt{u^T(\Sigma_\nu)^{-1}u}$. With this notation, we have the following bound.

Lemma 6 (Concentrability with Bellman Rank)**.** *For a mixture data-generation process and under the Bellman rank condition* (32)*, we have the upper bound*

$$K^\pi \leq \frac{1+m\lambda}{1+\lambda} \|\nu_\pi\|^2_{\Sigma_\nu^{-1}}, \tag{33}$$

Proof. Our proof exploits the upper bound (ii) from the claim (31) in Lemma 5. We first evaluate and redefine the ratio in this upper bound. Weak realizability coupled with the Bellman rank condition (32) implies that there exists some $Q^\pi_\star$ such that

$$0 = \langle f_j, \mathcal{B}^\pi Q^\pi_\star\rangle_\mu = p_j \mathbb{E}_{\mu_j} \mathcal{B}^\pi Q^\pi_\star = p_j \left\langle \nu_{\mu_j}, \, \xi_{Q^\pi_\star} \right\rangle, \qquad \text{for all } j = 1, \ldots, m, \text{ and}$$
$$0 = \langle \mathbb{1}, \mathcal{B}^\pi Q^\pi_\star\rangle_\pi = \mathbb{E}_\pi \mathcal{B}^\pi Q^\pi_\star = \left\langle \nu_\pi, \, \xi_{Q^\pi_\star} \right\rangle.$$

Therefore, we have the equivalences $\mathbb{E}_{\mu_j}\mathcal{B}^\pi Q = \left\langle \nu_{\mu_j}, \, (\xi_Q - \xi_{Q^\pi_\star})\right\rangle$ for all $j = 1, \ldots, m$, as well as $\mathbb{E}_\pi \mathcal{B}^\pi Q = \left\langle \nu_\pi, \, (\xi_Q - \xi_{Q^\pi_\star})\right\rangle$. Introducing the shorthand $\Delta_Q = \xi_Q - \xi_{Q^\pi_\star}$, we can bound the ratio as follows

$$\begin{aligned}
\sup_{Q \in \mathcal{C}^\pi_n} \left\{ \frac{(\langle \nu_\pi, \, \Delta_Q \rangle)^2}{\sum_{j=1}^m p_j^2 (\langle \nu_{\mu_j}, \, \Delta_Q \rangle)^2} \right\} &= \sup_{Q \in \mathcal{C}^\pi_n} \left\{ \frac{(\langle \nu_\pi, \, \Delta_Q \rangle)^2}{\Delta_Q^\top \Big(\sum_{j=1}^m p_j^2 \nu_{\mu_j} \nu_{\mu_j}^\top \Big) \Delta_Q} \right\} \\
&= \sup_{Q \in \mathcal{C}^\pi_n} \left\{ \frac{(\langle \nu_\pi, \, \Sigma_\nu^{-\frac{1}{2}} \widetilde{\Delta}_Q \rangle)^2}{\|\widetilde{\Delta}_Q\|_2^2} \right\} \qquad \text{where } \widetilde{\Delta}_Q = \Sigma_\nu^{\frac{1}{2}} \Delta_Q \\
&\leq \|\nu_\pi\|^2_{\Sigma_\nu^{-1}},
\end{aligned}$$

where the final step follows from the Cauchy–Schwarz inequality. □

Thus, when performing off-policy evaluation with a mixture distribution under the Bellman rank condition, the coefficient K^π is bounded by the alignment between the target policy π and the data-generating distribution μ, as measured in the the embedded space guaranteed by the Bellman rank condition. The structure of this upper bound is similar to a result that we derive in the sequel for linear approximation under Bellman closure (see Proposition 3).

6.6 Further comments on the prediction error test space

A few comments on the bound in Lemma 1: as in our previous results, the pre-factor $\frac{\|\epsilon\|^2_\mu + \lambda}{\|\mathbb{1}\|^2_\pi + \lambda}$ serves as a normalization factor. Disregarding this leading term, the second ratio measures how the prediction error $\epsilon = Q - Q^\pi_\star$ along μ transfers to π, as measured via the operator $\mathcal{I} - \gamma \mathbb{P}^\pi$. This interaction is complex, since it includes the *bootstrapping term* $-\gamma\mathbb{P}^\pi$. (Notably, such a term is not present for standard prediction or bandit problems, in which case $\gamma = 0$.) This term reflects the dynamics intrinsic to reinforcement learning, and plays a key role in proving "hard" lower bounds for offline RL (e.g., see the work [Zan20]).

Observe that the bound in Lemma 1 requires only weak realizability, and thus it always applies. This fact is significant in light of a recent lower bound [FKSLX21], showing that without Bellman

[7] If not, one can prove a result for a suitably regularized version.

closure, off-policy learning is challenging even under strong concentrability assumption (such as bounds on density ratios). Lemma 1 gives a sufficient condition without Bellman closure, but with a different measure that accounts for bootstrapping.

If, in fact, (weak) Bellman closure holds, then Lemma 1 takes the following simplified form:

Lemma 7 (OPC coefficient under Bellman closure)**.** *If $\mathcal{E}^\pi \subseteq \mathcal{F}^\pi$ and weak Bellman closure holds, then*

$$K^\pi \leq \max_{\epsilon \in \mathcal{E}^\pi} \left\{ \frac{\|\epsilon\|_\mu^2 + \lambda}{1+\lambda} \cdot \frac{\langle \mathbb{1}, \epsilon \rangle_\pi^2}{\langle \epsilon, \epsilon \rangle_\mu^2} \right\} \leq \max_{\epsilon \in \mathcal{E}^\pi} \left\{ \frac{\|\epsilon\|_\pi^2}{\|\epsilon\|_\mu^2} \right\}.$$

See Section 9.3 for the proof.

In such case, the concentrability measures the increase in the discrepancy $Q - Q'$ of the feasible predictors when moving from the dataset distribution μ to the distribution of the target policy π. In Section 4.3, we give another bound under weak Bellman closure, and thereby recover a recent result due to Xie et al. [XCJ+21]. Finally, in Section 5, we provide some applications of this concentrability factor to the linear setting.

6.7 From Importance Sampling to Bellman Closure

Let us show an application of Lemma 3 on an example with just two test spaces. Suppose that we suspect that Bellman closure holds, but rather than committing to such assumption, we wish to fall back to an importance sampling estimator if Bellman closure does not hold.

In order to streamline the presentation of the idea, let us introduce the following setup. Let π^b be a behavioral policy that generates the dataset, i.e., such that each state-action (s,a) in the dataset is sampled from its discounted state distribution d_{π^b}. Next, let the identifier o contain the trajectory from ν_{start} up to the state-action pair (s,a) recorded in the dataset. That is, each tuple (s,a,r,s^+,o) in the dataset $\mathcal{D}$ is such that $(s,a) \sim d_{\pi^b}$ and o contains the trajectory up to (s,a).

We now define the test spaces. The first one is denoted with $\mathcal{F}_\pi^{\text{IS}}$ and leverages importance sampling. It contains a single test function defined as the importance sampling estimator

$$\mathcal{F}_\pi^{\text{IS}} = \{f_\pi\}, \qquad \text{where } f_\pi(s,a,o) = \frac{1}{b_\pi} \prod_{(s_h,a_h)\in o} \frac{\pi(a_h \mid s_h)}{\pi^b(a_h \mid s_h)}. \tag{34}$$

The above product is over the random trajectory contained in the identifier o. The normalization factor $b_\pi \in \mathbb{R}$ is connected to the maximum range of the importance sampling estimator, and ensures that $\sup_{(s,a,o)} f_\pi(s,a,o) \leq 1$. The second test space is the prediction error test space $\mathcal{E}^\pi$ defined in Section 4.2.

With this choice, let us define three concentrability coefficients. $K^\pi_{(1)}$ arises from importance sampling, $K^\pi_{(2)}$ from the prediction error test space when Bellman closure holds and $K^\pi_{(3)}$ from the prediction error test space when just weak realizability holds. They are defined as

$$K^\pi_{(1)} \leq \sqrt{b_\pi \frac{(1+\lambda b_\pi)}{1+\lambda}} \qquad K^\pi_{(2)} \leq \max_{\epsilon \in \mathcal{E}^\pi_\star} \frac{\langle \mathbb{1}, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\pi^2}{\langle \epsilon, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\mu^2} \times \frac{\|\epsilon\|_\mu^2 + \lambda}{\|\,\mathbb{1}\,\|_\pi^2 + \lambda}, \qquad K^\pi_{(3)} \leq c_1 \frac{\|\mathcal{B}^\pi Q\|_\pi^2}{\|\mathcal{B}^\pi Q\|_\mu^2}.$$

Lemma 8 (From Importance Sampling to Bellman Closure)**.** *The choice $\mathcal{F}^\pi = \mathcal{F}^{IS}_\pi \cup \mathcal{E}^\pi$ for all $\pi \in \Pi$ ensures that with probability at least $1-\delta$, the oracle inequality (9) holds with $K^\pi \leq \min\{K^\pi_{(1)}, K^\pi_{(2)}, K^\pi_{(3)}\}$ if weak Bellman closure holds and $K^\pi \leq \min\{K^\pi_{(1)}, K^\pi_{(2)}\}$ otherwise.*

Proof. Let us calculate the off-policy cost coefficient associated with $\mathcal{F}^{\text{IS}}_\pi$. The unbiasedness of the importance sampling estimator gives us the following population constraint (here $\mu = d_{\pi^b}$)

$$|\langle f_\pi, \mathcal{B}^\pi Q\rangle_\mu| = |\mathbb{E}_\mu f_\pi \mathcal{B}^\pi Q| = \frac{1}{b_\pi}|\mathbb{E}_\pi \mathcal{B}^\pi Q| = \frac{1}{b_\pi}|\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle_\pi| \leq \frac{L}{\sqrt{n}}\sqrt{\|f_\pi\|_2^2 + \lambda}$$

The norm of the test function reads (notice that μ generates (s,a,o) here)

$$\|f_\pi\|^2_\mu = \mathbb{E}_\mu f_\pi^2 = \frac{1}{b_\pi^2}\mathbb{E}_\mu\left[\prod_{(s_h,a_h)\in o}\frac{\pi(a_h \mid s_h)}{\pi^b(a_h \mid s_h)}\right]^2 = \frac{1}{b_\pi^2}\mathbb{E}_\pi\left[\prod_{(s_h,a_h)\in o}\frac{\pi(a_h \mid s_h)}{\pi^b(a_h \mid s_h)}\right] \leq \frac{1}{b_\pi}.$$

Together with the prior display, we obtain

$$\frac{\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle_\pi^2}{b_\pi^2(\|f_\pi\|_2^2 + \lambda)} \leq \frac{\rho}{n}.$$

The resulting concentrability coefficient is therefore

$$K^\pi \leq \max_{Q\in\mathcal{C}^\pi_n}\frac{\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle_\pi^2}{1+\lambda} \times \frac{n}{\rho} \leq \max_{Q\in\mathcal{C}^\pi_n}\frac{\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle_\pi^2}{1+\lambda} \times \frac{b_\pi^2(\|f_\pi\|_2^2 + \lambda)}{\langle \mathbb{1}, \mathcal{B}^\pi Q\rangle_\pi^2} \leq b_\pi\frac{(1+\lambda b_\pi)}{1+\lambda}.$$

Chaining the above result with Lemmas 1 and 2, using Lemma 3 and plugging back into Theorem 3 yields the thesis. □

6.8 Implementation for Off-Policy Predictions

In this section, we describe a computationally efficient way in which to compute the upper/lower estimates (5). Given a finite set of $n_{\mathcal{F}}$ test functions, it involves solving a quadratic program with $2n_{\mathcal{F}} + 1$ constraints.

Let us first work out a concise description of the constraints defining membership in $\widehat{\mathcal{C}}^\pi_n$. Introduce the shorthand $n_f \overset{def}{=} \|f_j\|_n^2 + \lambda$. We then define the empirical average feature vector $\widehat{\phi}_f$, the empirical average reward $\widehat{r}_f$, and the average next-state feature vector $\widehat{\phi}^{+\pi}_f$ as

$$\widehat{\phi}_f = \frac{1}{\sqrt{n_f}}\sum_{(s,a,r,s^+)\in\mathcal{D}} f(s,a)\phi(s,a), \qquad \widehat{r}_f = \frac{1}{\sqrt{n_f}}\sum_{(s,a,r,s^+)\in\mathcal{D}} f(s,a)r,$$
$$\widehat{\phi}^{+\pi}_f = \frac{1}{\sqrt{n_f}}\sum_{(s,a,r,s^+)\in\mathcal{D}} f(s,a)\phi(s^+,\pi).$$

In terms of this notation, each empirical constraint defining $\widehat{\mathcal{C}}^\pi_n$ can be written in the more compact form

$$\frac{|\langle f, \delta^\pi Q\rangle_n|}{\sqrt{n_f}} = \left|\langle \widehat{\phi}_f - \gamma\widehat{\phi}^{+\pi}_f,\, w\rangle - \widehat{r}_f\right| \leq \sqrt{\frac{\rho}{n}}.$$

Then the set of empirical constraints can be written as a set of constraints linear in the critic parameter w coupled with the assumed regularity bound on w

$$\widehat{\mathcal{C}}_n^{\pi} = \Big\{ w \in \mathbb{R}^d \mid \|w\|_2 \le 1, \quad \text{and} \quad -\sqrt{\frac{\rho}{n}} \le \langle \widehat{\phi}_f - \gamma \widehat{\phi}_f^{+\pi},\, w \rangle - \widehat{r}_f \le \sqrt{\frac{\rho}{n}} \quad \text{for all } f \in \mathcal{F}^{\pi} \Big\}. \tag{35}$$

Thus, the estimates $\widehat{V}^{\pi}_{\text{min}}$ (respectively $\widehat{V}^{\pi}_{\text{max}}$) acan be computed by minimizing (respectively maximizing) the linear objective function $w \mapsto \langle [\mathbb{E}_{s \sim \nu_{\text{start}}} \mathbb{E}_{a \sim \pi} \phi(s,a)],\, w \rangle$ subject to the $2n_{\mathcal{F}} + 1$ constraints in equation (35). Therefore, the estimates can be computed in polynomial time for any test function with a cardinality that grows polynomially in the problem parameters.

6.9 Discussion of Linear Approximate Optimization

Here we discuss the presence of the supremum over policies in the coefficient $K^{\widetilde{\pi}}_{(1)}$ from equation (23). In particular, it arises because our actor-critic method iteratively approximates the maximum in the max-min estimate (6) using a gradient-based scheme. The ability of a gradient-based method to make progress is related to the estimation accuracy of the gradient, which is the Q estimates of the actor's current policy π_t; more specifically, the gradient is the Q function parameter w_t. In the general case, the estimation error of the gradient w_t depends on the policy under consideration through the matrix $\Sigma^{+\pi_t}_{\lambda,\text{Boot}}$, while it is independent in the special case of Bellman closure (as it depends on just Σ). As the actor's policies are random, this yields the introduction of a $\sup_{\pi \in \Pi}$ in the general bound. Notice the method still competes with the best comparator $\widetilde{\pi}$ by measuring the errors along the distribution of the comparator (through the operator $\mathbb{E}_{\widetilde{\pi}}$). To be clear, $\sup_{\pi \in \Pi}$ may not arise with approximate solution methods that do not rely only on the gradient to make progress (such as second-order methods); we leave this for future research. Reassuringly, when Bellman closure, the approximate solution method recovers the standard guarantees established in the paper [ZWB21].

7 General Guarantees

7.1 A deterministic guarantee

We begin our analysis stating a deterministic set of sufficient conditions for our estimators to satisfy the guarantees (8) and (9). This formulation is useful, because it reveals the structural conditions that underlie success of our estimators, and in particular the connection to weak realizability. In Section 7.2, we exploit this deterministic result to show that, under a fairly general sampling model, our estimators enjoy these guarantees with high probability.

In the previous section, we introduced the population level set $\mathcal{C}_n^\pi$ that arises in the statement of our guarantees. Also central in our analysis is the infinite data limit of this set. More specifically, for any fixed (ρ, λ), if we take the limit $n \to \infty$, then $\mathcal{C}_n^\pi$ reduces to the set of all solutions to the weak formulation (25)—that is

$$\mathcal{C}_\infty^\pi(\mathcal{F}^\pi) = \{Q \in \mathcal{Q}^\pi \mid \langle f, \mathcal{B}^\pi Q \rangle_\mu = 0 \quad \text{for all } f \in \mathcal{F}^\pi\}. \tag{36}$$

As before, we omit the dependence on the test function class $\mathcal{F}^\pi$ when it is clear from context. By construction, we have the inclusion $\mathcal{C}_\infty^\pi(\mathcal{F}^\pi) \subseteq \mathcal{C}_n^\pi(4\rho, \lambda; \mathcal{F}^\pi)$ for any non-negative pair (ρ, λ).

Our first set of guarantees hold when the random set $\widehat{\mathcal{C}}_n^\pi$ satisfies the *sandwich relation*

$$\mathcal{C}_\infty^\pi(\mathcal{F}^\pi) \subseteq \widehat{\mathcal{C}}_n^\pi(\rho, \lambda; \mathcal{F}^\pi) \subseteq \mathcal{C}_n^\pi(4\rho, \lambda; \mathcal{F}^\pi) \tag{37}$$

To provide intuition as to why this sandwich condition is natural, observe that it has two important implications:

(a) Recalling the definition of weak realizability (1), the weak solution $Q_\star^\pi$ belongs to the empirical constraint set $\widehat{\mathcal{C}}_n^\pi$ for any choice of test function space. This important property follows because $Q_\star^\pi$ must satisfy the constraints (25), and thus it belongs to $\mathcal{C}_\infty^\pi \subseteq \widehat{\mathcal{C}}_n^\pi$.

(b) All solutions in $\widehat{\mathcal{C}}_n^\pi$ also belong to $\mathcal{C}_n^\pi$, which means they approximately satisfy the weak Bellman equations in a way quantified by $\mathcal{C}_n^\pi$.

By leveraging these facts in the appropriate way, we can establish the following guarantee:

Proposition 4. *The following two statements hold.*

(a) Policy evaluation: If the set $\widehat{\mathcal{C}}_n^\pi$ satisfies the sandwich relation (37), then the estimates $(\widehat{V}_{min}^\pi, \widehat{V}_{max}^\pi)$ satisfy the width bound (8b). If, in addition, weak Bellman realizability for π is assumed, then the coverage (8a) condition holds.

(b) Policy optimization: If the sandwich relation (37) and weak Bellman realizability hold for all $\pi \in \Pi$, then any max-min (6) optimal policy $\widetilde{\pi}$ satisfies the oracle inequality (9).

See Section 8.1 for the proof of this claim.

In summary, Proposition 4 ensures that when weak realizability is in force, then the sandwich relation (37) is a sufficient condition for both the policy evaluation (8) and optimization (9) guarantees to hold. Accordingly, the next phase of our analysis focuses on deriving sufficient conditions for the sandwich relation to hold with high probabability.

7.2 Some high-probability guarantees

As stated, Proposition 4 is a "meta-result", in that it applies to any choice of set $\widehat{\mathcal{C}}_n^\pi \equiv \widehat{\mathcal{C}}_n^\pi(\rho, \lambda; \mathcal{F}^\pi)$ for which the sandwich relation (37) holds. In order to obtain a more concrete guarantee, we need to impose assumptions on the way in which the dataset was generated, and concrete choices of (ρ, λ) that suffice to ensure that the associated sandwich relation (37) holds with high probability. These tasks are the focus of this section.

7.2.1 A model for data generation

Let us begin by describing a fairly general model for data-generation. Any sample takes the form $z \stackrel{def}{=} (s, a, r, s^+, o)$, where the five components are defined as follows:

- the pair (s, a) index the current state and action.
- the random variable r is a noisy observation of the mean reward.
- the random state s^+ is the next-state sample, drawn according to the transition $\mathbb{P}(s, a)$.
- the variable o is an optional identifier.

As one example of the use of an identifier variable, if samples might be generated by one of two possible policies—say π_1 and π_2—the identifier can take values in the set $\{1, 2\}$ to indicate which policy was used for a particular sample.

Overall, we observe a dataset $\mathcal{D} = \{z_i\}_{i=1}^n$ of n such quintuples. In the simplest of possible settings, each triple (s, a, o) is drawn i.i.d. from some fixed distribution μ, and the noisy reward r_i is an unbiased estimate of the mean reward function $R(s_i, a_i)$. In this case, our dataset consists of n i.i.d. quintuples. More generally, we would like to accommodate richer sampling models in which the sample $z_i = (s_i, a_i, o_i, r_i, s_i^+)$ at a given time i is allowed to depend on past samples. In order to specify such dependence in a precise way, define the nested sequence of sigma-fields

$$\mathcal{F}_1 = \emptyset, \quad \text{and} \quad \mathcal{F}_i \stackrel{def}{=} \sigma\Big(\{z_j\}_{j=1}^{i-1}\Big) \qquad \text{for } i = 2, \ldots, n. \tag{38}$$

In terms of this filtration, we make the following definition:

Assumption 3 (Adapted dataset)**.** *An adapted dataset is a collection $\mathcal{D} = \{z_i\}_{i=1}^n$ such that for each $i = 1, \ldots, n$:*

- *There is a conditional distribution μ_i such that $(s_i, a_i, o_i) \sim \mu_i(\cdot \mid \mathcal{F}_i)$.*
- *Conditioned on (s_i, a_i, o_i), we observe a noisy reward $r_i = r(s_i, a_i) + \eta_i$ with $\mathbb{E}[\eta_i \mid \mathcal{F}_i] = 0$, and $|r_i| \leq 1$.*
- *Conditioned on (s_i, a_i, o_i), the next state s_i^+ is generated according to $\mathbb{P}(s_i, a_i)$.*

Under this assumption, we can define the (possibly) random reference measure

$$\mu(s, a, o) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^n \mu_i\big(s, a, o \mid \mathcal{F}_i\big). \tag{39}$$

In words, it corresponds to the distribution induced by first drawing a time index $i \in \{1, \ldots, n\}$ uniformly at random, and then sampling a triple (s, a, o) from the conditional distribution $\mu_i\big(\cdot \mid \mathcal{F}_i\big)$.

7.2.2 A general guarantee

Recall that there are three function classes that underlie our method: the test function class $\mathcal{F}$, the policy class Π, and the Q-function class $\mathcal{Q}$. In this section, we state a general guarantee (Theorem 3) that involves the metric entropies of these sets. In Section 7.2.3, we provide corollaries of this guarantee for specific function classes.

In more detail, we equip the test function class and the Q-function class with the usual sup-norm

$$\|f - \tilde{f}\|_\infty \stackrel{def}{=} \sup_{(s,a,o)} |f(s,a,o) - \tilde{f}(s,a,o)|, \quad \text{and} \quad \|Q - \tilde{Q}\|_\infty \stackrel{def}{=} \sup_{(s,a)} |Q(s,a) - \tilde{Q}(s,a)|,$$

and the policy class with the sup-TV norm

$$\|\pi - \widetilde{\pi}\|_{\infty,1} \stackrel{def}{=} \sup_s \|\pi(\cdot \mid s) - \widetilde{\pi}(\cdot \mid s)\|_1 = \sup_s \sum_a |\pi(a \mid s) - \widetilde{\pi}(a \mid s)|.$$

For a given $\epsilon > 0$, we let $\mathcal{N}_\epsilon(\mathcal{F})$, $\mathcal{N}_\epsilon(\mathcal{Q})$, and $\mathcal{N}_\epsilon(\Pi)$ denote the ϵ-covering numbers of each of these function classes in the given norms. Given these covering numbers, a tolerance parameter $\delta \in (0,1)$ and the shorthand $\phi(t) = \max\{t, \sqrt{t}\}$, define the radius function

$$\rho(\epsilon, \delta) \stackrel{def}{=} n\Big\{ \int_{\epsilon^2}^{\epsilon} \phi\big(\frac{\log N_u(\mathcal{F})}{n}\big) du + \frac{\log N_\epsilon(\mathcal{Q})}{n} + \frac{\log N_\epsilon(\Pi)}{n} + \frac{\log(n/\delta)}{n} \Big\}. \tag{40a}$$

In our theorem, we implement the estimator using a radius $\rho = \rho(\epsilon, \delta)$, where $\epsilon > 0$ is any parameter that satisfies the bound

$$\epsilon^2 \stackrel{(i)}{\leq} \bar{c}\, \frac{\rho(\epsilon,\delta)}{n}, \quad \text{and} \quad \lambda \stackrel{(i)}{=} 4 \frac{\rho(\epsilon,\delta)}{n}. \tag{40b}$$

Here $\bar{c} > 0$ is a suitably chosen but universal constant (whose value is determined in the proof), and we adopt the shorthand $\rho = \rho(\epsilon, \delta)$ in our statement below.

Theorem 3 (High-probability guarantees)**.** *Consider the estimates implemented using triple* $(\Pi, \mathcal{F}, \mathcal{Q})$ *that is weakly Bellman realizable (Assumption 1); an adapted dataset (Assumption 3); and with the choices* (40) *for* $(\epsilon, \rho, \lambda)$*. Then with probability at least* $1 - \delta$*:*

Policy evaluation: *For any* $\pi \in \Pi$*, the estimates* $(\widehat{V}^\pi_{min}, \widehat{V}^\pi_{max})$ *specify a confidence interval satisfying the coverage* (8a) *and width bounds* (8b)*.*

Policy optimization: *Any max-min policy* (6) $\widetilde{\pi}$ *satisfies the oracle inequality* (9)*.*

See Section 8.3 for the proof of the claim.

Choices of $(\rho, \epsilon, \lambda)$**:** Let us provide a few comments about the choices of $(\rho, \epsilon, \lambda)$ from equations (40a) and (40b). The quality of our bounds depends on the size of the constraint set $\mathcal{C}^\pi_n$, which is controlled by the constraint level $\sqrt{\frac{\rho}{n}}$. Consequently, our results are tightest when $\rho = \rho(\epsilon, \delta)$ is as small as possible. Note that ρ is an decreasing function of ϵ, so that in order to minimize it, we would like to choose ϵ as large as possible subject to the constraint (40b)(i). Ignoring the entropy integral term in equation (40b) for the moment—see below for some comments on it—these considerations lead to

$$n\epsilon^2 \asymp \log N_\epsilon(\mathcal{F}) + \log N_\epsilon(\mathcal{Q}) + \log N_\epsilon(\Pi). \tag{41}$$

This type of relation for the choice of ϵ in non-parametric statistics is well-known (e.g., see Chapters 13–15 in the book [Wai19] and references therein). Moreover, setting $\lambda \asymp \epsilon^2$ as in equation (40b)(ii) is often the correct scale of regularization.

Key technical steps in proof: It is worthwhile making a few comments about the structure of the proof so as to clarify the connections to Proposition 4 along with the weak formulation that underlies our methods. Recall that Proposition 4 requires the empirical $\widehat{\mathcal{C}}_n^\pi$ and population sets $\mathcal{C}_n^\pi$ to satisfy the sandwich relation (37). In order to prove that this condition holds with high probability, we need to establish uniform control over the family of random variables

$$\frac{\left| \langle f, \, \delta^\pi(Q) \rangle_n - \langle f, \, \mathcal{B}^\pi(Q) \rangle_\mu \right|}{\sqrt{\|f\|_n^2 + \lambda}}, \qquad \text{as indexed by the triple } (f, Q, \pi). \tag{42}$$

Note that the differences in the numerator of these variables correspond to moving from the empirical constraints on Q-functions that are enforced using the TD errors, to the population constraints that involve the Bellman error function.

Uniform control of the family (42), along with the differences $\|f\|_n - \|f\|_\mu$ uniformly over f, allows us to relate the empirical and population sets, since the associated constraints are obtained by shifting between the empirical inner products $\langle \cdot, \, \cdot \rangle_n$ to the reference inner products $\langle \cdot, \, \cdot \rangle_\mu$. A simple discretization argument allows us to control the differences uniformly in (Q, π), as reflected by the metric entropies appearing in our definition (40). Deriving uniform bounds over test functions f—due to the self-normalizing nature of the constraints—requires a more delicate argument. More precisely, in order to obtain optimal results for non-parametric problems (see Corollary 2 to follow), we need to localize the empirical process at a scale ϵ, and derive bounds on the localized increments. This portion of the argument leads to the entropy integral—which is localized to the interval $[\epsilon^2, \epsilon]$—in our definition (40a) of the radius function.

Intuition from the on-policy setting: In order to gain intuition for the statistical meaning of the guarantees in Theorem 3, it is worthwhile understanding the implications in a rather special case—namely, the simpler on-policy setting, where the discounted occupation measure induced by the target policy π coincides with the dataset distribution μ. Let us consider the case in which the identity function $\mathbb{1}$ belongs to the test class $\mathcal{F}^\pi$. Under these conditions, for any $Q \in \mathcal{C}_n^\pi$, we can write

$$\max_{Q \in \mathcal{C}_n^\pi} |\mathbb{E}_\pi \mathcal{B}^\pi Q| \overset{(i)}{=} \max_{Q \in \mathcal{C}_n^\pi} |\mathbb{E}_\mu \mathcal{B}^\pi Q| \overset{(ii)}{\leq} \sqrt{1+\lambda} \sqrt{\frac{\rho}{n}},$$

where equality (i) follows from the on-policy assumption, and step (ii) follows from the definition of the set $\mathcal{C}_n^\pi$, along with the condition that $\mathbb{1} \in \mathcal{F}^\pi$. Consequently, in the on-policy setting, the width bound (8b) ensures that

$$|\widehat{V}_{\min}^\pi - \widehat{V}_{\max}^\pi| \leq 2 \frac{\sqrt{1+\lambda}}{1-\gamma} \sqrt{\frac{\rho}{n}}. \tag{43}$$

In this simple case, we see that the confidence interval scales as $\sqrt{\rho/n}$, where the quantity ρ is related to the metric entropy via equation (40b). In the more general off-policy setting, the bound involves this term, along with additional terms that reflect the cost of off-policy data. We discuss these issues in more detail in Section 4. Before doing so, however, it is useful derive some specific corollaries that show the form of ρ under particular assumptions on the underlying function classes, which we now do.

7.2.3 Some corollaries

Theorem 3 applies generally to triples of function classes $(\Pi, \mathcal{F}, \mathcal{Q})$, and the statistical error $\sqrt{\frac{\rho(\epsilon,\delta)}{n}}$ depends on the metric entropies of these function classes via the definition (40a) of $\rho(\epsilon, \delta)$, and the choices (40b). As shown in this section, if we make particular assumptions about the metric entropies, then we can derive more concrete guarantees.

Parametric and finite VC classes: One form of metric entropy, typical for a relatively simple function class $\mathcal{G}$ (such as those with finite VC dimension) scales as

$$\log N_\epsilon(\mathcal{G}) \asymp d \, \log \big(\frac{1}{\epsilon}\big), \tag{44}$$

for some dimensionality parameter d. For instance, bounds of this type hold for linear function classes with d parameters, and for finite VC classes (with d proportional to the VC dimension); see Chapter 5 of the book [Wai19] for more details.

Corollary 1. *Suppose each class of the triple $(\Pi, \mathcal{F}, \mathcal{Q})$ has metric entropy that is at most polynomial (44) of order d. Then for a sample size $n \geq 2d$, the claims of Theorem 3 hold with $\epsilon^2 = d/n$ and*

$$\tilde{\rho}\big(\sqrt{\frac{d}{n}}, \delta\big) \stackrel{def}{=} c \, \Big\{ d \, \log \big(\frac{n}{d}\big) + \log \big(\frac{n}{\delta}\big) \Big\}, \tag{45}$$

where c is a universal constant.

Proof. Our strategy is to upper bound the radius ρ from equation (40a), and then show that this upper bound $\tilde{\rho}$ satisfies the conditions (40b) for the specified choice of ϵ^2. We first control the term $\log N_\epsilon(\mathcal{F})$. We have

$$\frac{1}{\sqrt{n}} \int_{\epsilon^2}^{\epsilon} \sqrt{\log N_u(\mathcal{F})} du \leq \sqrt{\frac{d}{n}} \int_0^\epsilon \sqrt{\log(1/u)} du \;=\; \epsilon \sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log(1/(\epsilon t))} dt \;=\; c\epsilon \log(1/\epsilon) \sqrt{\frac{d}{n}}.$$

Similarly, we have

$$\frac{1}{n} \int_{\epsilon^2}^{\epsilon} \log N_u(\mathcal{F}) du \leq \epsilon \frac{d}{n} \Big\{ \int_\epsilon^1 \log(1/t) dt + \log(1/\epsilon) \Big\} \;\leq\; c \, \epsilon \log(1/\epsilon) \frac{d}{n}.$$

Finally, for terms not involving entropy integrals, we have

$$\max \Big\{ \frac{\log N_\epsilon(\mathcal{Q})}{n}, \frac{\log N_\epsilon(\Pi)}{n} \Big\} \leq c \frac{d}{n} \log(1/\epsilon).$$

Setting $\epsilon^2 = d/n$, we see that the required conditions (40b) hold with the specified choice (45) of $\tilde{\rho}$. □

Richer function classes: In the previous section, the metric entropy scaled logarithmically in the inverse precision $1/\epsilon$. For other (richer) function classes, the metric entropy exhibits a polynomial scaling in the inverse precision, with an exponent $\alpha > 0$ that controls the complexity. More precisely, we consider classes of the form

$$\log N_\epsilon(\mathcal{G}) \asymp \Big(\frac{1}{\epsilon}\Big)^{\alpha}. \tag{46}$$

For example, the class of Lipschitz functions in dimension d has this type of metric entropy with $\alpha = d$. More generally, for Sobolev spaces of functions that have s derivatives (and the s^{th}-derivative is Lipschitz), we encounter metric entropies of this type with $\alpha = d/s$. See Chapter 5 of the book [Wai19] for further background.

Corollary 2. *Suppose that each function class* $(\Pi, \mathcal{F}, \mathcal{Q})$ *has metric entropy with at most* α*-scaling* (46) *for some* $\alpha \in (0, 2)$*. Then the claims of Theorem 3 hold with* $\epsilon^2 = (1/n)^{\frac{2}{2+\alpha}}$*, and*

$$\tilde{\rho}\big((1/n)^{\frac{1}{2+\alpha}}, \delta\big) = c\,\Big\{n^{\frac{\alpha}{2+\alpha}} + \log(n/\delta)\Big\}. \tag{47}$$

where c *is a universal constant.*

We note that for standard regression problems over classes with α-metric entropy, the rate $(1/n)^{\frac{2}{2+\alpha}}$ is well-known to be minimax optimal (e.g., see Chapter 15 in the book [Wai19], as well as references therein).

Proof. We start by controlling the terms involving entropy integrals. In particular, we have

$$\frac{1}{\sqrt{n}} \int_{\epsilon^2}^{\epsilon} \sqrt{\log N_u(\mathcal{F})}du \leq \frac{c}{\sqrt{n}} u^{1-\frac{\alpha}{2}}\Big|_0^{\epsilon} \;=\; \frac{c}{\sqrt{n}}\epsilon^{1-\frac{\alpha}{2}}.$$

Requiring that this term is of order ϵ^2 amounts to enforcing that $\epsilon^{1+\frac{\alpha}{2}} \asymp (1/\sqrt{n})$, or equivalently that $\epsilon^2 \asymp (1/n)^{\frac{2}{2+\alpha}}$.

If $\alpha \in (0, 1]$, then the second entropy integral converges and is of lower order. Otherwise, if $\alpha \in (1, 2)$, then we have

$$\frac{1}{n} \int_{\epsilon^2}^{\epsilon} \log N_u(\mathcal{F})du \leq \frac{c}{n} \int_{\epsilon^2}^{\epsilon} (1/u)^{\alpha} du \leq \frac{c}{n}(\epsilon^2)^{1-\alpha}.$$

Hence the requirement that this term is bounded by ϵ^2 is equivalent to $\epsilon^{2\alpha} \succsim (1/n)$, or $\epsilon^2 \succsim (1/n)^{1/\alpha}$. When $\alpha \in (1, 2)$, we have $\frac{1}{\alpha} > \frac{2}{2+\alpha}$, so that this condition is milder than our first condition.

Finally, we have $\max\big\{\frac{\log N_\epsilon(\mathcal{Q})}{n}, \frac{\log N_\epsilon(\Pi)}{n}\big\} \leq \frac{c}{n}\big(1/\epsilon\big)^{\alpha}$, and requiring that this term scales as ϵ^2 amounts to requiring that $\epsilon^{2+\alpha} \asymp (1/n)$, or equivalently $\epsilon^2 \asymp (1/n)^{\frac{2}{2+\alpha}}$, as before. ∎

8 Main Proofs

This section is devoted to the proofs of our guarantees for general function classes—namely, Proposition 4 that holds in a deterministic manner, and Theorem 3 that gives high probability bounds under a particular sampling model.

8.1 Proof of Proposition 4

Our proof makes use of an elementary simulation lemma, which we state here:

Lemma 9 (Simulation lemma). *For any policy π and function Q, we have*

$$\mathbb{E}_{S\sim\nu_{start}}(Q - Q^{\pi})(S,\pi) = \frac{\mathbb{E}_{\pi}\mathcal{B}^{\pi}Q}{1-\gamma} \tag{48}$$

See Section 8.2 for the proof of this claim.

8.1.1 Proof of policy evaluation claims

First of all, we have the elementary bounds

$$|\widehat{V}^{\pi}_{\min} - V^{\pi}| = |\min_{Q\in\widehat{\mathcal{C}}^{\pi}_n} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}| \leq \max_{Q\in\widehat{\mathcal{C}}^{\pi}_n}|\mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}|, \quad \text{and}$$
$$|\widehat{V}^{\pi}_{\max} - V^{\pi}| = |\max_{Q\in\widehat{\mathcal{C}}^{\pi}_n} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}| \leq \max_{Q\in\widehat{\mathcal{C}}^{\pi}_n}|\mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}|.$$

Consequently, in order to prove the bound (8b) it suffices to upper bound the right-hand side common in the two above displays. Since $\widehat{\mathcal{C}}^{\pi}_n \subseteq \mathcal{C}^{\pi}_n$, we have the upper bound

$$\begin{aligned}\max_{Q\in\widehat{\mathcal{C}}^{\pi}_n}|\mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}| &\leq \max_{Q\in\mathcal{C}^{\pi}_n}|\mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) - V^{\pi}| \\ &= \max_{Q\in\mathcal{C}^{\pi}_n}|\mathbb{E}_{S\sim\nu_{\text{start}}}[Q(S,\pi) - Q^{\pi}(S,\pi)]| \\ &\stackrel{\text{(i)}}{=} \frac{1}{1-\gamma}\max_{Q\in\mathcal{C}^{\pi}_n}\frac{\mathbb{E}_{\pi}\mathcal{B}^{\pi}Q}{1-\gamma}\end{aligned}$$

where step (i) follows from Lemma 9. Combined with the earlier displays, this completes the proof of the bound (8b).

We now show the inclusion $[\widehat{V}^{\pi}_{\min}, \widehat{V}^{\pi}_{\max}] \ni V^{\pi}$ when weak realizability holds. By definition of weak realizability, there exists some $Q^{\pi}_{\star} \in \mathcal{C}^{\pi}_{\infty}$. In conjunction with our sandwich assumption, we are guaranteed that $Q^{\pi}_{\star} \in \mathcal{C}^{\pi}_{\infty} \subseteq \widehat{\mathcal{C}}^{\pi}_n$, and consequently

$$\widehat{V}^{\pi}_{\min} = \min_{Q\in\widehat{\mathcal{C}}^{\pi}_n} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) \leq \min_{Q\in\mathcal{C}^{\pi}_{\infty}} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) \leq \mathbb{E}_{S\sim\nu_{\text{start}}}Q^{\pi}_{\star}(S,\pi) = V^{\pi}, \quad \text{and}$$
$$\widehat{V}^{\pi}_{\max} = \max_{Q\in\widehat{\mathcal{C}}^{\pi}_n} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) \geq \max_{Q\in\mathcal{C}^{\pi}_{\infty}} \mathbb{E}_{S\sim\nu_{\text{start}}}Q(S,\pi) \geq \mathbb{E}_{S\sim\nu_{\text{start}}}Q^{\pi}_{\star}(S,\pi) = V^{\pi}.$$

8.1.2 Proof of policy optimization claims

We now prove the oracle inequality (9) on the value $V^{\widetilde{\pi}}$ of a policy $\widetilde{\pi}$ that optimizes the max-min criterion. Fix an arbitrary comparator policy π. Starting with the inclusion $[\widehat{V}^{\widetilde{\pi}}_{\text{min}}, \widehat{V}^{\widetilde{\pi}}_{\text{max}}] \ni V^{\widetilde{\pi}}$, we have

$$V^{\widetilde{\pi}} \overset{(i)}{\geq} \widehat{V}^{\widetilde{\pi}}_{\text{min}} \overset{(ii)}{\geq} \widehat{V}^{\pi}_{\text{min}} = V^{\pi} - \Big(V^{\pi} - \widehat{V}^{\pi}_{\text{min}}\Big) \overset{(iii)}{\geq} V^{\pi} - \frac{1}{1-\gamma} \max_{Q \in \mathcal{C}^{\pi}_{n}} \frac{|\mathbb{E}_{\pi} \mathcal{B}^{\pi} Q|}{1-\gamma},$$

where step (i) follows from the stated inclusion at the start of the argument; step (ii) follows since $\widetilde{\pi}$ solves the max-min program; and step (iii) follows from the bound $|V^{\pi} - \widehat{V}^{\pi}_{\text{min}}| \leq \frac{1}{1-\gamma} \max_{Q \in \mathcal{C}^{\pi}_{n}} \frac{\mathbb{E}_{\pi} \mathcal{B}^{\pi} Q}{1-\gamma}$, as proved in the preceding section. This lower bound holds uniformly for all comparators π, from which the stated claim follows.

8.2 Proof of Lemma 9

For each $t = 1, 2, \ldots$, let $\mathbb{E}_t$ be the expectation over the state-action pair at timestep t upon starting from ν_{start}, so that we have $\mathbb{E}_{S \sim \nu_{\text{start}}}(Q - Q^{\pi})(S, \pi) = \mathbb{E}_0[Q - Q^{\pi}]$ by definition. We claim that

$$\mathbb{E}_0[Q - Q^{\pi}] = \sum_{\tau=1}^{t} \gamma^{\tau-1} \mathbb{E}_{\tau-1} \mathcal{B}^{\pi} Q + \gamma^{t} \mathbb{E}_t [Q - Q^{\pi}] \qquad \text{for all } t = 1, 2, \ldots. \tag{49}$$

For the base case $t = 1$, we have

$$\mathbb{E}_0[Q - Q^{\pi}] = \mathbb{E}_0[Q - \mathcal{T}^{\pi} Q] + \mathbb{E}_0[\mathcal{T}^{\pi} Q - \mathcal{T}^{\pi} Q^{\pi}] = \mathbb{E}_0[Q - \mathcal{T}^{\pi} Q] + \gamma \mathbb{E}_1[Q - Q^{\pi}], \tag{50}$$

where we have used the definition of the Bellman evaluation operator to assert that $\mathbb{E}_0[\mathcal{T}^{\pi} Q - \mathcal{T}^{\pi} Q^{\pi}] = \gamma \mathbb{E}_1[Q - Q^{\pi}]$. Since $Q - \mathcal{T}^{\pi} Q = \mathcal{B}^{\pi} Q$, the equality (50) is equivalent to the claim (49) with $t = 1$.

Turning to the induction step, we now assume that the claim (49) holds for some $t \geq 1$, and show that it holds at step $t + 1$. By a similar argument, we can write

$$\begin{aligned} \gamma^{t} \mathbb{E}_t[Q - Q^{\pi}] = \gamma^{t} \mathbb{E}_t[Q - \mathcal{T}^{\pi} Q + \mathcal{T}^{\pi} Q - \mathcal{T}^{\pi} Q^{\pi}] &= \gamma^{t} \mathbb{E}_t[Q - \mathcal{T}^{\pi} Q] + \gamma^{t+1} \mathbb{E}_{t+1}[Q - Q^{\pi}] \\ &= \gamma^{t} \mathbb{E}_t \mathcal{B}^{\pi} Q + \gamma^{t+1} \mathbb{E}_{t+1}[Q - Q^{\pi}]. \end{aligned}$$

By the induction hypothesis, equality (49) holds for t, and substituting the above equality shows that it also holds at time $t + 1$.

Since the equivalence (49) holds for all t, we can take the limit as $t \to \infty$, and doing so yields the claim.

8.3 Proof of Theorem 3

In the statement of the theorem, we require choosing $\epsilon > 0$ to satisfy the upper bound $\epsilon^2 \precsim \frac{\rho(\epsilon, \delta)}{n}$, and then provide an upper bound in terms of $\sqrt{\rho(\epsilon, \delta)/n}$. It is equivalent to instead choose ϵ to satisfy the lower bound $\epsilon^2 \succsim \frac{\rho(\epsilon, \delta)}{n}$, and then provide upper bounds proportional to ϵ. For the purposes of the proof, the latter formulation turns out to be more convenient and we pursue it here.

To streamline notation, let us introduce the shorthand $\langle f, \mathcal{D}^\pi(Q)\rangle \stackrel{def}{=} \langle f, \delta^\pi(Q)\rangle_n - \langle f, \mathcal{B}^\pi(Q)\rangle_\mu$. For each pair (Q, π), we then define the random variable

$$Z_n(Q,\pi) \stackrel{def}{=} \sup_{f \in \mathcal{F}^\pi} \frac{\left|\langle f, \mathcal{D}^\pi(Q)\rangle\right|}{\sqrt{\|f\|_n^2 + \lambda}}.$$

Central to our proof of the theorem is a uniform bound on this random variable, one that holds for all pairs (Q, π). In particular, our strategy is to exhibit some $\epsilon > 0$ for which, upon setting $\lambda = 4\epsilon^2$, we have the guarantees

$$\frac{1}{4} \leq \frac{\sqrt{\|f\|_n^2 + \lambda}}{\sqrt{\|f\|_\mu^2 + \lambda}} \leq 2 \qquad \text{uniformly for all } f \in \mathcal{F}, \text{ and} \tag{51a}$$

$$Z_n(Q,\pi) \leq \epsilon \quad \text{uniformly for all } (Q,\pi), \tag{51b}$$

both with probability at least $1 - \delta$. In particular, consistent with the theorem statement, we show that this claim holds if we choose $\epsilon > 0$ to satisfy the inequality

$$\epsilon^2 \geq \bar{c}\frac{\rho(\epsilon,\delta)}{n} \tag{52}$$

where $\bar{c} > 0$ is a sufficiently large (but universal) constant.

Supposing that the bounds (51a) and (51b) hold, let us now establish the set inclusions claimed in the theorem.

Inclusion $\mathcal{C}_\infty^\pi \subseteq \widehat{\mathcal{C}}_n^\pi(\epsilon)$: Define the random variable $M_n(Q,\pi) \stackrel{def}{=} \sup_{f\in\mathcal{F}^\pi} \frac{|\langle f, \mathcal{B}^\pi(Q)\rangle_\mu|}{\sqrt{\|f\|_n^2+\lambda}}$, and observe that $Q \in \mathcal{C}_\infty^\pi$ implies that $M_n(Q,\pi) = 0$. With this definition, we have

$$\sup_{f\in\mathcal{F}^\pi} \frac{\left|\langle f, \delta^\pi(Q)\rangle_n\right|}{\sqrt{\|f\|_n^2 + \lambda}} \stackrel{(i)}{\leq} M_n(Q,\pi) + Z_n(Q,\pi) \stackrel{(ii)}{\leq} \epsilon$$

where step (i) follows from the triangle inequality; and step (ii) follows since $M_n(Q,\pi) = 0$, and $Z_n(Q,\pi) \leq \epsilon$ from the bound (51b).

Inclusion $\widehat{\mathcal{C}}_n^\pi(\epsilon) \subseteq \mathcal{C}_n^\pi(4\epsilon)$ By the definition of $\mathcal{C}_n^\pi(4\epsilon)$, we need to show that

$$\bar{M}(Q,\pi) \stackrel{def}{=} \sup_{f\in\mathcal{F}^\pi} \frac{\left|\langle f, \mathcal{B}^\pi(Q)\rangle_\mu\right|}{\sqrt{\|f\|_\mu^2 + \lambda}} \leq 4\epsilon \qquad \text{for any } Q \in \widehat{\mathcal{C}}_n^\pi(\epsilon).$$

Now we have

$$\bar{M}(Q,\pi) \stackrel{(i)}{\leq} 2M_n(Q,\pi) \stackrel{(ii)}{\leq} 2\left\{\sup_{f\in\mathcal{F}^\pi} \frac{\left|\langle f, \delta^\pi(Q)\rangle_n\right|}{\sqrt{\|f\|_n^2 + \lambda}} + Z_n(Q,\pi)\right\} \stackrel{(iii)}{\leq} 2\{\epsilon + \epsilon\} \;=\; 4\epsilon,$$

where step (i) follows from the sandwich relation (51a); step (ii) follows from the triangle inequality and the definition of $Z_n(Q,\pi)$; and step (iii) follows since $Z_n(Q,\pi) \leq \epsilon$ from the bound (51b), and

$$\sup_{f\in\mathcal{F}^\pi} \frac{\left|\langle f, \delta^\pi(Q)\rangle_n\right|}{\sqrt{\|f\|_n^2 + \lambda}} \leq \epsilon, \qquad \text{using the inclusion } Q \in \widehat{\mathcal{C}}_n^\pi(\epsilon).$$

Consequently, the remainder of our proof is devoted to establishing the claims (51a) and (51b). In doing so, we make repeated use of some Bernstein bounds, stated in terms of the shorthand $\Psi_n(\delta) = \frac{\log(n/\delta)}{n}$.

Lemma 10. *There is a universal constant c such each the following statements holds with probability at least $1-\delta$. For any f, we have*

$$\Big| \|f\|_n^2 - \|f\|_\mu^2 \Big| \leq c \left\{ \|f\|_\mu \sqrt{\Psi_n(\delta)} + \Psi_n(\delta) \right\}, \tag{53a}$$

and for any (Q, π) and any function f, we have

$$\big| \langle f, \, \delta^\pi(Q) \rangle_n - \langle f, \, \mathcal{B}^\pi(Q) \rangle_\mu \big| \leq c \left\{ \|f\|_\mu \sqrt{\Psi_n(\delta)} + \|f\|_\infty \Psi_n(\delta) \right\}. \tag{53b}$$

These bounds follow by identifying a martingale difference sequence, and applying a form of Bernstein's inequality tailored to the martingale setting. See Section 8.6.3 for the details.

8.4 Proof of the sandwich relation (51a)

We claim that (modulo the choice of constants) it suffices to show that

$$\Big| \|f\|_n - \|f\|_\mu \Big| \leq \epsilon \qquad \text{uniformly for all } f \in \mathcal{F} \tag{54}$$

for some universal constant c'. Indeed, when this bound holds, we have

$$\|f\|_n + 2\epsilon \leq \|f\|_\mu + 3\epsilon \leq \frac{3}{2}\{\|f\|_\mu + 2\epsilon\}, \quad \text{and} \quad \|f\|_n + 2\epsilon \geq \|f\|_\mu + \epsilon \geq \frac{1}{2}\{\|f\|_\mu + 2\epsilon\},$$

so that $\frac{\|f\|_\mu + 2\epsilon}{\|f\|_n + 2\epsilon} \in \big[\frac{1}{2}, \frac{3}{2}\big]$. To relate this statement to the claimed sandwich, observe the inclusion $\frac{\|f\| + \sqrt{2}\epsilon}{\sqrt{\|f\|^2 + 4\epsilon^2}} \in [1, \sqrt{2}]$, where $\|f\|$ can be either $\|f\|_n$ or $\|f\|_\mu$. Combining this fact with our previous bound, we see that $\frac{\sqrt{\|f\|_n^2 + 4\epsilon^2}}{\sqrt{\|f\|_\mu^2 + 4\epsilon^2}} \in \Big[\frac{1}{\sqrt{2}}\frac{1}{2}, \frac{3\sqrt{2}}{2}\Big] \subset \big[\frac{1}{4}, 3\big]$, as claimed.

The remainder of our analysis is focused on proving the bound (54). Defining the random variable $Y_n(f) = \big| \|f\|_n - \|f\|_\mu \big|$, we need to establish a high probability bound on $\sup_{f \in \mathcal{F}} Y_n(f)$. Let $\{f^1, \ldots, f^N\}$ be an ϵ-cover of $\mathcal{F}$ in the sup-norm. For any $f \in \mathcal{F}$, we can find some f^j such that $\|f - f^j\|_\infty \leq \epsilon$, whence

$$\begin{aligned} Y_n(f) \leq Y_n(f^j) + \big| Y_n(f^j) - Y_n(f) \big| &\overset{(i)}{\leq} Y_n(f^j) + \big| \|f^j\|_n - \|f\|_n \big| + \big| \|f^j\|_\mu - \|f\|_\mu \big| \\ &\overset{(ii)}{\leq} Y_n(f^j) + \|f^j - f\|_n + \|f^j - f\|_\mu \\ &\overset{(iii)}{\leq} Y_n(f^j) + 2\epsilon, \end{aligned}$$

where steps (i) and (ii) follow from the triangle inequality; and step (iii) follows from the inequality $\max\{\|f^j - f\|_n, \|f^j - f\|_\mu\} \leq \|f^j - f\|_\infty \leq \epsilon$. Thus, we have reduced the problem to bounding a finite maximum.

Note that if $\max\{\|f^j\|_n, \|f^j\|_\mu\} \leq \epsilon$, then we have $Y_n(f^j) \leq 2\epsilon$ by the triangle inequality. Otherwise, we may assume that $\|f^j\|_n + \|f^j\|_n \geq \epsilon$. With probability at least $1-\delta$, we have

$$\Big|\|f^j\|_n - \|f\|_\mu\Big| = \frac{\Big|\|f^j\|_n^2 - \|f\|_\mu^2\Big|}{\|f^j\|_n + \|f^j\|_\mu} \overset{(i)}{\leq} \frac{c\big\{\|f^j\|_\mu\sqrt{\Psi_n(\delta)} + \Psi_n(\delta)\big\}}{\|f^j\|_\mu + \|f^j\|_n} \overset{(ii)}{\leq} c\Big\{\sqrt{\Psi_n(\delta)} + \frac{\Psi_n(\delta)}{\epsilon}\Big\},$$

where step (i) follows from the Bernstein bound (53a) from Lemma 10, and step (ii) uses the fact that $\|f^j\|_n + \|f^j\|_n \geq \epsilon$.

Taking union bound over all N elements in the cover and replacing δ with δ/N, we have

$$\max_{j\in[N]} Y_n(f^j) \leq c\Big\{\sqrt{\Psi_n(\delta/N)} + \frac{\Psi_n(\delta/N)}{\epsilon}\Big\}$$

with probability at least $1-\delta$. Recalling that $N = N_\epsilon(\mathcal{F})$, our choice (52) of ϵ ensures that $\sqrt{\Psi_n(\delta/N)} \leq c\ \epsilon$ for some universal constant c. Putting together the pieces (and increasing the constant $\bar{c}$ in the choice (52) of ϵ as needed) yields the claim.

8.5 Proof of the uniform upper bound (51b)

We need to establish an upper bound on $Z_n(Q,\pi)$ that that holds uniformly for all (Q,π). Our first step is to prove a high probability bound for a fixed pair. We then apply a standard discretization argument to make it uniform in the pair.

Note that we can write $Z_n(Q,\pi) = \sup_{f\in\mathcal{F}} \frac{V_n(f)}{\sqrt{\|f\|_n^2+\lambda}}$, where we have defined $V_n(f) \overset{def}{=} |\langle f, \mathcal{D}^\pi(Q)\rangle|$. Our first lemma provides a uniform bound on the latter random variables:

Lemma 11. *Suppose that $\epsilon^2 \geq \Psi_n\big(\delta/N_\epsilon(\mathcal{F})\big)$. Then we have*

$$V_n(f) \leq c\big\{\|f\|_\mu\epsilon + \epsilon^2\big\} \qquad \textit{for all } f\in\mathcal{F} \tag{55}$$

with probability at least $1-\delta$.

See Section 8.6.1 for the proof of this claim.

We claim that the bound (55) implies that, for any fixed pair (Q,π), we have

$$Y_n(Q,\pi) \leq c'\epsilon \qquad \text{with probability at least } 1-\delta.$$

Indeed, when Lemma 11 holds, for any $f\in\mathcal{F}$, we can write

$$\frac{V_n(f)}{\sqrt{\|f\|_n^2+\lambda}} = \frac{\sqrt{\|f\|_\mu^2+\lambda}}{\sqrt{\|f\|_n^2+\lambda}}\,\frac{V_n(f)}{\sqrt{\|f\|_\mu^2+\lambda}} \overset{(i)}{\leq} 3\,\frac{c\big\{\|f\|_\mu\epsilon+\epsilon^2\big\}}{\sqrt{\|f\|_\mu^2+\lambda}} \overset{(ii)}{\leq} c'\epsilon,$$

where step (i) uses the sandwich relation (51a), along with the bound (55); and step (ii) follows given the choice $\lambda = 4\epsilon^2$. We have thus proved that for any fixed (Q,π) and $\epsilon \geq \Psi_n\big(\delta/N_\epsilon(\mathcal{F})\big)$, we have

$$Z_n(Q,\pi) \leq c'\epsilon \qquad \text{with probability at least } 1-\delta. \tag{56}$$

Our next step is to upgrade this bound to one that is uniform over all pairs (Q,π). We do so via a discretization argument: let $\{Q^j\}_{j=1}^J$ and $\{\pi^k\}_{k=1}^K$ be ϵ-coverings of $\mathcal{Q}$ and Π, respectively.

Lemma 12. *We have the upper bound*

$$\sup_{Q,\pi} Z_n(Q,\pi) \leq \max_{(j,k)\in[J]\times[K]} Z_n(Q^j,\pi^k) + 4\epsilon. \tag{57}$$

See Section 8.6.2 for the proof of this claim.

If we replace δ with $\delta/(JK)$, then we are guaranteed that the bound (56) holds uniformly over the family $\{Q^j\}_{j=1}^J \times \{\pi^k\}_{k=1}^K$. Recalling that $J = N_\epsilon(\mathcal{Q})$ and $K = N_\epsilon(\Pi)$, we conclude that for any ϵ satisfying the inequality (52), we have $\sup_{Q,\pi} Z_n(Q,\pi) \leq \tilde{c}\epsilon$ with probability at least $1-\delta$. (Note that by suitably scaling up ϵ via the choice of constant $\bar{c}$ in the bound (52), we can arrange for $\tilde{c} = 1$, as in the stated claim.)

8.6 Proofs of supporting lemmas

In this section, we collect together the proofs of Lemmas 11 and 12, which were stated and used in Section 8.5.

8.6.1 Proof of Lemma 11

We first localize the problem to the class $\mathcal{F}(\epsilon) = \{f \in \mathcal{F} \mid \|f\|_\mu \leq \epsilon\}$. In particular, if there exists some $\tilde{f} \in \mathcal{F}$ that violates (55), then the rescaled function $f = \epsilon\tilde{f}/\|\tilde{f}\|_\mu$ belongs to $\mathcal{F}(\epsilon)$, and satisfies $V_n(f) \geq c\epsilon^2$. Consequently, it suffices to show that $V_n(f) \leq c\epsilon^2$ for all $f \in \mathcal{F}(\epsilon)$.

Choose an ϵ-cover of $\mathcal{F}$ in the sup-norm with $N = N_\epsilon(\mathcal{F})$ elements. Using this cover, for any $f \in \mathcal{F}(\epsilon)$, we can find some f^j such that $\|f - f^j\|_\infty \leq \epsilon$. Thus, for any $f \in \mathcal{F}(\epsilon)$, we can write

$$V_n(f) \leq V_n(f^j) + V_n(f - f^j) \quad \leq \underbrace{V_n(f^j)}_{T_1} + \underbrace{\sup_{g\in\mathcal{G}(\epsilon)} V_n(g)}_{T_2}, \tag{58}$$

where $\mathcal{G}(\epsilon) \overset{def}{=} \{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}, \|f_1 - f_2\|_\infty \leq \epsilon\}$. We bound each of these two terms in turn. In particular, we show that each of T_1 and T_2 are upper bounded by $c\epsilon^2$ with high probability.

Bounding T_1: From the Bernstein bound (53b), we have

$$V_n(f^k) \leq c\big\{\|f^k\|_\mu\sqrt{\Psi_n(\delta/N)} + \|f^k\|_\infty \Psi_n(\delta/N)\big\} \qquad \text{for all } k \in [N]$$

with probability at least $1-\delta$. Now for the particular f^j chosen to approximate $f \in \mathcal{F}(\epsilon)$, we have

$$\|f^j\|_\mu \leq \|f^j - f\|_\mu + \|f\|_\mu \leq 2\epsilon,$$

where the inequality follows since $\|f^j - f\|_\mu \leq \|f^j - f\|_\infty \leq \epsilon$, and $\|f\|_\mu \leq \epsilon$. Consequently, we conclude that

$$T_1 \leq c\Big\{2\epsilon\sqrt{\Psi_n(\delta/N)} + \Psi_n(\delta/N)\Big\} \;\leq\; c'\epsilon^2 \qquad \text{with probability at least } 1-\delta.$$

where the final inequality follows from our choice of ϵ.

Bounding T_2: Define $\mathcal{G} \stackrel{def}{=} \{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}\}$. We need to bound a supremum of the process $\{V_n(g), g \in \mathcal{G}\}$ over the subset $\mathcal{G}(\epsilon)$. From the Bernstein bound (53b), the increments $V_n(g_1) - V_n(g_2)$ of this process are sub-Gaussian with parameter $\|g_1 - g_2\|_\mu \leq \|g_1 - g_2\|_\infty$, and sub-exponential with parameter $\|g_1 - g_2\|_\infty$. Therefore, we can apply a chaining argument that uses the metric entropy $\log N_t(\mathcal{G})$ in the supremum norm. Moreover, we can terminate the chaining at 2ϵ, because we are taking the supremum over the subset $\mathcal{G}(\epsilon)$, and it has sup-norm diameter at most 2ϵ. Moreover, the lower interval of the chain can terminate at $2\epsilon^2$, since our goal is to prove an upper bound of this order. Then, by using high probability bounds for the suprema of empirical processes (e.g., Theorem 5.36 in the book [Wai19]), we have

$$T_2 \leq c_1 \int_{2\epsilon^2}^{2\epsilon} \phi\Big(\frac{\log N_t(\mathcal{G})}{n}\Big) dt + c_2 \big\{\epsilon \sqrt{\Psi_n(\delta)} + \epsilon \Psi_n(\delta)\big\} + 2\epsilon^2$$

with probability at least $1 - \delta$. (Here the reader should recall our shorthand $\phi(s) = \max\{s, \sqrt{s}\}$.)

Since $\mathcal{G}$ consists of differences from $\mathcal{F}$, we have the upper bound $\log N_t(\mathcal{G}) \leq 2 \log N_{t/2}(\mathcal{F})$, and hence (after making the change of variable $u = t/2$ in the integrals)

$$T_2 \leq c_1' \int_{\epsilon^2}^{\epsilon} \phi\Big(\frac{\log N_u(\mathcal{F})}{n}\Big) du + c_2 \big\{\epsilon \sqrt{\Psi_n(\delta)} + \epsilon \Psi_n(\delta)\big\} \;\leq\; \tilde{c}\epsilon^2,$$

where the last inequality follows from our choice of ϵ.

8.6.2 Proof of Lemma 12

By our choice of the ϵ-covers, for any (Q, π), there is a pair (Q^j, π^k) such that

$$\|Q^j - Q\|_\infty \leq \epsilon, \quad \text{and} \quad \|\pi^k - \pi\|_{\infty,1} = \sup_s \|\pi^k(\cdot \mid s) - \pi(\cdot \mid s)\|_1 \leq \epsilon.$$

Using this pair, an application of the triangle inequality yields

$$\big|Z_n(Q, \pi) - Z_n(Q^j, \pi^k)\big| \leq \underbrace{\big|Z_n(Q, \pi) - Z_n(Q, \pi^k)\big|}_{T_1} + \underbrace{\big|Z_n(Q, \pi^k) - Z_n(Q^j, \pi^k)\big|}_{T_2}$$

We bound each of these terms in turn, in particular proving that $T_1 + T_2 \leq 24\epsilon$. Putting together the pieces yields the bound stated in the lemma.

Bounding T_2: From the definition of Z_n, we have

$$T_2 = \big|Z_n(Q, \pi^k) - Z_n(Q^j, \pi^k)\big| \leq \sup_{f \in \mathcal{F}} \frac{|\langle f, \mathcal{D}^{\pi^k}(Q - Q^j)\rangle|}{\sqrt{\|f\|_n^2 + \lambda}}.$$

Now another application of the triangle inequality yields

$$\begin{aligned}
|\langle f, \mathcal{D}^{\pi^k}(Q - Q^j)\rangle| &\leq |\langle f, \delta^{\pi^k}(Q - Q^j)\rangle_n| + ||\langle f, \mathcal{B}^{\pi^k}(Q - Q^j)\rangle|_\mu \\
&\leq \|f\|_n \|\delta^{\pi^k}(Q - Q^j)\|_n + \|f\|_\mu \|\mathcal{B}^{\pi^k}(Q - Q^j)\|_\mu \\
&\leq \max\{\|f\|_n, \|f\|_\mu\} \Big\{\|\delta^{\pi^k}(Q - Q^j)\|_\infty + \|\mathcal{B}^{\pi^k}(Q - Q^j)\|_\infty\Big\}
\end{aligned}$$

where step (i) follows from the Cauchy–Schwarz inequality. Now in terms of the shorthand $\Delta \overset{def}{=} Q - Q^j$, we have

$$\|\mathcal{B}^{\pi^k}(Q - Q^j)\|_\infty = \sup_{(s,a)} \Big| \Delta(s,a) - \gamma \mathbb{E}_{s^+ \sim \mathbb{P}(s,a)} \big[\Delta(s^+, \pi) \big] \Big| \leq 2\|\Delta\|_\infty \leq 2\epsilon. \tag{59a}$$

An entirely analogous argument yields

$$\|\delta^{\pi^k}(Q - Q^j)\|_\infty \leq 2\epsilon \tag{59b}$$

Conditioned on the sandwich relation (51a), we have $\sup_{f \in \mathcal{F}} \frac{\max\{\|f\|_n, \|f\|_\mu\}}{\sqrt{\|f\|_n^2 + \lambda}} \leq 4$. Combining this bound with inequalities (59a) and (59b), we have shown that $T_2 \leq 4\big\{2\epsilon + 2\epsilon\big\} = 16\epsilon$.

Bounding T_1: In this case, a similar argument yields

$$|\langle f, \, (\mathcal{D}^{\pi} - \mathcal{D}^{\pi^k})(Q) \rangle| \leq \max\{\|f\|_n, \|f\|_\mu\} \, \Big\{ \|(\delta^\pi - \delta^{\pi^k})(Q)\|_n + \|(\mathcal{B}^\pi - \mathcal{B}^{\pi^k})(Q)\|_\mu \}.$$

Now we have

$$\begin{aligned} \|(\delta^\pi - \delta^{\pi^k})(Q)\|_n &\leq \max_{i=1,\dots,n} \Big| \sum_{a'} \big(\pi(a' \mid s_i) - \pi^k(a' \mid s_i) \big) Q(s_i^+, a') \Big| \\ &\leq \max_s \sum_{a'} |\pi(a' \mid s) - \pi^k(a \mid s)| \, \|Q\|_\infty \\ &\leq \epsilon. \end{aligned}$$

A similar argument yields that $\|(\mathcal{B}^\pi - \mathcal{B}^{\pi^k})(Q)\|_\mu| \leq \epsilon$, and arguing as before, we conclude that $T_1 \leq 4\{\epsilon + \epsilon\} = 8\epsilon$.

8.6.3 Proof of Lemma 10

Our proof of this claim makes use of the following known Bernstein bound for martingale differences (cf. Theorem 1 in the paper [BLL+11]). Recall the shorthand notation $\Psi_n(\delta) = \frac{\log(n/\delta)}{n}$.

Lemma 13 (Bernstein's Inequality for Martingales)**.** *Let $\{X_t\}_{t \geq 1}$ be a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t \geq 1}$. Suppose that $|X_t| \leq 1$ almost surely, and let $\mathbb{E}_t$ denote expectation conditional on $\mathcal{F}_t$. Then for all $\delta \in (0,1)$, we have*

$$\Big| \frac{1}{n} \sum_{t=1}^n X_t \Big| \leq 2 \Big[\Big(\frac{1}{n} \sum_{t=1}^n \mathbb{E}_t X_t^2 \Big) \Psi_n(2\delta) \Big]^{1/2} + 2\Psi_n(2\delta) \tag{60}$$

with probability at least $1 - \delta$.

With this result in place, we divide our proof into two parts, corresponding to the two claims (53b) and (53a) stated in Lemma 10.

Proof of the bound (53b): Recall that at step i, the triple (s, a, o) is drawn according to a conditional distribution $\mu_i(\cdot \mid \mathcal{F}_i)$. Similarly, we let d_i denote the distribution of (s, a, r, s^+, o) conditioned on the filtration $\mathcal{F}_i$. Note that μ_i is obtained from d_i by marginalizing out the pair (r, s^+). Moreover, by the tower property of expectation, the Bellman error is equivalent to the average TD error.

Using these facts, we have the equivalence

$$\begin{aligned}
\langle f, \delta^\pi Q\rangle_{d_i} &= \mathbb{E}_{d_i}\big\{f(s,a,o)[Q(s,a) - r - \gamma Q(s^+, \pi)]\big\} \\
&= \mathbb{E}_{(s,a,o)\sim\mu_i}\big\{f(s,a,o)\mathbb{E}_{r\sim R(s,a), s^+\sim\mathbb{P}(s,a)}[Q(s,a) - r - \gamma Q(s^+,\pi)]\big\} \\
&= \mathbb{E}_{(s,a,o)\sim\mu_i}\big\{f(s,a,o)[Q(s,a) - (\mathcal{T}^\pi Q)(s,a)]\big\} \\
&= \langle f, \mathcal{B}^\pi Q\rangle_{\mu_i}.
\end{aligned}$$

As a consequence, we can write $\langle f,\, \delta^\pi(Q)\rangle_n - \langle f,\, \mathcal{B}^\pi(Q)\rangle_\mu = \frac{1}{n}\sum_{i=1}^n W_i$ where

$$W_i \stackrel{def}{=} f(s_i,a_i,o_i)[Q(s_i,a_i) - r_i - \gamma Q(s_i^+,\pi)] - \mathbb{E}_{d_i}\big\{f(s,a,o)[Q(s,a) - r - \gamma Q(s^+,\pi)]\big\}$$

defines a martingale difference sequence (MDS). Thus, we can prove the claim by applying a Bernstein martingale inequality.

Since $\|r\|_\infty \le 1$ and $\|Q\|_\infty \le 1$ by assumption, we have $\|W_i\|_\infty \le 3\|f\|_\infty$, and

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{d_i}[W_i^2] \le 9\,\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mu_i}[f^2(s_i,a_i,o_i)] \;=\; 9\|f\|_\mu^2.$$

Consequently, the claimed bound (53b) follows by applying the Bernstein bound stated in Lemma 13.

Proof of the bound (53a): In this case, we have the additive decomposition

$$\|f\|_n^2 - \|f\|_\mu^2 = \frac{1}{n}\sum_{i=1}^n \Big\{\underbrace{f^2(s_i,a_i,o_i) - \mathbb{E}_{\mu_i}[f^2(s,a,o)]}_{W_i'}\Big\},$$

where $\{W_i'\}_{i=1}^n$ again defines a martingale difference sequence. Note that $\|W_i'\|_\infty \le 2\|f\|_\infty^2 \le 2$, and

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mu_i}[(W_i')^2] \stackrel{(i)}{\le} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mu_i}\big[f^4(S,A,O)\big] \;\le\; \|f\|_\infty^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mu_i}\big[f^2(S,A,O)\big] \;\stackrel{(ii)}{\le}\; \|f\|_\mu^2,$$

where step (i) uses the fact that the variance of f^2 is at most the fourth moment, and step (ii) uses the bound $\|f\|_\infty \le 1$. Consequently, the claimed bound (53a) follows by applying the Bernstein bound stated in Lemma 13.

9 Proofs for Section 4 and Section 6.5

In this section, we collect together the proofs of results stated without proof in Section 4 and Section 6.5.

9.1 Proof of Proposition 1

Proof. Since $f^* \in \mathcal{F}^\pi$, we are guaranteed that the corresponding constraint must hold. It reads as

$$|\mathbb{E}_\mu \frac{1}{b_\pi} \frac{d_\pi}{\mu} \mathcal{B}^\pi Q|^2 = \frac{1}{b_\pi^2} |\mathbb{E}_\pi \mathcal{B}^\pi Q|^2 \overset{(iii)}{\leq} \big(\frac{1}{b_\pi^2} \|\frac{d_\pi}{\mu}\|_\mu^2 + \lambda\big)\frac{\rho}{n}.$$

where step (iii) follows from the definition of population constraint. Re-arranging yields the upper bound

$$\frac{|\mathbb{E}_\mu \frac{d_\pi}{\mu} \mathcal{B}^\pi Q|^2}{(1+\lambda)\frac{\rho}{n}} \leq \frac{\big(\|\frac{d_\pi}{\mu}\|_\mu^2 + b_\pi^2 \lambda\big)\frac{\rho}{n}}{(1+\lambda)\frac{\rho}{n}} = \frac{\mathbb{E}_\pi\left[\frac{d_\pi(S,A)}{\mu(S,A)}\right] + b_\pi^2\lambda}{1+\lambda},$$

where the final step uses the fact that

$$\|\frac{d_\pi}{\mu}\|_\mu^2 = \mathbb{E}_\mu \frac{d_\pi^2(S,A)}{\mu^2(S,A)} = \mathbb{E}_\pi \frac{d_\pi(S,A)}{\mu(S,A)}$$

Thus, we have established the bound (i) in our claim (12).

The upper bound (ii) follows immediately since $\mathbb{E}_\pi \frac{d_\pi(s,a)}{\mu(s,a)} \leq \sup_{(s,a)} \frac{d_\pi(s,a)}{\mu(s,a)} \leq b_\pi$.

□

9.2 Proof of Lemma 1

Some simple algebra yields

$$\mathcal{B}^\pi Q - \mathcal{B}^\pi Q_\star^\pi = [Q - \mathcal{T}^\pi Q] - [Q_\star^\pi - \mathcal{T}^\pi Q_\star^\pi] = (\mathcal{I} - \gamma \mathbb{P}^\pi)(Q - Q_\star^\pi) = (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon.$$

Taking expectations under π and recalling that $\langle f, \mathcal{B}^\pi Q_\star^\pi \rangle_\pi = 0$ for all $f \in \mathcal{F}^\pi$ yields

$$\langle f, \mathcal{B}^\pi Q \rangle_\pi = \langle f, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\pi.$$

Notice that for any $Q \in \mathcal{Q}^\pi$ there exists a test function $\epsilon = Q - Q_\star^\pi \in \mathcal{E}^\pi$, and the associated population constraint reads

$$\frac{\big|\langle \epsilon, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\mu\big|}{\sqrt{\|\epsilon\|_\mu^2 + \lambda}} \leq \sqrt{\frac{\rho}{n}}.$$

Consequently, the off-policy cost coefficient can be upper bounded as

$$K^\pi \leq \max_{\epsilon \in \mathcal{E}_\star^\pi} \Big\{ \frac{\rho}{n} \, \frac{\langle \mathbb{1}, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\pi^2}{1+\lambda} \Big\} \leq \max_{\epsilon \in \mathcal{E}_\star^\pi} \Big\{ \frac{\|\epsilon\|_\mu^2 + \lambda}{\|\,\mathbb{1}\,\|_\pi^2 + \lambda} \, \frac{\langle \mathbb{1}, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\pi^2}{\langle \epsilon, (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon \rangle_\mu^2} \Big\},$$

as claimed in the bound (14).

9.3 Proof of Lemma 7

If weak Bellman closure holds, then we can write

$$\mathcal{B}^\pi Q = Q - \mathcal{T}^\pi Q = Q - \mathcal{P}^\pi(Q) \in \mathcal{E}^\pi.$$

For any $Q \in \mathcal{Q}^\pi$, the function $\epsilon = Q - \mathcal{P}^\pi(Q)$ belongs to $\mathcal{E}^\pi$, and the associated population constraint reads $\frac{|\langle \epsilon, \epsilon \rangle_\mu|}{\sqrt{\|\epsilon\|_\mu^2 + \lambda}} \leq \sqrt{\frac{\rho}{n}}$. Consequently, the off-policy cost coefficient is upper bounded as

$$K^\pi \leq \max_{\epsilon \in \mathcal{E}^\pi} \Big\{ \frac{n}{\rho} \, \frac{v\langle \mathbb{1}, \epsilon \rangle_\pi^2}{1+\lambda} \Big\} \leq \max_{\epsilon \in \mathcal{E}^\pi} \Big\{ \frac{\|\epsilon\|_\mu^2 + \lambda}{1+\lambda} \, \frac{\langle \mathbb{1}, \epsilon \rangle_\pi^2}{\langle \epsilon, \epsilon \rangle_\mu^2} \Big\} \; \leq \; \max_{\epsilon \in \mathcal{E}^\pi} \Big\{ \frac{\langle \mathbb{1}, \epsilon \rangle_\pi^2}{\langle \epsilon, \epsilon \rangle_\mu^2} \Big\},$$

where the final inequality follows from the fact that $\|\epsilon\|_\mu \leq 1$.

9.4 Proof of Lemma 2

We split our proof into the two separate claims.

Proof of the bound (16a): When the test function class includes $\mathcal{F}_\pi^{\mathcal{B}}$, then any Q feasible must satisfy the population constraints

$$\frac{\langle \mathcal{B}^\pi Q', \mathcal{B}^\pi Q \rangle_\mu}{\sqrt{\|\mathcal{B}^\pi Q'\|_\mu^2 + \lambda}} \leq \sqrt{\frac{\rho}{n}}, \qquad \text{for all } Q' \in Q^\pi.$$

Setting $Q' = Q$ yields $\frac{\|\mathcal{B}^\pi Q\|_\mu^2}{\sqrt{\|\mathcal{B}^\pi Q\|_\mu^2 + \lambda}} \leq \sqrt{\frac{\rho}{n}}$. If $\|\mathcal{B}^\pi Q\|_\mu^2 \geq \lambda$, then the claim holds, given our choice $\lambda = c\frac{\rho}{n}$ for some constant c. Otherwise, the constraint can be weakened to $\frac{\|\mathcal{B}^\pi Q\|_\mu^2}{\sqrt{2\|\mathcal{B}^\pi Q\|_\mu^2}} \leq \sqrt{\frac{\rho}{n}}$, which yields the bound (16a).

Proof of the bound (16b): We now prove the sequence of inequalities stated in equation (16b). Inequality (i) follows directly from the definition of K^π and Lemma 2. Turning to inequality (ii), an application of Jensen's inequality yields

$$\langle \mathbb{1}, \mathcal{B}^\pi Q \rangle_\pi^2 = [\mathbb{E}_\pi \mathcal{B}^\pi Q]^2 \leq \mathbb{E}_\pi [\mathcal{B}^\pi Q]^2 = \|\mathcal{B}^\pi Q\|_\pi^2.$$

Finally, inequality (iii) follows by observing that

$$\sup_{Q \in \mathcal{Q}^\pi} \frac{\|\mathcal{B}^\pi Q\|_\pi^2}{\|\mathcal{B}^\pi Q\|_\mu^2} = \sup_{Q \in \mathcal{Q}^\pi} \frac{\mathbb{E}_\pi[(\mathcal{B}^\pi Q)(s,a)]^2}{\mathbb{E}_\mu[(\mathcal{B}^\pi Q)(s,a)]^2} = \sup_{Q \in \mathcal{Q}^\pi} \frac{\mathbb{E}_\mu \Big[\frac{d_\pi(s,a)}{\mu(s,a)}\Big] [(\mathcal{B}^\pi Q)(s,a)]^2}{\mathbb{E}_\mu[(\mathcal{B}^\pi Q)(s,a)]^2} \leq \sup_{(s,a)} \frac{d_\pi(s,a)}{\mu(s,a)}.$$

10 Proofs for the Linear Setting

We now prove the results stated in Section 5. Throughout this section, the reader should recall that Q takes the linear function $Q(s,a) = \langle w, \phi(s,a)\rangle$, so that the bulk of our arguments operate directly on the weight vector $w \in \mathbb{R}^d$.

Given the linear structure, the population and empirical covariance matrices of the feature vectors play a central role. We make use of the following known result (cf. Lemma 1 in the paper [ZJZ21]) that relates these objects:

Lemma 14 (Covariance Concentration). *There are universal constants* (c_1, c_2, c_3) *such that for any* $\delta \in (0,1)$, *we have*

$$c_1 \mathbb{E}_\mu \phi\phi^\top \preceq \frac{1}{n}\sum_{i=1}^n \phi_i\phi_i^\top + \frac{c_2}{n}\log\frac{nd}{\delta} I \preceq c_3 \mathbb{E}_\mu \phi\phi^\top + \frac{c_4}{n}\log\frac{nd}{\delta} I. \tag{61}$$

with probability at least $1-\delta$.

10.1 Proof of Proposition 2

Under weak realizability, we have

$$\langle f_j, \mathcal{B}^\pi Q_\star^\pi\rangle_\mu = 0 \qquad \text{for all } j = 1,\ldots,d. \tag{62}$$

Thus, at (s,a) the Bellman error difference reads

$$\begin{aligned}
\mathcal{B}^\pi Q(s,a) - \mathcal{B}^\pi Q_\star^\pi(s,a) &= [Q - \mathcal{T}^\pi Q](s,a) - [Q_\star^\pi - \mathcal{T}^\pi Q_\star^\pi](s,a) \\
&= [Q - Q_\star^\pi](s,a) - \gamma \mathbb{E}_{s^+ \sim \mathbb{P}(s,a)}[Q - Q_\star^\pi](s^+, \pi) \\
&= \big\langle w - w_\star^\pi,\, \phi(s,a) - \gamma\phi^{+\pi}(s,a)\big\rangle
\end{aligned} \tag{63}$$

To proceed we need the following auxiliary result:

Lemma 15 (Linear Parameter Constraints). *With probability at least* $1-\delta$, *there exists a universal constant* $c_1 > 0$ *such that if* $Q \in \mathcal{C}_n^\pi$ *then* $\|w - w_\star^\pi\|^2_{\Sigma^{+\pi}_{\lambda,Boot}} \leq c_1 \frac{d\rho}{n}$.

See Section 10.2 for the proof.

Using this lemma, we can bound the OPC coefficient as follows

$$\begin{aligned}
K^\pi \overset{(i)}{\leq} \frac{n}{\rho} \max_{Q\in\mathcal{C}_n^\pi} \langle \mathbb{1}, \mathcal{B}^\pi Q - \mathcal{B}^\pi Q_\star^\pi\rangle_\pi^2 &\overset{(ii)}{\leq} \frac{n}{\rho}\, [\mathbb{E}_\pi(\phi - \gamma\phi^{+\pi})^\top (w - w_\star^\pi)]^2 \\
&\overset{(iii)}{\leq} \frac{n}{\rho}\, \|\mathbb{E}_\pi \phi - \gamma\phi^{+\pi}\|^2_{(\Sigma^{+\pi}_{\lambda,\mathrm{Boot}})^{-1}} \|w - w_\star^\pi\|^2_{\Sigma^{+\pi}_{\lambda,\mathrm{Boot}}} \\
&\leq c_1 d \|\mathbb{E}_\pi \phi - \gamma\phi^{+\pi}\|^2_{(\Sigma^{+\pi}_{\lambda,\mathrm{Boot}})^{-1}}.
\end{aligned}$$

Here step (i) follows from the definition of off-policy cost coefficient, (ii) leverages the linear structure and (iii) is Cauchy-Schwartz.

10.2 Proof of Lemma 15

Under the event of Theorem 3, the statement of Eq. (51a) holds, and in particular

$$\frac{1}{c_1(\sqrt{\|f\|_\mu^2+\lambda})} \geq \frac{1}{\sqrt{\|f\|_n^2+\lambda}} \geq \frac{1}{c_2(\sqrt{\|f\|_\mu^2+\lambda})}.$$

Thus, the j constraint reads

$$\frac{L}{\sqrt{n}} \gtrsim \frac{\langle f_j, \mathcal{B}^\pi Q\rangle_\mu}{\sqrt{\|f\|_n^2+\lambda}} = \frac{\langle f_j, \mathcal{B}^\pi Q\rangle_\mu}{\sqrt{\widehat{\lambda}_j+\lambda}}$$

where the last step follows from

$$\|f_j\|_{\mathcal{D}}^2 = \frac{1}{n}\sum_{(s,a,r,s^+)\in\mathcal{D}} (f_j(s,a))^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{u}_j^\top \phi_i)^2 = \widehat{u}_j^\top \widehat{\Sigma}\widehat{u}_j = \widehat{\lambda}_j.$$

Now, squaring and summing over the constraints and using Eq. (63) yields

$$\begin{aligned}
d\frac{L^2}{n} &\gtrsim \sum_{j=1}^m \langle \frac{\widehat{u}_j^\top \phi}{\sqrt{\widehat{\lambda}_j+\lambda}}, (\phi-\gamma\phi^{+\pi})^\top(w-w_\star^\pi)\rangle_\mu^2 \\
&= \sum_{j=1}^m \Big[\frac{\widehat{u}_j^\top}{\sqrt{\widehat{\lambda}_j+\lambda}} \mathbb{E}_\mu \phi(\phi-\gamma\phi^{+\pi})^\top(w-w_\star^\pi)\Big]^2 \\
&= \sum_{j=1}^m \Big[\frac{\widehat{u}_j^\top}{\sqrt{\widehat{\lambda}_j+\lambda}} \underbrace{(\Sigma-\gamma\Sigma^{+\pi})(w-w_\star^\pi)}_{\overset{def}{=}y}\Big]^2 \\
&= y^\top\Big(\sum_{j=1}^m \frac{\widehat{u}_j\widehat{u}_j^\top}{\widehat{\lambda}_j+\lambda}\Big)y \\
&= y^\top\Big(\widehat{\Sigma}+\lambda I\Big)^{-1} y \\
&\gtrsim y^\top \Sigma_\lambda^{-1} y.
\end{aligned}$$

The last inequality holds via Lemma 14 (*Covariance Concentration*) with probability at least $1-\delta$ since λ is a large enough regularizer. Let us complete the quadratic form:

$$\|y+\lambda(w-w_\star^\pi)\|_{\Sigma_\lambda^{-1}}^2 \leq (\|y\|_{\Sigma_\lambda^{-1}} + \lambda\|(w-w_\star^\pi)\|_{\Sigma_\lambda^{-1}})^2 \lesssim \|y\|_{\Sigma_\lambda^{-1}}^2 + \lambda.$$

Therefore, adding λ to both sides of the prior display and noticing that $\lambda \lesssim \frac{L^2}{n}$ gives

$$\begin{aligned}
d\frac{L^2}{n} &\gtrsim \|y+\lambda(w-w_\star^\pi)\|_{\Sigma_\lambda^{-1}}^2 \\
&= (w-w_\star^\pi)(\Sigma_\lambda-\gamma\Sigma^{+\pi})^\top\Big(\Sigma_\lambda^{-1}\Big)(\Sigma_\lambda-\gamma\Sigma^{+\pi})(w-w_\star^\pi) \\
&= (w-w_\star^\pi)(\Sigma_{\lambda,\mathrm{Boot}}^{+\pi})(w-w_\star^\pi) \\
&= \|(w-w_\star^\pi)\|_{\Sigma_{\lambda,\mathrm{Boot}}^{+\pi}}^2.
\end{aligned}$$

10.3 Proof of Proposition 3

Under weak Bellman closure, we have

$$\mathcal{B}^\pi Q = Q - \mathcal{T}^\pi Q = \phi^\top (w - \mathcal{P}^\pi(w)). \tag{64}$$

With a slight abuse of notation, let $\mathcal{P}^\pi(w)$ denote the weight vector that defines the action-value function $\mathcal{P}^\pi(Q)$. We introduce the following auxiliary lemma:

Lemma 16 (Linear Parameter Constraints with Bellman Closure)**.** *With probability at least* $1-\delta$*, if* $Q \in \mathcal{C}_n^\pi$ *then* $\|w - \mathcal{P}^\pi(w)\|_{\Sigma_\lambda}^2 \leq c_1 \frac{d\rho}{n}$.

See Section 10.4 for the proof. Using this lemma, we can bound the OPC coefficient as follows

$$\begin{aligned} K^\pi &\overset{(i)}{\leq} \frac{n}{\rho} \max_{Q \in \mathcal{C}_n^\pi} \langle \mathbb{1}, \mathcal{B}^\pi Q \rangle_\pi^2 \overset{(ii)}{\leq} \frac{n}{\rho} \left[\mathbb{E}_\pi(\phi)^\top (w - \mathcal{P}^\pi(w)) \right]^2 \\ &\overset{(iii)}{\leq} \frac{n}{\rho} \|\mathbb{E}_\pi \phi\|_{(\Sigma_\lambda)^{-1}}^2 \|w - \mathcal{P}^\pi(w)\|_{\Sigma_\lambda}^2 \\ &\leq c_1 d \|\mathbb{E}_\pi \phi\|_{(\Sigma_\lambda)^{-1}}^2. \end{aligned}$$

Here step (i) follows from the definition of off-policy cost coefficient, (ii) leverages the linear structure and (iii) is Cauchy-Schwartz.

10.4 Proof of Section 10.4

Under the event of Theorem 3, the statement of Eq. (51a) holds, and in particular

$$\frac{1}{c_1(\sqrt{\|f\|_\mu^2 + \lambda})} \geq \frac{1}{\sqrt{\|f\|_n^2 + \lambda}} \geq \frac{1}{c_2(\sqrt{\|f\|_\mu^2 + \lambda})}.$$

Thus, the j constraint reads

$$\frac{L}{\sqrt{n}} \gtrsim \frac{\langle f_j, \mathcal{B}^\pi Q \rangle_\mu}{\sqrt{\|f\|_n^2 + \lambda}} = \frac{\langle f_j, \mathcal{B}^\pi Q \rangle_\mu}{\sqrt{\widehat{\lambda}_j + \lambda}}$$

where the last step follows from

$$\|f_j\|_{\mathcal{D}}^2 = \frac{1}{n} \sum_{(s,a,r,s^+) \in \mathcal{D}} (f_j(s,a))^2 = \frac{1}{n} \sum_{i=1}^n (\widehat{u}_j^\top \phi_i)^2 = \widehat{u}_j^\top \widehat{\Sigma} \widehat{u}_j = \widehat{\lambda}_j.$$

Now, squaring and summing over the constraints and using Eq. (64) yields

$$
\begin{aligned}
d\frac{L^2}{n} &\gtrsim \sum_{j=1}^{m} \langle \frac{\widehat{u}_j^\top \phi}{\sqrt{\widehat{\lambda}_j + \lambda}}, \phi^\top (w - \mathcal{P}^\pi(w)) \rangle_\mu^2 \\
&= \sum_{j=1}^{m} \Big[\frac{\widehat{u}_j^\top}{\sqrt{\widehat{\lambda}_j + \lambda}} \mathbb{E}_\mu \phi \phi^\top (w - \mathcal{P}^\pi(w)) \Big]^2 \\
&= \sum_{j=1}^{m} \Big[\frac{\widehat{u}_j^\top}{\sqrt{\widehat{\lambda}_j + \lambda}} \underbrace{\Sigma(w - \mathcal{P}^\pi(w))}_{\overset{def}{=} y} \Big]^2 \\
&= y^\top \Big(\sum_{j=1}^{m} \frac{\widehat{u}_j \widehat{u}_j^\top}{\widehat{\lambda}_j + \lambda} \Big) y \\
&= y^\top \Big(\widehat{\Sigma} + \lambda I \Big)^{-1} y \\
&\gtrsim y^\top \Sigma_\lambda^{-1} y.
\end{aligned}
$$

The last inequality holds via Lemma 14 (*Covariance Concentration*) with probability at least $1 - \delta$ since λ is a large enough regularizer. Let us complete the quadratic form:

$$
\| y + \lambda (w - \mathcal{P}^\pi(w)) \|_{\Sigma_\lambda^{-1}}^2 \leq (\|y\|_{\Sigma_\lambda^{-1}} + \lambda \|(w - \mathcal{P}^\pi(w))\|_{\Sigma_\lambda^{-1}})^2 \lesssim \|y\|_{\Sigma_\lambda^{-1}}^2 + \lambda.
$$

Therefore, adding λ to both sides of the prior display and noticing that $\lambda \lesssim \frac{L^2}{n}$ gives

$$
\begin{aligned}
d\frac{L^2}{n} &\gtrsim \| y + \lambda (w - \mathcal{P}^\pi(w)) \|_{\Sigma_\lambda^{-1}}^2 \\
&= (w - \mathcal{P}^\pi(w)) \Sigma_\lambda^\top \Big(\Sigma_\lambda^{-1} \Big) \Sigma_\lambda (w - \mathcal{P}^\pi(w)) \\
&= (w - \mathcal{P}^\pi(w)) (\Sigma_\lambda) (w - \mathcal{P}^\pi(w)) \\
&= \|(w - \mathcal{P}^\pi(w))\|_{\Sigma_\lambda}^2.
\end{aligned}
$$

11 Proof of Theorem 2

In this section, we prove the guarantee on our actor-critic procedure stated in Theorem 2.

11.1 Adversarial MDPs

We now introduce sequence of adversarial MDPs $\{\mathcal{M}_t\}_{t=1}^T$ used in the analysis. Each MDP $\mathcal{M}_t$ is defined by the same state-action space and transition law as the original MDP $\mathcal{M}$, but with the reward functions R perturbed by R_t—that is

$$\mathcal{M}_t \overset{def}{=} \langle \mathcal{S}, \mathcal{A}, R + R_t, \mathbb{P}, \gamma \rangle. \tag{65}$$

For an arbitrary policy π, we denote with Q_t^π and with A_t^π the action value function and the advantage function on $\mathcal{M}_t$; the value of π from the starting distribution ν_{start} is denoted by V_t^π. We immediately have the following expression for the value function, which follows because the dynamics of $\mathcal{M}_t$ and $\mathcal{M}$ are identical and the reward function of $\mathcal{M}_t$ equals that of $\mathcal{M}$ plus R_t

$$V_t^\pi \overset{def}{=} \frac{1}{1-\gamma} \mathbb{E}_\pi \Big[R + R_t \Big]. \tag{66}$$

Consider the action value function $\underline{\widehat{Q}}_{\pi_t}$ returned by the critic, and let the reward perturbation $R_t = \mathcal{B}^{\pi_t} \underline{\widehat{Q}}_{\pi_t}$ be the Bellman error of the critic value function $\underline{\widehat{Q}}_{\pi_t}$. The special property of $\mathcal{M}_t$ is that the action value function of π_t on $\mathcal{M}_t$ equals the critic lower estimate $\underline{\widehat{Q}}_{\pi_t}$.

Lemma 17 (Adversarial MDP Equivalence). *Given the perturbed MDP $\mathcal{M}_t$ from equation (65) with $R_t \overset{def}{=} \mathcal{B}^{\pi_t} \underline{\widehat{Q}}_{\pi_t}$, we have the equivalence*

$$Q_t^{\pi_t} = \underline{\widehat{Q}}_{\pi_t}.$$

Proof. We need to check that $\underline{\widehat{Q}}_{\pi_t}$ solves the Bellman evaluation equations for the adversarial MDP, ensuring that $\underline{\widehat{Q}}_{\pi_t}$ is the action-value function of π_t on $\mathcal{M}_t$. Let $\mathcal{T}_t^{\pi_t}$ be the Bellman evaluation operator on $\mathcal{M}_t$ for policy π_t. We have

$$\underline{\widehat{Q}}_{\pi_t} - \mathcal{T}_t^{\pi_t}(\underline{\widehat{Q}}_{\pi_t}) = \underline{\widehat{Q}}_{\pi_t} - \mathcal{T}^{\pi_t}(\underline{\widehat{Q}}_{\pi_t}) - R_t = \mathcal{B}^{\pi_t}\underline{\widehat{Q}}_{\pi_t} - \mathcal{B}^{\pi_t}\underline{\widehat{Q}}_{\pi_t} = 0.$$

Thus, the function $\underline{\widehat{Q}}_{\pi_t}$ is the action value function of π_t on $\mathcal{M}_t$, and it is by definition denoted by $Q_t^{\pi_t}$. □

This lemma shows that the action-value function $\underline{\widehat{Q}}_{\pi_t}$ computed by the critic is equivalent to the action-value function of π_t on $\mathcal{M}_t$. Thus, we can interpret the critic as performing a model-based pessimistic estimate of π_t; this view is useful in the rest of the analysis.

11.2 Equivalence of Updates

The second step is to establish the equivalence between the update rule (22), or equivalently as the update (67a), to the exponentiated gradient update rule (67b).

Lemma 18 (Equivalence of Updates)**.** *For linear Q-functions of the form $Q_t(s,a) = \langle w_t,\, \phi(s,a)\rangle$, the parameter update*

$$\pi_{t+1}(a \mid s) \propto \exp(\phi(s,a)^\top(\theta_t + \eta w_t)), \tag{67a}$$

is equivalent to the policy update

$$\pi_{t+1}(a \mid s) \propto \pi_t(a \mid s)\exp(\eta Q_t(s,a)), \qquad \pi_1(a \mid s) = \frac{1}{|\mathcal{A}_s|}. \tag{67b}$$

Proof. We prove this claim via induction on t. The base case ($t = 1$) holds by a direct calculation. Now let us show that the two update rules update π_t in the same way. As an inductive step, assume that both rules maintain the same policy $\pi_t \propto \exp(\phi(s,a)^\top\theta_t)$ at iteration t; we will show the policies are still the same at iteration $t+1$. At any (s,a), we have

$$\begin{aligned}\pi_{t+1}(a \mid s) \propto \exp(\phi(s,a)^\top(\theta_t + \eta w_t)) &\propto \exp(\phi(s,a)^\top\theta_t)\exp(\eta\phi(s,a)^\top w_t) \\ &\propto \pi_t(a \mid s)\exp(\eta Q_t(s,a)).\end{aligned}$$

□

Recall that θ_t is the parameter associated to π_t and that w_t is the parameter associated to $\underline{\widehat{Q}}_{\pi_t}$. Using Lemma 18 together with Lemma 17 we obtain that the actor policy π_t satisfies through its parameter θ_t the mirror descent update rule (67b) with $Q_t = \underline{\widehat{Q}}_{\pi_t} = Q_t^{\pi_t}$ and $\pi_1(a \mid s) = 1/|\mathcal{A}_s|,\ \forall(s,a)$. In words, the actor is using Mirror descent to find the best policy on the sequence of adversarial MDPs $\{\mathcal{M}_t\}$ implicitly identified by the critic.

11.3 Mirror Descent on Adversarial MDPs

Our third step is to analyze the behavior of mirror descent on the MDP sequence $\{\mathcal{M}_t\}_{t=1}^T$, and then translate such guarantees back to the original MDP $\mathcal{M}$. The following result provides a bound on the average of the value functions $\{V^{\pi_t}\}_{t=1}^T$ induced by the actor's policy sequence. This bound involves a form of optimization error[8] given by

$$\mathcal{E}_{opt}(T) = 2\sqrt{\frac{2\log|\mathcal{A}|}{T}},$$

as is standard in mirror descent schemes. It also involves the *perturbed rewards* given by $R_t \overset{def}{=} \mathcal{B}^{\pi_t}Q_t^{\pi_t}$.

Lemma 19 (Mirror Descent on Adversarial MDPs)**.** *For any positive integer T, applying the update rule (67b) with $Q_t = Q_t^{\pi_t}$ for T rounds yields a sequence such that*

$$\frac{1}{T}\sum_{t=1}^{T}\Big[V^{\widetilde{\pi}} - V^{\pi_t}\Big] \leq \frac{1}{1-\gamma}\left\{\mathcal{E}_{opt}(T) + \frac{1}{T}\sum_{t=1}^{T}\Big[-\mathbb{E}_{\widetilde{\pi}}R_t + \mathbb{E}_{\pi_t}R_t\Big]\right\}, \tag{68}$$

valid for any comparator policy $\widetilde{\pi}$.

[8] Technically, this error should depend on $|\mathcal{A}_s|$, if we were to allow the action spaces to have varyign cardinality, but we elide this distinction here.

See Section 11.6 for the proof.

To be clear, the comparator policy $\widetilde{\pi}$ need belong to the soft-max policy class. Apart from the optimization error term, our bound (68) involves the behavior of the perturbed rewards R_t along the comparator $\widetilde{\pi}$ and π_t, respectively. These correction terms arise because the actor performs the policy update using the action-value function $Q_t^{\pi_t}$ on the perturbed MDPs instead of the real underlying MDP.

11.4 Pessimism: Bound on $\mathbb{E}_{\pi_t} R_t$

The fourth step of the proof is to leverage the pessimistic estimates returned by critic to simplify equation (68). Using Lemma 9 and the definition of adversarial reward R_t we can write

$$\widehat{V}^{\pi}_{\min} - V^{\pi_t} = \frac{1}{1-\gamma}\langle \mathbb{1}, \mathcal{B}^{\pi_t}\underline{\widehat{Q}}_{\pi_t}\rangle_{\pi_t} = \frac{1}{1-\gamma}\mathbb{E}_{\pi_t}\mathcal{B}^{\pi_t}\underline{\widehat{Q}}_{\pi_t} = \frac{1}{1-\gamma}\mathbb{E}_{\pi_t} R_t.$$

Since weak realizability holds, Theorem 3 guarantees that $\widehat{V}^{\pi}_{\min} \leq V^{\pi}$ uniformly for all $\pi \in \Pi$ with probability at least $1-\delta$. Coupled with the prior display, we find that

$$\mathbb{E}_{\pi_t} R_t \leq 0. \tag{69}$$

Using the above display, the result in Eq. (68) can be further upper bounded and simplified.

11.5 Concentrability: Bound on $\mathbb{E}_{\widetilde{\pi}} R_t$

The term $\mathbb{E}_{\widetilde{\pi}} R_t$ can be interpreted as an approximate concentrability factor for the approximate algorithm that we are investigating.

Bound under only weak realizability: Lemma 15 gives with probability at least $1-\delta$ that any surviving Q in $\mathcal{C}_n^{\pi_t}$ must satisfy: $\|w - w_\star^{\pi_t}\|^2_{\Sigma^{+\pi_t}_{\lambda,\text{Boot}}} \lesssim \frac{d\rho}{n}$ where $w_\star^{\pi_t}$ is the parameter associated to the weak solution $Q_\star^{\pi_t}$. Such bound must apply to the parameter $w_t \in \widehat{\mathcal{C}}_n^{\pi_t}$ identified by the critic.[9].

We are now ready to bound the remaining adversarial reward along the distribution of the comparator $\widetilde{\pi}$.

$$\begin{aligned}
|\mathbb{E}_{\widetilde{\pi}} R_t| &= |\mathbb{E}_{\widetilde{\pi}} \mathcal{B}^{\pi_t}\underline{\widehat{Q}}_{\pi_t}| \\
&\overset{(i)}{=} |\mathbb{E}_{\widetilde{\pi}}(\phi - \gamma\phi^{+\pi_t})^\top (w_t - w_\star^{\pi_t})| \\
&\leq \|\mathbb{E}_{\widetilde{\pi}}[\phi - \gamma\phi^{+\pi_t}]\|_{(\Sigma^{+\pi_t}_{\lambda,\text{Boot}})^{-1}} \|w_t - w_\star^{\pi_t}\|_{\Sigma^{+\pi_t}_{\lambda,\text{Boot}}} \\
&\leq c\sqrt{\frac{d\rho}{n}} \sup_{\pi\in\Pi}\Big\{\|\mathbb{E}_{\widetilde{\pi}}[\phi - \gamma\phi^{+\pi}]\|_{(\Sigma^{+\pi}_{\lambda,\text{Boot}})^{-1}}\Big\}.
\end{aligned} \tag{70}$$

Step (i) follows from the expression (63) for the weak Bellman error, along with the definition of the weak solution $Q_\star^{\pi_t}$.

[9] We abuse the notation and write $w \in \widehat{\mathcal{C}}_n^{\pi}$ in place of $Q \in \widehat{\mathcal{C}}_n^{\pi}$

Bound under weak Bellman closure: When Bellman closure holds we proceed analogously. The bound in Lemma 16 ensures with probability at least $1-\delta$ that $\|w - \mathcal{P}^{\pi_t}(w)\|_{\Sigma_\lambda}^2 \leq c\,\frac{d\rho}{n}$ for all $w \in \mathcal{C}_n^{\pi_t}$; as before, this relation must apply to the parameter chosen by the critic $w_t \in \widehat{\mathcal{C}}_n^{\pi_t}$. The bound on the adversarial reward along the distribution of the comparator $\widetilde{\pi}$ now reads

$$\begin{aligned} |\mathbb{E}_{\widetilde{\pi}} R_t| \;=\; |\mathbb{E}_{\widetilde{\pi}} \mathcal{B}^{\pi_t} \underline{\widehat{Q}}_{\pi_t}| &\overset{\text{(i)}}{=} |\mathbb{E}_{\widetilde{\pi}} \phi^\top (w_t - \mathcal{P}^{\pi_t}(w_t))| \\ &\leq \|\mathbb{E}_{\widetilde{\pi}} \phi\|_{\Sigma_\lambda^{-1}} \|w_t - \mathcal{P}^{\pi_t}(w_t)\|_{\Sigma_\lambda} \\ &\leq c\, \|\mathbb{E}_{\widetilde{\pi}} \phi\|_{\Sigma_\lambda^{-1}} \sqrt{\frac{d\rho}{n}}. \end{aligned} \tag{71}$$

Here step (i) follows from the expression (64) for the Bellman error under weak closure.

11.6 Proof of Lemma 19

We now prove our guarantee for a mirror descent procedure on the sequence of adversarial MDPs. Our analysis makes use of a standard result on online mirror descent for linear functions (e.g., see Section 5.4.2 of Hazan [Haz21]), which we state here for reference. Given a finite cardinality set $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$, and a distribution ν over $\mathcal{X}$, we define $f(\nu) \overset{def}{=} \sum_{x \in \mathcal{X}} \nu(x) f(x)$. The following result gives a guarantee that holds uniformly for any sequence of functions $\{f_t\}_{t=1}^T$, thereby allowing for the possibility of adversarial behavior.

Proposition 5 (Adversarial Guarantees for Mirror Descent)**.** *Suppose that we initialize with the uniform distribution $\nu_1(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$, and then perform T rounds of the update*

$$\nu_{t+1}(x) \propto \nu_t(x) \exp(\eta f_t(x)), \quad \textit{for all } x \in \mathcal{X}, \tag{72}$$

using $\eta = \sqrt{\frac{\log|\mathcal{X}|}{2T}}$. If $\|f_t\|_\infty \leq 1$ for all $t \in [T]$ then we have the bound

$$\frac{1}{T} \sum_{t=1}^{T} \Big[f_t(\widetilde{\nu}) - f_t(\nu_t) \Big] \leq \mathcal{E}_{opt}(T) \overset{def}{=} 2\sqrt{\frac{2\log|\mathcal{X}|}{T}}. \tag{73}$$

where $\widetilde{\nu}$ is any comparator distribution over $\mathcal{X}$.

We now use this result to prove our claim. So as to streamline the presentation, it is convenient to introduce the advantage function corresponding to π_t. It is a function of the state-action pair (s, a) given by

$$A_t^{\pi_t}(s,a) \overset{def}{=} Q_t^{\pi_t}(s,a) - \mathbb{E}_{a^+ \sim \pi_t(\cdot|s)} Q_t^{\pi_t}(s, a^+).$$

In the sequel, we omit dependence on (s,a) when referring to this function, consistent with the rest of the paper.

From our earlier observation (66), recall that the reward function of the perturbed MDP $\mathcal{M}_t$ corresponds to that of $\mathcal{M}$ plus the perturbation R_t. Combining this fact with a standard simulation lemma (e.g., [K+03]) applied to $\mathcal{M}_t$, we find that

$$V^{\widetilde{\pi}} - V^{\pi_t} = V_t^{\widetilde{\pi}} - V_t^{\pi_t} + \frac{1}{1-\gamma}\Big[-\mathbb{E}_{\widetilde{\pi}} R_t + \mathbb{E}_{\pi_t} R_t \Big] \;=\; \frac{1}{1-\gamma}\Big[\mathbb{E}_{\widetilde{\pi}} A_t^{\pi_t} - \mathbb{E}_{\widetilde{\pi}} R_t + \mathbb{E}_{\pi_t} R_t \Big]. \tag{74a}$$

Now for any given state s, we introduce the linear objective function

$$f_t(\nu) \stackrel{def}{=} \mathbb{E}_{a\sim\nu} Q_t^{\pi_t}(s,a) \; = \; \sum_{a\in\mathcal{A}} \nu(a) Q_t^{\pi_t}(s,a),$$

where ν is a distribution over the action space. With this choice, we have the equivalence

$$\mathbb{E}_{a\sim\widetilde{\pi}} A_t^{\pi_t}(s,a) = f_t(\widetilde{\pi}(\cdot \mid s)) - f_t\big(\pi_t(\cdot \mid s)\big),$$

where the reader should recall that we have fixed an arbitrary state s. Consequently, applying the bound (73) with $\mathcal{X} = \mathcal{A}$ and these choices of linear functions, we conclude that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{a\sim\widetilde{\pi}} A_t^{\pi_t}(s,a) \leq \mathcal{E}_{opt}(T). \tag{74b}$$

This bound holds for any state, and also for any average over the states.

We now combine the pieces to conclude. By computing the average of the bound (74a) over all T iterations, we find that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\Big[V^{\widetilde{\pi}} - V^{\pi_t}\Big] &\leq \frac{1}{1-\gamma}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\widetilde{\pi}} A_t^{\pi_t} + \frac{1}{T}\sum_{t=1}^{T}\Big[-\mathbb{E}_{\widetilde{\pi}} R_t + \mathbb{E}_{\pi_t} R_t\Big]\right\} \\ &\leq \frac{1}{1-\gamma}\left\{\mathcal{E}_{opt}(T) + \frac{1}{T}\sum_{t=1}^{T}\Big[-\mathbb{E}_{\widetilde{\pi}} R_t + \mathbb{E}_{\pi_t} R_t\Big]\right\},\end{aligned}$$

where the final inequality follow from the bound (73), applied for each s. We have thus established the claim.