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

Audrey Huang
Computer Science
University of Illinois at Urbana-Champaign
audreyh5@illinois.edu

Nan Jiang
Computer Science
University of Illinois at Urbana-Champaign
nanjiang@illinois.edu

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).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In comparison, little is known about off-policy function estimation, and the guarantees are generally less desirable. Not only do 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 defeats 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 specific distributions [KL02; AYBBLSW19] (see details in Appendix A).

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, prior works provide very limited understanding in how choice of regularization affects learning guarantees, and have considered only naïve forms of regularization (state-action-*independent* and typically under d^D), under which different forms of regularization are essentially treated equally under a coarse-grained theory [YNDLS20]. In contrast, we provide much

[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.

more fine-grained characterization of the effects of regularization, which leads to novel insights about how to design better regularizers, and existing DICE estimators are subsumed as special cases of our method when we choose very simple regularizers (see Remark 3 in Section 5).

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 $(s_i, a_i) \sim d^D$, $r \sim R(\cdot|s_i, a_i)$, 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 provided in Appendix G.

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 $f(s,\pi) := \mathbb{E}_{a\sim\pi(\cdot|s)}[q(s,a)]$. Elementwise multiplication between two vectors u and v of the same dimension is $u \circ v$, and elementwise division is u/v.

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.,

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

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) can 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 $\widehat{q} \approx q^\pi$, they only 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 the latter is possible when $w^\pi \in \mathcal{W}$, 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 realizability 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 solving 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 its 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:

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

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 the inclusion of 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 its derivative $f'_{s,a}$ 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 and 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

[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 $\|\widehat{q} - q^\pi\|_{2,\nu}$.

in $[0,1]$, 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 (bounded $\|w^*_f\|_\infty$) 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 show in Section 6 that this idea is empirically effective.

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 $\mathcal{W}, \mathcal{Q}$: $\widehat{w} = \arg\min_{w\in\mathcal{W}} \max_{q\in\mathcal{Q}} \widehat{L}_f^w(q,w)$, 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 the 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 linear 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.

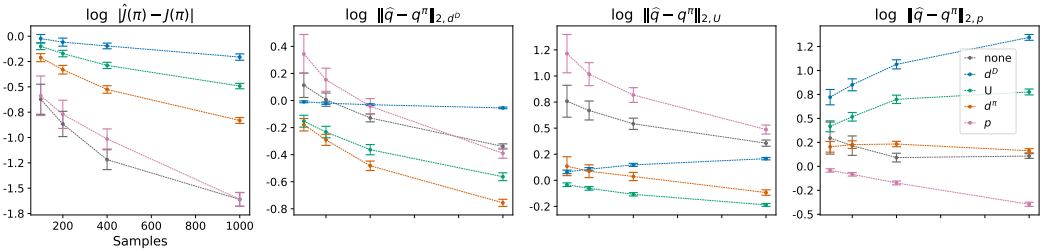

Figure 1: Error of off-policy return and function estimation as a function of sample size. Legend shows regularizing distribution ν and header shows error-measuring distribution ν' (see text). Error bars show 95% confidence intervals calculated from 1000 runs.

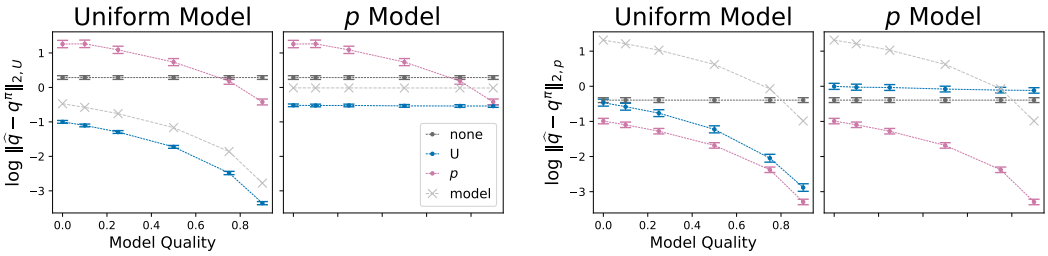

Figure 2: Estimation error when the regularizer incorporates a model $\widetilde{q}$, where x-axes represent the parameter m that controls the quality of $\widetilde{q}$. The "model" line shows performance of $\widetilde{q}$. Sample size is 500 and the results are from 500 runs.

$\nu' = d^D$), 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.

Acknowledgments and Disclosure of Funding

The authors thank Jinglin Chen, Wenhao Zhan, and Jason Lee for valuable discussions during the early phase of the project, and Anonymous Reviewer e62P for insightful feedback that helped improve this paper during the review process. NJ acknowledges funding support from ARL Cooperative Agreement W911NF-17-2-0196, NSF IIS-2112471, NSF CAREER IIS-2141781, and Adobe Data Science Research Award.

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).

[AVDB18] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. "A rewriting system for convex optimization problems". In: *Journal of Control and Decision* 5.1 (2018), pp. 42–60.

[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.

[DB16] Steven Diamond and Stephen Boyd. "CVXPY: A Python-embedded modeling language for convex optimization". In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[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.

[KU20] Nathan Kallus and Masatoshi Uehara. "Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes." In: *J. Mach. Learn. Res.* 21.167 (2020), pp. 1–63.

[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]

A Further Discussions

Motivating Examples Below, we give a thorough discussion, through four examples of how off-policy function estimation used in downstream learning algorithms. We highlight the discrepancies between what these algorithms assume about the function estimates and what existing work is able to achieve, and demonstrate how our work closes these gaps.

Batch Learning: [LSAB19] design an off-policy policy gradient algorithm that requires estimating the density-ratio w^π to correct the offline data distribution to the on-policy distribution. In their convergence analysis, they assume access to a blackbox w^π estimator that is accurate under d^D, and refer to [LSAB19] as a possible method. However, as per Proposition 1, [LLTZ18] and existing works do not provide desirable guarantees for such a task.

Online Reinforcement Learning: The seminal paper of [KL02] designs the CPI algorithm for on-policy policy improvement, which inspired popular empirical algorithms such as TRPO and PPO. CPI requires an oracle for estimating the advantage function ($\approx$ value function up to offset) accurately under the on-policy distribution, i.e., distribution induced by the current policy (see their Sec 7.1). While this is easy to do by simple squared-loss regression onto on-policy trajectories, it can be sample-inefficient as it fails to leverage off-policy data collected by previous policies. On the other hand running something like TD on all data considers a distribution different from the on-policy one. Our method offers a direct solution: use all data in the Bellman error part of the objective, and only use on-policy trajectories in the regularizer.

Online Reinforcement Learning: [AYBBLSW19] designs a no-regret policy optimization algorithm assuming access to value-function estimation oracles. In their Theorem 5.1, they assume that the oracle outputs an estimate of q^π that is accurate under $\nu = d^{\pi^*}$. While d^{π^*} is obviously not accessible to us and our method does not apply as-is, one might use our theoretical insights to design heuristics, such as up-weighting high-reward states in the offline distribution, as a way to mimic d^{π^*}.

Model selection in Offline Return Estimation: Model selection in offline return estimation: Hyperparameter tuning is a huge practical hurdle in offline return estimation [Pai+20], i.e., all OPE estimators for return estimation (except for importance sampling which has exponential variance) require some form of function approximation, and it is hard to choose the right function class with offline data alone. To address this issue, [ZJ21] proposes a model selection process over candidate function estimates of q^π, which must be provided by base algorithms that perform function estimation.

Function Estimation & Downstream Tasks Online algorithms using off-policy function estimation as a subroutine, such as [KL02; AYBBLSW19], may require the estimates to be accurate on unknown distributions such as d^π or d^{π^*} (where π^* is the optimal policy), which may not be immediately accessible to the user. The user may be able to use domain knowledge to "guess" a distribution ν close to or covering the unknown distribution of interest. Then our guarantees for function estimation over ν could similarly, with a change of distribution, be converted to guarantees on the true distribution of interest. To this end, an important avenue of future work involves a thorough investigation of how our off-policy function estimation method interacts with such downstream learning algorithms, their assumptions, and their guarantees, as well as how our method can tailored to improve downstream tasks.

Faster rate One weakness of our result is the $O(n^{-1/4})$ slow rate of estimation. While $O(n^{-1/2})$ generalization error bounds for related stochastic saddle point exist [ZHWZ21], they only apply to strongly-convex-strongly-concave problems, whereas our problem is strongly-convex-non-strongly-concave (L_f^q is affine in w and L_f^w is affine in q), making the result not directly applicable. One immediate idea is to introduce dual regularization to make our objectives also strongly concave in the discriminator. However, while primal regularization does not change the feasible space and guarantees that the learned function will be q^π (or w^π, respectively), dual regularization *does* change the optimal solution, introducing a bias. This leads to a trade-off between the improvement in error bounds due to strong concavity and the additional bias, and our preliminary investigation shows that an optimal trade-off between the two sources of errors still leads to an $O(n^{-1/4})$ rate. Therefore, improving the rate (if it is possible at all) will require novel technical tools for the generalization analyses of strongly-convex-non-strongly-concave stochastic saddle point problems, which will be an interesting future direction.

On a related note, while the rate for estimating q^π and w^π is only $O(n^{-1/4})$, we can combine them in a doubly robust form to get $O(n^{-1/2})$ rate for return estimation by careful choices of the regularizing distributions ν and η; see Appendix F for details.

Comparison to off-policy learning As mentioned earlier, our results are enabled by technical tools adapted from [ZHHJL22], whose work focuses on off-policy policy learning and learns w^π for a near-optimal π that is accurate under d^D as an intermediate step. While most of our surprising observations are in the value-function learning scenario (Section 4), comparing our guarantee for learning w^π (Section 5) to that of [ZHHJL22] still yields interesting observations about the difference between off-policy evaluation and learning. Most notably, we do not need to control the strength of regularization in Eq.(3), since the feasible space is a singleton and there is no objective before we introduce $\mathbb{E}_\nu[f(q)]$. In contrast, the feasible space is not a singleton in [ZHHJL22] (it is the space of all possible occupancies) and there is already a return optimization objective, so [ZHHJL22] need to carefully control the strength of their regularization. As a consequence, [ZHHJL22] obtain $O(n^{-1/6})$ rate, showing how off-policy learning is potentially more difficult than off-policy function estimation. Another interesting difference is related to our exact characterization of w_f^* and q_f^*: [ZHHJL22] do not have a closed-form expression for their optimal dual solution. Such a lack of direct characterization leads to requiring additional assumptions to guarantee the boundedness of such variables (see their Assumptions 11 and 12), which is not a problem in our setting. Finally, our analyses lead to novel algorithmic ideas such as using state-action-dependent regularizers and incorporating approximate models in the regularizers, which are potentially also useful for policy learning.

B Proofs for Section 4

B.1 Proof of Theorem 2

From Assumption 1 and Lemma 7, we know that the regularization function $\mathbb{E}_\nu[f_{s,a}(q(s,a))]$ is an M-strongly convex function in q on the $\|\cdot\|_{2,\nu}$ norm. Now consider $L_f^q(q, w_f^*)$, the Lagrangian function (4) at the optimal discriminator w_f^*. Since $L_f^q(q, w_f^*)$ is composed of the regularization function plus terms that are linear in q, $L_f^q(q, w_f^*)$ is also an M-strongly convex function in q.

As (q^π, w_f^*) is the saddle point solution of L_f^q, we know $q^\pi = \arg\min_q L_f^q(q, w_f^*)$. Then from the strong convexity of L_f^q,

$$
\begin{aligned}
||\widehat{q} - q^\pi||_{2,\nu} &\leq \sqrt{\frac{2\left(L_f^q(\widehat{q}, w_f^*) - L_f^q(q^\pi, w_f^*)\right)}{M^q}} \\
&\leq \sqrt{\frac{4\epsilon_{stat}^q}{M^q}}, \qquad \text{(Lemma 9)}
\end{aligned}
$$

where ϵ_{stat}^q is given in Lemma 8.

We provide the helper lemmas and their proofs below:

Lemma 7. *Suppose $f_{s,a} : \mathbb{R} \to \mathbb{R}$ is M-strongly convex. Then $\mathbb{E}_\nu[f_{s,a}(q(s,a))] : \mathbb{R}^{|SA|} \to \mathbb{R}$ is M-strongly convex on $\|\cdot\|_\nu$.*

Proof. From the strong convexity of $f_{s,a}$, for any $x, y \in \mathbb{R}$,

$$
f_{s,a}(x) - f_{s,a}(y) \leq f'_{s,a}(x)(x-y) - \frac{M}{2}(x-y)^2
$$

Then for $q, q' \in \mathbb{R}^{|SA|}$,

$$
\begin{aligned}
&\mathbb{E}_\nu[f_{s,a}(q(s,a))] - \mathbb{E}_\nu[f_{s,a}(q'(s,a))] \\
&\quad\leq \mathbb{E}_\nu[f'_{s,a}(q(s,a))(q(s,a) - q'(s,a))] - \mathbb{E}_\nu[\frac{M}{2}(q(s,a) - q'(s,a))^2] \\
&\quad\leq \mathbb{E}_\nu[f'_{s,a}(q(s,a))(q(s,a) - q'(s,a))] - \left(\min_{s,a} \frac{M}{2}\right) \mathbb{E}_\nu[(q(s,a) - q'(s,a))^2] \\
&\quad= \langle \nabla_q \mathbb{E}_\nu[f_{s,a}(q(s,a))], q - q' \rangle - \frac{M}{2}\mathbb{E}_\nu[(q(s,a) - q'(s,a))^2]
\end{aligned}
$$

since $\nabla_q \mathbb{E}_\nu[f_{s,a}(q(s,a))] = \nu \circ f'_{s,a}(q)$, which gives our result. □

Lemma 8. *Suppose Assumption 3 holds. Then for all $(q,w) \in \mathcal{Q} \times \mathcal{W}$, w.p. $\geq 1-\delta$,*

$$|\widehat{L}^q_f(q,w) - L^q_f(q,w)| \leq \epsilon^q_{stat},$$

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

Proof. From the linearity of the expectation, it is clear that $L^q_f(q,w) = \mathbb{E}[\widehat{L}^q_f]$. Let $l_i = w(s_i,a_i)\left(r(s_i,a_i) + \gamma q(s'_i,\pi) - q(s_i,a_i)\right)$. From Assumption 3,

$$\begin{aligned}|l_i| &\leq \|w\|_\infty + (1+\gamma)\|w\|_\infty\|q\|_\infty \\ &\leq C^q_{\mathcal{W}} + (1+\gamma)C^q_{\mathcal{W}}C^q_{\mathcal{Q}}\end{aligned}$$

Then using Hoeffding's inequality with union bound, for all $q, w \in \mathcal{Q} \times \mathcal{W}$, w.p. $\geq 1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n l_i - \mathbb{E}_{d^D}[l_i]\right| \leq \left(C^q_{\mathcal{W}} + (1+\gamma)C^q_{\mathcal{W}}C^q_{\mathcal{Q}}\right)\sqrt{\frac{2\log\frac{2|\mathcal{W}||\mathcal{Q}|}{\delta}}{n}} = \epsilon^q_{stat}$$

□

Lemma 9. *Under Assumptions 1, 2, 3, w.p. $\geq 1-\delta$,*

$$L^q_f(\widehat{q}, w^*_f) - L^q_f(q^\pi, w^*_f) \leq 2\epsilon^q_{stat}.$$

where ϵ^q_{stat} is given in Lemma 8.

Proof. Let $\widehat{w}(q) := \arg\max_{w\in\mathcal{W}} \widehat{L}^q_f(q,w)$. We decompose the error as follows:

$$\begin{aligned}
L^q_f(q^\pi, w^*_f) - L^q_f(\widehat{q}, w^*_f) &= L^q_f(q^\pi, w^*_f) - L^q_f(q^\pi, \widehat{w}(q^\pi)) && (1)\ \geq 0 \\
&+ L^q_f(q^\pi, \widehat{w}(q^\pi)) - \widehat{L}^q_f(q^\pi, \widehat{w}(q^\pi)) && (2)\ \geq -\epsilon^q_{stat} \\
&+ \widehat{L}^q_f(q^\pi, \widehat{w}(q^\pi)) - \widehat{L}^q_f(\widehat{q}, \widehat{w}(\widehat{q})) && (3)\ \geq 0 \\
&+ \widehat{L}^q_f(\widehat{q}, \widehat{w}(\widehat{q})) - \widehat{L}^q_f(\widehat{q}, w^*_f) && (4)\ \geq 0 \\
&+ \widehat{L}^q_f(\widehat{q}, w^*_f) - L^q_f(\widehat{q}, w^*_f) && (5)\ \geq -\epsilon^q_{stat}
\end{aligned}$$

Combining the terms gives the result, and we provide a brief justification for each inequality below. Terms (2) and (5) follow from Lemma 8.

Term (1) ≥ 0 since (q^π, w^*_f) is the saddlepoint solution.

Term (3) ≥ 0, since $\widehat{q} = \arg\min_{q\in\mathcal{Q}} \widehat{L}^q_f(q, \widehat{w}(q))$, and $q^\pi \in \mathcal{Q}$.

Term (4) ≥ 0 because $w^*_f \in \mathcal{W}$. □

B.2 Proof of Lemma 3

Since strong duality holds, the saddle point (q^π, w^*_f) satisfies the KKT conditions. Then from stationarity, for all (s,a),

$$0 = \nu(s,a)f'_{s,a}(q^\pi(s,a)) + \gamma\sum_{s',a'} P^\pi(s,a|s',a')d^D(s',a')w^*_f(s',a') - d^D(s,a)w^*_f(s,a).$$

Writing this in matrix form, letting $f'(q^\pi)$ be shorthand for $[f'_{s,a}(q^\pi(s,a))]_{s,a} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, w^*_f must satisfy the equality:

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

B.3 Proof of Proposition 4

Rearranging the closed form of w_f^* from Lemma 3 and taking the absolute value of both sides,

$$\begin{aligned} d^D \circ |w_f^*| &= |(I - \gamma \widetilde{P}^\pi)^{-1} (\nu \circ f'(q^\pi))| \\ &\leq \|f'(q^\pi)\|_\infty |(I - \gamma \widetilde{P}^\pi)^{-1} \nu| \\ &= \frac{1}{1-\gamma} \|f'(q^\pi)\|_\infty \cdot d_\nu^\pi \end{aligned}$$

Then dividing both sides by d^D element-wise, this implies

$$\begin{aligned} |w_f^*| &\leq \frac{1}{1-\gamma} \|f'(q^\pi)\|_\infty \cdot (d_\nu^\pi / d^D) \\ &\leq \frac{1}{1-\gamma} \|f'(q^\pi)\|_\infty \cdot \|d_\nu^\pi / d^D\|_\infty \end{aligned}$$

As the above inequality holds for all (s, a),

$$||w_f^*||_\infty \leq \frac{1}{1-\gamma} \|f'(q^\pi)\|_\infty \cdot \|d_\nu^\pi / d^D\|_\infty.$$

C Proofs for Section 5

C.1 Proof of Lemma 5

From the KKT stationarity conditions:

$$0 = d^D(s,a) \left(\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[q_f^*(s', \pi) \right] - q_f^*(s,a) \right) - \nu(s,a) f'_{s,a}(w^\pi(s,a))$$

or in matrix form, letting $f'(w^\pi)$ be shorthand for $[f'_{s,a}(w^\pi(s,a))]_{s,a} \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$,

$$\eta \circ f'(w^\pi) = d^D \circ (I - \gamma P^\pi) q_f^*$$

Then q_f^* must satisfy

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

C.2 Proof of Theorem 6

The proof is of a similar nature as the proof of Theorem 2 (Appendix B.1). From Assumption 4 and Lemma 7, we know that that $L_f^w(w, q_f^*)$ is an M-strongly convex function in w on the $|| \cdot ||_{2,\eta}$ norm. Since (w^π, q_f^*) is the saddle point solution of L_f^w, from strong convexity we know that the error of $\widehat{w}$ is bounded as

$$\begin{aligned} ||\widehat{w} - w^\pi||_{2,d^D} &\leq \sqrt{\frac{2\left(L_f^w(w^\pi, q_f^*) - L_f^w(\widehat{w}, q_f^*)\right)}{M^w}} \\ &\leq \sqrt{\frac{4\epsilon_{stat}^w}{M^w}} \qquad \text{(Lemma 11)}, \end{aligned}$$

where ϵ_{stat}^w is given in Lemma 10.

Remark 5. In Theorem 6 of the main text, there is an additional $O(C_f^w/\sqrt{n})$ term in the statistical error ϵ_{stat}^w, which would arise if the regularization function $\mathbb{E}_\eta[f_{s,a}(w(s,a))]$ were to be estimated from samples. However, we state early on in the paper that we assume the regularizer can be calculated exactly, as sampling is a trivial extension. Correspondingly, the correct expression for the statistical error is:

$$\epsilon_{stat}^w = (1+\gamma) C_{\mathcal{W}}^w C_{\mathcal{Q}}^w \sqrt{2 \log \tfrac{4|\mathcal{Q}||\mathcal{W}|}{\delta} / n} + (1-\gamma) C_{\mathcal{Q}}^w \sqrt{2 \log \tfrac{4|\mathcal{Q}|}{\delta} / n_0},$$

and, to remain consistent with the rest of the paper, we provide the proof and lemma for this ϵ_{stat}^w below.

Lemma 10. *Suppose Assumption 6 holds. Then for all $(w, q) \in \mathcal{W} \times \mathcal{Q}$, w.p. $\geq 1 - \delta$,*

$$|\widehat{L}_f^w(w, q) - L_f^w(w, q)| \leq \epsilon_{stat}^w,$$

where $\epsilon_{stat}^w = (1+\gamma) C_{\mathcal{W}}^w C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + (1-\gamma) C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{Q}|}{\delta}}{n_0}}$.

Proof. Let $l_i = w(s_i, a_i)(\gamma q(s_i', \pi) - q(s_i, a_i))$. Using Assumption 6,

$$\begin{aligned} |l_i| &\leq (1+\gamma)||w||_\infty ||q||_\infty \\ &\leq (1+\gamma) C_{\mathcal{W}}^w C_{\mathcal{Q}}^w \end{aligned}$$

Then using Hoeffding's inequality with union bound, w.p. $\geq 1 - \delta/2$ we have that for all $w, q \in \mathcal{W} \times \mathcal{Q}$,

$$\left| \frac{1}{n} \sum_{i=1}^n l_i - \mathbb{E}_{d^D}[l_i] \right| \leq (1+\gamma) C_{\mathcal{W}}^w C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{W}||\mathcal{Q}|}{\delta}}{n}}$$

Similarly, for all $q \in \mathcal{Q}$, w.p. $\geq 1 - \delta/2$,

$$\left| \frac{1}{n_0} \sum_{i=1}^{n_0} q(s_{0,i}, \pi) - \mathbb{E}_{\mu_0}[q(s_{0,i}, \pi)] \right| \leq C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{Q}|}{\delta}}{n_0}}$$

Since $L_f^w(w, q) = \mathbb{E}_\eta[f_{s,a}(w(s,a))] + \mathbb{E}_{d^D}[l_i] + \mathbb{E}_{\mu_0}[q(s_0, \pi)]$, but the first term can be calculated exactly, taking a union bound over the above two inequalities, we have that w.p. $\geq 1 - \delta$,

$$|\widehat{L}_f^w(w, q) - L_f^w(w, q)| \leq (1+\gamma) C_{\mathcal{W}}^w C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + (1-\gamma) C_{\mathcal{Q}}^w \sqrt{\frac{2 \log \frac{4|\mathcal{Q}|}{\delta}}{n_0}}$$

□

Lemma 11. *Under Assumptions 4, 5, 6, w.p. $\geq 1 - \delta$,*

$$L_f^w(w_f^*, q_f^*) - L_f^w(\widehat{w}, q_f^*) \leq 2\epsilon_{stat}^w$$

Proof of Lemma 11 Letting $\widehat{q}(w) = \arg\max_{q \in \mathcal{Q}} \widehat{L}_f^w(w, q)$, we decompose the error as follows:

$$\begin{aligned} L_f^w(\widehat{w}, q_f^*) - L_f^w(w^\pi, q_f^*) = {} & L_f^w(\widehat{w}, q_f^*) - \widehat{L}_f(\widehat{w}, q_f^*) && (1) \ \geq -\epsilon_{stat}^w \\ & + \widehat{L}_f^w(\widehat{w}, q_f^*) - \widehat{L}_f^w(\widehat{w}, \widehat{q}(\widehat{w})) && (2) \ \geq 0 \\ & + \widehat{L}_f^w(\widehat{w}, \widehat{q}(\widehat{w})) - \widehat{L}_f^w(w^\pi, \widehat{q}(w^\pi)) && (3) \ \geq 0 \\ & + \widehat{L}_f^w(w^\pi, \widehat{q}(w^\pi)) - L_f^w(w^\pi, \widehat{q}(w^\pi) && (4) \ \geq -\epsilon_{stat}^w \\ & + L_f^w(w^\pi, \widehat{q}(w^\pi)) - L_f^w(w^\pi, q_f^*) && (5) \ \geq 0 \end{aligned}$$

Combining the inequalities gives the result. We give a brief justification for each term below. Terms (1) and (4) follow from Lemma 10.

Term (2) ≥ 0, since $q_f^* \in \mathcal{Q}$.

Term (3) ≥ 0 since $w^\pi \in \mathcal{W}$ and $\widehat{w} = \arg\max_{w \in \mathcal{W}} \widehat{L}_f^w(w, \widehat{q}(w))$.

Term (5) ≥ 0 since (w^π, q_f^*) is a saddle point solution.

D Additional Details of the Experiments

D.1 Derivation

We now derive the system of equations for our value function estimation experiments in Section 6. Letting the regularization function be $f_{s,a}(x) = \frac{1}{2}x^2$ for all (s, a), the objective is

$$\min_q \max_w L_f^q(q, w) = \frac{1}{2}\mathbb{E}_\nu[q^2(s,a)] + \mathbb{E}_{d^D}\left[w(s,a)\left(r(s,a) + \gamma q(s', \pi) - q(s,a)\right)\right], \tag{10}$$

Letting $\mathbb{E}_n$ denote the empirical average over $\mathcal{D}$ for clarity, with empirical samples and the linear classes $\mathcal{Q}, \mathcal{W}$, the objective becomes:

$$\min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} \widehat{L}_f^q(q, w) = \frac{1}{2}\mathbb{E}_\nu[\alpha^\top \phi(s,a)\phi(s,a)^\top \alpha] + \beta^\top \Big(\mathbb{E}_n\left[\phi(s,a)r(s,a)\right] + \mathbb{E}_n\left[\gamma\phi(s,a)\phi(s',\pi)^\top - \phi(s,a)\phi(s,a)^\top\right]\alpha\Big)$$

Since $\beta \in \mathbb{R}^d$, $\max_{w\in\mathcal{W}} \widehat{L}_f^q(q, w) = +\infty$ for any q, unless α sets the the second term to 0. This is satisfied by α such that

$$\mathbb{E}_n\left[\phi(s,a)\phi(s,a)^\top - \gamma\phi(s,a)\phi(s',\pi)^\top\right]\alpha = \mathbb{E}_n\left[\phi(s,a)r(s,a)\right].$$

However, there may in general be infinite feasible α depending on the linear features and samples. For our specific linear parameterization of $\mathcal{Q}, \mathcal{W}$, the constraints form an underdetermined $d \times k$ system of equations, which has infinite solutions.

This is where the regularization term $\mathbb{E}_\nu[\alpha^\top \phi(s,a)\phi(s,a)^\top \alpha]$ comes into play. For any regularizing distribution ν, our method will output a solution that minimizes this term, i.e. that minimizes the norm of $q = \Phi^\top \alpha$ on ν. If $\nu = 0$, for example, the algorithm will output any feasible point; if $\nu = 1/|\mathcal{SA}|$, the algorithm will output q with smallest L2 norm.

Connection to LSTDQ When using the same linear class for $\mathcal{W}$ and $\mathcal{Q}$, the solution to the constraints in Eq.(3) (i.e., ignoring the regularization objective)—if the solution is unique given matrix invertibility—coincides with LSTDQ [UHJ20]. As mentioned in Section 2, LSTDQ enjoys function-estimation guarantees under matrix invertibility. In fact, we believe it is possible to extend the analysis even when $\mathcal{Q}$ and $\mathcal{W}$ use different features of dimensions d and k, respectively; as long as $k \geq d$ and the matrix in Eq.(3) has full row-rank[9] (i.e., *overdetermined*), similar guarantees for LSTDQ should still hold, though we are not aware of an explicit documentation of this fact. In contrast, our setup is more challenging as we are in the regime of $k < d$, and the constraints in Eq.(3) is *underdetermined*, nullifying the guarantees of LSTDQ. In such cases, the use of regularization is important for guaranteeing function estimation, as also shown in our experiments.

D.2 Experimental Setup

Feature Design In total, the tabular environment has 400 state-action values, and we design Φ to aggregate states that correspond to unique entries (within 3 decimal places) of q^π. In Figure 1, $\widetilde{\Phi}$ is composed of the set of features given by

$$\{(I - \gamma\widetilde{P}^\pi)^{-1}(\nu \circ q^\pi)/d^D, (I - \gamma\widetilde{P}^\pi)^{-1}(\nu \circ q^\pi)\}_{\nu\in\mathcal{V}}.$$

The first of these two entries is the closed-form solution of w_f^* given in Lemma 3, and satisfies the realizability requirements of all methods; the second is included for optimization stability.

In Figure 2, we use a model with constant value equal to the average value of q^π on the support of p, i.e. $\overline{q} = 1/|\mathcal{SA}| \sum_{s,a} q^\pi(s,a) \cdot \mathbb{1}_{\{p>0\}}$. To maintain realizability when the model is included in the regularization function, $\widetilde{\Phi}$ is composed of the set

$$\Big\{(I - \gamma\widetilde{P}^\pi)^{-1}(\nu \circ q^\pi),\ (I - \gamma\widetilde{P}^\pi)^{-1}(\nu \circ q^\pi \circ \mathbb{I}(\widetilde{q} > 0)),\ (I - \gamma\widetilde{P}^\pi)^{-1}(\nu \circ \mathbb{I}(\widetilde{q} > 0))\Big\}_{\nu\in\mathcal{V}}$$

[9] In the finite-sample regime, one needs to lower-bound the smallest singular value of such matrices instead of imposing full-rankness [PKBK22].

The reason why this preserves realizability is as follows. When ν is the regularization distribution, and the input model is $\widetilde{q} = (mq^\pi + (1-m)\overline{q}) \circ \mathbb{1}(p > 0))$ for some constant $\overline{q}$, the closed-form solution w_f^* can be expanded as

$$\begin{aligned} w_f^* &= (I - \gamma \widetilde{P}^\pi)^{-1}(\nu \circ (q^\pi - \widetilde{q})) \\ &= (I - \gamma \widetilde{P}^\pi)^{-1}(\nu \circ q^\pi) - m \cdot (I - \gamma \widetilde{P}^\pi)^{-1}(\nu \circ \mathbb{1}(p > 0) \circ q^\pi) \\ &\quad - (1-m)\overline{q} \cdot (I - \gamma \widetilde{P}^\pi)^{-1}(\nu \circ \mathbb{1}(p > 0)), \end{aligned}$$

which implies w_f^* can be expressed as a linear combination of the three previously defined features.

Solver We solve the linear system using CVXPY with optimizer SCS [DB16; AVDB18].

Environment The Gridwalk is a 10x10 environment with 4 actions corresponding to cardinal directions. The objective is to reach the goal state (lower right corner). In each state, the agent receives a reward inversely proportional to its distance from a goal state. Each trajectory terminates after 100 steps. The initial states are randomly distributed over the upper half of the grid.

The target policy is defined to be a deterministic optimal policy that always moves towards the goal by first going right, and then down. To create a strong shift, the behavioral policy is designed to largely explore only the bottom left portion of the grid, providing poor coverage over the target policy and starting states. Specifically, letting the following probabilities refer to distributions over actions [RIGHT, DOWN, LEFT, UP], the target policy π has distribution $[1, 0, 0, 0]$ over actions until it hits the right wall, then $[0, 1, 0, 0]$. The behavior policy takes $[0.1, 0.4, 0.5, 0]$ until it hits the right wall, then takes $[0, 0.5, 0.5, 0]$.

E Approximation and Optimization Error

The main results of this paper (Theorems 2, 6) utilize assumptions on realizability (Assumption 2, 5), as well as (implicit) assumptions of perfect optimization. In this section, we analyze how approximation errors, i.e. when the saddle point solution is not contained in $\mathcal{Q} \times \mathcal{W}$, and optimization errors affect our error bounds. Due to the similarity in proofs between value function and weight learning, we provide them only for value function learning; analogous methods can be used to derive similar results for weight learning.

E.1 Finite-sample Guarantees

First, we relax the realizability requirements of Assumption 2. Define the approximation errors:

$$\begin{aligned} \epsilon_{approx,q} &= \min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a))] + \mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))]| \\ \epsilon_{approx,w} &= \min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} |\mathbb{E}_{d^D}[(w(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]| \\ \epsilon_{approx} &:= \epsilon_{approx,q} + \epsilon_{approx,w}. \end{aligned}$$

$\epsilon_{approx,q}$ is composed of the worst-case weighted combination of Bellman errors of the best candidate $q \in \mathcal{Q}$, as well as the difference between the regularization function at q and q^π. The error $\epsilon_{approx,w}$ measures the distance between the best candidate $w \in \mathcal{W}$ and the saddle point solution w_f^* by projecting the difference onto the worst-case Bellman error $\mathcal{T}^\pi q - q$.

Remark 6. To increase intuition of $\epsilon_{approx,q}$, we can relax the difference in regularization terms as $\mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))] \leq C_{f'}^q ||q^\pi - q||_{2,\nu}$, which is also the norm upon which the $\widehat{q}$ estimation guarantee is given (Theorem 2) . Reflecting the nature of the value function estimation task, this states that, even if there is a candidate $q \in \mathcal{Q}$ with low Bellman error (e.g. if data is sparse), $\epsilon_{approx,q}$ will still be large if q is far from q^π on the desired distribution ν.

Next, we can also relax the (implicit) assumptions that we obtain the true optima of (5). Let $(\widehat{q}, \widehat{w})$ be the approximate solutions of (5) found by the algorithm. As before, define $\widehat{w}(q) := \arg\max_{w \in \mathcal{W}} \widehat{L}(w, q)$ to be the true empirical maximizer for any $q \in \mathcal{Q}$. Note that since we allow for optimization error, it is not necessarily the case that $\widehat{q} = \arg\min_{q \in \mathcal{Q}} \widehat{L}_f^q(q, \widehat{w}(q))$ and

$\widehat{w} = \widehat{w}(\widehat{q}) = \arg\max_{w \in \mathcal{W}} \widehat{L}_f^q(\widehat{q}, w)$. Correspondingly, define the following optimization errors:

$$\begin{aligned}
\epsilon_{opt,w} &\geq \widehat{L}_f^q(\widehat{q}, \widehat{w}(\widehat{q})) - \widehat{L}_f^q(\widehat{q}, \widehat{w}) \\
\epsilon_{opt,q} &\geq \widehat{L}_f^q(\widehat{q}, \widehat{w}(\widehat{q})) - \min_{q \in \mathcal{Q}} \widehat{L}_f^q(q, \widehat{w}(q)) \\
\epsilon_{opt} &:= \epsilon_{opt,q} + \epsilon_{opt,w}.
\end{aligned}$$

$\epsilon_{opt,w}$ states that the estimate $\widehat{w}$ should not be too far from the best discriminator in $\mathcal{W}$ for $\widehat{q}$, while $\epsilon_{opt,q}$ states that the estimate $\widehat{q}$ should not be too far from the minimax solution.

Using the above definitions, we provide the following generalization of Theorem 2, which accounts for approximation and optimization errors.

Theorem 12. *Under Assumptions 1 and 3, with probability at least* $1 - \delta$*,*

$$||\widehat{q} - q^\pi||_{2,\nu} \leq \sqrt{\frac{4\epsilon_{stat}^q + 2\epsilon_{approx} + 2\epsilon_{opt}}{M^q}},$$

where ϵ_{stat}^q *is given in Theorem 2.*

E.2 Proof of Theorem 12

The proof takes the same overall steps as the proof of Theorem 2 (Appendix B.1), but relies on Lemma 13 to incorporate the approximation and optimization errors:

$$\begin{aligned}
||\widehat{q} - q^\pi||_{2,\nu} &\leq \sqrt{\frac{2\left(L_f^q(\widehat{q}, w_f^*) - L_f^q(q^\pi, w_f^*)\right)}{M^q}} \\
&\leq \sqrt{\frac{4\epsilon_{stat}^q + 2\epsilon_{approx,q} + 2\epsilon_{approx,w} + 2\epsilon_{opt,q} + 2\epsilon_{opt,w}}{M^q}}. \qquad \text{(Lemma 13)}
\end{aligned}$$

Below, we state and prove the helper lemma, which bounds the difference between the Lagrangian objective (4) at the saddle point (q^π, w_f^*) and the point $(\widehat{q}, w_f^*)$:

Lemma 13. *Under Assumptions 1 and 3, w.p.* $\geq 1 - \delta$*,*

$$L_f^q(\widehat{q}, w_f^*) - L_f^q(q^\pi, w_f^*) \leq 2\epsilon_{stat}^q + \epsilon_{approx,q} + \epsilon_{approx,w} + \epsilon_{opt,q} + \epsilon_{opt,w}.$$

Proof. With some abuse of notation (as $\widetilde{q}, \widetilde{w}$ previously referred to models used with the regularizer), for brevity in this section, let $\widetilde{q}$ be the minimizer of $\epsilon_{approx,q}$ and $\widetilde{w}$ be the minimizer of $\epsilon_{approx,w}$. That is,

$$\begin{aligned}
\widetilde{q} &= \arg\min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a))] + \mathbb{E}_\nu[f(q(s,a)) - f(q^\pi(s,a))]| \\
\widetilde{w} &= \arg\min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} |\mathbb{E}_{d^D}[(w(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]|.
\end{aligned}$$

Decompose the error as follows:

$$\begin{aligned}
L_f^q(q^\pi, w_f^*) - L_f^q(\widehat{q}, w_f^*) = {} & L_f^q(q^\pi, w_f^*) - L_f^q(q^\pi, \widehat{w}(\widetilde{q})) && (1)\ \geq 0 \\
& + L_f^q(q^\pi, \widehat{w}(\widetilde{q})) - L_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) && (2)\ \geq -\epsilon_{approx,q} \\
& + L_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) - \widehat{L}_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) && (3)\ \geq -\epsilon_{stat} \\
& + \widehat{L}_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) - \widehat{L}_f^q(\widehat{q}, \widehat{w}) && (4)\ \geq -\epsilon_{opt,q} \\
& + \widehat{L}_f^q(\widehat{q}, \widehat{w}) - \widehat{L}_f^q(\widehat{q}, \widetilde{w}) && (5)\ \geq -\epsilon_{opt,w} \\
& + \widehat{L}_f^q(\widehat{q}, \widetilde{w}) - L_f^q(\widehat{q}, \widetilde{w}) && (6)\ \geq -\epsilon_{stat} \\
& + L_f^q(\widehat{q}, \widetilde{w}) - L_f^q(\widehat{q}, w_f^*) && (7)\ \geq -\epsilon_{approx,w}
\end{aligned}$$

First, (1) holds because (q^π, w_f^*) is the saddle point solution of L_f^q over all $q, w \in \mathbb{R} \times \mathbb{R}$. The statistical errors in (3) and (6) follow from Lemma 8.

Next, we justify the optimization errors. For (4),

$$\widehat{L}_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) - \widehat{L}_f^q(\widehat{q}, \widehat{w}) \geq \widehat{L}_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) - \widehat{L}_f^q(\widehat{q}, \widehat{w}(\widehat{q})) \geq \min_{q \in \mathcal{Q}} \widehat{L}_f^q(q, \widehat{w}(q)) - \widehat{L}_f^q(\widehat{q}, \widehat{w}(\widehat{q})) \geq -\epsilon_{opt,q}.$$

For (5),

$$\widehat{L}_f^q(\widehat{q}, \widehat{w}) - \widehat{L}_f^q(\widehat{q}, \widetilde{w}) \geq \widehat{L}_f^q(\widehat{q}, \widehat{w}) - \max_{w \in \mathcal{W}} \widehat{L}_f^q(\widehat{q}, w) \geq -\epsilon_{opt,w}$$

Finally, we justify the approximation errors, starting with (2). Note that for any $q, w \in \mathcal{Q} \times \mathcal{W}$,

$$\begin{aligned}
&|L_f^q(q^\pi, w) - L_f^q(q, w)| \\
&\quad = |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a) - \mathcal{T}^\pi q^\pi(s,a) + q^\pi(s,a))] \\
&\qquad + \mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))]| \\
&\quad = |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a))] + \mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))]| \\
&\quad \leq \max_{w \in \mathcal{W}} |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a))] + \mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))]|.
\end{aligned}$$

Then since $\widetilde{q}$ was chosen to minimize the above expression,

$$\begin{aligned}
&L_f^q(q^\pi, \widehat{w}(\widetilde{q})) - L_f^q(\widetilde{q}, \widehat{w}(\widetilde{q})) \\
&\quad \geq -\max_{w \in \mathcal{W}} |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi \widetilde{q}(s,a) - \widetilde{q}(s,a))] + \mathbb{E}_\nu[f_{s,a}(\widetilde{q}(s,a)) - f_{s,a}(q^\pi(s,a))]| \\
&\quad = -\min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_{d^D}[w(s,a)(\mathcal{T}^\pi q(s,a) - q(s,a))] + \mathbb{E}_\nu[f_{s,a}(q(s,a)) - f_{s,a}(q^\pi(s,a))]| \\
&\quad = -\epsilon_{approx,q}.
\end{aligned}$$

Next we justify (8). For any $w \in \mathcal{W}$ and $q \in \mathcal{Q}$,

$$\begin{aligned}
|L_f^q(q, w) - L_f^q(q, w_f^*)| &= |\mathbb{E}_{d^D}[(w(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]| \\
&\leq \max_{q \in \mathcal{Q}} |\mathbb{E}_{d^D}[(w(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]|.
\end{aligned}$$

Then since $\widetilde{w}$ was chosen to minimize the RHS of the above inequality,

$$\begin{aligned}
L_f^q(\widehat{q}, \widetilde{w}) - L_f^q(\widehat{q}, w_f^*) &\geq -\max_{q \in \mathcal{Q}} |\mathbb{E}_{d^D}[(\widetilde{w}(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]| \\
&= -\min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} |\mathbb{E}_{d^D}[(w(s,a) - w_f^*(s,a))(\mathcal{T}^\pi q(s,a) - q(s,a))]| \\
&= -\epsilon_{approx,w}.
\end{aligned}$$

Combining these inequalities gives the lemma statement. □

F Off-Policy Return Estimation

Section 4 demonstrates how q-value estimates $\widehat{q}$ can be obtained, and Section 5 demonstrates how weight estimates $\widehat{w}$ can be obtained. The estimates $\widehat{q}$ and/or $\widehat{w}$ can additionally be used for downstream off-policy evaluation (OPE) of the policy's value $J(\pi)$, which can be equivalently defined in the following three ways:

$$J(\pi) = (1-\gamma)\mathbb{E}_{s_0 \sim \mu_0}[q^\pi(s_0, \pi)] \qquad \text{("value function-based")}$$

$$J(\pi) = \mathbb{E}_{(s,a) \sim d^D, r \sim R(\cdot|s,a)}[w^\pi(s,a) \cdot r] \qquad \text{("weight-based")}$$

$$\begin{aligned}
J(\pi) &= (1-\gamma)\mathbb{E}_{s_0 \sim \mu_0}[q^\pi(s_0, \pi)] \qquad \text{("doubly robust")} \\
&\quad + \mathbb{E}_{(s,a) \sim d^D, r \sim R(\cdot|s,a), s' \sim P(\cdot|s,a)}[w^\pi(s,a)(r + q^\pi(s', \pi) - q^\pi(s,a))]
\end{aligned}$$

With finite samples and estimates $\widehat{q}$ and $\widehat{w}$ approximating q^π and w^π, respectively, their corresponding off-policy estimators are:

$$\widehat{J}^q(\pi) = (1-\gamma)\frac{1}{n_0}\sum_{i=1}^{n_0} \widehat{q}(s_{0,i}, \pi)$$

$$\widehat{J}^w(\pi) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(s_i, a_i) r_i$$

$$\widehat{J}^{dr}(\pi) = (1-\gamma)\frac{1}{n_0}\sum_{j=1}^{n_0} \widehat{q}(s_{0,j}, \pi) + \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(s_i, a_i)\left(r_i + \widehat{q}(s_i', \pi) - \widehat{q}(s_i, a_i)\right)$$

While the OPE estimator $\widehat{J}^{dr}(\pi)$ utilizes both the weights and value functions, $\widehat{J}^{w}(\pi)$ and $\widehat{J}^{q}(\pi)$ utilize only one or the other. As a result, when $\widehat{q}$ and $\widehat{w}$ are estimated as in Sections 4 and 5, respectively, $\widehat{J}^{w}(\pi)$ and $\widehat{J}^{q}(\pi)$ both inherit their $O(n^{-1/4})$ sample complexities:

Corollary 14. *Suppose Assumptions 1, 2, and 3 hold, and let* $\widehat{q} = \arg\min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} \widehat{L}^{q}_{f}(q,w)$. *Then with probability* $\geq 1-2\delta$,

$$|\widehat{J}^{q}(\pi) - J(\pi)| \leq \epsilon^{q}_{eval} + \sqrt{\mathcal{C}_{\mu^{\pi}_{0}/\nu}} \cdot \epsilon^{q}_{est},$$

where $\epsilon^{q}_{eval} = (1-\gamma)C^{q}_{\mathcal{Q}}\sqrt{2\log\frac{2|\mathcal{Q}|}{\delta}/n_0}$, $\mathcal{C}_{\mu^{\pi}_{0}/\nu} = ||\mu^{\pi}_{0}/\nu||_{\infty}$, *and* ϵ^{q}_{stat} *is as in Theorem 2.*

Corollary 15. *Suppose Assumption 4, 5, and 6 hold, and let* $\widehat{w} = \arg\min_{w\in\mathcal{W}} \max_{q\in\mathcal{Q}} \widehat{L}^{w}_{f}(q,w)$. *Then with probability* $\geq 1-2\delta$,

$$|\widehat{J}^{w}(\pi) - J(\pi)| \leq \epsilon^{w}_{eval} + \sqrt{\mathcal{C}_{d^{D}/\eta}} \cdot \epsilon^{w}_{est},$$

where $\epsilon^{w}_{eval} = C^{w}_{\mathcal{W}}\sqrt{\frac{2\log\frac{2|\mathcal{W}|}{\delta}}{n}}$, $\mathcal{C}_{d^{D}/\eta} = \|{}^{d^{D}}/\eta\|_{\infty}$, *and* ϵ^{w}_{est} *is as in Theorem 6.*

However, when $\widehat{q}$ and $\widehat{w}$ are used together in the doubly robust estimator $\widehat{J}^{dr}$, their estimation error becomes multiplicative, and $\widehat{J}^{dr}(\pi)$ can achieve the $O(n^{-\frac{1}{2}})$ fast rate of convergence. In Theorem 16 below, we present two versions this guarantee. The first requires no additional assumptions beyond $d^{D} > 0$, which we already make (see footnote 5), but involves the largest singular value of $I - \gamma P^{\pi}$, which may be difficult to characterize. The second utilizes an additional assumption, and replaces the singular value with an occupancy ratio, stated below. The assumption requires that all next states s' are also present as states s in transitions of d^{D} (a condition which may reasonably hold in practice), and is also made by [UIJKSX21].

Assumption 7 (Next State Coverage)**.** Let $d^{D}(s) = \sum_{a} d^{D}(s,a)$ be the marginal distribution of states s in d^{D}, and $d^{D}_{s'}(s) := \sum_{s',a'} P(s|s',a')d^{D}(s',a')$ be the marginal distribution of next states s'. Suppose

$$\mathcal{C}_{s'/s} := ||d^{D}_{s'}(\cdot)/d^{D}(\cdot)||_{\infty} < \infty$$

Theorem 16. *Suppose Assumption 1, 2, 3, 4, 5, and 6 hold. Let* $\widehat{w}$ *and* $\widehat{q}$ *be estimated from:*

$$\widehat{q} = \arg\min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} \widehat{L}^{q}_{f}(q,w)$$
$$\widehat{w} = \arg\min_{w\in\mathcal{W}} \max_{q\in\mathcal{Q}} \widehat{L}^{w}_{f}(q,w).$$

Then with probability $\geq 1-3\delta$,

$$|\widehat{J}^{dr}(\pi) - J(\pi)| \leq \epsilon^{dr}_{eval} + \sigma_{max}(I-\gamma P^{\pi}) \cdot \sqrt{\mathcal{C}_{d^{D}/\eta}\mathcal{C}_{d^{D}/\nu}} \cdot \epsilon^{w}_{est} \cdot \epsilon^{q}_{est},$$

If Assumption 7 additionally holds, with probability $\geq 1-3\delta$,

$$|\widehat{J}^{dr}(\pi) - J(\pi)| \leq \epsilon^{dr}_{eval} + \left(1+\gamma\sqrt{\mathcal{C}_{s'/s}\mathcal{C}_{\pi/\pi^{D}}}\right) \cdot \sqrt{\mathcal{C}_{d^{D}/\eta}\mathcal{C}_{d^{D}/\nu}} \cdot \epsilon^{w}_{est} \cdot \epsilon^{q}_{est},$$

where $\epsilon^{dr}_{eval} = (1-\gamma)C^{q}_{\mathcal{Q}}\sqrt{2\log\frac{2|\mathcal{Q}|}{\delta}/n_0} + C^{w}_{\mathcal{W}}(1+(1+\gamma)C^{q}_{\mathcal{Q}})\sqrt{2\log\frac{2|\mathcal{W}||\mathcal{Q}|}{\delta}/n}$, σ_{max} *denotes the largest singular value, and* ϵ^{q}_{est} *and* ϵ^{w}_{est} *are as in Theorems 2 and 6.*

As the evaluation error ϵ^{dr}_{eval} in Theorem 16 is $O(n^{-1/2})$, the sample complexity of doubly robust estimation is rate-limited by $\epsilon^{w}_{est} \cdot \epsilon^{q}_{est}$, the product of weight and value function estimation errors. If both functions can be estimated at an $O(n^{-1/4})$ rate, as is true of our method, then $\widehat{J}^{dr}(\pi)$ attains the overall $O(n^{-1/2})$ fast rate. Finally, while Theorem 16 assumes for simplicity that the same $\mathcal{Q}, \mathcal{W}$ classes are used in both of its optimization problems, it can easily be extended to the case where different pairs of function classes are used as long as the required assumptions hold.

Remark 7 (Comparison to Related Work)**.** [YNDLS20] conduct experiments comparing off-policy evaluation using $\widehat{J}^q(\pi), \widehat{J}^w(\pi), \widehat{J}^{dr}(\pi)$, and generally observe that $\widehat{J}^{dr}(\pi)$ has higher variance and worse performance than either $\widehat{J}^q(\pi)$ or $\widehat{J}^w(\pi)$. Though at first glance this may appear to contradict Theorem 16, that is actually not the case; in fact, our theoretical analysis provides insight into why [YNDLS20] may observe such a phenomenon. In contrast to Theorem 16, when using $\widehat{J}^{dr}(\pi)$ [YNDLS20] utilize saddle point predictions $(\widehat{q}, \widehat{w})$ from either *only* value function learning or *only* weight learning, e.g. $(\widehat{q}, \widehat{w}) = \arg\min_{q \in \mathcal{Q}} \arg\max_{w \in \mathcal{W}} \widehat{L}^q_f(q, w)$ that approximates (q^π, w^*_f). Continuing with this example (and the same applies to weight learning), it is clear from our analysis that $\widehat{w}$ estimated in such a manner may not approximate w^π at all, leading to increased estimation error of $\widehat{J}^{dr}(\pi)$ over $\widehat{J}^q(\pi)$. First, the closed-form solution we have derived for w^*_f in (Lemma 3) shows that w^*_f may have a significantly different magnitude from w^π. Second, even if ν and f were chosen such that $w^*_f \approx w^\pi$, as per the reasons stated in Section 4.1, we are not even guaranteed to output $\widehat{w}$ close to w^*_f since L^q_f is not regularized in w. In order to obtain the estimation benefits of doubly robust estimation, our analysis shows that $\widehat{q}$ and $\widehat{w}$ should be separately estimated from their respective optimization problems, then combined in $\widehat{J}^{dr}(\pi)$. This is in accordance with similar results from [KU20] and [UIJKSX21].

F.1 Proof of Corollary 14

Let $\widetilde{J}(\pi) = (1-\gamma)\mathbb{E}_{\mu_0}[\widehat{q}(s, \pi)]$. We decompose the error as

$$|\widehat{J}(\pi) - J(\pi)| \leq |\widehat{J}(\pi) - \widetilde{J}(\pi)| + |\widetilde{J}(\pi) - J(\pi)|$$

First we bound $|\widehat{J}(\pi) - \widetilde{J}(\pi)|$. Using Hoeffding's with union bound, for all $q \in \mathcal{Q}$, w.p. $\geq 1 - \delta$,

$$\left| \frac{1}{n_0} \sum_{i=1}^{n} q(s_{0,i}, \pi) - \mathbb{E}_{\mu_0}[q(s, \pi)] \right| \leq (1-\gamma) C^q_{\mathcal{Q}} \sqrt{\frac{2 \log \frac{2|\mathcal{Q}|}{\delta}}{n_0}} := \epsilon^q_{eval},$$

which implies $|\widehat{J}(\pi) - \widetilde{J}(\pi)| \leq \epsilon^q_{eval}$. For the second term, let $C_{\mu^\pi_0/\nu} = ||\mu^\pi_0/\nu||_\infty$. Then w.p. $\geq 1 - \delta$

$$\begin{aligned}
|\widetilde{J}(\pi) - J(\pi)| &= (1-\gamma)|\langle \mu^\pi_0, \widehat{q} - q^\pi \rangle| \\
&\leq (1-\gamma)||\widehat{q} - q^\pi||_{1,\mu^\pi_0} \\
&\leq (1-\gamma)||\widehat{q} - q^\pi||_{2,\mu^\pi_0} \\
&= (1-\gamma)\sqrt{\mathcal{C}_{\mu^\pi_0/\nu}}||\widehat{q} - q^\pi||_{2,\nu} \\
&\leq (1-\gamma)\sqrt{\mathcal{C}_{\mu^\pi_0/\nu}}\epsilon^q_{est}
\end{aligned}$$

using Theorem 2 in the last line.

F.2 Proof of Corollary 15

Let $\widetilde{J}(\pi) = \mathbb{E}_{d^D}[\widehat{w}(s, a) r(s, a)]$. We decompose the error as

$$|\widehat{J}^w(\pi) - J(\pi)| \leq |\widehat{J}^w(\pi) - \widetilde{J}(\pi)| + |\widetilde{J}(\pi) - J(\pi)|$$

For the first term, using Hoeffding's with union bound, w.p. $\geq 1 - \delta$, for all $w \in \mathcal{W}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} w(s_i, a_i) r_i - \mathbb{E}_{d^D}[w(s, a) r(s, a)] \right| \leq C^w_{\mathcal{W}} \sqrt{\frac{2 \log \frac{2|\mathcal{W}|}{\delta}}{n}} := \epsilon^w_{eval}$$

which implies $|\widehat{J}(\pi) - \widetilde{J}(\pi)| \leq \epsilon^w_{eval}$. For the second term,

$$\begin{aligned}
|\widehat{J}(\pi) - J(\pi)| &= |\langle \widehat{w} \cdot d^D, r\rangle - \langle w^\pi \cdot d^D, r\rangle| \\
&\leq ||d^D \cdot (\widehat{w} - w^\pi)||_1 ||r||_\infty \\
&\leq ||d^D \cdot (\widehat{w} - w^\pi)||_1 = ||\widehat{w} - w^\pi||_{d^D,1} \\
&\leq ||\widehat{w} - w^\pi||_{d^D,2} \\
&\leq \sqrt{\mathcal{C}_{d^D/\eta}} ||\widehat{w} - w^\pi||_{2,\eta} \\
&\leq \sqrt{\mathcal{C}_{d^D/\eta}} \epsilon^w_{est}
\end{aligned}$$

w.p. $\geq 1 - \delta$, using Theorem 6 in the last line. Taking a union bound over both terms gives the stated result.

F.3 Proof of Theorem 16

Let $\widetilde{J}(\pi) = (1-\gamma)\mathbb{E}_{\mu_0^\pi}[\widehat{q}(s,a)] + \mathbb{E}_{d^D}[\widehat{w}(s,a)(r + \widehat{q}(s',\pi) - \widehat{q}(s,a))]$. Again we decompose the error as:

$$|\widehat{J}^{dr}(\pi) - J(\pi)| \leq |\widehat{J}^{dr}(\pi) - \widetilde{J}(\pi)| + |\widetilde{J}(\pi) - J(\pi)|.$$

For the first term, since $\mathbb{E}[\widehat{J}^{dr}(\pi)] = \widetilde{J}(\pi)$, w.p. $\geq 1 - \delta$ we have that $\forall\, q, w \in \mathcal{Q} \times \mathcal{W}$,

$$|\widehat{J}^{dr}(\pi) - \widetilde{J}(\pi)| \leq (1-\gamma) C^q_{\mathcal{Q}} \sqrt{\frac{2 \log \frac{2|\mathcal{Q}|}{\delta}}{n_0}} + C^w_{\mathcal{W}}(1 + (1+\gamma)C^q_{\mathcal{Q}}) \sqrt{\frac{2 \log \frac{2|\mathcal{W}||\mathcal{Q}|}{\delta}}{n}} := \epsilon^{dr}_{eval}$$

For the second term,

$$\begin{aligned}
|\widetilde{J}(\pi) - J(\pi)| &= |(1-\gamma)\langle \widehat{q}, \mu_0^\pi\rangle + \langle \widehat{w} \cdot d^D, r + \gamma P^\pi \widehat{q} - \widehat{q}\rangle - (1-\gamma)\langle q^\pi, \mu_0^\pi\rangle| \\
&= |(1-\gamma)\langle \widehat{q}, \mu_0^\pi\rangle + \langle \widehat{w} \cdot d^D, r + \gamma P^\pi \widehat{q} - \widehat{q}\rangle - (1-\gamma)\langle q^\pi, \mu_0^\pi\rangle - \langle \widehat{w} \cdot d^D, r + \gamma P^\pi q^\pi - q^\pi\rangle| \\
&= |\langle \widehat{q} - q^\pi, (1-\gamma)\mu_0^\pi + (\gamma P^{\pi,\top} - I)(d^D \cdot \widehat{w})\rangle| \\
&= \left|\left\langle \widehat{q} - q^\pi, (I - \gamma P^{\pi,\top})(d^D \cdot w^\pi - d^D \cdot \widehat{w})\right\rangle\right| \\
&\leq ||(I - \gamma P^\pi)(\widehat{q} - q^\pi)||_{2,d^D} ||\widehat{w} - w^\pi||_{2,d^D}
\end{aligned}$$

where the last equality is due to the fact that $(1-\gamma)\mu_0^\pi = (I - \gamma P^\pi)(d^D \cdot w^\pi)$, and the final inequality is from Cauchy-Schwarz. We can automatically bound the $||\widehat{w} - w^\pi||_{2,d^D}$ term using Theorem 6, and it remains to bound $||(I - \gamma P^\pi)(\widehat{q} - q^\pi)||_{2,d^D}$. We will consider two cases, first when $d^D > 0$ thus Diag(d^D) is invertible, and second, when Assumption 7 is satisfied.

In the first case, let $D = \text{Diag}(d^D)$, which by assumption is invertible. Then

$$\begin{aligned}
||(I - \gamma P^\pi)(\widehat{q} - q^\pi)||^2_{2,d^D} &= (\widehat{q} - q^\pi)^\top (I - \gamma P^\pi)^\top D (I - \gamma P^\pi)(\widehat{q} - q^\pi) \\
&= ||D^{1/2}(I - \gamma P^\pi)(\widehat{q} - q^\pi)||^2_2 \\
&= ||D^{1/2}(I - \gamma P^\pi) D^{-1/2} D^{1/2}(\widehat{q} - q^\pi)||^2_2 \\
&\leq ||D^{1/2}(\widehat{q} - q^\pi)||^2_2 ||D^{1/2}(I - \gamma P^\pi) D^{-1/2}||^2_2 \\
&= ||\widehat{q} - q^\pi||^2_{d^D,2} ||I - \gamma P^\pi||^2_2
\end{aligned}$$

in the last line using the fact that the eigenvalues of a matrix A and $L^{-1}AL$ are the same for any invertible matrix L. Thus, denoting the largest singular value of a matrix by σ_{max},

$$|\widetilde{J}(\pi) - J(\pi)| \leq \sigma_{max}\left(I - \gamma P^\pi\right) ||\widehat{w} - w^\pi||_{2,d^D} ||\widehat{q} - q^\pi||_{2,d^D}$$

Using Theorem 6 and Theorem 2 in the last line to control the errors of $\widehat{w}$ and $\widehat{q}$ in the last line, followed by a union bound over the three inequalities, gives the result.

For the second case, we can directly apply Lemma 17:

$$\begin{aligned}|\widetilde{J}(\pi) - J(\pi)| &\leq ||(I - \gamma P^\pi)(\widehat{q} - q^\pi)||_{2,d^D} ||\widehat{w} - w^\pi||_{2,d^D} \\ &\leq \left(||\widehat{q} - q^\pi||_{2,d^D} + \gamma ||P^\pi(\widehat{q} - q^\pi)||_{2,d^D}\right) ||\widehat{w} - w^\pi||_{2,d^D} \\ &\leq \left(1 + \gamma\sqrt{\mathcal{C}_{s'/s}\mathcal{C}_{\pi/\pi^D}}\right) ||\widehat{q} - q^\pi||_{2,d^D} ||\widehat{w} - w^\pi||_{2,d^D},\end{aligned}$$

and again applying Theorem 6 and Theorem 2 gives the result.

Lemma 17 uses Assumption 7 to bound the distance in value functions under the transition operator, and is stated and proved below.

Lemma 17. *Under Assumption 7,*

$$||P^\pi(\widehat{q} - q^\pi)||_{2,d^D} \leq \sqrt{\mathcal{C}_{s'/s}\mathcal{C}_{\pi/\pi^D}} ||\widehat{q} - q^\pi||_{2,d^D}.$$

Proof. Define $||P^\pi||_{2,d^D} := \sup_{x \neq 0} ||P^\pi x||_{2,d^D} / ||x||_{2,d^D}$. Then

$$||P^\pi(\widehat{q} - q^\pi)||_{2,d^D} \leq ||P^\pi||_{2,d^D} ||\widehat{q} - q^\pi||_{2,d^D}.$$

It remains to bound $||P^\pi||_{2,d^D}$. For any x,

$$\begin{aligned}||P^\pi x||_{2,d^D}^2 &= \mathbb{E}_{(s,a)\sim d^D}\left[\left(\mathbb{E}_{(s',a')\sim P^\pi(\cdot|s,a)}[x(s',a')]\right)^2\right] \\ &\leq \mathbb{E}_{(s,a,s',a')\sim d^D \times P^\pi}[x(s',a')^2] \\ &\leq \max_{s,a}\left|\frac{d^D_{s'}(s)\pi(a|s)}{d^D(s)\pi^D(a|s)}\right| \mathbb{E}_{(s,a)\sim d^D}[x(s,a)^2] \\ &= \mathcal{C}_{s'/s}\mathcal{C}_{\pi/\pi^D} ||x||_{2,d^D}^2\end{aligned}$$

This implies that $||P^\pi||_{2,d^D} \leq \sqrt{\mathcal{C}_{s'/s}\mathcal{C}_{\pi/\pi^D}}$, which gives the stated result. □

G Infinite Function Classes

Our results for finite function classes can be easily extended to infinite function classes using covering numbers. We show that our method value function estimation under infinite function classes achieves the same $\widetilde{O}(n^{-1/4})$ rate as it does under finite function classes (Theorem 2). The same results also apply to weight function learning using similar proof techniques.

G.1 Finite-sample Guarantees with Infinite Function Classes

First, we define the covering functions used in our results and analysis:

Definition 1 (Covering Number)**.** *For a function class $\mathcal{F}$, the covering number $\mathcal{N}_\infty(\epsilon, \mathcal{F})$ is defined to be the minimum cardinality of a set $\overline{\mathcal{F}} \subseteq \mathcal{F}$, such that for any $f \in \mathcal{F}$, there exists $\overline{f} \in \overline{\mathcal{F}}$ with $\|f - \overline{f}\|_\infty \leq \epsilon$.*

Our guarantee for value function learning under infinite function classes is stated below, showing that we achieve the same rate as we do with finite classes.

Theorem 18. *Suppose Assumptions 1, 2, 3 hold. Then, with probability at least $1 - \delta$, for $\epsilon = \frac{B}{2A\sqrt{n}}$,*

$$||\widehat{q} - q^\pi||_{2,\nu} \leq 2\sqrt{\frac{2B}{M^q}}\left(\frac{2\log\frac{2\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}\right)^{-1/4},$$

where $\mathcal{N}_\infty(\epsilon, \mathcal{Q})$ and $\mathcal{N}_\infty(\epsilon, \mathcal{W})$ are as per Definition 1, and $A = 1 + (1+\gamma)C^q_{\mathcal{Q}} + 2(1+\gamma)C^q_{\mathcal{W}}$ and $B = C^q_{\mathcal{W}}(1 + (1+\gamma)C^q_{\mathcal{Q}})$.

The proof is given below.

G.2 Proof of Theorem 18

The statistical error of estimating $\widehat{L}(q,w)$ under infinite function classes is the main technical detail of this proof. Given that, the stated bound on $\|\widehat{q}-q^\pi\|_\nu$ can be derived using the same methods (leveraging strong convexity and Lemma 9) as were used in the proofs for value function estimation under finite function classes, i.e. for Theorem 2 (in Appendix B.1) and for Theorem 12 (in Appendix E.2).

The bound on this statistical error is stated then proved below:

Lemma 19 (Statistical Error under Infinite Function Classes)**.** *Suppose Assumption 3 holds. Then, setting $\epsilon = \frac{B}{2A\sqrt{n}}$, for any $(q,w)\in\mathcal{Q}\times\mathcal{W}$ with probability at least $1-\delta$,*

$$|L(q,w)-\widehat{L}(q,w)| \le 2B\sqrt{\frac{2\log\frac{2\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}},$$

where $\mathcal{N}_\infty(\epsilon,\mathcal{Q})$ and $\mathcal{N}_\infty(\epsilon,\mathcal{W})$ are as per Definition 1, and $A = 1+(1+\gamma)C^q_{\mathcal{Q}}+2(1+\gamma)C^q_{\mathcal{W}}$ and $B = C^q_{\mathcal{W}}(1+(1+\gamma)C^q_{\mathcal{Q}})$.

Proof of Lemma 19. First, because the regularization term computes $\mathbb{E}_\nu[\cdot]$ exactly (not from samples), it has no effect on our bound. Formally, define the unregularized population Lagrangian to be

$$L_0(q,w) = \mathbb{E}_{d^D}[r(s,a)+\gamma q(s',\pi)-q(s,a)],$$

and its empirical version to be $\widehat{L}_0(q,w)$. Then the LHS of Lemma 19 is equivalent to

$$\begin{aligned}|L(q,w)-\widehat{L}(q,w)| &= |L_0(q,w)+\mathbb{E}_\nu[f_{s,a}(q(s,a))]-\widehat{L}_0(q,w)-\mathbb{E}_\nu[f_{s,a}(q(s,a))]|\\ &= |L_0(q,w)-\widehat{L}_0(q,w)|,\end{aligned}$$

so it suffices to bound the statistical error of estimating the unregularized Lagrangian $\widehat{L}_0$.

For some (later to-be-specified) $\epsilon>0$, let $\overline{\mathcal{Q}}$ be a minimal ϵ-covering of $\mathcal{Q}$ in the infinity norm as per Definition 1, that is, $|\overline{\mathcal{Q}}| = \mathcal{N}_\infty(\epsilon,\mathcal{Q})$. Let $\overline{\mathcal{W}}$ be defined similarly for $\mathcal{W}$. Then for any $(q,w)\in\mathcal{Q}\times\mathcal{W}$, let $(\overline{q},\overline{w})\in\overline{\mathcal{Q}}\times\overline{\mathcal{W}}$ be such that $\|q-\overline{q}\|_\infty\le\epsilon$ and $\|w-\overline{w}\|_\infty\le\epsilon$. By triangle inequality,

$$|L_0(q,w)-\widehat{L}_0(q,w)| \le |L_0(q,w)-\widehat{L}_0(q,w)-L(\overline{q},\overline{w})-\widehat{L}(\overline{q},\overline{w})| + |L(\overline{q},\overline{w})-\widehat{L}(\overline{q},\overline{w})|$$

Next, define $\ell_{sas'}(q,w) := w(s,a)(r(s,a)+\gamma q(s',\pi)-q(s,a))$ such that $L_0(q,w) = \mathbb{E}_{d^D}[\ell_{sas'}(q,w)]$ and $\widehat{L}_0(q,w) = \frac{1}{n}\sum_{i=1}^n \ell_{s_ia_is_i'}(q,w)$. Then we can further upper bound the above as:

$$|L_0(q,w)-\widehat{L}_0(q,w)| \le 2\max_{s,a,s'}\underbrace{|\ell_{sas'}(q,w)-\ell_{sas'}(\overline{q},\overline{w})|}_{\text{(T1)}} + + \underbrace{|L(\overline{q},\overline{w})-\widehat{L}(\overline{q},\overline{w})|}_{\text{(T2)}}.$$

Term (T1) can be controlled using the ϵ-covering definition, and (T2) can be controlled using standard concentration methods. Their respective bounds are provided below, with proofs in the next subsection:

Lemma 20 (Bound for T1)**.** *Let $\overline{\mathcal{Q}}$ and $\overline{\mathcal{W}}$ be ϵ-coverings of $\mathcal{Q}$ and $\mathcal{W}$, respectively, satisfying Definition 1. Then for any $(q,w)\in\mathcal{Q}\times\mathcal{W}$, there exists $(\overline{q},\overline{w})\in\overline{\mathcal{Q}}\times\overline{\mathcal{W}}$ such that $\|q-\overline{q}\|_\infty\le\epsilon$ and $\|w-\overline{w}\|_\infty\le\epsilon$, and if Assumption 3 holds,*

$$\max_{sas'}|\ell_{sas'}(q,w)-\ell_{sas'}(\overline{q},\overline{w})| \le A\epsilon,$$

with $A = 1+(1+\gamma)C^q_{\mathcal{Q}}+2(1+\gamma)C^q_{\mathcal{W}}$.

Lemma 21 (Bound for T2)**.** *Let $\overline{\mathcal{Q}}$ and $\overline{\mathcal{W}}$ be minimal ϵ-coverings of $\mathcal{Q}$ and $\mathcal{W}$, respectively, as in Definition 1, that is, $|\overline{\mathcal{Q}}| = \mathcal{N}_\infty(\epsilon,\mathcal{Q})$ and $|\overline{\mathcal{W}}| = \mathcal{N}_\infty(\epsilon,\mathcal{W})$. Then if Assumption 3 holds, for any $(\overline{q},\overline{w})\in\overline{\mathcal{Q}}\times\overline{\mathcal{W}}$ w.p. $\ge 1-\delta$,*

$$|L_0(\overline{q},\overline{w})-\widehat{L}_0(\overline{q},\overline{w})| \le B\sqrt{\frac{2\log\frac{\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}},$$

where $B = \mathcal{C}^q_{\mathcal{W}}(1+(1+\gamma)\mathcal{C}^q_{\mathcal{Q}})$.

Putting these two bounds together, letting A be as in Lemma 20 and B be as in Lemma 21, we have

$$|L_0(q,w) - \widehat{L}_0(q,w)| \leq 2A\epsilon + B\sqrt{\frac{\log \frac{2\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}}.$$

Choosing $\epsilon = \frac{B}{2A\sqrt{n}}$ gives the final bound:

$$\begin{aligned}|L(q,w) - \widehat{L}(q,w)| = |L_0(q,w) - \widehat{L}_0(q,w)| &\leq \frac{B}{\sqrt{n}} + B\sqrt{\frac{\log \frac{2\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}} \\ &\leq 2B\sqrt{\frac{\log \frac{2\mathcal{N}_\infty(\epsilon,\mathcal{Q})\mathcal{N}_\infty(\epsilon,\mathcal{W})}{\delta}}{n}}.\end{aligned}$$

□

G.3 Proofs for Helper Lemmas

The proofs of Lemmas 20 and 21 are given below:

Proof of Lemma 20. For any s, a, s', (since this tuple is fixed, going forward, we drop the s, a, s' subscript from ℓ for brevity)

$$\begin{aligned}|\ell_{sas'}(q,w) - \ell_{sas'}(\overline{q},\overline{w})| &= |\ell(q,w) - \ell(q,\overline{w}) + \ell(q,\overline{w}) - \ell(\overline{q},\overline{w})| \\ &\leq \underbrace{|\ell(q,w) - \ell(q,\overline{w})|}_{\text{T3}} + \underbrace{|\ell(q,\overline{w}) - \ell(\overline{q},\overline{w})|}_{\text{T4}}\end{aligned}$$

(T3) expresses the error from the covering approximation for w, while (T4) expresses this for q. First, to bound (T3),

$$\begin{aligned}|\ell(q,w) - \ell(q,\overline{w})| &= |(w(s,a) - \overline{w}(s,a))(r(s,a) + \gamma q(s',\pi) - q(s,a)| \\ &\leq \|w - \overline{w}\|_\infty(\|r\|_\infty + (1+\gamma)\|q\|_\infty) \\ &\leq \epsilon(1 + (1+\gamma)C^q_{\mathcal{Q}}).\end{aligned}$$

To bound (T4),

$$\begin{aligned}|\ell(q,\overline{w}) - \ell(\overline{q},\overline{w})| &= |\overline{w}(s,a)(\gamma q(s',\pi) - q(s,a) - \gamma\overline{q}(s',\pi) + \overline{q}(s,a))| \\ &\leq 2(1+\gamma)\|\overline{w}\|_\infty\|q - \overline{q}\|_\infty \\ &\leq 2(1+\gamma)C^q_{\mathcal{W}}\epsilon.\end{aligned}$$

Since these two inequalities hold for any sas', combining them directly gives lemma statement. □

Proof of Lemma 21. This is a straightforward application of Hoeffding's with union bound over $\overline{\mathcal{Q}}, \overline{\mathcal{W}}$, akin to the proof of Lemma 8 (which is over $\mathcal{Q}, \mathcal{W}$). □