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,

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

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 $x, y \in \mathbb{R}$,

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

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

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

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. $\geq 1-\delta$,*

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

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

Proof. From the linearity of the expectation, it is clear that $L_f^q(q,w) = \mathbb{E}[\widehat{L}_f^q]$. 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,

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

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

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

□

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

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

where ϵ_{stat}^q is given in Lemma 8.

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

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

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

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

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 $L_f^w(w, q) = \mathbb{E}_\eta[f_{s,a}(w(s,a))] + \mathbb{E}_{d^{\mathcal{D}}}[l_i] + \mathbb{E}_{\mu_0}[q(s_0, \pi)]$, 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

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

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

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

[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:

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

$\epsilon_{opt,w}$ 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,

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

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:

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

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* $\widehat{q} = \arg\min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} \widehat{L}_f^q(q, w)$. *Then with probability* $\geq 1 - 2\delta$,

$$|\widehat{J}^{q}(\pi) - J(\pi)| \leq \epsilon^q_{eval} + \sqrt{\mathcal{C}_{\mu_0^\pi/\nu}} \cdot \epsilon^q_{est},$$

where $\epsilon^q_{eval} = (1-\gamma) C^q_{\mathcal{Q}} \sqrt{2 \log \frac{2|\mathcal{Q}|}{\delta} / n_0}$, $\mathcal{C}_{\mu_0^\pi/\nu} = ||\mu_0^\pi/\nu||_\infty$, *and* ϵ^q_{stat} *is as in Theorem 2.*

Corollary 15. *Suppose Assumption 4, 5, and 6 hold, and let* $\widehat{w} = \arg\min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} \widehat{L}_f^w(q, w)$. *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:*

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

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 $\widetilde{J}(\pi) = \mathbb{E}_{d^\mathcal{D}}[\widehat{w}(s,a) r(s,a)]$. 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 $w \in \mathcal{W}$,

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

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 $\text{Diag}(d^D)$ 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,

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

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 $A = 1 + (1+\gamma)C_{\mathcal{Q}}^q + 2(1+\gamma)C_{\mathcal{W}}^q$ and $B = C_{\mathcal{W}}^q(1 + (1+\gamma)C_{\mathcal{Q}}^q)$.

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 $1 - \delta$,*

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

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

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

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 $\|q - \overline{q}\|_\infty \leq \epsilon$ and $\|w - \overline{w}\|_\infty \leq \epsilon$. By triangle inequality,

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

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

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

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 $\|q - \overline{q}\|_\infty \leq \epsilon$ and $\|w - \overline{w}\|_\infty \leq \epsilon$, and if Assumption 3 holds,*

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

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. $\geq 1 - \delta$,*

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

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 $\mathcal{Q}, \mathcal{W}$). □