Bellman Residual Orthogonalization for Offline Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

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 Appendix A.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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 Appendix A.1, and compare with the weight learning literature at the population level in Appendix A.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 Appendix A.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 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.

[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.

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 $\sup_{(s,a)} |Q^\pi(s, a)| \le 1$. 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 $V^\pi = \mathbb{E}_{S_0 \sim \nu_{\text{start}}}[V^\pi(S_0)]$. The *discounted occupancy measure* associated with a policy π is defined as $d_\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h[(S_h, A_h) = (s, a)]$. 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 $\sup_{(s,a,o)} |f(s, a, o)| \le 1$, 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*

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

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 Appendix A.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

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

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

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

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

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

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 $\mathcal{C}_n^\pi$ of the data-dependent set $\widehat{\mathcal{C}}_n^\pi$. More precisely, we define the population set

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

where $\langle f,\, g \rangle_\mu \stackrel{def}{=} \int f(s,a) g(s,a) d\mu$ is the inner product induced by a distribution[2] μ over (s,a). As before, we use the shorthand notation $\mathcal{C}_n^\pi$ 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 $\widehat{\mathcal{C}}_n^\pi$, the set $\mathcal{C}_n^\pi$ 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:

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

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

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:

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

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.

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 Appendix B 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 $(\widehat{V}_{min}^\pi, \widehat{V}_{max}^\pi)$ 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

[2] See Section B.2.1 for a precise definition of the relevant μ for a fairly general sampling model.

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

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

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

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

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

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

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 Appendix A.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 $\|f^*\|_\infty \leq 1$. Then the OPC coefficient satisfies*

$$K^\pi \stackrel{(i)}{\leq} \frac{\mathbb{E}_\pi\left[\frac{d_\pi(S,A)}{\mu(S,A)}\right] + b_\pi^2\lambda}{1+\lambda} \stackrel{(ii)}{\leq} \frac{b_\pi\left(1+b_\pi\lambda\right)}{1+\lambda} \tag{12}$$

The proof is in Appendix D.1. Since $\lambda = \lambda_n \to 0$ as n increases, the OPC coefficient is bounded by a multiple of the expected ratio $\mathbb{E}_\pi\left[\frac{d_\pi(S,A)}{\mu(S,A)}\right]$. 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 $\epsilon = Q - Q_\star^\pi$, where $Q_\star^\pi$ is a Q-function from the definition of weak realizability. The Bellman error can be re-written as $\mathcal{B}^\pi Q = \mathcal{B}^\pi Q - \mathcal{B}^\pi Q_\star^\pi = (\mathcal{I} - \gamma \mathbb{P}^\pi)\epsilon$. When realizability holds, in the linear setting and at the population level, the LSTD solution seeks to satisfy the projected Bellman equations

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

In the linear case, $\mathcal{E}_\star^\pi$ 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 $\mathcal{F}^\pi = \mathcal{E}_\star^\pi = \{Q - Q_\star^\pi\}$. Since $\mathcal{E}_\star^\pi$ is unknown (as it depends on the weak solution $Q_\star^\pi$), we choose instead the larger class

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

which contains $\mathcal{E}_\star^\pi$. The resulting approach can be seen as performing a projection of the Bellman operator $\mathcal{B}^\pi Q$ into the error space $\mathcal{E}_\star^\pi$, 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 $\mathcal{E}_\star^\pi$ 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*

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

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 $\mathcal{E}_\star^\pi$ 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 Appendix D.2 for the proof, and Appendix A.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

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

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 $\mathcal{F}_\pi^{\mathcal{B}}$, we have*

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

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

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

See Appendix D.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 $\{\mathcal{F}_m^\pi\}_{m=1}^M$ of M different test function classes, and define the union $\mathcal{F}^\pi = \bigcup_{m=1}^M \mathcal{F}_m^\pi$. For each $m = 1, \ldots, M$, let K_m^π be the OPC coefficient defined by the function class $\mathcal{F}_m^\pi$ 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)**.** $K^\pi(\mathcal{F}) \leq \min_{m=1,\ldots,M} K_m^\pi$.

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]

[3] For space reasons, we defer to Appendix A.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.

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 $g_w(s,a) \stackrel{def}{=} \langle w,\, \phi(s,a)\rangle \;=\; \sum_{j=1}^d w_j \phi_j(s,a)$. The class of *linear functions* takes the form

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

Throughout our analysis, we assume that $\|\phi(s,a)\|_2 \le 1$ 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 $\widehat{V}^\pi_{\text{min}}$ and $\widehat{V}^\pi_{\text{max}}$ from equation (5) can be computed by solving a quadratic program, with two linear constraints for each test function. (See Appendix A.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 $\phi_i \stackrel{def}{=} \phi(s_i, a_i)$. 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

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

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 Appendix A.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.

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* $\phi^{+\pi}(s,a) \stackrel{def}{=} \mathbb{E}_{s^+\sim\mathbb{P}(s,a)}\phi(s^+,\pi)$, along with the population covariance matrix $\Sigma \stackrel{def}{=} \mathbb{E}_\mu\big[\phi(s,a)\phi^\top(s,a)\big]$, and its λ-regularized version $\Sigma_\lambda \stackrel{def}{=} \Sigma + \lambda I$. We also define the matrices

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

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 Appendix E.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 Appendix E.3 for the proof.

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

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

5.2 Actor-critic scheme for policy optimization

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

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

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]

- 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 $\widehat{V}^{\pi_t}_{\text{min}}$ 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:

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

Moreover, in making the following assertion, we assume that every weak solution $Q^{\pi}_{\star}$ can be evaluated against the distribution of a comparator policy $\widetilde{\pi} \in \Pi$, i.e., $\langle \mathbb{1}, \mathcal{B}^{\pi} Q^{\pi}_{\star} \rangle_{\widetilde{\pi}} = 0$ 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 $\widetilde{\pi} \in \Pi$,*

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

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

See Appendix F for the proof. Whenever Bellman closure holds, the result automatically inherits the more favorable concentrability coefficient $K^{\widetilde{\pi}}_{(2)}$, 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 $K^{\widetilde{\pi}}_{(1)}$. 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 Appendix A.9 for a more detailed discussion of this issue.

References

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

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

[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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations include the choice of the test functions
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The work is theoretical in nature, no negative societal impacts anticipated
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]