Beyond the Return: Off-policy Function Estimation under User-specified Error-measuring Distributions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Off-policy evaluation often refers to two related tasks: estimating the expected return of a policy and estimating its value function (or other functions of interest, such as density ratios). While recent works on marginalized importance sampling (MIS) show that the former can enjoy provable guarantees under realizable function approximation, the latter is only known to be feasible under much stronger assumptions such as prohibitively expressive discriminators. In this work, we provide guarantees for off-policy function estimation under only realizability, by imposing proper regularization on the MIS objectives. Compared to commonly used regularization in MIS, our regularizer is much more flexible and can account for an arbitrary user-specified distribution, under which the learned function will be close to the groundtruth. We provide *exact* characterization of the optimal dual solution that needs to be realized by the discriminator class, which determines the data-coverage assumption in the case of value-function learning. As another surprising observation, the regularizer can be altered to relax the data-coverage requirement, and completely eliminate it in the ideal case with strong side information.

1 Introduction

Off-policy evaluation (OPE) often refers to two related tasks in reinforcement learning (RL): estimating the expected return of a target policy using a dataset collected from a different *behavior* policy, versus estimating the policy's value function (or other functions of interest, such as density ratios). The former is crucial to hyperparameter tuning and verifying the performance of a policy before real-world deployment in offline RL [VLJY19; Pai+20; ZJ21]. The latter, on the other hand, plays an important role in (both online and offline) training, often as the subroutine of actor-critic-style algorithms [LP03; LSAB19], but is also generally more difficult than the former: if an accurate value function is available, one could easily estimate the return by plugging in the initial distribution.

Between the two tasks, the theoretical nature of off-policy return estimation is relatively well understood, especially in terms of the function-approximation assumptions needed for sample-complexity guarantees. Among the available algorithms, importance sampling (IS) and its variants [PSS00; TTG15; JL16] do not require any function approximation, but incur *exponential-in-horizon* variance; Fitted-Q Evaluation [EGW05; LVY19] can enjoy polynomial sample complexity under appropriate coverage assumptions, but the guarantee relies on the strong Bellman-completeness assumption on the function class; marginalized importance sampling (MIS) methods, which have gained significant attention recently [LLTZ18; XMW19; UHJ20; NCDL19], use two function classes to simultaneously approximate the value and the density-ratio (or weight) function and optimize minimax objectives. Notably, it is the only family of methods known to produce accurate return estimates with a polynomial sample complexity, when the function classes only satisfy the relatively weak realizability assumptions (i.e., they contain the true value and weight functions).

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

In comparison, little is known about off-policy function estimation, and the guarantees are generally less desirable. Not only the limitations of IS and FQE on return estimation carry over to this more challenging task, but MIS also loses its major advantage over FQE: despite the somewhat misleading impression left by many prior works, that MIS can handle function estimation the same way as return estimation,[1] MIS for function estimation often requires unrealistic assumptions such as prohibitively expressive discriminators. For concreteness, a typical guarantee for function estimation from MIS looks like the following (see e.g., Theorem 4 of [LLTZ18] and Lemmas 1 and 3 of [UHJ20]):

Proposition 1 (Function estimation guarantee for MIS, informal)**.** *Suppose the offline data distribution d^D satisfies $d^D(s,a) > 0, \forall s, a$. Given value-function class $\mathcal{Q}$ with $q^\pi \in \mathcal{Q}$ and weight class $\mathcal{W} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $q^\pi = \arg\min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} L(w,q)$ for some appropriate population loss function L.*[2]

To enable the identification of the value function q^π, the result requires the discriminator class $\mathcal{W}$ to be the *space of all possible functions over the state-action space* ($\mathcal{W} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$). In the finite-sample regime, using such a class incurs a sample complexity that depends on the size of the state-action space, which completely beats the purpose of function approximation.

In addition, these results only hold asymptotically, where the function of interest can be exactly identified in a point-wise manner. Such an overly strong guarantee is unrealistic in the finite-sample regime, where one can only hope to approximate the function well in an average sense *under some distribution*, i.e., finite-sample performance guarantees should ideally bound $\|\widehat{q} - q^\pi\|_{2,\nu}$ for the learned $\widehat{q}$, where $\|\cdot\|_{2,\nu}$ is ν-weighted 2-norm. Such fine-grained analyses are non-existent in MIS. Even in the broader literature, such results not only require Bellman-completeness-type assumptions [UIJKSX21], they also come with some fixed ν (which is not necessarily d^D; see Section 2) and the user has no freedom in choosing ν. This creates a gap in the literature, as downstream learning algorithms that use off-policy function estimation as a subroutine often assume the estimation to be accurate under certain specific distributions [KL02; AYBBLSW19].

To summarize, below are two important open problems on off-policy function estimation:

1. *Is it possible to obtain polynomial*[3] *sample complexity for off-policy function estimation, using function classes that only satisfy realizability-type assumptions?*
2. *Can we specify a distribution ν to the estimation algorithm, such that the learned function will be close to the groundtruth under ν?*

In this work, we answer both open questions in the positive. By imposing proper regularization on the MIS objectives, we provide off-policy function estimation guarantees under only realizability assumptions on the function classes. Compared to commonly used regularization in MIS [NCDL19; ND20; YNDLS20], our regularizer is much more flexible and can account for an arbitrary user-specified distribution ν, under which the learned function will be close to the groundtruth. We provide *exact* characterization of the optimal dual solution that needs to be realized by the discriminator, which determines the data-coverage assumption in value-function learning. As another surprising observation, the regularizer can be altered to relax the data-coverage requirement, and in the ideal case *completely eliminate* it when strong side information is available. Proof-of-concept experiments are also conducted to validate our theoretical predictions.

2 Related Works

Regularization in MIS The use of regularization is very common in the MIS literature, especially in DICE algorithms [NCDL19; NDKCLS19; YNDLS20]. However, most prior works that consider regularization use tabular derivations and seldom provide finite-sample function-approximation guarantees on even return estimation, let alone function estimation. (An exception is the work of [UIJKSX21], who analyze related estimators under Bellman-completeness-type assumptions; see the next paragraph.) More importantly, while some existing DICE estimators are subsumed as our special cases when we choose very simple regularizers (see Remark 3 in Section 5), prior works provide very limited understanding in how the choice of regularizers affects learning guarantees, and hence have only considered these naïve forms of regularization (typically state-action-*independent* and

[1] For example, [LSAB19] assumed a weight estimation oracle and cited [LLTZ18] as a possible instance.

[2] Concretely, one can choose L as Eq.(4) without the regularization term, which recovers MQL in [UHJ20].

[3] By "polynomial", we mean polynomial in the horizon, the statistical capacities and the boundedness of the function classes, and the parameter that measures the degree of data coverage.

under d^D)—as different forms of regularization are essentially treated equally under a coarse-grained theory [YNDLS20]. In contrast, we provide much more fine-grained characterization of the effects of regularization, which leads to novel insights about how to design better regularizers.

Fitted-Q Evaluation (FQE) Outside the MIS literature, one can obtain return *and* value-function estimation guarantees via FQE [DJW20; CJ19; LVY19; UIJKSX21]. However, it is well understood that FQE and related approaches require Bellman-completeness-type assumptions, such as the function class being *closed* under the Bellman operator. Even putting aside the difference between completeness vs. realizability, we allow for a user-specified error-measuring distribution, which is not available in FQE or any other existing method. The only distribution these methods are aware of is the data distribution d^D, and even so, FQE and variants rarely provide guarantees on $\|\widehat{q} - q^\pi\|_{2,d^D}$, but often on the Bellman error (e.g., $\|\widehat{q} - \mathcal{T}^\pi \widehat{q}\|_{2,d^D}$) instead [UIJKSX21], and obtaining guarantees on a distribution of interest often requires multiple indirect translations and loose relaxations.

LSTDQ Our analyses focus on general function approximation. When restricted to linear classes, function estimation guarantees for q^π under d^D can be obtained by LSTDQ methods [LP03; BY09; DNP14] when the function class only satisfies realizability of q^π [PKBK22]. However, this requires an additional matrix invertibility condition (see Assumption 3 of [PKBK22]), and it is still unclear what this condition corresponds to in general function approximation.[4] Moreover, many general methods—including MIS [UHJ20] and other minimax methods [ASM08; XCJMA21]—coincide with LSTDQ in the linear case, so the aforementioned results can be viewed as a specialized analysis leveraging the properties of linear classes.

PRO-RL [ZHHJL22] Our key proof techniques are adapted from [ZHHJL22], whose goal is offline policy learning. They learn the importance weight function w^π for a near-optimal π, and provide $\|\widehat{w} - w^\pi\|_{2,d^D}$ guarantees as an intermediate result. Despite using similar technical tools, our most interesting and surprising results are in the value-function estimation setting, which is not considered by [ZHHJL22]. Our novel algorithmic insights, such as incorporating error-measuring distributions and approximate models in the regularizers, are also potentially useful in [ZHHJL22]'s policy learning setting. Our analyses also reveal a number of important differences between OPE and offline policy learning, which will be discussed in Appendix A.

3 Preliminaries

We consider off-policy evaluation (OPE) in Markov Decision Processes (MDPs). An MDP is specified by its state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ ($\Delta(\cdot)$ is the probability simplex), reward function $R : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$, discount factor $\gamma \in [0, 1)$, and an initial state distribution $\mu_0 \in \Delta(\mathcal{S})$. We assume $\mathcal{S}$ and $\mathcal{A}$ are finite and discrete, but their cardinalities can be arbitrarily large. Given a target policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, a random trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$ can be generated as $s_0 \sim \mu_0, a_t \sim \pi(\cdot|s_t), r_t \sim R(\cdot|s_t, a_t), s_{t+1} \sim P(\cdot|s_t, a_t), \forall t \geq 0$; we use $\mathbb{E}_\pi$ and $\mathbb{P}_\pi$ to refer to expectation and probability under such a distribution. The expected discounted return (or simply return) of π is $J(\pi) := \mathbb{E}_\pi[\sum_t \gamma^t r_t]$. The Q-value function of π is the unique solution of the Bellman equations $q^\pi = \mathcal{T}^\pi q^\pi$, with the Bellman operator $\mathcal{T}^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined as $\forall q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, (\mathcal{T}^\pi q)(s, a) := \mathbb{E}_{r\sim R(\cdot|s,a)}[r] + \gamma(P^\pi q)(s, a)$. Here $P^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is the state-action transition operator of π, defined as $(P^\pi q)(s, a) := \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')}[q(s', a')]$. Functions over $\mathcal{S} \times \mathcal{A}$ (such as q) are also treated as $|\mathcal{S} \times \mathcal{A}|$-dimensional vectors interchangeably.

In OPE, we want to estimate q^π and other functions of interest based on a historical dataset collected by a possibly different policy. As a standard simplification, we assume that the offline dataset consisting of n i.i.d. tuples $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ sampled as $(s_i, a_i) \sim d^D$, $r = R(s_i, a_i)$, and $s'_i \sim P(\cdot|s_i, a_i)$. We call d^D the (offline) data distribution. As another function of interest, the (marginalized importance) weight function w^π is defined as $w^\pi(s, a) := d^\pi(s, a)/d^D(s, a)$, where $d^\pi(s, a) = (1 - \gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}_\pi[s_t = s, a_t = a]$ is the discounted state-action occupancy of π. For technical convenience we assume $d^D(s, a) > 0\ \forall s, a$, so that quantities like w^π are always well defined and finite.[5] Similarly to q^π, w^π also satisfies a recursive equation, inherited from the Bellman

[4] It is hinted by [UHJ20] that the invertibility is related to a loss minimization condition in MIS, but the connection only holds for return estimation.

[5] It will be trivial to remove this assumption at the cost of cumbersome derivations. Also, these density ratios can still take prohibitively large values even if they are finite, and we will need to make additional boundedness assumptions to enable finite-sample guarantees anyway, so their finiteness does not trivialize the analyses.

flow equation for d^π: $d^\pi = (1-\gamma)\mu_0^\pi + \gamma \widetilde{P}^\pi d^\pi$, where $(s,a) \sim \mu_0^\pi \Leftrightarrow s \sim \mu_0, a \sim \pi(\cdot|s)$ is the initial state-action distribution, and $\widetilde{P}^\pi = (P^\pi)^\top$ is the transpose of the transition matrix.

Function Approximation We will use function classes $\mathcal{Q}$ and $\mathcal{W}$ to approximate q^π and w^π, respectively. We assume finite $\mathcal{Q}$ and $\mathcal{W}$, and extension to infinite classes under appropriate complexity measures (e.g., covering number) is routine and orthogonal to the main insights of the paper.

Additional Notation $\|\cdot\|_{2,\nu} := \sqrt{\mathbb{E}_\nu[(\cdot)^2]}$ is the weighted 2-norm of a function under distribution ν. We also use a standard shorthand $f(s,\pi) := \mathbb{E}_{a\sim\pi(\cdot|s)}[q(s,a)]$. $u \circ v$ between two vectors u and v of the same dimension is elementwise multiplication, and u/v is elementwise division.

4 Value-function Estimation

In this section we show how to estimate $\widehat{q} \approx q^\pi$ with guarantees on $\|\widehat{q} - q^\pi\|_{2,\nu}$ for a user-specified ν, and identify the assumptions under which provable sample-complexity guarantees can be obtained. We begin with the familiar Bellman equations, that q^π is the unique solution to:

$$\mathbb{E}_{r\sim R(\cdot|s,a)}[r] + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[q(s',\pi)] - q(s,a) = 0,\ \forall s,a \in \mathcal{S}\times\mathcal{A}. \tag{1}$$

While the above set of equations uniquely determines $q = q^\pi$, this is only true if we can enforce *all* the $|\mathcal{S}\times\mathcal{A}|$ constraints, which is intractable in large state-space problems. In fact, even estimating (a candidate q's violation of) a single constraint is infeasible as that requires sampling from the same state multiple times, which is related to the infamous double-sampling problem [Bai95].

To overcome this challenge, prior MIS works often relax Eq.(1) by taking a *weighted combination* of these equations, e.g.,

$$\mathbb{E}_{d^D}\left[w(s,a)\left(r(s,a) + \gamma q(s',\pi) - q(s,a)\right)\right] = 0,\ \forall w \in \mathcal{W}. \tag{2}$$

In words, instead of enforcing $|\mathcal{S}\times\mathcal{A}|$ equations, we only enforce their linear combinations; the linear coefficients are $d^D(s,a)\cdot w(s,a)$, and w belongs to a class $\mathcal{W}$ with limited statistical capacity to enable sample-efficient estimation. While each constraint in Eq.(2) now be efficiently checked on data, this comes with a big cost that a solution to Eq.(2) is not necessarily q^π. Prior works handle this dilemma by aiming lower: instead of learning $\widehat{q} \approx q^\pi$, the problem becomes tractable if we only aim to learn $\widehat{q}$ that can approximate the policy's return, i.e. $\mathbb{E}_{s\sim\mu_0}[\widehat{q}(s,\pi)] \approx J(\pi) = \mathbb{E}_{s\sim\mu_0}[q^\pi(s,\pi)]$. While [UHJ20] show that this is possible under $w^\pi \in \mathcal{W}$, they also show explicit counterexamples where $\widehat{q} \neq q^\pi$ even with infinite data. As a result, how to estimate $\widehat{q} \approx q^\pi$ under comparable assumptions (instead of the prohibitive $\mathcal{W} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as in Proposition 1) is still an open problem.

4.1 Estimator

We now describe our approach to this problem. Recall that the goal is to obtain error bounds for $\|\widehat{q} - q^\pi\|_{2,\nu}$ for some distribution $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$ specified by the user. Note that we do *not* require information about r and s' that are generated after $(s,a) \sim \nu$ and only care about the (s,a) marginal itself, so the user can pick ν in an arbitrary manner without knowing the transition and the reward functions of the MDP. We assume that ν is given in a way that we can take expectation $\mathbb{E}_{(s,a)\sim\nu}[(\cdot)]$, and extension to the case where ν is given via samples is straightforward.

To achieve this goal, we first turn Eq.(1) into an equivalent *constrained convex program*: given a collection of strongly convex and differentiable functions $\{f_{s,a} : \mathbb{R} \to \mathbb{R}\}_{s,a}$—we will refer to the collection as f and discuss its choice later—consider

$$\begin{aligned} &\min_q\ \mathbb{E}_{(s,a)\sim\nu}[f_{s,a}(q(s,a))] \\ &\text{s.t. } \mathbb{E}_{r\sim R(\cdot|s,a)}[r] + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[q(s',\pi)] - q(s,a) = 0,\ \forall s,a \in \mathcal{S}\times\mathcal{A}. \end{aligned} \tag{3}$$

The constraints here are the same as Eq.(1). Since Eq.(1) uniquely determines $q = q^\pi$, the feasible space of Eq.(3) is a *singleton*, so we can impose any objective function on top of these constraints (here we use $\mathbb{E}_{(s,a)\sim\nu}[f_{s,a}(q(s,a))]$) and it will not change the optimal solution (which is always q^π, the only feasible point). As we will see, however, $f = \{f_{s,a} : \mathbb{R} \to \mathbb{R}\}_{s,a}$ will serve as an important regularizer in the function-approximation setting and is crucial to our estimation guarantees.

Remark 1 ((s,a)-dependence of f)**.** Regularizers in prior works are (s,a)-*independent* [NCDL19; YNDLS20; ZHHJL22]. As we will see in Section 4.3, allowing for (s,a)-dependence is very important for designing regularizers with improved guarantees and performances.

We now rewrite (3) in its Lagrangian form, with $d^D \circ w$ serving the role of dual variables:

$$\min_q \max_w L_f^q(q, w) := \mathbb{E}_\nu[f_{s,a}(q(s,a))] + \mathbb{E}_{d^D}\left[w(s,a)\left(r(s,a) + \gamma q(s', \pi) - q(s,a)\right)\right]. \quad (4)$$

Finally, our actual estimator approximates Eq.(4) via finite-sample approximation of the population loss L_f^q, and searches over restricted function classes $\mathcal{Q}$ and $\mathcal{W}$ for q and w, respectively:

$$(\widehat{q}, \widehat{w}) = \arg\min_{q \in \mathcal{Q}} \arg\max_{w \in \mathcal{W}} \widehat{L}_f^q(q, w), \quad (5)$$

where $\widehat{L}_f^q(q, w) = \mathbb{E}_\nu[f_{s,a}(q(s,a))] + \frac{1}{n}\sum_{i=1}^n w(s_i, a_i)\left(r_i + \gamma q(s_i', \pi) - q(s_i, a_i)\right)$.

Intuition for identification Before giving the detailed finite-sample analysis, we provide some high-level intuitions for why we can obtain the desired guarantee on $\|\widehat{q} - q^\pi\|_{2,\nu}$. Note that Eq.(4) is structurally similar to Eq.(2), and we still cannot verify the Bellman equation for q^π in a per-state-action manner, so the caveat of Eq.(2) seems to remain; why can we identify q^π under ν?

The key here is to show that it suffices to check the loss function L_f^q only under a special choice of w (as opposed to all of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$). Importantly, this special w is *not* $w = w^\pi$;[6] rather, it is the saddle point of our regularized objective L_f^q: let (q^π, w_f^*) be a saddle point of L_f^q (we will give the closed form of w_f^* later). As long as $w_f^* \in \mathcal{W}$—even if $\mathcal{W}$ is extremely "simple" and contains nothing but w_f^*—we can identify q^π.

To see that, it is instructive to consider the special case of $\mathcal{W} = \{w_f^*\}$ and the limit of infinite data. In this case, our estimator becomes $\arg\min_{q\in\mathcal{Q}} L_f^q(q, w_f^*)$. By the definition of saddle point:

$$L_f^q(q^\pi, w_f^*) \le L_f^q(q, w_f^*), \ \forall q.$$

While this shows that q^π is a minimizer of the loss, it does not imply that it is a unique minimizer. However, identification immediately follows from the convexity brought by regularization: since $f : \mathbb{R} \to \mathbb{R}$ is strongly convex, $q \to \mathbb{E}_\nu[f_{s,a}(q(s,a))]$ as a mapping from $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ to $\mathbb{R}$ is strongly convex under $\|\cdot\|_{2,\nu}$ (see Lemma 7 in Appendix B for a formal statement and proof), and $L_f^q(q, w_f^*)$ inherits such convexity since the other terms are affine in q. It is then obvious that q^π is the unique minimizer of $L_f^q(q^\pi, w_f^*)$ up to $\|\cdot\|_{2,\nu}$, that is, any minimizer of L_f^q must agree with q^π on (s,a) pairs supported on ν. Our finite-sample analysis below shows that the above reasoning is robust to finite-sample errors and having functions other than w_f^* in $\mathcal{W}$.

4.2 Finite-sample Guarantees

In this subsection we state the formal guarantee of our estimator for q^π and the assumptions under which the guarantee holds. We start with the condition on the regularization function f:

Assumption 1 (Strong convexity of f)**.** Assume $f_{s,a} : \mathbb{R} \to \mathbb{R}$ is nonnegative, differentiable, and M^q-strongly convex for each $s \in \mathcal{S}, a \in \mathcal{A}$. In addition, assume both $f_{s,a}$ and $f_{s,a}'$ take finite values for any finite input.

This assumption can be concretely satisfied by a simple choice of $f_{s,a}(x) = \frac{1}{2}x^2$, which is independent of (s,a) and yields $M^q = 1$. Alternative choices of f will be discussed in Section 4.3. Next are the realizability and boundedness of $\mathcal{W}$ and $\mathcal{Q}$:

Assumption 2 (Realizability)**.** Suppose $w_f^* \in \mathcal{W}$, $q^\pi \in \mathcal{Q}$.

Assumption 3 (Boundedness of $\mathcal{W}$ and $\mathcal{Q}$)**.** Suppose $\mathcal{W}$ and $\mathcal{Q}$ are bounded, that is, $C_\mathcal{Q}^q := \max_{q\in\mathcal{Q}} \|q\|_\infty < \infty, \quad C_\mathcal{W}^q := \max_{w\in\mathcal{W}} \|w\|_\infty < \infty.$

As a remark, Assumption 2 implicitly assumes the existence of w_f^*. As we will see in Section 4.3, the existence and finiteness of w_f^* is automatically guaranteed given the finiteness of $f_{s,a}'$ (Assumption 1) and $d^D(s,a) > 0 \ \forall s, a$. More importantly, Assumptions 2 3 together imply that $\|q^\pi\|_\infty \le C_\mathcal{Q}^q$ and $\|w_f^*\|_\infty \le C_\mathcal{W}^q$, which puts constraints on how small $C_\mathcal{Q}^q$ and $C_\mathcal{W}^q$ can be. For example, it is common to assume that $C_\mathcal{Q}^q = \frac{1}{1-\gamma}$, i.e., the maximum possible return when rewards are bounded in $[0, 1]$,

[6] In fact, w^π should not appear in our analysis at all: w^π is defined w.r.t. the initial distribution of the MDP, μ_0, which has nothing to do with our goal of bounding $\|\widehat{q} - q^\pi\|_{2,\nu}$.

and this way $\|q^\pi\|_\infty \le C^q_{\mathcal{Q}}$ will hold automatically. The magnitude of $\|w^*_f\|_\infty$ and $C^q_{\mathcal{W}}$, however, is more nuanced and interesting, and we defer the discussion to Section 4.3.

Now we are ready to state the main guarantee for identifying q^π. All proofs of this section can be found in Appendix B.

Theorem 2. *Suppose Assumptions 1, 2, 3 hold. Then, with probability at least* $1-\delta$*,*

$$||\widehat{q} - q^\pi||_{2,\nu} \le 2\sqrt{\frac{\epsilon^q_{stat}}{M^q}},$$

where $\epsilon^q_{stat} = \left(C^q_{\mathcal{W}} + (1+\gamma)C^q_{\mathcal{W}}C^q_{\mathcal{Q}}\right)\sqrt{\frac{2\log\frac{2|\mathcal{W}||\mathcal{Q}|}{\delta}}{n}}$.

Theorem 2 shows the desired bound on $\|\widehat{q} - q^\pi\|_{2,\nu}$, which depends on the magnitude of functions in $\mathcal{W}$ and $\mathcal{Q}$ as well as their logarithmic cardinalities, which are standard measures of statistical complexity for finite classes. One notable weakness is the $O(n^{-1/4})$ slow rate; this is due to translating the $\epsilon^q_{stat} = O(n^{-1/2})$ deviation between L and $\widehat{L}$ into $\|\widehat{q} - q^\pi\|_{2,\nu}$ via a convexity argument, which takes square root of the error. The possibility of and obstacles to obtaining an $O(n^{-1/2})$ rate will be discussed in Section 7.

4.3 On the Closed Form of w^*_f and the Data Coverage Assumptions

One unusual aspect of our guarantees in Section 4.2 is that we do not make any explicit data coverage assumptions, yet such assumptions are known to be necessary even for return estimation (typically the boundedness of $w^\pi = d^\pi/d^D$). Indeed, our data-coverage assumption is implicit in Assumptions 2 and 3, which require $\|w^*_f\|_\infty \le C^q_{\mathcal{W}} < \infty$. If data fails to provide sufficient coverage, $\|w^*_f\|_\infty$ will be large and our bound in Theorem 2 will suffer due to a large value of $C^q_{\mathcal{W}}$.

To make the data coverage assumption explicit, we provide the closed-form expression of w^*_f:

Lemma 3. *The saddle point of* (4) *is* $(q^\pi, w^*_f) = \arg\min_q \arg\max_w L^q_f(q,w)$*, where*

$$w^*_f = (I - \gamma\widetilde{P}^\pi)^{-1}\left(\nu \circ f'(q^\pi)\right)/d^D. \tag{6}$$

Here $f'(q^\pi)$ *is the shorthand for* $[f'_{s,a}(q^\pi(s,a))]_{s,a} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

The closed-form expression in Eq.(6) looks very much like a density ratio: if we replace $\nu \circ f'(q^\pi)$ with μ^π_0, we have $(I-\gamma\widetilde{P}^\pi)^{-1}\mu^\pi_0 = d^\pi/(1-\gamma)$, and the expression would be the ratio between d^π and d^D (up to a horizon factor). Therefore, w^*_f can be viewed as the density ratio of π against d^D when π starts from the "fake" initial distribution $\nu \circ f'(q^\pi)$. However, $\nu \circ f'(q^\pi)$ is in general not a valid distribution, as it is not necessarily normalized or even non-negative, making $\|w^*_f\|_\infty$ difficult to intuit. Below we give relaxations of $\|w^*_f\|_\infty$, which are more interpretable and give novel insights into how to relax the data-coverage assumption via tweaking f.

Proposition 4. $\|w^*_f\|_\infty \le \frac{1}{1-\gamma}\cdot\|d^\pi_\nu/d^D\|_\infty \cdot \|f'(q^\pi)\|_\infty$*, where* d^π_ν *is the discounted state-action occupancy of* π *under* ν *as the initial state-action distribution.*

The proposition states that $\|w^*_f\|_\infty$ can be bounded if data provides sufficient coverage over d^π_ν, and if $f'(q^\pi)$ is bounded. The former shows that d^D needs to cover not only ν, but also state-action pairs reachable by π starting from ν. The latter is easily satisfied, and can be bounded again for concrete choices of f, e.g. $\|f'(q^\pi)\|_\infty \le \|q^\pi\|_\infty \le \frac{1}{1-\gamma}$ for $f_{s,a}(x) = \frac{1}{2}x^2$.

Designing f to relax the coverage assumption Lemma 3 shows that the coverage assumption ($\|w^*_f\|_\infty$) depends on f (or rather its derivative f'), which opens up the possibility of properly designing f to relax it. In fact, we could completely eliminate the coverage assumption if we could set $f'(q^\pi) = \mathbf{0}$, but that would require unrealistically strong side information.

As a concrete example, consider $f_{s,a}(x) = \frac{1}{2}(x - q^\pi(s,a))^2$, and it is easy to verify that $f'_{s,a}(q^\pi(s,a)) = x - q^\pi(s,a)|_{x=q^\pi(s,a)} = 0$. Compared to $f_{s,a}(x) = \frac{1}{2}x^2$, the new f essentially adds a 1st-order term $q^\pi(s,a)\cdot x$ to change w^*_f, while leaving the convexity required by Assumption 1 intact, which only depends on the 2nd-order term $\frac{1}{2}x^2$. Of course, this is not a viable choice of f in practice as it requires knowledge of q^π, which is precisely our learning target.

While the reason $f_{s,a}(x) = \frac{1}{2}(x - q^\pi(s,a))^2$ can eliminate the coverage requirement is obvious retrospectively (q^π already minimizes $\mathbb{E}_\nu[f(q)]$ even without any data), our analyses apply much more generally and characterize the effects of arbitrary f on the coverage assumption. Inspired by this example, we can consider practically feasible choices such as $f_{s,a}(x) = \frac{1}{2}(x - \widetilde{q}(s,a))^2$, where $\widetilde{q}$ is an approximation of q^π obtained by other means, e.g. a guess based on domain knowledge. If $\widetilde{q} \approx q^\pi$, our estimator enjoys significantly relaxed coverage requirements. But even if $\widetilde{q}$ is a poor approximation of q^π, it does not affect our estimation guarantees as long as the condition implied by Proposition 4 is satisfied. (In fact, $f_{s,a}(x) = \frac{1}{2}x^2$ is a special case of $\widetilde{q} \equiv 0$.) Such a use of approximate models is similar to how doubly robust estimators [DLL11; JL16; TB16] enjoy reduced variance given an accurate model, and remain unbiased even if the approximate model is arbitrarily poor. We will also empirically evaluate the effectiveness of this idea in Section 6.

5 Weight-function Estimation

Similar to value-function estimation, our methodology can also be applied to estimate the weight function w^π. Due to the similarity with Section 4 in the high-level spirit, we will be concise in this section and only explain in detail when there is a conceptual difference from Section 4. Some notations (such as the function classes $\mathcal{W}$ and $\mathcal{Q}$) will be abused, but we emphasize that this section considers a different learning task than Section 4, so they should be viewed as different objects (e.g., the realizability assumptions for $\mathcal{W}$ and $\mathcal{Q}$ below will be different from those in Section 4).

As before, we assume that the user provides a distribution[7] $\eta \in \Delta(\mathcal{S} \times \mathcal{A})$ and our goal is to develop an estimator with guarantees on $\|\widehat{w} - w^\pi\|_{2,\eta}$. Analogous to Section 4, consider

$$\begin{aligned} &\min_w \mathbb{E}_{(s,a)\sim\eta}[f_{s,a}(w(s,a))] \\ &\text{s.t. } d^D(s,a)w(s,a) = (1-\gamma)\mu_0^\pi(s,a) + \gamma \textstyle\sum_{s',a'} P^\pi(s,a|s',a')d^D(s',a')w(s,a),\ \forall s,a. \end{aligned} \tag{7}$$

Here $f = \{f_{s,a}\}_{s,a}$ will need to satisfy similar assumptions as in Section 4. The constraints are the Bellman flow equations with a change of variable $d(s,a) = d^D(s,a) \cdot w(s,a)$. Their unique solution is $d(s,a) = d^\pi(s,a)$ (and hence $w(s,a) = d^\pi(s,a)/d^D(s,a)$), thus the feasible space is again a singleton, and the objective does not alter the optimal solution. We then use dual variables q to rewrite (7) in its Lagrangian form: $\min_w \max_q L_f^w(q,w) :=$

$$\mathbb{E}_\eta\left[f_{s,a}(w(s,a))\right] + (1-\gamma)\mathbb{E}_{\mu_0}\left[q(s,\pi)\right] + \mathbb{E}_{d^D}\left[w(s,a)(\gamma q(s',\pi) - q(s,a))\right] \tag{8}$$

We approximate the saddle-point solutions by optimizing the empirical loss $\widehat{L}_f^w$ over restricted function classes $\mathcal{W}, \mathcal{Q}$: $(\widehat{w}, \widehat{q}) = \arg\max_{w\in\mathcal{W}} \arg\min_{q\in\mathcal{Q}} \widehat{L}_f^w(q,w)$, where $\widehat{L}_f^w(q,w) := \mathbb{E}_\eta\left[f_{s,a}(w(s,a))\right] + (1-\gamma)\frac{1-\gamma}{n_0}\sum_{j=1}^{n_0} q(s_j,\pi) + \frac{1}{n}\sum_{i=1}^n w(s_i,a_i)\left(\gamma q(s_i',\pi) - q(s_i,a_i)\right)$, and $\{s_j\}_{j=1}^{n_0}$ is a separate dataset sampled i.i.d. from μ_0 to provide information about initial distribution.

We provide the closed-form expression for the saddle point of L_f^w below, which resembles the Q-function for a proxy reward function $f'(w^\pi) \circ \eta/d^D$.

Lemma 5. *The closed form solutions of* (8) *are* $(w^\pi, q_f^*) = \arg\min_w \arg\max_q L_f^w(q,w)$*, where*

$$q_f^* = (I - \gamma P^\pi)^{-1}(f'(w^\pi) \circ \eta/d^D). \tag{9}$$

Remark 2 (Data Coverage Assumption)**.** As we will see, the only data coverage assumption we need is the boundedness of $w^\pi = d^\pi/d^D$. Since w^π is the function of interest and practical algorithms can only output functions of well-bounded ranges, such an assumption is an essential part of the learning task itself and hardly an additional requirement. Moreover, unlike Section 4, changing f here will not affect the data-coverage assumption, though it still alters q_f^*, and a properly chosen f (e.g., with $f'(w^\pi) \approx \mathbf{0}$) can still result in a q_f^* with small magnitude and thus make learning easier.

Remark 3 (Connection to DualDICE)**.** We can recover DualDICE [NCDL19] by choosing $f_{s,a}(x) = \frac{1}{2}x^2$ and $\nu = d^D$. Despite producing the same estimator, the derivations and assumptions under which the two works analyze the estimator are different. Their Theorem 2 only provides return

[7] Recall we assume $d^D(s,a) > 0\ \forall s,a$ for technical convenience. When this is not the case, η should be supported on d^D, as the target function w^π is only defined on the support of d^D.

estimation guarantees, and depends on an implicit assumption of highly expressive function classes[8] similar to Proposition 1. Moreover, they do not characterize how the choice of f can affect the learning guarantees (their f is (s,a)-independent). This is one of the main insights of our paper and leads to the discovery of more practical regularizers, e.g. $f_{s,a}(x) = \frac{1}{2}(x - \tilde{w}(s,a))^2$ with model $\widetilde{w}$.

Below we present the assumptions, then learning guarantee for $\widehat{w}$.

Assumption 4 (Strongly Convex Objective)**.** Suppose for all s, a, $f_{s,a}$ is differentiable, non-negative, and M^w-strongly convex. Further, suppose $f_{s,a}$ and its derivative take finite values on any finite inputs, and let $C_f^w := \max_{w \in \mathcal{W}} ||f(w)||_\infty$.

Assumption 5 (Realizability)**.** Suppose $w^\pi \in \mathcal{W}$, $q_f^* \in \mathcal{Q}$.

Assumption 6 (Bounded $\mathcal{W}$ and $\mathcal{Q}$)**.** Let $C_{\mathcal{W}}^w := \max_{w \in \mathcal{W}} ||w||_\infty$ and $C_{\mathcal{Q}}^w := \max_{q \in \mathcal{Q}} ||q||_\infty$. Suppose $\mathcal{W}$ and $\mathcal{Q}$ are bounded function classes, that is, $C_{\mathcal{W}}^w < \infty$ and $C_{\mathcal{Q}}^w < \infty$.

Theorem 6. *Suppose Assumptions 4, 5, 6, hold. Then w.p.* $\geq 1-\delta$, $\|\widehat{w} - w^\pi\|_{2,\eta} \leq 2\sqrt{\frac{\epsilon_{stat}^w}{M^w}}$, *where* $\epsilon_{stat}^w = \left(C_f^w + (1+\gamma)C_{\mathcal{W}}^w C_{\mathcal{Q}}^w\right)\sqrt{\frac{2\log\frac{4|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + (1-\gamma)C_{\mathcal{Q}}^w\sqrt{\frac{2\log\frac{4|\mathcal{Q}|}{\delta}}{n_0}}$.

6 Experiments

We now provide experimental results to verify our theoretical predictions and insights. As [YNDLS20] have performed extensive experiments on return estimation with simple regularization ($f_{s,a}(x) = \frac{1}{2}x^2$), we focus on the task of q^π estimation, and the following two questions unique to our work:

Q1. When the goal is to minimize $\|\widehat{q} - q^\pi\|_{2,\nu}$, how much benefit does regularizing with ν bring in practice, compared to regularizing with other distributions (or no regularization at all)?

Q2. Can incorporating (even relatively poor) models in regularization (e.g., $f_{s,a}(x) = \frac{1}{2}(x - \widetilde{q}(s,a))^2$ from Section 4.3) improve estimation?

Setup We study these questions in a large tabular Gridwalk environment [NCDL19; YNDLS20], with a deterministic target policy π that is optimal, and a behavior policy that provides limited coverage over the target policy; see Appendix D for further details. To mimic the identification challenges associated with restricted function classes, we use a linear function class $\mathcal{Q} = \{\Phi^\top \alpha : \alpha \in \mathbb{R}^d\}$ and discriminator class $\mathcal{W} = \{\widetilde{\Phi}^\top \beta : \beta \in \mathbb{R}^k\}$, where $k < d \ll |\mathcal{S} \times \mathcal{A}|$. The features $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times d}$, $\widetilde{\Phi} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times k}$ are chosen to satisfy the realizability assumptions of all estimators. Under linear classes, our estimator (Eq.(5)) becomes a convex optimization problem with d variables and k linear constraints, and can be solved by standard packages. This allows us to avoid difficult minimax optimization—which is still an open problem in the MIS literature—and focus on the statistical behaviors of our estimators, which is what our theoretical predictions are about.

Remark 4. When no regularization is used, our estimator coincides with MQL [UHJ20]. If we further had $\widetilde{\Phi} = \Phi$, the estimator would coincide with LSTDQ. While Section 2 mentioned that LSTDQ enjoys function-estimation guarantees [PKBK22] (and folklore suggests they extend to $\widetilde{\Phi} \neq \Phi$), the guarantee only holds in the regime of $k \geq d$, i.e., the k linear constraints are *over-determined*. In our case, however, we have *under-determined* constraints ($k < d$), creating a more challenging learning task (which our theory can handle) where LSTDQ's guarantees do not apply.

Choice of Distributions We consider a set of diverse distributions $\mathcal{V} = \{d^D, \mu_0^\pi, d^\pi, U, p\}$, where U is uniform over $\mathcal{S} \times \mathcal{A}$ and $p \propto (d^\pi \circ \mathbb{I}[w^\pi > 50])$. The distribution p isolates the least-covered states reached by π, which makes learning an accurate Q-function on ν a harder task.

Results for Q1 We use a default regularizer $f = \frac{1}{2}x^2$ with different regularizing distributions $\nu \in \mathcal{V}$, and measure $\|\widehat{q} - q^\pi\|_{2,\nu'}$ for different ν'. The results are shown in Figure 1, which exhibit the expected trend: for example, regularizing with $\nu = p$ performs poorly when the error is measured under $\nu' = d^D$ and U due to the large mismatch between ν and ν'. However, when $\nu' = p$ (rightmost panel), regularizing with $\nu = p$ significantly outperforms others. Similar behaviors can also be observed on U, though they are certainly not absolute (e.g., $\nu = d^D$ does not do very well on

[8] In our notation, they measure the approximation error of $\mathcal{W}$ as $\max_{w' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \min_{w \in \mathcal{W}} \|w - w'\|$, essentially requiring $\mathcal{W}$ (and similarly $\mathcal{Q}$) to closely approximate every function over $\mathcal{S} \times \mathcal{A}$. However, we suspect that they could have measured realizability errors instead without changing much of their proofs.

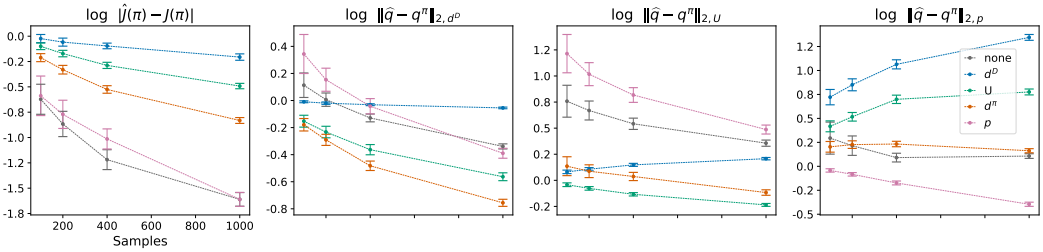

Figure 1: Error of off-policy return and function estimation as a function of sample size. Legend shows regularizing distribution ν and header shows error-measuring distribution ν' (see text). Error bars show 95% confidence intervals calculated from 1000 runs.

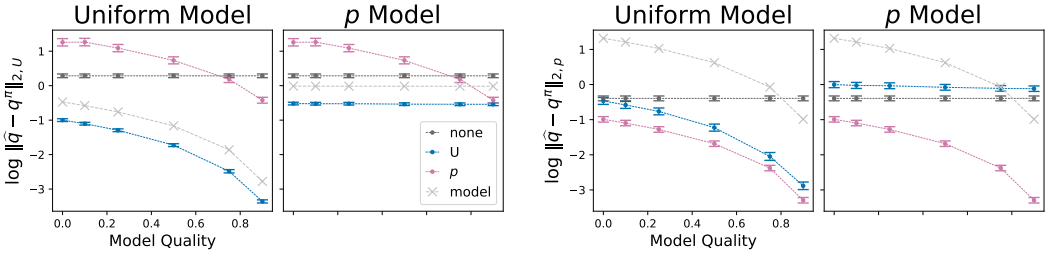

Figure 2: Estimation error when the regularizer incorporates a model $\widetilde{q}$, where x-axes represent the parameter m that controls the quality of $\widetilde{q}$. Sample size is 500 and the results are from 500 runs.

$\nu' = d^D$), which suggests potential directions for more refined and accurate theory. Moreover, using no regularization ("none") generally does not perform well for any ν', but still manages to achieve a high accuracy for return estimation $J(\pi)$, which is consistent with prior theory [e.g., UHJ20] that return estimation does not require regularization.

Results for Q2 We now use $f_{s,a}(x) = \frac{1}{2}(x - \widetilde{q}(s,a))^2$ with different $\widetilde{q}$ to verify how the quality of $\widetilde{q}$ affect estimation accuracy. We first consider a "uniform model" $\widetilde{q} = mq^\pi + (1-m)\overline{q}$, where $\overline{q}$ is a constant and $m \in [0,1]$ controls the quality $\widetilde{q}$. As shown from Panels 1 & 3 in Figure 2, our estimator's accuracy generally improves with a better $\widetilde{q}$ (i.e., as m increases). Moreover, equipping $\widetilde{q}$ with an appropriate regularizing distribution ν (e.g., $\nu = U$ for both panels) can significantly outperform no regularization, even with a very poor $\widetilde{q}$ (e.g., $m = 0.1$). It also outperforms the model prediction itself (i.e., $\widehat{q} = \widetilde{q}$), showing that the improvement is not due to our estimator simply taking predictions from $\widehat{q}$, but using the regularization to better assist the identification of q^π from data.

The previous model's quality is uniform across $\mathcal{S} \times \mathcal{A}$. We then consider a scenario where $\widetilde{q}$ is zeroed out outside p's support, making it only a good approximation of q^π on p. In this case, we see that regularization cannot benefit much from the model when the error is measured on $\nu' = U$ (Panel 2), but when $\nu' = \nu = p$ (Panel 4), regularization can still bring benefits, as expected from our theory.

7 Discussion and Conclusion

In this paper we showed that proper regularization can yield function-estimation guarantees for MIS methods under only realizable function approximation. Compared to prior works, our regularizer is more flexible and can accommodate a user-specified error-measuring distribution. Further theoretical investigation provides fine-grained characterization of how the choice of regularization affects learning guarantees, which leads to the discovery of regularizers that incorporate approximate models (such as $\widetilde{q}$). While the superiority of such regularizers is perhaps obvious retrospectively, it is not allowed in the prior works' derivation that assumes (s,a)-independent regularization, and our theoretical results provide a deep understanding for even more general regularization schemes. In Appendix A, we provide further discussions on two topics: (1) the barriers to obtaining a faster $O(n^{-1/2})$ rate, and (2) comparison to [ZHHJL22] reveals interesting differences between off-policy function estimation and policy learning, and insights in this paper may also be useful for the policy learning task.

References

[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". In: *Machine Learning* (2008).

[AYBBLSW19] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. "Politex: Regret bounds for policy iteration using expert prediction". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3692–3702.

[Bai95] Leemon Baird. "Residual algorithms: Reinforcement learning with function approximation". In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.

[BY09] Dimitri P Bertsekas and Huizhen Yu. "Projected equation methods for approximate solution of large linear systems". In: *Journal of Computational and Applied Mathematics* 227.1 (2009), pp. 27–50.

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

[DJW20] Yaqi Duan, Zeyu Jia, and Mengdi Wang. "Minimax-optimal off-policy evaluation with linear function approximation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2701–2709.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. "Doubly Robust Policy Evaluation and Learning". In: *Proceedings of the 28th International Conference on Machine Learning*. 2011, pp. 1097–1104.

[DNP14] Christoph Dann, Gerhard Neumann, and Jan Peters. "Policy evaluation with temporal differences: A survey and comparison". In: *The Journal of Machine Learning Research* 15 (2014), pp. 809–883.

[EGW05] Damien Ernst, Pierre Geurts, and Louis Wehenkel. "Tree-based batch mode reinforcement learning". In: *Journal of Machine Learning Research* 6 (2005), pp. 503–556.

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

[KL02] Sham Kakade and John Langford. "Approximately Optimal Approximate Reinforcement Learning". In: *Proceedings of the 19th International Conference on Machine Learning*. Vol. 2. 2002, pp. 267–274.

[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* 31 (2018), pp. 5356–5366.

[LP03] Michail G Lagoudakis and Ronald Parr. "Least-squares policy iteration". In: *The Journal of Machine Learning Research* 4 (2003), pp. 1107–1149.

[LSAB19] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. "Off-policy policy gradient with state distribution correction". In: *arXiv preprint arXiv:1904.08473* (2019).

[LVY19] Hoang Le, Cameron Voloshin, and Yisong Yue. "Batch Policy Learning under Constraints". In: *International Conference on Machine Learning*. 2019, pp. 3703–3712.

[NCDL19] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. "Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections". In: *arXiv preprint arXiv:1906.04733* (2019).

[ND20] Ofir Nachum and Bo Dai. "Reinforcement Learning via Fenchel-Rockafellar Duality". In: *arXiv preprint arXiv:2001.01866* (2020).

[NDKCLS19] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. "AlgaeDICE: Policy Gradient from Arbitrary Experience". In: *arXiv preprint arXiv:1912.02074* (2019).

[Pai+20] Tom Le Paine, Cosmin Paduraru, Andrea Michi, Caglar Gulcehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. "Hyperparameter Selection for Offline Reinforcement Learning". In: *arXiv preprint arXiv:2007.09055* (2020).

[PKBK22] Juan C Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. "A Sharp Characterization of Linear Estimators for Offline Policy Evaluation". In: *arXiv preprint arXiv:2203.04236* (2022).

[PSS00] Doina Precup, Richard S Sutton, and Satinder P Singh. "Eligibility Traces for Off-Policy Policy Evaluation". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. 2000, pp. 759–766.

[TB16] Philip Thomas and Emma Brunskill. "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning". In: *Proceedings of the 33rd International Conference on Machine Learning*. 2016.

[TTG15] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. "High Confidence Off-Policy Evaluation". In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. 2015.

[UHJ20] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. "Minimax Weight and Q-Function Learning for Off-Policy Evaluation". In: *Proceedings of the 37th International Conference on Machine Learning*. 2020, pp. 1023–1032.

[UIJKSX21] 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". In: *arXiv preprint arXiv:2102.02981* (2021).

[VLJY19] Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. "Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning". In: *arXiv preprint arXiv:1911.06854* (2019).

[XCJMA21] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. "Bellman-consistent Pessimism for Offline Reinforcement Learning". In: *arXiv preprint arXiv:2106.06926* (2021).

[XMW19] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. "Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling". In: *arXiv preprint arXiv:1906.03393* (2019).

[YNDLS20] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. "Off-policy evaluation via the regularized lagrangian". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6551–6561.

[ZHHJL22] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. "Offline reinforcement learning with realizability and single-policy concentrability". In: *arXiv preprint arXiv:2202.04634* (2022).

[ZHWZ21] Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. "Generalization bounds for stochastic saddle point problems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 568–576.

[ZJ21] Siyuan Zhang and Nan Jiang. "Towards Hyperparameter-free Policy Selection for Offline Reinforcement Learning". In: *Advances in Neural Information Processing Systems* 34 (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] We discuss the slow rate of our guarantee in Appendix A.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (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)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(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)? [No] Our experiments are on relatively simple synthetic environments, and we did not particularly track the use of computation.

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]